Towards Ontological Foundations of Research Information Systems

Dissertation
zur Erlangung des akademischen Grades eines
Doktors der Philosophie
der Philosophischen Fakultät III
der Universität des Saarlandes

vorgelegt von

Jörg, Brigitte

aus

Kempten (Allgäu)

Der Dekan: Univ. Prof. Dr. Roland Brünken

Erstberichterstatter: Herr Univ.-Prof. Dr. H. Zimmermann

Zweitberichterstatter: Herr Univ.-Prof. Dr. H. Uszkoreit

Drittberichterstatter: Herr Univ.-Prof. Dr. Keith G. Jeffery

Tag der Disputation: 3. Juni 2013

Table of Contents

List of Figures

List of Tables

Sample Descriptions in Formal Notations

Glossary

ADL	*A*rchitecture *D*efinition *L*anguages
ARIS	*I*ntegrated *S*ystems *Ar*chitecture
BFO	*B*asic *F*ormal *O*ntology
BIBO	*Bi*bliographic *O*ntology
BWW	*B*unge-*W*and-*W*eber
CASRAI	*C*onsortia *A*dvancing *S*tandards in *R*esearch *A*dministration
CERIF	*C*ommon *E*uropean *R*esearch *I*nformation *F*ormat
CIM	*C*omputer *I*ndependent *M*odel
CLARIN	*C*ommon *L*anguage *R*esources and *T*echnology *In*frastructure
CMS	*C*ontent *M*anagement *S*ystems
COAR	*C*onfederation of *O*pen *A*ccess *R*epositories
CODASYL	*Co*mmittee on *Da*ta *S*ystems *L*anguages
CORDIS	*C*ommunity *R*esearch and *D*evelopment *I*nformation *S*ervice
CRIS	*C*urrent *R*esearch *I*nformation *S*ystem
CWA	*C*losed-*W*orld *A*ssumption
CV	*C*urriculum *V*itae
DAML-OIL	DARPA Agent Markup Language – Ontology Inference Layer
DBMS	Database Management Systems
DCAM	DCMI Abstract Model
DCMI	*D*ublin *C*ore *M*etadata *I*nitiative
DELOS	*D*igital *L*ibrary *R*eference *M*odel
DFG	*G*erman *R*esearch *F*oundation
DOI	*D*igital *O*bject *I*dentifier
DOLCE	*D*escriptive *O*ntology for *L*inguistic and *C*ognitive *E*ngineering
DTD	*D*ocument *T*ype *D*efinition
DL	*D*escription *L*ogics
DSSP	*D*ata*S*pace *S*upport *P*latforms
EC	*E*uropean *C*ommission
EDMS	*E*lectronic *D*ocument *M*anagement *S*ystems
ESF	*E*uropean *S*cience *F*oundation
EXI	*E*fficient *X*ML *I*nterchange Format
ERM	*E*ntity *R*elationship *M*odel
ERP	*E*nterprise *R*esource *P*lanning
ESFRI	*E*uropean *S*trategy *F*orum on *R*esearch *I*nfrastructures
FERON	*F*ield-*e*xtensible *R*esearch *O*ntology *(this work)*
FET	*F*uture and *E*merging *T*echnologies
FOAF	*F*riend-*o*f-*a*-*F*riend
FP7	*7*th *F*ramework *P*rogramme
FRBR	*F*unctional *R*equirements for *B*ibliographic *R*ecords
GDI	*G*eneral *D*efinition of *I*nformation
GI	*G*eneric *I*dentifier
HLT	*H*uman *L*anguage *T*echnology
HTML	*H*ypertext *M*arkup *L*anguage
HTTP	*H*ypertext *T*ransfer *P*rotocol
ICP	*I*nternational *C*ataloguing *P*rinciples
ICSU	*I*nternational *C*ouncil for *S*cience
IETF	*I*nternet *E*ngineering *T*ask *F*orce
IFLA	*I*nternational *F*ederation of *L*ibrary *A*ssociations and *I*nstitutions
IR	*I*nstitutional *R*epository
IRI	*I*nternationalized *R*esource *I*dentifiers
ISBD	*I*nternational *S*tandard *B*ibliographic *D*escription
ISM	*I*mplementation *S*pecific *M*odel
ISO	*I*nternational *O*rganization for *S*tandardization
JISC	*J*oint *I*nformation *S*ystems *C*ommittee
KB	*K*nowledge *B*ase
KOS	*K*nowledge *O*rganisation *S*ystems
LD	*L*inked *D*ata
LIS	*L*ibrary and *I*nformation *S*cience
LMS	*L*earning *M*anagement *S*ystem
LO	*L*earning *O*bject
LOD	*L*inked *O*pen *D*ata
LOM	*L*earning *O*bject *M*etadata

LT	*L*anguage *T*echnology
MARC	The MARC formats are standards for the representation and communication of bibliographic and related information in machine-readable form.
MARTIF	*M*achine *R*eadable *T*erminology *I*nterchange *F*ormat
MDA	*M*odel-*d*riven *A*rchitectures
MDE	*M*odel-*d*riven *E*ngineering
MDD	*M*odel-*d*riven *D*evelopment
MIS	*M*anagement *I*nformation *S*ystems
MODS	*M*etadata *O*bject *D*escription *S*chema
MOF	*M*eta-*O*bject-*F*acility
NLP	*N*atural *L*anguage *P*rocessing
NSF	*N*ational *S*cience *F*oundation
OA	*O*pen *A*ccess
OAI	*O*pen *A*ccess *I*nitiative
OAI-PMH	*O*pen *A*ccess *I*nitiative – *P*rotocol for *M*etadata *H*arvesting
OAIS	*O*pen *A*rchival *I*nformation *S*ystem
OECD	*O*rganisation for *E*conomic *C*o-operation and *D*evelopment
OMG	*O*pen *M*anagement *G*roup
OODBS	*O*bject-*O*riented *D*atabase *S*ystem
OPAC	*O*nline *P*ublic *A*ccess *C*atalog
ORCID	*O*pen *R*esearcher and *C*ontributor *ID*
ORE	*O*bject *R*euse and *E*xchange
OWA	*O*pen-*W*orld *A*ssumption
OWL	*W*eb *O*ntology *L*anguage
PIM	*P*latform *I*ndependent *M*odel
PSM	*P*latform *S*pecific *M*odel
QA	*Q*uestion-*A*nswering
R&D	*R*esearch and *D*evelopment
RDA	*R*esource *D*escription and *A*ccess
RDF	Resource Description Framework
RELAX-NG	*RE*gular *LA*nguage for *X*ML *N*ext *G*eneration
REST	*R*epresentatioal *S*tate *T*ransfer
RFC	*R*equest *f*or *C*omments
RDF	*R*esource *D*escription *F*ramework
RIM	*R*esearch *I*nformation *M*anagement
RIS	*R*esearch *I*nformation *S*ystem
SBVR	*S*emantics of *B*usiness *V*ocabulary and *R*ules
SKOS	*S*imple *K*nowledge *O*rganisation *S*ystem
SOA	*S*ervice-*o*riented *A*rchitectures
SOAP	*S*imple *O*bject *A*ccess *P*rotocol
SPARQL	*S*PARQL *P*rotocol *a*nd *R*DF *Q*uery *L*anguage
SQL	*S*tructured *Q*uery *L*anguage
STA	*S*cientific and *T*echnological *A*ctivities
STET	*S*cientific and *T*echnical *E*ducation and *T*raining
SUMO	*S*uggested *U*pper *M*erged *O*ntology
SUO	*S*tandard *U*pper *O*ntology
SURF	The Dutch higher education and research partnership for network services and information and communication technology (ICT).
SW	*S*emantic *W*eb
TAG	*T*echnical *A*rchitecture *G*roup
TEI	*T*ext *E*ncoding *I*nitiative
OLAC	*O*pen *L*anguage *A*rchive *C*ommunity
UNESCO	*U*nited *N*ations *E*ducational, *S*cientific and *C*ultural *O*rganization
UNISIST	Unescos World Scientific Information Programme
URI	*U*niform *R*esource *I*dentifier
UML	*U*nified *M*odeling *L*anguage
VIAF	*V*irtual *I*nternational *A*uthority *F*ile
VIVO	An interdisciplinary network of scientists
W3C	*W*orld *W*ide *W*eb *C*onsortium
WWW	*W*orld *W*ide *W*eb
XML	*E*xtensible *M*arkup *L*anguage
XSD	*X*ML *S*chema *D*efinition *L*anguage

Abstract

Despite continuous advancements in information system technologies it is still not simple to receive relevant answers to Science-related queries. Getting answers requires a gathering of information from heterogeneous systems, and the volume of responses that semantically do not match with the queried intensions overwhelms users. W3C initiatives with extensions such as the Semantic Web and the Linked Open Data Web introduced important technologies to overcome the issues of semantics and access by promoting standard representation formats – formal ontologies – for information integration. These are inherent in architectural system styles, where increased openness challenges the traditional closed-world and often adhocly designed systems. However, technology on its own is not meaningful and the information systems community is increasingly becoming aware of foundations and their importance with guiding system analyses and conceptual design processes towards sustainable and more integrative information systems. As a contribution, this work develops a formal ontology FERON – *F*ield-*e*xtensible *R*esearch *On*tology – following the foundations as introduced by Mario Bunge and applied to information systems design by Wand and Weber, i.e. Bunge-Wand-Weber (BWW). Nevertheless, FERON is not aimed at the modelling of an information system as such, but at the description of a perceived world – *the substantial things* – that an information system ought to be able to model. FERON is a formal description of the Research domain – a formal ontology according to latest technological standards. Language Technology was chosen as a subdomain to demonstrate its field extensibility. The formal FERON ontology results from a hybrid modelling approach; it was first described top-down based on a many years activity of the author and then fine-tuned bottom-up through a comprehensive analysis and re-use of openly available descriptions and standards. The entire FERON design process was accompanied by an awareness of architectural system levels and system implementation styles, but was at first aimed at a human domain understanding, which according to the *General Definition of Information* (GDI) is achievable through well-formed meaningful data.

German Summary

Trotz kontinuierlich verbesserter Informationssystemtechnologien ist es nicht einfach möglich, relevante Antworten auf forschungsverwandte Suchanfragen zu erhalten. Dies liegt unter anderem daran, dass Informationen in verschiedenen Systemen bereitgestellt werden, und dass die Beschreibung der bereitgestellten Informationen nicht mit den Beschreibungen der gestellten Fragen übereinstimmen. Neuere Technologien wie das Semantische Web oder Linked Open Data ermöglichen zwar verbesserte Beschreibungen und Zugriffe – jedoch sind die Technologien an sich auch nicht bedeutungsvoll. Weitergehende, fundierende Ansätze zur Beschreibung von Informationenen finden daher zunehmend Anerkennung und Zuspruch in der wissenschaftlichen Gemeinde, diese beinflussen konsequenterweise die Systemanalyse sowie das Systemdesign. Die vorliegende Arbeit entwickelt eine formale Ontologie einer Forschungswelt die disziplinenübergreifend skaliert, namentlich FERON – *F*ield-*e*xtensible *R*esearch *On*tology, basierend auf den Ansätzen der Bunge-Wand-Weber (BWW) Ontologie.

Der Titel der Arbeit "Towards Ontological Foundations of Research Information Systems" übersetzt: „Zur ontologischen Fundierung von Forschungsinformationssystemen". Im Titel ist ontologisch zuallererst im philosophischen Sinne zu verstehen, und nicht zu verwechseln mit der dann resultierenden Ontologie im technologischen Sinne einer formalen Beschreibung der wahrgenommenen *Forschungswelt* – namentlich FERON. Eine Klärung der Begriffe *Ontologie*, *Konzept*, *Entität*, *Daten* und *Information* zum Verständnis der vorliegenden Arbeit wird in Kapitel 2.5 versucht, ein Verständnis wurde als kritisch für die Qualität der resultierenden formalen Ontologie FERON, aber auch als hilfreich für den Leser vorweggenommen, insbesondere weil die genannten Begriffe über Disziplinen hinweg oftmals sehr unterschiedlich wahrgenommen werden. Die Analyse und Modellierung von FERON basiert auf der Bedeutung dieser grundlegenden Begriffe wie die philosophische und wissenschaftliche Literatur verschiedener Disziplinen sie belegt.

Die vorliegende Arbeit entwickelt FERON, und modelliert eine *Welt* der Forschung in disziplinenübergreifender Weise mittels neuester technologischer Standards – formal in RDF/OWL. Die fachspezifische Erweiterbarkeit ist durch Eingliederung von Beschreibungen des Gebietes Sprachtechnologie demonstriert. Die Modellierung wurde durchgehend von der Theorie Mario Bunges begleitet, welche Wand und Weber für eine Anwendung während der Systemanalyse und Systemgestaltung interpretierten und welche im Kapitel 3.1.1 vorgestellt wird. Die *Idee* ist als *Bunge-Wand-Weber Ontologie* (BWW) zunehmend bekannt und demgemäße *ontologische* Ansichten sind teilweise in formalen Beschreibungssprachen und

Werkzeugen eingebunden, und damit bei der Modellierung explizit nutzbar. Neben BWW werden kurz die Fundierungsansätze von DOLCE, SUMO und Cyc vorgestellt und deren Relevanz für FERON verdeutlicht.

Eine fehlende Fundierung in der Disziplin *Informationssysteme* wurde lange Zeit als wesentliche Ursache für die vermisste wissenschaftliche Akzeptanz der Disziplin betrachtet; größtenteils wurden Informationssysteme pragmatisch und adhoc entwickelt und skalierten daher nicht konsistent. Zunehmend wird jedoch eine theoretische und insbesondere die ontologische Fundierung von Informationssystemen als wertvoll anerkannt – von der Idee bis hin zur Implementierung aber auch während der Umgestaltungsphasen. Konzepte fundierter Informationssysteme im funktional-technischen Sinne sind als modellgetriebene Architektur bekannt und werden hier durch die Ansätze von Zachmann und Scheer verdeutlicht. In der kurzen Geschichte IT-basierter Informationssysteme wurden phasenweise immer wieder strukturell unterschiedliche Modelle angewandt. Diese werden daher im Kapitel 3.2 Modellierungsgrammatiken untersucht und deren Unterschiede dargestellt – namentlich das Entity-Relationship-Modell, semantische Netzwerke, das relationale Modell, hierarchische Modelle und objekt-orientierte Modelle. Darüberhinaus sind insbesondere formale Ontologien durch die Web Standardisierungsaktivitäten und W3C Empfehlungen ein rasant wachsendes Segment, verstärkt durch politische Entscheidungen für offene Daten und implizierend offene Systeme.

Im Vergleich zu traditionellen und weitestgehend geschlossenen sogenannten *closed-world* Systemen sind hinsichtlich der Modellierung bestimmte Aspekte zu beachten. Diese unterliegen im Gegensatz zu offenen Systemen dem Paradigma des *kompletten Wissens* und sind sozusagen *vorschreibend*; im System aktuell nicht vorhandene Information wird als *nicht existent* interpretiert. Dahingegen gehen offene *open-world* Systeme davon aus, dass nicht vorhandene Information aktuell unbekannt ist – und die bekannte Information nicht vorschreibt sondern *beschreibt*. Weitere Unterschiede die es bezüglich der Modellierung zu beachten gilt, befassen sich mit zeitlich geprägten Verknüpfungen – über sogenannte *Links* oder *Relationships* – aber auch mit Entitäten und deren Identitäten. Da FERON keine Ontologie eines Informationssystems selbst modelliert, sondern eine Welt für eine mögliche Umsetzung in einem Informationssystem bechreibt sind weitergehende Modellierungsaspekte in Kapitel 3.3 lediglich erklärt und es wird auf Beispiele verwiesen.

In der vorliegenden Arbeit wird keine explizite Anwendung empfohlen, weil ein Informationssystem immer derjenigen Form entsprechen sollte, welche einer bestimmten Funktion folgt, und weil die Vorwegnahme von Funktionen eine Dimension darstellt die weit über das Maß der vorliegenden Arbeit hinaus geht.

FERON beschreibt eine Welt der Forschung; vorhandene Modellierungsansätze von Forschungsinformationssystemem werden mit Kapitel 4.1 den Ansätzen verwandter Arten gegenübergestellt – nämlich, wissenschaftlichen Repositorien, Datenrepositorien, Digitalen Bibliotheken, Digitalen Archiven und Lehre Systemen. Die untersuchten Modelle offenbaren neben inhaltlichen Unterschieden auch die Verschiedenheit der Modellierungsansätze von z.B. Referenzmodellen gegenüber formalen Datenmodellen oder offenen Weltbeschreibungen, und damit auch die einhergehende Schwierigkeit von Integration. Insbesondere *formale Ontologien* erlauben über die traditionellen Ansätze hinweg, automatische Schlußfolgerungen und Beweisführungen, welche jedoch hier nicht weitergehend erörtert werden. FERON war von Anfang an für den menschlichen Leser konzipiert, wenn auch formal beschrieben.

Der Modellierungsansatz in FERON ist hybrid und wird in Kapitel 7 erläutert. Eine hybride Modellierung war möglich durch eine mehr als zehn-jährige Erfahrung und Tätigkeit der Autorin in diesem Bereich, auch belegt durch zahlreiche Peer-Review Publikationen. Der erste Entwurf von FERON erfolgte demgemäß zuallererst im *Top-Down* Verfahren (Figure 29), bevor mittels umfassender Analyse (dokumentiert in den Kapiteln 5 und 6) von verfügbaren Domänenbeschreibungen sukszessive eine *Bottom-Up* Anpassung von FERON vorgenommen wurde (Figure 68), welche bereits standardisierte und bereits definierte Beschreibungen und Eigenschaften wenn möglich integrierte (Figure 67). FERON ist eine ontologisch fundierte, formale Beschreibung – eine formale Ontologie – einer Forschungswelt zur vereinfachten, konsistenten Umsetzung von standardisierten, integrativen Forschungsinformationssystemen oder Fachinformationssystemen. Substantielle Entitäten wurden grundsätzlich erkannt, und deren Eigenschaften sowie Verknüpfungen formal beschrieben (Kapitel 7): *Ressource* unterschieden nach *Nicht-Informations-Ressource* und *Informations-Ressource*. Erstere unterscheidet nach *Agent (Person, Organisationseinheit)*, *Aktivität (Methode, Projekt, Bildung, Ereignis)*, *Förderung (Programm, Einkommen)*, *Messung* und *Infrastruktur (Werkzeug, Dienst, Einrichtung)*, zweitere nach *Publikation*, *Literatur*, *Produkt (Daten), Wissensorganisationssystem*, auch bekannt als *KOS* (Knowledge Organisation System), wie in der nachfolgenden Graphik (Figure 1) demonstriert.

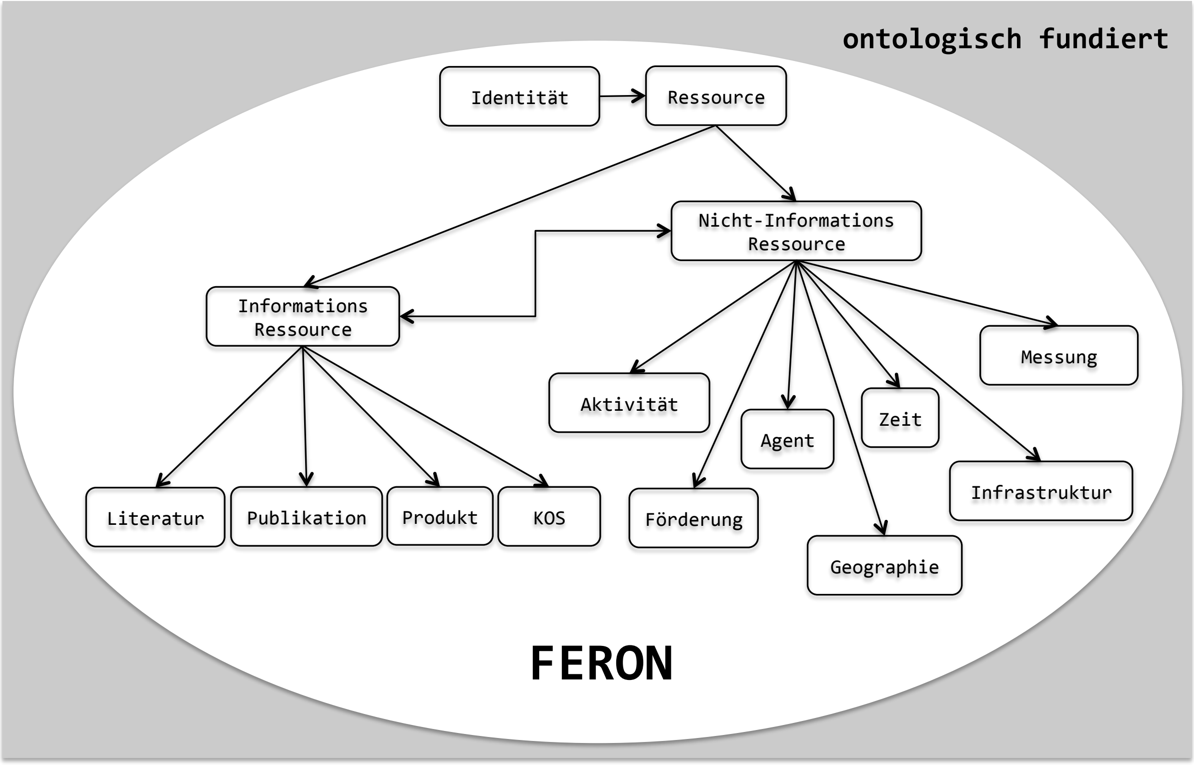


Figure 1: FERON (deutsch: Auf Disziplinen erweiterbare Forschungsontologie)

Kapitel 7 präsentiert FERON und dessen formale Einbindung von übergreifenden Eigenschaften wie *Sprache*, *Zeit*, *Geographie*, *zeitlich geprägte Verknüpfung*, *ontologische Verpflichtung*, *Namensraum*, *Klasse*, *Eigenschaft*, *funktionales Schema*, *Entität und Identität*. Seine inherente Struktur erlaubt eine einfache Disziplinen- oder Domänenerweiterung. Die Sprachtechnologie (englisch: *Language Technology* – abgekürzt *LT*) wird als Gebiet zur Demonstration der Erweiterung von FERON formal eingebunden, und mit Kapitel 6 insbesondere seine substantiell fach-spezifischen Entitäten wie *Methode*, *Projekt*, *Daten*, *Service*, *Infrastruktur*, *Messung*, aber auch *KOS* untersucht.

Eine Erweiterung der Ontologie FERON für explizit-funktionale Anforderungen an ein Informationssystem, oder für weitergehende disziplinen-spezifische Eigenschaften, z.B. einer linguistisch verbesserten Anwendung für sprachtechnologische Weiterverarbeitung, ist möglich, erfordert jedoch tiefergehendes Fachwissen.

Ziel der Arbeit war es zuallererst, das Verständnis für die Domäne *Forschung* zu verbessern – mit weiterreichendem Blick auf eine allgemeine integrative system-technische Entwicklung zur Verbesserung von Informationszugriff und Informationsqualität. Daneben wurden historische, gesellschaftliche aber auch politische Faktoren beobachtet, welche helfen, die wachsenden Anforderungen jenseits der Technologie zu bewältigen.

FERON ist als formales Model *FERON.owl* valide und wird mit der vorliegenden Arbeit sozusagen als Template zur weiteren Befüllung bereitgestellt. Darauf basierend sind formale Restriktionen sowie disziplinen-spezifische und terminologische Erweiterungen direkt möglich. Daten-*Instanzen* wie in den präsentierten Beispielen sind mittels *FERON.pprj* verfügbar.

Acknowledgements

I wish to thank very much Professor Dr. Harald H. Zimmermann for his availability and for his patience over the many years, and especially for his encouragement to continue and finish this work at times when it did not feel like a goal that could be easily achieved. I will keep in my memory especially the discussions we had about different classification systems and their individual notations. They contributed very much to my basic understanding of inherent heterogeneous structures in knowledge organisation systems and the need for identifiers with interlinkage between systems. I want to furthermore mention the Nietzsche seminars and the collaborative work while setting up an interlinked structured XML-driven Web portal with expandable information items, from where a lot of my motivation originated and when actually it was then, that I decided to study Information Systems and to continue in this direction.

I wish to thank very much Professor Dr. Hans Uszkoreit, at first for the idea of this thesis. Furthermore, for enabling me, to create an environment where I could combine my interests with the work requirements to a large extent. It started during the writing of my Master's thesis and continued with my later work and engagement in euroCRIS activities. DFKI's LT Lab directed by Professor Uszkoreit was always a highly dynamic and forward thinking open environment. In particular, I will remember discussions we had about ontology, knowledge, and scientific information that contributed much to my understanding of these concepts, and which are so much in the center of my interest and this work. I want to also mention DFKI as an institution and a recognised center of excellence with engagements in European and international activities, supplying not only a very professional and dedicated work environment with highly motivated people, but which contributed a lot to my understanding of Research and the scientific domain in a close relationship to the commercial sector, and thus prepared me adequately for my future career. Within DFKI, I want to especially mention two colleagues: Stephan Busemann who discussed with me LT-related matters and who referred me to relevant LT literature whilst being aware that my background was not LT; Hans-Ulrich Krieger for very exciting discussions related to time modeling and reasoning, contributing substantially to my understanding of logics, where his knowledge over inferencing and reasoning is astonishing and goes far beyond the work here.

I wish to thank very much Professor Dr. Keith Jeffery for highly inspiring and enlightening discussions in the context of information systems, for his openness with knowledge and experience, for his stressing of the importance of history, and for his advocacy of relational structures to reveal the contrasts in related approaches. I learned a lot about the Research ecosystem from the work within euroCRIS and the euroCRIS Board, where it was a pleasure to work under Keith as the President for more then eight years. Keith's encouragement "nobody ever said it will be easy" often kept me going, when I was in doubt about the continuation of this thesis.

Last but not least, I want to thank my family – especially my parents, my sister and my brothers for always being there when I needed them – and my friends for understanding the importance that this work had for me, and which made me a rare visitor for some time.

I am back now.

1 Introduction

The last century imposed dramatic changes to Science [Kuhn 1962] and the speed of change is still ever increasing. Derek J. De Solla Price was amongst the first to investigate and measure statistically scientific activity in terms of “manpower, literature, talent, and expenditure on a national and on an international scale” to learn about the changes treating Science as a “measurable entity” [de Solla Price 1963, preface]. He had recognised that the growing gap between activities and expenditures, will require increased governance: “In a saturation economy of science it is obvious that the proper deployment of resources becomes much more important than the expensive attempts to increase them” (p. 112). Science is transforming into a global business, turning knowledge into a commercial good “universities are being challenged to show their values as the commercial sector emerges into the same arena”[1]. The pursuit of new knowledge is often approached through applications “translation of research findings or knowledge into new or improved products and services is increasingly seen as an integral part of the research process” [EC Report 2010, pp. 24 ff.]; knowledge has become “democratized in the sense that more people are aware of the issues and are social actors in the application of knowledge”, but at the same time knowledge has become much more complex “reflected in the emergence of new disciplines, new methodologies and ways of thinking, transforming societies and the way in which knowledge is created and used”; the boundaries between basic or applied research are ‘blurring’ (p. 24). Traditionally, knowledge was produced in a “disciplinary, primarily cognitive context” [Gibbons et al. 1994, p.1], whilst now it is embedded “in a broader, transdisciplinary social and economic context”, and terms such as ‘applied science’, ‘technological research’, or ‘R&D’ (p. 2) are considered inadequate to describe this ‘new mode’ of knowledge production [Gibbons et al. 1994]. “Interdisciplinary thinking is rapidly becoming an integral feature of research as a result of four powerful ‘drivers’: the inherent complexity of nature and society, the desire to explore problems and questions that are not confined to a single discipline, the need to solve societal problems, and the power of new technologies” [Statement of the US Committee on Facilitating Interdisciplinary Research (2004); cited in the EC Report 2010, p. 25].

The global research community is at a crossroads – enforced by “fiscal responsibility, limited research funds, greater number of students, increased competition, the changing nature of science in its relationship to society and global economies, and the growing internationalization of science. These changes have occured just as new and emerging

[1] New platform is positive sign for research: http://www.researchinformation.info/news/news_story.php?news_id=711 (by Neil Jacobs, published in January 2011)

information and communication technologies have been impacting on organizations" [Zimmermann 2002, p. 11], and has inevitably lead to questions of governance. Where changes in information and communication processes are induced by technology, it is also clear, that effectively it needs a 'push' from policy [Hornbostel 2002, p. 29] for wider implementation and uptake. The European Commission revealed "for researchers receiving funding [...] open-access publishing "will be the norm" in the forthcoming framework programme Horizon 2020"[2]. Access to scholarly information had been dominated for almost half a century by Thomson ISI[3] [Jeffery 2010 p. 7]; its shortcomings have stimulated several initiatives to improve the access [Hornbostel 2006, pp. 30–31][4]. Amongst them are Current Research Information Systems (*CRISs*) describing the wider scientific context, to cope with an increased complexity even beyond academia [Cox et al. 2011][5], [Baker 2012]. New services such as CiteSeer[6], GoogleScholar[7], or Microsoft Academic Search[8] emerged and enable access through full-text-indexing of open or subscription-based scholarly material.

Research information is a multi-valued asset required by various stakeholders. It has therefore often been, and still is maintained in distributed and often home-grown systems. In spite of improved technologies, commercialisation and governmental support, it is still not easy to get relevant answers to science-related questions, and information overload remains a significant issue. Users are still overwhelmed by a volume of responses that to a high degree does not match with their queried intensions. Furthermore, queries are required to being replicated across systems, and need particular amendments within each. New technologies and Web standardisation initiatives with extensions such as the Semantic Web [Brickley & Guha 2000] and the Linked Open Data Web [Bizer et al. 2009] introduced important steps to

[2] Open Access Push from Brussels: http://www.insidehighered.com/news/2012/05/17/european-union-links-research-grants-open-access (Last visit: May 20th, 2012)

[3] Now called Thomson Reuters – Web of Science: http://thomsonreuters.com/products_services/science/science_products/a-z/web_of_science/ (Last visit: May 20th, 2012)

[4] „In Europe, the invention of a European Social Science Citation Index was discussed - Recently, competitive products like SCOPUS have been launched. - Instead of citation analysis the power of evidence of „usage metrics" is being examined (WEB-logs, link resolver, download metrics et cetera). - With open-access publications the markets of scientific journals are changing as are the techniques of quality assurance (open peer review). - Self-archiving now supplements publishing in traditional peer-reviewed journals rather than replacing it. - Repositories for the self-archiving of scientific publications have been created by many research institutions[2]. Different initiatives are attempting to establish reference systems for these documents, and to develop techniques to harvest metadata and methods to mark the relevance of these publications. - CRISs that cross-link the print material to either entities provide new options with regard to utilising the collection of publications for the scientists' purpose (cp. Jeffery & Asserson 2004). [Hornbostel 2006, pp. 31-32]

[5] Group of Eight view of measuring the impact of research [in Australia]: http://theconversation.edu.au/group-of-eight-view-of-measuring-the-impact-of-research-4818 (Last visit: March 20th, 2012) The US is funding STAR METRICS – an initiative to monitor the impact of federal science investments on employment, knowledge generation, and health outcomes: http://www.nih.gov/news/health/jun2010/od-01.htm (Last visit: March 20th, 2012)

[6] CiteSeer – Scientific Literature Digital Library and Search Engine: http://citeseerx.ist.psu.edu/ (Last visit: July 20th, 2012)

[7] Google Scholar: http://scholar.google.com/ (Last visit: July 20th, 2012)

[8] Microsoft Academic Search: http://academic.research.microsoft.com/ (Last visit: July 20th, 2010)

overcome these issues by promoting standardised representation formats – formal ontologies – to support conceptual descriptions enhancing information integration. These are increasingly employed during information system analysis and with design [Gruber 1993], [Guarino 1998], and inherent in recent architectural styles through openness, and where they challenge traditional, rather monolithic container systems.

Conceptual models have finally been recognised and acknowledged as being key with achieving interoperability, information sharing and re-use [Hevner et al. 2004], [Siau 2002], [Wand and Weber 1990], and are thus critical for efficiency, cost-savings, i.e. governance. Conceptual agreement can only be achieved through conceptual understanding, and which is supported by formal foundations [Shanks et al. 2003] and [Herrera et al. 2005] [Guarino & Guizzardi 2006]. However, founded and formally integrative domain descriptions or models are still rare. This work aims at the human understanding of the Research domain in general and Language Technology in particular. It develops an ontologically founded, formal domain description; a formal field-agnostic Research ontology open for field and other extensions: FERON – *F*ield-*E*xtensible *R*esearch *O*ntology. Its extensibility is demonstrated with the field of Language Technology (LT).

1.1 Motivation

The first aim of this work is to support a human understanding of the Research domain in general and Language Technology in particular towards improved system integration and interoperability, and thus information access and information quality. The conviction is such that this is only achievable by a thorough domain analysis driven by foundations and through the inclusion of a reflection of multiple perspectives. This is possible, because the author of this thesis has more than ten years of experience in the field of Research Information Systems and Modeling; and this contribution profits strongly from two major activities in which the author has been very actively involved throughout these years. First, the modeling of an ontology-driven scientific information system in the field of Language Technology (LT) – in 2001 initially setup and since then further developed within the LT Lab at DFKI – namely LT World[9]. Second, the lead of a European task group and involved Board membership for continued development of a generic standard model for Current Research Information

[9] Language Technology World: http://www.lt-world.org/ (Last visit: March 20th, 2012)

Systems (CRISs) – namely CERIF[10] – a EU recommendation to Member States, since 2002 in the responsibility of euroCRIS, applying relational techniques for domain grounding, conceptual modeling or system setup, as well as hierarchical methods for integration and exchange, and increasingly in need of formal semantics with applications and setups. Many of these activities are documented in peer-reviewed publications.

Where LT World's underlying system ontology has been developed pragmatically to provide structured access for the LT community through a single public entry point, and to support the maintenance and continuous updates of a very comprehensive range of LT information through ongoing developments, required extensions and changes, the more generic CRIS model has been politically and historically grounded in European contexts, aimed at a large-scale cross-nation-standardised integration of research information towards interoperability, sharing and re-use – where the decision of federation or centralisation has very often been a topic, and where both kinds are finally under way at national levels within Europe and of increased interest internationally.

This work challenges both activities in that a generic Research domain description is formally ontologised, and the field – that is LT – descriptions meaningfully integrated, while the entire analysis and modeling processes are guided through ontological foundations.

1.2 Contribution

This work develops FERON – the *F*ield-*E*xtensible *R*esearch *O*ntology. FERON is modeled in RDF/OWL and employs frames for better human readability. The entire analysis and the design processes were guided by ontological foundations, and some of them become obvious from the modelled committing constructs. The conceptual FERON domain descriptions were selected or derived from openly available not always formal domain descriptions where possible; hence FERON represents a 'perceived' world. FERON allows for additional sub-domain integration and thus cross-community discussion. The formal OWL description can be employed and extended for machine processing with reasoning or inferencing, not least within the both modelled domains. FERON as a formal domain description is available with this work – it is a valid FERON.owl file. FERON will certainly contribute to continued CERIF / CRIS activities in European and increasingly international contexts towards the interoperation of Research Information systems with increased quality.

[10] CERIF – the Common European Research Information Format – a EU recommendation to Member States in the responsibility of euroCRIS: http://www.eurocris.org/Index.php?page=CERIFreleases&t=1 (Last visit: March 20th, 2012)

1.3 Structure

The work is introduced with a chapter of Basic Notions (2) terms such as Information, Data and Metadata, Entity, Concept and Ontology. Basically, these are intended for guidance of the modeling work and for the human reader. It is recommended to read this chapter in advance because the notions are recalled throughout the thesis. The work starts with a chapter on Conceptual Modeling (3), where the importance of ontological foundations is stressed and well-known approaches are presented, discussed and compared. To support the understanding of strengths and weaknesses of different modeling grammars, the work presents the most significant approaches within the short history of Information Technology-supported system design and discusses common issues. Chapter (4) Information Systems and Architectures presents systems from intersecting domains and architectural styles, to recall the awarenes of incremental steps towards system implementation, integration and sustainability. Chapter (5) Analysis of Research Entities investigates identified substantial domain entities by an in-depth analysis of openly available formats or descriptions. The analysis is continued in chapter (6) for the domain and integration of entities from the field of Language Technology. Chapter (7) presents the resulting *f*ield-*e*xtensible *R*esearch *O*ntology – *FERON*. With chapter (8) the work concludes and provides some thoughts over extensions.

Where possible exact page information is cited, except from the references to entire manifestations of a work. Because most of the cited articles were downloaded from the Web, exact page information is not always preserved in the pdf document. In cases where no exact page numbers are available, the page numbers of the downloaded article itself are given. Cited words are indicated with single quotes, cited sentences or paragraphs with double quotes. Formal construct or concept replications are presented in italics. British English is used for the thesis writing, but spelling differences in citations of American English cannot be avoided.

2 Basic Notions

"every science presupposes some metaphysics" [Mario Bunge 1977]

"Man kann auch nicht verlangen, dass Alles definiert werde, wie man auch vom Chemiker nicht verlangen kann, dass er alle Stoffe zerlege." [Gottlob Frege 1892][11]

"Doch ein Begriff muß bey dem Worte seyn"
(Student to Mephistopheles in Goethe's Faust Part 1, Scene III)

The basic notions section is considered important for an understanding of the difficulties with the task to be solved in this work, and it is recommended to read them in advance. The goal is not to define the selected concepts entirely, but to refer to existing theories and descriptions that are considered valuable and relevant contributions to their understanding. Throughout the work, during analyses, statements and with presentations of results, a conceptual term compliance is attempted in the spirit of Chalmers (1999, pp. 104-105), who claims: "Observation statements must be expressed in the language of some theory [...] the statements and the concepts figuring in them, will be as precise and informative as the theory in whose language they are formed is precise and informative" [cited from Capurro & Hjørland 2003].

The provided notions are considered especially important, because concepts behind terms such as Information, Data, Entity, Concept, or Ontology are by nature not only highly ambiguous, but even more so, when used across communities. The provision of basic notions prevents from a 'conceptual chaos'[12].

[11] In his article *Über Begriff und Gegenstand* Gottlob Frege explains the difference between concept and object in logical terms in the philosophy of language. He clearly states that his description is not meant to be a definition "Eine Definition zur Einführung eines Namens für Logischeinfaches ist nicht möglich." [Frege 1892]

[12] A.M. Schrader (1983) studied about 700 definitions of *information science* and its antecedents from 1900 until 1981 and found: "[T]he literature of information science is characterized by a conceptual chaos. This conceptual chaos issues from a variety of problems in the definitional literature of information science: uncritical citing of previous definitions; conflating of study and practice; obsessive claims to scientific status; a narrow view of technology; disregard for literature without the science or technology label; inappropriate analogies; circular definition; and the multiplicity of vague, contradictory and sometimes bizarre notions of the nature of the term "information" (Schrader, 1983, p. 99)". [Capurro & Hjørland 2003]

2.1 Information

Information in the center of this work is mostly concerned with the semantics or meaning in information systems. There is “not yet consensus on the definition of semantic information” and even the very concept of *information* itself is not expected to be singly and unifiedly defined but only relatively to a well-specified context of application [Floridi 2005, pp.1-2][13]. Therefore, the multiple discussions about the *information* term[14] and its meaning as such will not be deeply discussed in this work, but instead there will be explanations and definitions presented that support the understanding of *information*, and that allow for a formal and meaningful description of the two domains of interest – Research in general and Language Technology in particular – towards a meaningful *information*-integration and thus inter-system-operation.

The Stanford Encyclopedia of Philosophy formalises (semantic) *information*:

σ is an instance of information, understood as semantic content, if and only if:

- (GDI.1) σ consists of one or more *data*;
- (GDI.2) the data in σ are *well-formed*;
- (GDI.3) the well-formed data in σ are *meaningful*;

Figure 2: The General Definition of Information (GDI) [15]

[13] Claude Edwood Shannon (1916-2001): An American mathematician known as the father of the Information Theory published his paper: “A Mathematical Theory Of Communication.” in 1949. His mathematical theory of communication assumes that “semantic aspects of communication are irrelevant to the engineering problem”. [Shannon 1948] is therefore irrelevant in our work. Claude Shannon furthermore remarked: The word “information” has been given different meanings by various writers in the general field of information theory. It is likely that at least a number of these will prove sufficiently useful in certain applications to deserve further study and permanent recognition. *It is hardly to be expected that a single concept of information would satisfactorily account for the numerous possible applications of this general field.* (from “The Lattice Theory of Information”, in Shannon [1993] pp. 180-183, first sentence, italics added)” [Floridi 2005, p. 2]

[14] There have been various attemps towards a unified definition of information. For reference, see (Braman 1989; Losee 1997; Machlup 1983; NATO 1974, 1975, 1983; Schrader 1984; Wellisch 1972; Wersig and Neveling 1975) [Floridi 2005, p. 361]. Further pointers are [Capurro & Hjørland 2003], [Spree 2002], [Uszkoreit 1999, chapter: “Das Wesen der Information”] and additional statements from the German speaking community defining the information term as e.g.: “Information ist geglückter Transfer von Wissen” (Harald H. Zimmermann); “Information ist Wissen in Aktion” (R. Kuhlen); “Information ist die Verringerung von Ungewißheit” (G. Wersig). A reference to definitions and discussions is provided to indicate the semantic ambiguitites inherent in the information term, for reflection over and awareness of the current context – namely: information in information systems.

[15] Stanford Encyclopedia of Philosophy: http://plato.stanford.edu/entries/information-semantic/ (Last visit: May 20th, 2012)

In this work – in addition, the understanding of *information* reflects the property of being informative[16] and thus contextual-evident of the recorded data within a system upon request or view, but without initial communicative intent. Accordingly, the aim is to achieve a maximum in re-use and sharing of *information* by multiple stakeholders. As a consequence this thesis supports the GDI; it is considered inline with Buckland’s concept of “information as thing”[17] – a tangible entity; because “ultimately information systems, including “expert systems” and information retrieval systems, can deal directly with information ***only*** in this sense”, and which Buckland distinguishes from “information as knowledge” or “information as process”, in that both are mental activities and therefore intangible in information systems, as indicated in the matrix of Figure 3 [Buckland 1991, p. 352].

	INTANGIBLE	**TANGIBLE**
ENTITY	2. Information-as-knowledge Knowledge	3. Information-as-thing Data, document
PROCESS	1. Information-as-process Becoming informed	4. Information processing Data processing

Figure 3: Four Aspects of information [Buckland 1991, p. 352]

Recorded data in information systems are representations of things; tangible and processable; initially unintended but informative or meaningful. Furthermore, *information-as-thing* as well as *information-as-process* is finally situational; “whether any particular object, document, data or event is going to be informative depends on the circumstances, just as the ‘relevance’ of a document or a fact is ***situational*** depending on the inquiry and on the expertise of the inquirer (Wilson 1973).” Situational is equal to explicitly contextual and consequently, *information-as-thing* is enabled through meaningful data, in support of a more advanced information processing, inferencing and views. [Floridi 2005, pp. 252-253] explains, that

[16] “Natural Sign” is the long-established technical term in philosophy and semiotics for things that are informative but without communicative intent (Clarke, 1987; Eco, 1976). [Buckland 1991]

[17] “The term “information” is also used attributively for objects, such as data and documents, that are referred to as “information” because they are regarded as being informative, as “having the quality of imparting knowledge or commmunicating information; instructive. (Oxford English Dictionary, 1989, vol.7, p 946).” [Buckland 1991, p. 351] Wiener asserted that “Information is information, not material nor energy.” Norbert Wiener (1894-1964), American mathematician known as the originator of Cybernetics published a paper [Wiener 1948]: “Cybernetics: Or Control and Communication in the Animal and the Machine.” From an information systems perspective, information requires material for its encoding and subsequent communication and processing, and this work supports Rolf Landauer’s view, who claims, that „information is physical”, advocating the impossibility of physically disembodied information, through the equation “representation = physical implementation” [Floridi 2005, p. 355].

over the last three decades Buckland's information concept of *objective semantic information* in terms of *data* + *meaning*, where "the various, mathematical, syntactical or pragmatical senses in which one may speak of information are not strictly relevant and can be disregarded, has gained sufficient consensus to become an operational standard [[18]] in fields such as Information Science; Information Systems Theory, Methodology, Analysis and Design; Information (Systems) Management; Database Design; and Decision Theory, since these deal with data and information as reified entities." The GDI concept is increasingly applied in the Sciences; a recent article in Nature underscores the thesis, that "there is no meaningful information without data and conversely, data cannot be generated or valued without prior knowledge" [Mons et al. 2011, p. 1]. Such a view is also reflected in ongoing research documentation activities, and e.g. in a public EC-initiated consultation[19], where "scientific information" refers to both, (1) scientific (and scholarly, adademic) publications published in peer-reviewed journals and (2) research data. Increasingly, research funding organisations require the deposit of data underlying the reported and published research results.

2.2 Data and Metadata

I*nformation* and *data* are interrelated. The Diaphoric Definition of Data (DDD) explains: "A datum is a putative fact regarding some difference or lack of uniformity within some context."[20] [Buckland 1991, p. 353] defines: ""Data," as the plural form of the Latin word "datum," means "things that have been given." It is, therefore, an apt term for the sort of information-as-thing that has been processed in some way for use." [Borgman 2011, p. 5] holds, that *data* "may exist only in the eye of the beholder […] but not perceived as such by the recipients"; and are at best, "alleged evidence," and considers data, a difficult concept to define; even more in the context of data sharing, where the term *data* is "meant to be broadly inclusive" because it refers to forms that require computational machinery and software, or with 'datasets' that are often related to grouping, content, relatedness or purpose. The Oxford English Dictionary defines data as "things known or assumed as facts; facts collected together for reference or information, reasoning or calculation." [Mons et. al. 2011, p. 281]

[18] [Floridi 2005] here refers to the General Definition of Information (GDI).

[19] European Commission Consultation on scientific information in the digital age "On-line survey on scientific information in the digital age": http://ec.europa.eu/research/consultations/scientific_information/consultation_en.htm (Last visit: November 30th, 2011)

[20] Semantic Conceptions of Information (Stanford Encyclopedia of Philosophy): http://plato.stanford.edu/entries/information-semantic/ (Last visit: May 20th, 2012)

introduce the chicken and egg metaphor to explain the paradox: “If we assume data to be the eggs, which need brooding (curation) to become chickens (articles), and we require the mating of complementary units of information to generate yet more fertile eggs, we have a reasonable frame or reference.” [CCSDS Blue Book 2002, p. 8] define *data* as “[a] reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing”, which is very much inline with the GDI and Buckland’s definition, if data are well formed, and if suitable is understood as meaningful or informative.

In the vicinity of *data* there is always *metadata*, often simply described as “data about data”. A popular metadata initiative – the Dublin Core Metadata Intitiative (DCMI) states on its website: “Metadata articulates a context for objects of interest -- “resources” such as MP3 files, library books, or satellite images – in the form of “resource descriptions””. As a tradition, resource description dates back to the earliest archives and library catalogs. The modern “metadata” field that gives rise to Dublin Core and other recent standards emerged with the Web revolution of the mid-1990s.”[21] In this context, the arguments of [Haase 2004, pp. 205-206] are very relevant; he identified the growing importance of quality metadata, and where quality is with precision and semantic grounding. His ‘Metadata Twist’ forecasts that economic significance of *metadata* will increase while with technological progress the average value of content will decrease. [Haase 2004, p. 204] defines *metadata* as “any data which conveys knowledge about an item without requiring examination of the item itself”; *metadata* is therefore different from *data* in that it carries knowledge or context – *purpose* – and it may even be considered *information* in the previously presented sense if the linkage to *data* and meaning is preserved, and if it is well-formed. Contextual views over *metadata* are increasingly relevant because “[m]etadata is not simply a description of the information contained in a work or web page; the choice of a metadata scheme also significies community membership” [Marshal & Shipman 2003, p. 62]. It is therefore important that *metadata* schemes allow for contextual representations and, in order to being informative preserve linkages to their underlying or originating datasets[22] [Asserson and Jeffery 2003, pp. 31–33], [Gartner 2008, p. 15], [Ducloy et al. 2010, p. 1], [PMSEIC 2006, p. 29].

[21] Dublin Core Metadata Inititiative: http://dublincore.org/metadata-basics/ (Last visit: March 25th, 2011)

[22] „The next era in research articles will take content beyond what is provided by the author, linking to relevant data and other information from external sources to provide even greater added value to researchers. Different disciplines might focus on different content types, such as telescopic data for astronomers or molecular images for biologists, but across the scientific community authors are increasingly adding supporting content that can bring further depth and context to an article. But, for this to happen, the right dots must be connected – giving researchers the content they need and helping them to find the proper context for the content.“ http://www.researchinformation.info/features/feature.php?feature_id=274 (Last visit: March 25th, 2011)

2.3 Entity

From reading Quine (1948) it is understood that “to be is to be the value of a variable” and what in information systems is situational or constructed towards a view or at a time. Chen, who developed the Entity Relationship Model for a unified view of data, calls an *entity* “a ‘thing’ which can be distinctly identified”, and presents as examples, person, company, or event [Chen 1976, p. 10]. Furthermore, he calls a relationship “an association among entities” providing as example a “father-son” linkage between two person entities, noting that “some people may view something (e.g. marriage) as an entity while other people may view it as a relationship”, and in his view, the decision to be taken is with the enterprise administrator. Understanding *entities* as things, implies their formal description through properties by which they are represented and identified within a system. Their representation requires decisions about composition and granularity, and thus necessity and contingency; and refers to notions of intension or extension with *concepts* through time. [Welty & Fikes 2006, p. 230] e.g., introduce a dimension, where “[t]he problem of diachronic identity becomes trivial since entities are four dimensional, and the notion of change is accounted for simply by giving different properties to different temporal parts of an entity so that Leibniz’s law always holds. This approach has the problem however, that determining what is an entity is rather arbitrary; in fact any mere collection of matter over time can be an entity” [Welty & Fikes 2006, p. 231]. The *entity* and thus identity issue is of particular relevance with representation, integration and exchange of *information* between systems, and obviously crucial in the semantic web or a world of linked data, where different *information* providers contribute *data*, that assumingly represent the *same* entities [Bizer et. al. 2009, p. 7] while supplying their provenance or legacy semantics. “As a practical matter some consensus is needed to agree on what to collect and store in retrieval-based information systems, in archives, data bases, libraries, museums, and office files [...] In the provision of access to information by means of formal information systems, the question of whether or not two pieces of information are the same (or, at least, equivalent) is important.” [Buckland 1991, p. 357]

2.4 Concept

In the best case, information systems are built following conceptual models, and where the relevant conceptual entities are represented by linguistic expressions. When dealing with the meaning of linguistic expressions, two different views seem relevant for understanding “what it means”. [Carnap 1947] proposed a new approach which he had called the method of extension and intension[23]: “The meaning of any expression is analyzed into two meaning components, the intension, which is apprehended by the understanding of the expression, and the extension, which is determined by empirical investigation.” His approach layed the foundation of a modal logic, that is, a theory for concepts like necessity and contingency, possibility and impossibility “which philosophers and logicians will find valuable in solving many puzzling problems” [Carnap 1947, *introduction*]. [Brachman 1976, p. 138] distinguishes equally between “an *extension*, or the members of a particular world designated by the expression, and an *intension*, an abstraction of the properties of those individuals which acts in such a way as to select from any possible world the set of individuals that are described by the language expression”. While examining *concepts* from the perspective of semantic-net authors based on [Quillian 1969], who hoped to represent with so-called semantic networks anything expressible in natural language, [Brachman 1976] found that “[a]uthors have invariably relied on readers’ intuitions about what concepts are, without discussing their implemented structure in any detail[24] [...], while it is well understood what a class is, it is never clear what a “concept” is” [Brachman 1976, p. 130]. He proposed first an extensional approach towards defining *concepts*: “At best, we can infer from the way concepts seem to be implemented in existing nets that they can be defined as groups of features or properties, or occasionally as predicates”, but later admits that it is the intensional side of concept nodes that would allow a program to determine what relationships were entailed by the assertion of a particular relation: “The formal notion of intension is precisely what is needed to firm up our representations enough to perform this kind of task” [Brachman 1976, p. 138-139]. This work follows Brachman and agrees, that without an investigation of extensional occurences, an intensional *concept* description is difficult, but, in the end, intensional descriptions are most generic and thus appropriate for representing multiple extensions. Furthermore, Brachman’s view is considered inline with the conclusion provided by [Di Nitto & Rosenblum 1999, p. 9]; where they investigated architectural description languages in support of networked system design understanding intension as top-down approach while extension refers to bottom-up approaches: “The lesson we learned from our experience is that the top-down approach adopted by the software architecture community in

[23] “After giving a detailed critical discussion of the traditional method, according to which any expression of language (a word, a phrase, or a sentence) is regarded as a name of one unique entity (a thing, a property, a class, a relation, a proposition, a fact, etc.), Mr. Carnap concludes that the various forms of this method of the name-relation lead to numerous difficulties and complications” [Carnap 1947].

[24] “Recall that Quilian expected to be able to represent uniformly *anything* that could be expressed in natural language. This has led to the assumption that virtually anything can be defined in terms of nodes and links; there consequently have existed networks that have purported to represent “facts”, “meanings of sentences”, “propositions”, “actions”, “events”, “properties”, “assertions”, “predicates”, “objects”, “classes”, “sets”, “relations”, among other things.” [Brachman 1976, p. 130]

the development of languages and tools seems in many ways to ignore the results that practitioners have achieved (in a bottom up way) in the definition of middlewares."

Conceptualisation is related to ontology. [Gruber 1993, p. 1] regards conceptualisation as "an abstract, simplified view of the world that we wish to represent for some purpose". [Guarino & Giaretta 1995, p. 907-908] analysed Gruber's ontology definition "an explicit specification of a conceptualization", and discovered substantial differences in their understanding of 'conceptualisation'; they assign the properties of formal ontology to "not so much the bare existence of certain objects, but rather the rigorous description of their forms of being, i.e. their structural features". Where Gruber refers to "a set of extensional relations describing a particular state of affairs" (Genesereth & Nilsson 1987), Guarino and Giaretta approach intensionally and propose a revised definition. Within their glossary they informally define 'conceptualization' as "an intensional semantic structure which encodes the implicit rules constraining the structure of a piece of reality". This approach follows the understanding of Brachman in that conceptual modeling requires both, the extensions and the intensions. Furthermore it agrees with [Lindland et al. 1994, p. 1] in that conceptual modeling is closely linked to linguistic expressions due to "statements in some language", while at the same time it supports [Buitelaar 2009, p. 1] claiming that ontologies, and thus, their *concepts* as such are "logical theories and independent of natural language", but represented through linguistic expressions. For reference in this very context and to further demonstrate the difficulty with appropriate linguistic expressions for *concepts* – which are of course necessary – a reference to [Kripke 1980, p. 1-2] is provided, he discusses the theory of names, requiring notions of "identity across possible worlds", which he calls "rigid designator" if in every possible world it designates the same object, "nonrigid or accidental designator if that is not the case", and who maintains, that "*names* are rigid designators". The conclusion is therefore, that *concepts* also refer to identity and thus entity, and also become more meaningful with data.

This work will not further investigate Kripkes thesis, nor the notion of identity for which it refers to the above notion of *entity*, and leaves out the notion of language in the sense of natural language in a linguistic or grammatical sense, which is not a focus of this work. This work is concerned with ontological (*conceptual*) foundations improving the (semantic) clarity of the Research domain in general and Language Technology in particular – towards information system setup, for advanced information processing, information integration and information exchange – through provision of formal conceptual domain descriptions. Therefore, it distinguishes the language in linguistic expressions of *concepts* representing the entities in information systems from the 'natural language' as applied or recorded through speech or in written texts. In this sense, it deals with so-called *structured information* in information systems rather than *unstructured information* as inherent in natural language.

2.5 Ontology

In his paper *Toward Principles for the Design of Ontologies Used for Knowledge Sharing*, [Gruber 1993] approaches the area of developing formal ontologies from an engineering perspective, to support knowledge sharing activities. He discusses design decisions and representation choices and their validation against some design criteria. His understanding of *ontology* is "an explicit specification of a conceptualization" (p. 908), with the term borrowed from philosphy, meaning "a systematic account of Existence" and applied to Artificial Intelligence (AI) where what 'exists' is that which can be represented. For Gruber: "Formally, an ontology is the statement of a logical theory.[25]" A comprehensive analysis of ontology from a debate in AI has been presented by [Guarino & Giaretta 1995] – they distinguish (the below bullet points are cited) between ontology as:

(1) a philosophical discipline

(2) an informal conceptual system

(3) a formal semantic account

(4) a specification of a 'conceptualization'

(5) a representation of a conceptual system via logic theory

  (5.1.) characterized by specific formal properties

  (5.2.) characterized by its specific purpose

(6) the vocabulary used by a logical theory

(7) ontology as a meta-level specification of a logical theory

[Guarino & Giaretta 1995, pp. 1–2] themselves interprete (1)[26] radically different from (2-7); consider (2+3) a conceptual semantic entity; either informal (unspecified) or formal (expressed in terms of suitable formal structures at the semantic level); while (5-7) appear as specific syntactic views; and (4) – it refers to [Gruber 1993] – may collapse into (5.1) when

[25] "Ontologies are often equated with taxonomic hierarchies of classes, but class definitions, and the subsumption relation need not be limited to these forms. Ontologies are also not limited to *conservative definitions*, that is, definitions in the traditional logic sense that only introduce terminology and do not add any knowledge about the world (Enderton 1972). To specify a conceptualization one needs to state axioms that *do* constrain the possible interpretations for the defined terms." [Gruber 1993, p. 909]

[26] The philosophical discipline 'Ontology' is usually identified by a capitalized term, namely the branch of philosophy, which deals with the nature and the organisation of reality. Ontology as such is usually contrasted with Epistemology, which deals with the nature and sources of our knowledge (Cocchiarella 1991) [Guarino & Giaretta 1995 p. 908].

intended as a vocabulary; as will (6); finally (7) is interpreted as specifying the "architectural components" used within a particular domain theory. They analysed the *ontology* definition by Gruber (1993) and discovered substantial differences in their understanding of *ontology*: "[W]e cannot see a particular theory as a specification of a conceptualization, since conceptualizations can be only partially characterized. What we can specify is a set of conceptualizations, i.e. an ontological commitment." Furthermore, they refer to the definition of Nino Cocciarella, which they consider particularly pregnant and argue, that in practice *formal ontology "*can be intended as the theory of the distinctions, which can be applied independently of the state of the world [...]", and they distinguish between various kinds of "symbol-level artifacts" and their "conceptual (or semantical) counterparts", to suggest conceptualisation denotes "a semantic structure which reflects a particular conceptual system", and ontological theory denotes "a logical theory intended to express ontological knowledge (interpretation)" [Guarino & Giaretta 1995, pp. 3-5] that can be read, sold or physically shared. Additionally, they note, that the linguistic terms used to denote relevant relations cannot be thought of as mere comments or informal extra-information, but suggest that formal structures used for conceptualisation should somehow account for their meaning, which "cannot coincide with an extensional relation". In their glossary (p. 6), they define ontology in one sense as a "synonym of *conceptualization*" (see above), in another sense as "a logical theory which gives an explicit, partial account of a *conceptualization*". For [Sowa 1999, preface], *ontology* is a theory or technique applied for knowledge representation, besides logic and computation; "*ontology* defines the kinds of things that exist in the application domain."

[Buitelaar et al. 2009, p. 1] consider *ontologies* as "logical theories and independent of natural language", but argue that a grounding in natural language is needed as well as an association of rich linguistic information, to support ontology engineering, population and verbalisation. [Jarrar & Meersmann 2002, p. 1238] present "a database-inspired approach towards engineering of formal ontologies, implemented as shared resources to express agreed formal semantics for a real world domain" addressing issues like knowledge reusability, shareability, scalability of the engineering process and methodology, efficient and effective ontology storage and management, and coexistence of heterogeneous rule systems that surround an ontology, mediating between it and application agents – which they call DOGMA. They argue, that "correct understanding of ontologies must reconcile that they are repositories of (in principle) language- and task-independent knowledge, while an effective use by e.g. software agents naturally requires interaction with *some necessarily lexical* representation."

The fact that many authors explain or define the term *ontology* or investigate and refer to available definitions of *ontology*, indicates the difficulties and ambiguities inherent in its concept. However, it seems clear, that such differences are related to technologies and backgrounds and to the tasks being achieved. [Guarino 1998, p. 3] aimed to overcome confusions with *ontology*, *ontological commitment*, and *conceptualisation*; [Guarino & Giaretta 1995] approach *ontology* intensionally while [Smith 2003] explains *ontology* philosophically and [Jarrar & Meersmann 2002] take a bottom-up approach; [Noy 2004] provides a survey of ontology-based methods from the viewpoint of semantic-integration mostly towards automated reasoning, [Gruber 1993a, p. 202] understands, that *ontologies* are "also like conceptual schemata in database systems". [Welty 2003] investigates the meaning of *ontology* across disciplines in his editorial of the AI Magazine. The latest W3C overview document of the Web Ontology Language (OWL 2) gives the following definition: "Ontologies are formalized vocabularies of terms, often covering a specific domain and shared by a community of users. They specify the definitions of terms by describing their relationships with other terms in the ontology." [W3C 2009]

With the growing popularity of the Semantic Web, the production of ontologies proliferated substantially, as well the number of kinds differing in abstractness, completeness, granularity, formality, structure, thematic range, form, syntax and semantics. It has been demonstrated that the term *ontology* has been used extensively and its meaning has been increasingly broadened. [Guarino 1998, p. 3] provides a comprehensive list of research and application fields, where the importance of ontologies has been recognised: knowledge engineering, knowledge representation, qualitative modelling, language engineering, database design, information modelling, information integration, object-oriented analysis, information retrieval and extraction, management and organisation, agent-based system design. In addition to these research fields, he recognised application areas like enterprise integration, natural language translation, medicine, mechanical engineering, standardisation of product knowledge, electronic commerce, geographic information systems, legal information systems, and biological information systems. For a reference to the identified research fields and application areas he uses the generic term *information systems*, in its broadest sense.

Disparate backgrounds, languages, tools, and techniques have been identified as the major barriers to effectively communicate among people, organisations, and/or software systems [Uschold & Gruninger 1996, *abstract*]. The basic notions support the understanding of the employed concepts and are considered valuable for the current discourse towards overcoming the barrier of understanding the Research domain in general, and LT in particular.

3 Conceptual Modeling

"A little Semantics goes a long way"
(James Hendler 1997)[27]

The need for improved understanding and formal descriptions over domains is increasing, especially with new technologies such as the Semantic Web and its extension the Linked Open Web, but also with more open data and access initiatives, and thus sharing or re-use, i.e. information system integration at the large scale. [Chen et al. 1999, p. 287] proposed to overcome integration problems through conceptual models. "Conceptual Modeling is the activity of formally describing some aspects of the physical and social world around us for purpose of understanding and communication" [Mylopoulos 1992]. While the importance of conceptual modeling is finally being accepted as "an important aspect of systems analysis" [Wand et al. 1995, p. 285], the field has long been criticised for its lack of foundations with modeling methods [Hevner et al. 2004, p. 2], [Siau 2002, p. 106], [Wand & Weber 1990, p 1282]. Although the need for foundations had been recognised in the late 50s [Wyssusek 2006, p. 63] and advocated[28] with emerging modelling grammars, conceptual modeling had been approached mostly pragmatically: "technical questions are predominate, and fundamental considerations and reflections beyond current trends are missing" [Schütte & Rotthowe 1998, p. 242]. It is only recently that theoretical foundations in conceptual modeling and during system analysis and design have started, and the most prominent approaches refer to ontologies. Contrary to Wyssusek[29] there is a conviction [Shanks et al. 2003, p. 85], [Herrera et al. 2005, p. 572-573] that "[t]heories of ontology lead to improved conceptual models and help to ensure they are indeed faithful representations of their focal domains", and that "the ontology-driven approach to conceptual modeling is well and alive, and that it dramatically improves the quality of information systems" [Guarino & Guizzardi 2006, p. 1].

[27] The phrase has become one of the slogans of the Semantic Web movement. Its brief history is revealed on James A. Hendler's website: http://www.cs.rpi.edu/~hendler/LittleSemanticsWeb.html (Last visit: April 2nd, 2012)

[28] [Woods 1975, p. 36-37] aimed at triggering foundational discussions. [Brachman 1976, p. 128] attributed the failure of semantic networks to missing foundations. [Chen 1976, p. 10] introduced "Multilevel Views on Data", [Thomas 2006, p. 7] refers to foundational reference models in business process modeling.

[29] Wyssusek concluded, that "the project of developing theoretical foundations of conceptual modelling on the basis of philosophical ontology is neither feasable nor defensible. Yet this conclusion does not mean that Wand and Weber's work has been erroneous. Rather the project of ontology-based conceptual modelling appears to be impossible in principle" [Wyssusek 2006, p. 74]. Wyssusek based his arguments on the analysis of the BWW approach.

This work aims at a meaningful formal representation of two domains – namely Research in general and LT in particular towards their integration in information systems. The ontological foundations are intended for guiding the entire analysis and design processes with FERON, a discipline-agnostic field-extensible Research ontology (chapter 7). Next, relevant ontological foundations are presented; that is, mostly the BWW ontology, but insight is also provided to the Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE), the Suggested Upper Merged Ontology (SUMO), and in very brief, the Cyc knowledge base. Subsequent, popular modelling grammars are introduced to reveal their inherent structural differences before highlighting common modeling issues and their varied methods.

3.1 Ontological Foundations

Foundations or *foundational ontologies* in particular support variously[30] during information system analysis and design, and have thus been successfully incorporated with tools and with evaluations of modeling languages and frameworks. The most prominent approach with ontological foundations[31] in conceptual modelling refers to the BWW ontology (Wand & Weber 1988), and it will be introduced first. Furthermore, top-level ontologies such as DOLCE and SUMO are investigated and the Cyc knowledge base briefly introduced.

3.1.1 Bunge-Wand-Weber Ontology (BWW)

Wand and Weber (1988) aimed at theoretical foundations of processes in systems analysis and divided the study of information systems into three dimensions. The first is strongly related to behavioral sciences, i.e. how systems are used and deployed, the second is rather referred to engineering and computer science, i.e. software and communication, and the third is concerned with foundations: "[t]he process by which systems are analyzed, designed and

[30] [Gemino & Wand 2005, p. 302] suggest employment of conceptual models in early design phases to ensure a sound model balance. [Mylopoulos 1992, p. 2] recognises their support with communication between stakeholders, and (Rolland & Cauvet 1992) see their usefulness with mediation between users' requirements and system design. [Wand et. al. 1995, p. 285] suggested a usage of ontology, concept theory, and speech act theory for developing enterprise systems. [Gruber 1993] and a little later [Guarino 1998, p. 3] recognised their importance with knowledge engineering, database design and integration, information retrieval and extraction.

[31] This work does not investigate *model theory,* which refers to Mathematics and Logics or Computer Science. "Model theory is a systematic method for evaluating the truth of a statement in terms of a model." [Sowa 2007, p. 86] "In accordance with TARSKIS semantic model theory many researchers use the logical model term. This definition talks about a model, if the interpretation of a mathematical structure were true for all axioms and derivation rules of the structure (see also Bung67, ElNa94). This understanding of the model does not assume a relation to the reality but analyzes the interrelation of structures." [Schütte & Rotthowe 1998, p. 243]

constructed" [Wand & Weber 1990a, p. 123-124][32]. Accordingly, their first foundational steps require the perception of "an information system as an abstract concept (as opposed to a physical artifact or the way it is used) [while anticipating it is built] to provide information that otherwise would have required the effort of observing or predicting some reality" (p. 124). Information systems are therefore either "a representation of a real world system" or "a representation of some perceived reality" (p.124). In [Wand & Weber 1990, p. 1282], they distinguish and define the two kinds as follows:

(1) Information systems are themselves models of the real world and ontology identifies the basic things in the real world that information systems ought to be able to model.

(2) Information systems are also things in the real world and ontology provides a basis for modeling information systems themselves.

The distinction is relevant for this work which aims at an information system of the first kind: through ontological analysis design a world of Research in general and within it a world of LT in particular – perceived through, what Wand and Weber call a "collection of interacting things" (p. 126) – with an awareness that information as a *thing* becomes tangible only in information systems [Buckland 1991, p. 352, (see Figure 3)]. The identification of basic things reduces complexity and thus supports with decouplings of the analysis and design processes "to construct a better artefact" (p. 126); hence acknowledging ontology as "the foundations on which to build a theory of good decomposition of systems" (p. 126) to capture its main aspects with its descriptions "the system's structure (statics) and its behaviour (dynamics)" (p. 125) for either of the above two kinds.

Wand and Weber were motivated by an understanding that the formalisation of system concepts could be based on a theory of reality representation. They employed Mario Bunge's ontology as their "main source of constructs" (p. 124) and extended it to the then so-called Bunge-Wand-Weber (BWW) ontology, which is since been widely used.[33] They present their ontology adoption as follows: "Formalization of structure begins by adapting the fundamental concept of a substantial individual. A substantial individual can be *composite* – that is, composed of other individuals, which comprise its *composition* – or it can be *simple*. All substantial individuals possess *properties*. Indeed, properties represent our knowledge about

[32] In their definitions they understand ‚analysis' to deal with modelling reality, whereas the design process aims at constructing a model of implemented representation. [Wand & Weber 1990, p. 125]

[33] [Evermann 2009, p. 1] claims that the BWW's "specification in natural language is the key inhibitor to its wider use". [Rosemann et al. 2004, p. 119 ff.] report of about 25 papers that applied the BWW ontology for analysis of use with modeling grammars, suggesting a "more rigorous process" in applications.

substantial individuals. For composite individuals, properties can be *hereditary* or *emergent*. [...] Since we always observe substantial individuals with their properties, the concept of a *thing* is more useful. A thing X is a substantial individual x with its properties: X = <x, p(x)>.

Things are perceived (in sciences, ontology, or in other symbolic representations) as concepts or models. This approach is formalized via the notion of a conceptual schema or a *model thing*. A model thing is defined by a *frame of reference*,[34] M, and a set of functions that represent its properties. Together they comprise a *functional schema*. A functional schema of a thing, X, is a certain non-empty set, M, together with a finite sequence, F, of functions on M: X_m = <M, F>. It is an ontological postulate that every thing can be modelled as a functional schema." [Wand & Weber 1990a, p. 126-130] [Evermann & Wand 2001, p. 356-357] introduced the basic concepts of the BWW-ontology as presented in Table 1.

Table 1: Basic concepts of the BWW-ontology [Evermann & Wand 2001, pp. 356–357].

BWW Concepts	**Explanations** *(all text is cited)*
Substantial Things	The world is made up of *substantial things* that possess *properties*. Things change by acquiring or losing properties. Things are not destroyed or created. Rather, they come into being (or disappear) through acquisition or loss of properties, or via composition or decomposition.
Property	A property can either be *intrinsic* – possessed by the thing itself (e.g. color), or *mutual* – possessed jointly by two or more things (e.g. distance).
Composite Thing	Things can combine to form a *composite thing*. There exist basic things that cannot be decomposed. Composite things posses *emergent properties* that are not possessed by any component. For example, a computer possesses processing power, not possessed by any individual component.
Law	A law is a relationship between properties. In particular, a law can be specified in terms of *precedence of properties*: Property A *preceedes* property B iff whenever a thing possesses B, it possesses A.
Class	A BWW-*class* is the set of things that have one common property, a kind is the set of things that have two or more common properties and a natural kind is a kind where some of the properties are related by laws. Examples are respectively the set of red things, the set of red and heavy things and the set of things that are red and heavy whose color and weight are related by a law. It is important to note that in our ontology, classes, kinds and natural kinds are defined over an existing set of things. In this sense, the things are the *primary* construct, not the class or natural kind. It follows that there can be no classes without members.
Attributes (State Functions)	*Attributes* are representations of the properties of a thing as perceived by an observer. They can be thought of as functions of time (and other conditions of observation) e.g. specifying the color of thing x at time t. Such functions are called *state functions*.

[34] "A frame of reference here is more general than in physics, as it includes a point of view, conditions of observation, and time (if applicable)." [Wand & Weber 1990a, p. 126-130]

Functional Schema	A set of attributes used to describe a set of things with common properties is called a *functional schema*. Depending on which aspects one is interested in, there can be different schemas describing the same thing. The *state* of a thing is a *complete* assignment of values to all state functions in the functional schema.
Event	A change of a state is termed an *event*.
Lawful	A thing is always in a *lawful* state, one that is allowed by the laws by which it abides. A state may be *stable* or *unstable*. If a thing is in an unstable state, it will spontaneously undergo a transition to another state until it reaches a stable state.
Interaction	Two things are said to interact if the presence of one of them affects the states the other traverses. Interactions are manifested by mutual properties. For example, if one thing hits another, this will change the combined speed of the pair.

[Evermann 2009] developed an UML and OWL formalisation of the BWW upper level ontology to trigger its usage in the Semantic Web community, and reported about a ERM version of Bunge's ontology by (Rosemann & Green 2002). During the research activities for this work, Everman's formal models as well as the Rosemann and Green model, were not easily retrievable and accessible on the Web and will therefore not be further investigated. Nevertheless, are they of upmost importance by being explicitly expressed, and thus available for the wider communication, mutual understanding, distribution, awareness, disambiguation, and for further clarification. With this work, the Bunge constructs as inherent in modeling grammars and tools were applied. BWW foundations are reflected in W3C's Web Ontology Language (OWL) specification (introduced in section 3.2.6 Formal Ontology) and enabled through e.g. the openly available Protégé[35] ontology modeling tool. To achieve the goal of this work, almost the entire set of the above named basic BWW concepts was applied with FERON, except *Event*, which is considered as belonging to a subsequent level concerned with process modeling and therefore concrete application setup. An *event* in the described sense implies timely state changes and requires rule definitions. Information systems of the mentioned second kind employ such rules. In the current context they are understood as to being situational within the world of the information system and as such depend on legacy, user needs and organisational workflows, i.e. system requirements; and are therefore not relevant for this work.

The goal of this work is to identify the basic things of a perceived Research world and a perceived LT world in particular – agnostic of situational change triggering events or views, i.e. application rules (although potential *lawful functions* or *functional schemata* to represent

[35] Protégé has been developed and is conituously maintained and extended at Stanford; it is the most popular tool used in the scientific community for ontology modeling: http://protege.stanford.edu/ (Last visit: April 2nd, 2012)

these states in information systems through time are enabled in FERON). Rules depend on laws, *lawfully* applied upon things through *events* ensured by modeled potential *state functions* with modeled potential *functional schemata*. *Interaction* as well is related to state changes upon lawful rules triggered by events applied upon things, ensured and thus represented through modeled potential *state functions* in *functional schemata*; enabled or ensured through FERON, but not explicitly defined upon *things*. FERON may thus be considered as a lawful graph – in its entirety perceived as a frame – to describe a Research world in general and an LT world in particular. The concepts *Lawful* and *Interaction* are discussed with respect to applications and with examples, but not explicitly modelled.

3.1.2 Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE)

DOLCE[36] is one module of the WonderWeb Foundational Ontologies Library, which was dedicated to serve as a starting point for building new ontologies. The library's mission is very much inline with this work in that "one of the most important and critical questions when starting a new ontology is determining what things there are in the domain to be modeled" [Masolo et al. 2001, p. 2]. DOLCE has never been intended as a "'universal' standard ontology, but rather as a *reference module*" – but, being a foundational ontology, it has been "ultimately devoted to facilitate mutual understanding and inter-operability among people and machines" (p. 3). DOLCE applies basic philosophical categories with ontology design and ontological choices. Its domain of discourse is the *particular*[37] distinguishing two kinds: *abstract* and *spatio-temporal-particular*; the first "do not have spatial nor temporal qualities, and they are not qualities themselves"; the latter are considered a "[d]ummy class for optimizing some property universes". The DOLCE-Lite.owl[38] (2005) ontology models *abstract* and *spatio-temporal-particular* as subclasses of particular (Figure 5), i.e. formally the both are kinds of a particular. A *particular* is assumed to not have instances, whereas *universals* do. DOLCE [Masolo et al. 2001] descriptions talk about *universals*, however, the

[36] DOLCE was developed under the IST theme between 01/02 and 06/06 in the EU-funded project WonderWeb (http://wonderweb.semanticweb.org/) with the objectives to develop a framework of techniques and methodologies for semantic integration, migration, reconciliation and sharing of ontologies towards building the Semantic Web. DOLCE was intended as "a *library* of foundational ontologies, systematically related to each other in a way that makes the rationales and alternatives underlying different ontological choices as explicit as possible." [Masolo et al. 2003, p. 2] This work refers to DOLCE version 397, classified; The DOLCE and DnS ontologies. OWL engineering by Aldo Gangemi. Ontology, Online: http://www.loa.cnr.it/ontologies/DOLCE-Lite.owl (Last visit: April 2nd, 2012)

[37] "The extensional coverage of DOLCE is as large as possible, sinces it ranges on 'possibilia', i.e. all possible individuals that can be postulated by means of DOLCE axioms. Possibilia include physical objects, substances, processes, qualities, conceptual regions, non-physical objects, collections and even arbitrary sums of objects." (Src: DOLCE-Lite.owl)

[38] http://www.loa.istc.cnr.it/ontologies/DOLCE-Lite.owl (Last visit: December 28th, 2011)

modeled class in DOLCE-Lite.owl is labelled *abstract* and it seems the two are used ambiguously and not clearly distinguished and sufficiently explained. [Masolo et al. 2003, p. 9] admit that “characterization of the concept of universal is still very vague since it does not clarify whether sets, predicates, and abstracts should be included among the universals“, and “if abstracts are entities not extended in space-time, they can differ from universals in many aspects” (p. 9).[39] They try to characterise ontological distinctions between *universals* and *particulars* by the means of a primitive *instantiation*: “particulars are entities that *cannot* have instances; universals are entities that *can* have instances”. The graphical representation in Figure 4 subsumes under *Abstract* a *Fact*, *Set*, or *Region*, and it is slightly different from the latest DOLCE-Lite.owl file, which has been visualized with OntoGraf in Figure 5.

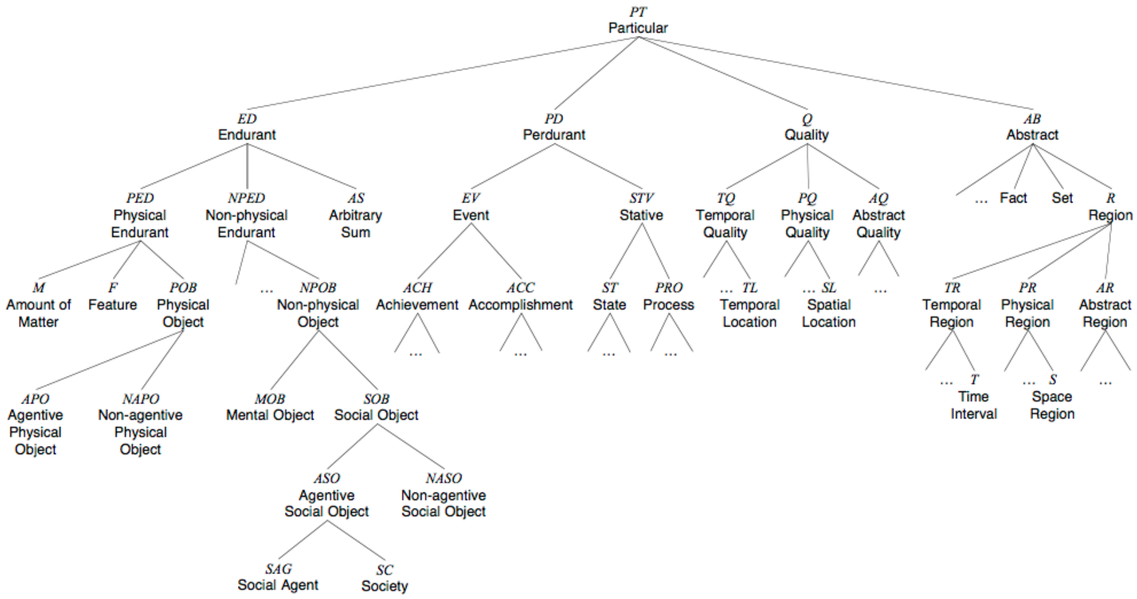


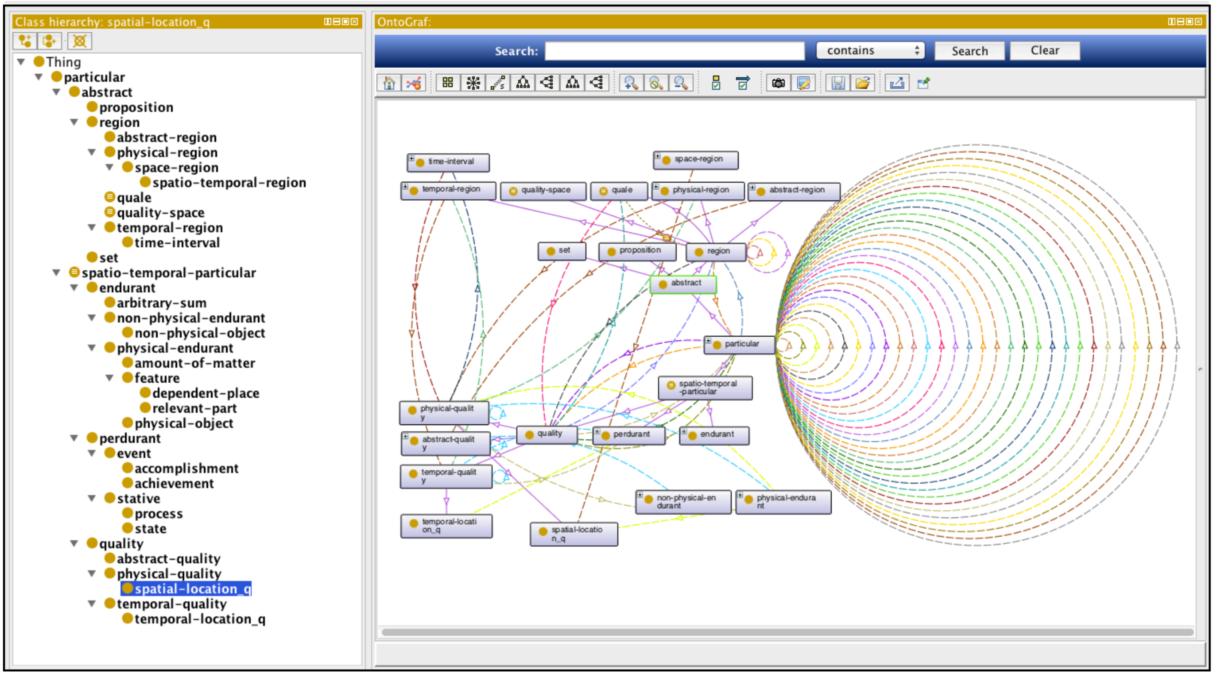
Figure 4: Taxonomy of DOLCE basic categories [Masolo et al. 2003, p. 14]

DOLCE distinguishes between *endurant* and *perdurant* as *spatio-temporal particulars* according to their timely behavior; and the DOLCE authors had been aware of the then ongoing debates and were “sympathetic” with a proposal by Peter Simons treating endurants and perdurants as equivalent classes through abstraction. DOLCE has a clear cognitive *bias* and aims to capture the ontological categories underlying natural language and common sense; not committing to the metaphysical assumptions of an intrinsic nature of the world. [Gangemi et al. 2002] view DOLCE’s categories, being at a “*mesoscopic* level” in that they are “just *descriptive* notions that assist in making *already formed* conceptualizations explicit.

[39] “Properties and relations (corresponding to predicates in a logical language) are usually considered as universals.“ [Gangemi et al. 2002, p. 2]

They do not provide therefore a *prescriptive* (or 'revisionary') framework to conceptualize entities. In other words, our [DOLCE's] categories describe entities in a post-hoc way, reflecting more or less the surface structure of language and cognition" (p. 167).

This becomes clear with the alignment of an explicit concept, e.g. person, in DOLCE, which requires the decision as to where the concept will be subsumed; Person is not an abstract, because person features birthdate and birthplace; that is time and space, i.e. person is categorised as a *spatio-temporal particular*. Navigating down, this taken branch requires the next decision as to whether person is an endurant, a perdurant, or a quality; and which for a person cannot be disjointly featured.[40] Metaphysically, person may well be perceived as a quality but it may not be strictly categorized as such in the sense of DOLCE. Intensionally, person may be perceived in its whole entirety as an endurant, but extensionally (e.g. in the role of a student during a time) it may well be perceived as a perdurant.



- Thing
  - particular
    - abstract
      - proposition
      - region
        - abstract-region
        - physical-region
          - space-region
            - spatio-temporal-region
        - quale
        - quality-space
        - temporal-region
          - time-interval
      - set
    - spatio-temporal-particular
      - endurant
        - arbitrary-sum
        - non-physical-endurant
          - non-physical-object
        - physical-endurant
          - amount-of-matter
          - feature
            - dependent-place
            - relevant-part
          - physical-object
      - perdurant
        - event
          - accomplishment
          - achievement
        - stative
          - process
          - state
      - quality
        - abstract-quality
        - physical-quality
          - spatial-location_q
        - temporal-quality
          - temporal-location_q

Figure 5: DOLCE-Lite.owl graph with property range indications (OntoGraf)

DOLCE seems very useful for categorizing particulars for machine processing, however it does not explicitly support in achieving the goal of this work – namely, formally describing a world of Research in general and LT in particular – addressed first, at human readibility.

Figure 5, indicates the property ranges with DOLCE's classes; that is, multiple recursive properties with *particular* through the right hand-side circles. Each arrow represents a property under the domain-range space of *particular*, such as *atomic-part*, *atomic-part-of*,

[40] In DOLCE-Lite.owl, "[qualities can be seen as the basic entities we can perceive or measure [...] Perdurants (AKA occurrences) comprise what are variously called events, processes, phenomena, activities and states [...] The main characteristic of endurants is that all of them are independent essential wholes".

boundary, *boundary-of*, *exact location*, *generic-constituent*, *has-quality*, and many more – which will not be further investigated. DOLCE is a useful contribution to reflections about the existence and the behavior of entities and their properties in a timely and spatial context, and highly relevant for design decisions with information system modelling. DOLCE was not intended as a universal standard ontology but should be used – as it has been intended – in the sense of a reference module for continued communication and mutual understanding.

3.1.3 Suggested Upper Merged Ontology (SUMO)

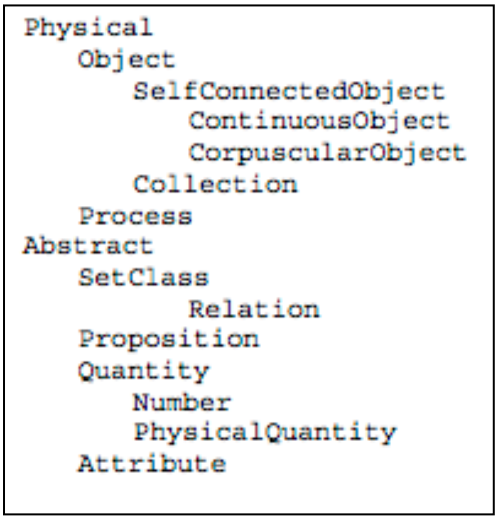
The Suggested Upper Merged Ontology (SUMO)[41] has been suggested as an *upper level ontology* providing general-purpose foundational terms for more specific domain ontologies in support of e.g. automated natural language understanding and for integration of software. SUMO has been created with extensive input from the Suggested Upper Ontology (SUO), as a result from convening "a diverse group of collaborators from the fields of engineering, philosophy, and information science", and through "merging publicly available ontological content"[42]. Once the content had been translated, i.e. *syntactically merged* – a *semantic merge* combined John Sowa's, Russel's, and Norvig's upper level ontologies into "a single conceptual structure" and one other class containing everything else [Niles & Pease 2001, pp. 2-4]. During the semantic merging processes of foundational with lower-level contents, four important cases have been distinguished:

- *nothing maps:* once decided useful, it is a matter of finding a place, and it may involve the creation of intermediate levels
- *concept/axiom out of place in schema:* may require removal of concept
- *perfect overlap:* has with SUMO often occured with mereotopoligical theories
- *partial overlap:* has been considered the biggest challenge

SUMO was investigated at the top level (Figure 6), which talks in *concepts*; the top one is *Entity* subsumed by *Physical* and *Abstract*.

[41] SUMO consisted of ~ 20.000 terms and ~70.000 axioms with all domain ontologies combined. Although the site itself gives a "Last modified" information – this does not refer to most provided ontologies – it is therefore not clear, if ontologies are regularly updated – all those ontologies extending SUMO are available under the GNU Public License: http://www.ontologyportal.org/ (Last visit: July 19th, 2011).

[42] "This content included the ontologies available on the Ontolingua server, John Sowa's upper level ontology, the ontologies developed by ITBM-CNR, and various mereotopological theories, among other sources." [Niles & Pease 2001, p. 3]



```
Physical
   Object
      SelfConnectedObject
         ContinuousObject
         CorpuscularObject
      Collection
   Process
Abstract
   SetClass
         Relation
   Proposition
   Quantity
      Number
      PhysicalQuantity
   Attribute
```

Figure 6: SUMO Top Level [Niles & Pease 2001, p. 5]

[Niles & Pease 2001] admit there was a heated debate behind SUMO's disjointness of *Object* and *Process* among *endurantists*, who adopt a 3D orientiation, and *perdurantists*, in favour of a 4D orientation, finally resulting into adoption of "a 3D orientation by making 'Object' and 'Process' disjoint siblings of the parent node 'Physical' [...] to incorporate content from process-related ontologies" (p. 5). The concept *Process* as well as the concept *Object* have not been further defined in the mentioned work, but *Process* has been referred to as one of the most challenging continuing tasks, being in need of more guidance. An *Object* is formally divided into two disjoint concepts, namely *SelfConnectedObject* and *Collection*; the former's parts all being mediately or immediately connected with each other, where *Collection* consists of disconnected parts, and "the relation between these parts and their corresponding 'Collection' is known as 'member'[43] (p. 6). At the highest level, SUMO distinghishes *Physical* and *Abstract* subsuming four disjoint concepts: *Set*, *Proposition*, *Quantity*, *Attribute*. In SUMO, a *Set* is an ordinary set-theoretic notion, where a *Class* "is understood as a 'Set' with a property or conjunction of properties that constitute the conditions for membership in the 'Class', and a 'Relation' is a 'Class' of ordered tuples [...] immediately subsumed by 'Class' because we restrict 'Relations' to those ordered tuples that express intensional content [...] The concept of 'Proposition' is understood as "the notion of semantic or informational content [...] The class of 'Attributes' includes all qualities, properties, etc. that are not reified as 'Objects' [...] Finally, 'Quantity' under 'Abstract' is divided into 'Number' and 'PhysicalQuantity'. The former is understood as a count independent of an implied or explicit

[43] "Note that this ‚member' predicate is different from the ‚instance' and ‚element' predicates, which relate things to the ‚Classes' or ‚Sets' to which they belong. Unlike ‚Classes' and ‚Sets', ‚Collections' have a position in space-time, and ‚members' can be added and substracted without thereby changing the identity of the ‚Collection'. Some examples of ‚Collections' are toolkits, football teams, flocks of sheep." [Niles & Pease 2001, p. 7].

measurement system, and the latter is taken to be a complex consisting of a 'Number' and a particular unit of measure." (p. 6)

Investigating the SUMO top-level ontology reveals its inherent upper *intensional* approaches derived from underlying semantic merges with Sowa's, Russel's and Norwig's ontologies (which here, will not be investigated individually). SUMO comes very close to a conceptual structure in the sense of the goal of this work – identifying entities (*physical* or *abstract)*, their potential *(object)* and composite *(selfconnected, collection)* structure and involved activity *(process)*, as well as their properties *(setclass, proposition, quanity, attribute)*, and potential intensional interaction *(relations)* at an upper level. SUMO is written in the so-called SUO-KIF language, where the defined concepts are formally declared. [Niles & Pease 2001, p. 5] introduce quantifier and relationship specifications such as *forall*, =>, <=>, *exists*, *and*, *or*, *instance-of*, *part-of*, *subclass-of*, *connected*, *immediately-connected*, *equal*, *member*. SUMO is free and owned by the IEEE. The SUMO ontology is organised modular and divided into self-contained subontologies (Figure 7).

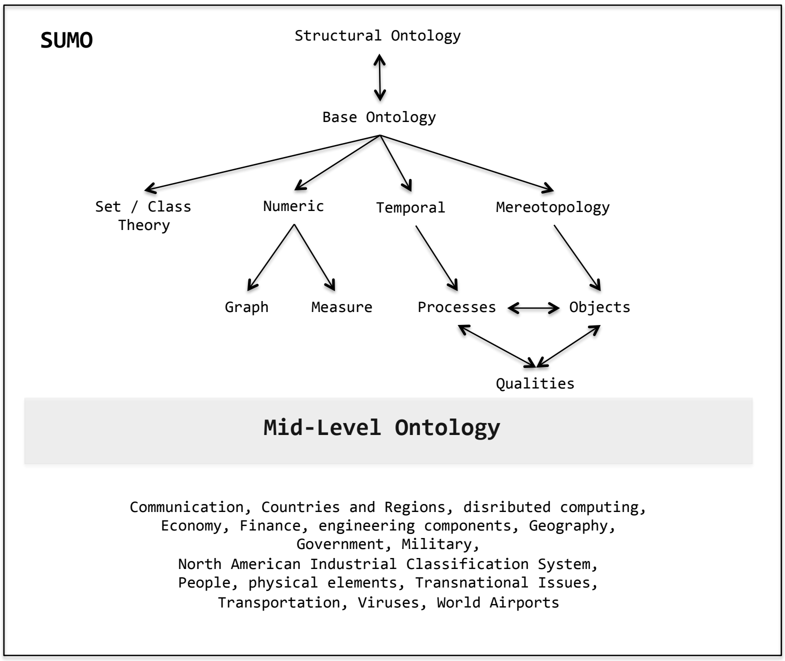

SUMO
Structural Ontology
Base Ontology
Set / Class Theory
Numeric
Temporal
Mereotopology
Graph
Measure
Processes
Objects
Qualities
Mid-Level Ontology
Communication, Countries and Regions, disributed computing,
Economy, Finance, engineering components, Geography,
Government, Military,
North American Industrial Classification System,
People, physical elements, Transnational Issues,
Transportation, Viruses, World Airports

Figure 7: SUMO ontology as presented at the public portal[44]

[Niles & Pease 2001, p. 8] admit, it may be unattainable to reach the intended goal of "a single, consistent, and comprehensive ontology" and suggest "the best we can do is to make clear the various representational choices and bundle them up in consistent and independent packages and, where possible, state mappings between corresponding packages." The

[44] Suggested Upper Merged Ontology – Ontology Portal: http://www.ontologyportal.org/ (Last visit: July 19th, 2011)

concepts in the Top Level SUMO harmonize well with the BWW ontology constructs, and a rough mapping is provided within this work in Table 2.

3.1.4 Cyc

The Cyc knowledge base (KB) is a formalised representation of a vast quantity of fundamental human knowledge: facts, rules of thumb, and heuristics for reasoning about the objects and events of everyday life. The medium of representation is the formal language CycL[45]. Cyc is not a frame-based system: the Cyc team thinks of the KB instead as a sea of assertions, with each assertion being no more 'about' one of the terms involved than another. According to CyC documents: The Cyc KB is divided into many (currently thousands of) 'microtheories', each of which is essentially a bundle of assertions that share a common set of assumptions; some microtheories are focused on a particular domain of knowledge, a particular level of detail, a particular interval in time, etc. The microtheory mechanism allows Cyc to independently maintain assertions, which are prima facie contradictory, and enhances the performance of the Cyc system by focusing the inferencing process. [Marshal & Shipman 2003, p. 59] consider Cyc as an infrastructure for knowledge acquisiton, representation, and utilisation across diverse use contexts. A version of the Cyc knowledge base has recently been made available as OpenCyc – *the world's largest and most complete general knowledge base and commonsense reasoning engine*. During the analysis of Research Entities in this work within section 5.2, the Cyc ontology will be consulted – a „rich and diverse collection of its real-world concepts"[46] to complement the conceptual perceptions. Cyc is available in UMBEL[47], through which it is intended as a basis for the construction of domain ontologies.

3.1.5 Summary

For this work, the BWW *ontology* is considered most important with respect to guiding the entire activity and current task, not least also, because it is increasingly supported by a growing community and seemingly implemented with latest modeling grammars and tools.

[45] What's in Cyc: http://cyc.com/cyc/technology/whatiscyc_dir/whatsincyc, Foundations of Knowledge Representation in CyC: http://www.cyc.com/doc/tut/ppoint/why_use_logic_files/v3_document.htm (Last visit: January 15th, 2012)

[46] The full OpenCyc content is now available both as downloadable OWL ontologies as well as via semantic web endpoints (i.e., permanent URIs). These URIs return RDF representations of each Cyc concept as well as a human-readable version when accessed via a Web Browser. OpenCyc for the Semantic Web: http://sw.opencyc.org/ (Last visit: January 15th, 2012)

[47] Upper Mapping and Binding Exchange Layer (UMBEL): http://www.umbel.org/ See related information at: http://www.w3.org/2005/Incubator/lld/wiki/Vocabulary_and_Dataset#Bibliographic_Ontology_.28BIBO.29

The Investigation of DOLCE and SUMO revealed the differences in approaches towards so-called upper or foundational ontologies driven by specific application needs. In Table 2 a comparison of the approaches is presented through a rough mapping, whilst being aware that the task of this work is not the mapping of foundational ontologies, but rather their application with system analyses and design.

Table 2: Comparing DOLCE and SUMO with BWW

BWW	Explaining Notes	SUMO	DOLCE
Substantial Things	SUMO's entities may be seen as Substantial Things, as well as DOLCE's Particulars.	**Entity** **Physical** **Abstract**	**Particular** **Spacio-Temporal** **Abstract**
Property	Abstract in SUMO as well as in DOLCE may be viewed as property, as well as their subclasses.	**Abstract** **Set/Class** **Proposition** **Quantity** **Attribute**	**Abstract** **Set** **Proposition** **Region** **Quality-Space**
Composite Thing	A SUMO Object may well be understood as BWW's composite thing, variously connected or a *member* in collections. The debated Process may thus be understood as such as well. The corresponding DOLCE particular Endurant may be considered a composite thing, as well as a perceived Quality. Composite things posses *emergent properties* that are not possessed by any component. For example, a computer possesses processing power, not possessed by any individual component.	**Object** **SelfConnected** **Continuous** **Corpuscular** **Collection** **Process**	**Endurant** **Arbitrary-Sum** **Non-Physical** **Physical** **Quality** **Perdurant** **Event** **Stative**

Law	The laws underlying SUMO and DOLCE, were not deeply investigated, but only possible examples are given.	**SUO-KIF Syntax**
		forall;
		=>;
		exists;
		subclass-of
			DOLCE Syntax
			property-domain;
			property-range;
			inverse;
			super;
			subclass

Class	The SUMO Set / Class may well be understood as a BWW *Class*. In DOLCE, the Class concept is not explicitly defined, only the Set, in the mathematical sense.	**Set/Class**	**Set**
Attributes (State Functions)	In SUMO, Relation is restricted to ordered tuples expressing intensional content; these may well be mapped to state functions aka attributes. A number as a physical quantity may well be seen as an attribute or function. DOLCE's Quale – *an atomic region* may be seen as a functional state, however, missing time and space; the same for the mentioned in SUMO.	**(Relation)** **(Number)** **(PhysicalQuantity)**	**(Quale)**

Functional Schema	The SUMO Set/Class as well as DOLCE's regions, may well be seen as a *functional schema* describing a thing, however, the BWW's concept are considered much wider, being temporally as well as geographically bounded.	**(Set / Class)**	**(Set)** **(Abstract-Region)** **Physical-Region** **Temporal-Region**
Event	SUMO does not have a corresponding concept, where the DOLCE Event may be seen as such.		**Event**
Lawful	The lawfulness in SUMO and DOLCE is assumingly validated, however it will not be investigated.		
Interaction	Interaction in SUMO and DOLCE may be defined, but will not be further investigated.		

3.2 Modeling Grammars

Conceptual modeling languages or grammars provide the constructs to represent phenomena and the rules how these interact. Their strength and weakness is measured in terms of abilities to generate representation scripts, extensions and means of disambiguation [Wand et al. 1995, p. 285], [Shanks et al. 2003, p. 87, Oei et al. 1992], and a selection decision is often related to the quality of the resulting model or artefacts. "The degree of correspondence between reality and the modelling grammar has an important impact on the quality of the resulting models [*cites* Schütte & Rotthowe 1998, p. 246] and, thus, on the quality of the subsequent artefacts derived from these models" [Gehlert & Esswein 2006, p. 119]. Influential modeling grammars in the history of modeling are presented, where ontology is increasingly (though only recently) recognised important with information system and database design [Wand & Weber 1990; Guarino 1998; Wyssusek 2006; Wand & Weber 2006; Guizzardi & Halpin 2008]. It is expected, that the growing need for interoperation and data exchange will further push the need for formalised semantics – and thus ontologies[48]. Process models will not be analysed here, because they are considered more relevant with concrete engineering and application workflows, and go beyond the scope of this work in that they are rather prescriptive, while this work aims at being descriptive in providing formal models of a perceived world of Research and LT.

[48] "The Semantic Web gets down to business": (Online article, February 2011 in Computerworld) http://www.computerworld.com/s/article/9209118/The_semantic_Web_gets_down_to_business (Last visit: May 1st, 2012) "Ontotext, Structured Dynamics form Strategic Partnership": http://www.pr.com/press-release/272196 (Last visit: May 1st, 2012)

3.2.1 The Entity-Relationship Model (ERM)

The Entity-Relationship Model [Chen 1976, p. 9] aimed to overcome disadvantages inherent in the network, the relational and entity set models through achieving a high degree of data independence based on set theory and relation theory, offering a *unified view over data*, from which the three existing models could still be derived; the ERM could thus be considered as a generalization or extension of these models (p. 10). For the study of a *data model* Chen stressed the importance of identifying "levels of logical views of data" with which a model should be concerned (see also Figure 8):

(1) Information concerning entities and relationships which exist in our minds.

(2) Information structure: organization of information in which entities and relationships are represented by data.

(3) Access-path-independent data structure: the data structures which are not involved with search schemes, indexing schemes, etc.

(4) Access-path-dependent data structure.

(1) refers to Buckland's understanding of intangible information at this level, which is thus *information-as-knowledge* and furthermore inline with MDE's *CIM* view (see 4.2.3) or the ARIS *conceptual model* and what [Guizzardi & Halpin 2008, pp. 1–2] call "the existence of things in the world regardless of their (possibly) multiple representations", and where they cite (Mealy 1967, p. 525) to claim, that "*This is an issue of ontology, or the question of what exists*" (p. 2). Chen calls it "conceptual objects in our mind" [Chen 1976, p. 14]. The second level (2) is about representations of these conceptual objects i.e. entities, and refers to MDE's *PIM* view (see 4.2.3) and, what Buckland understands as *tangible* or *information-as-thing*, i.e. what is valid according to the GDI. [Chen 1976] was convinced that the granularity of representations influences design processes, and that considerations of relations, relationship relations, attributes and value sets, datatypes, keys, and identification, as well as the model's constructs support data integrity, and thus incorporation of "important semantic information about the real world“ (p. 9). The third level (3) is about the range of potential system features, and considered comparable to MDE's *PSM* view (see 4.2.3); platform-specific, and what Buckland calls *information processing*. Buckland does not cover level (4), which refers to MDE's *ISM* view (see 4.2.3), thus being implementation, application, interface, user, or even query specific. In ERMs, an attribute is defined as a function that maps an entity in an entity set to a single value in a value set, and at level (2), the values of a primary key are used to represent entities converging to objects, where non-key value sets (domains) are functionally

dependent on primary-key value sets. “The entity-relatioship model adopts a top-down approach in utilizing semantic information to organize data in entity/relationship relations” (p. 29), where the relational model during a normalisation starts with arbitrary relations, and “may be viewed as a bottom-up approach” (p. 29). Figure 8 as in [Chen 1976, p. 11] compares the ERM with the network model, where the relational and entity set model uses multiple levels of logical views.

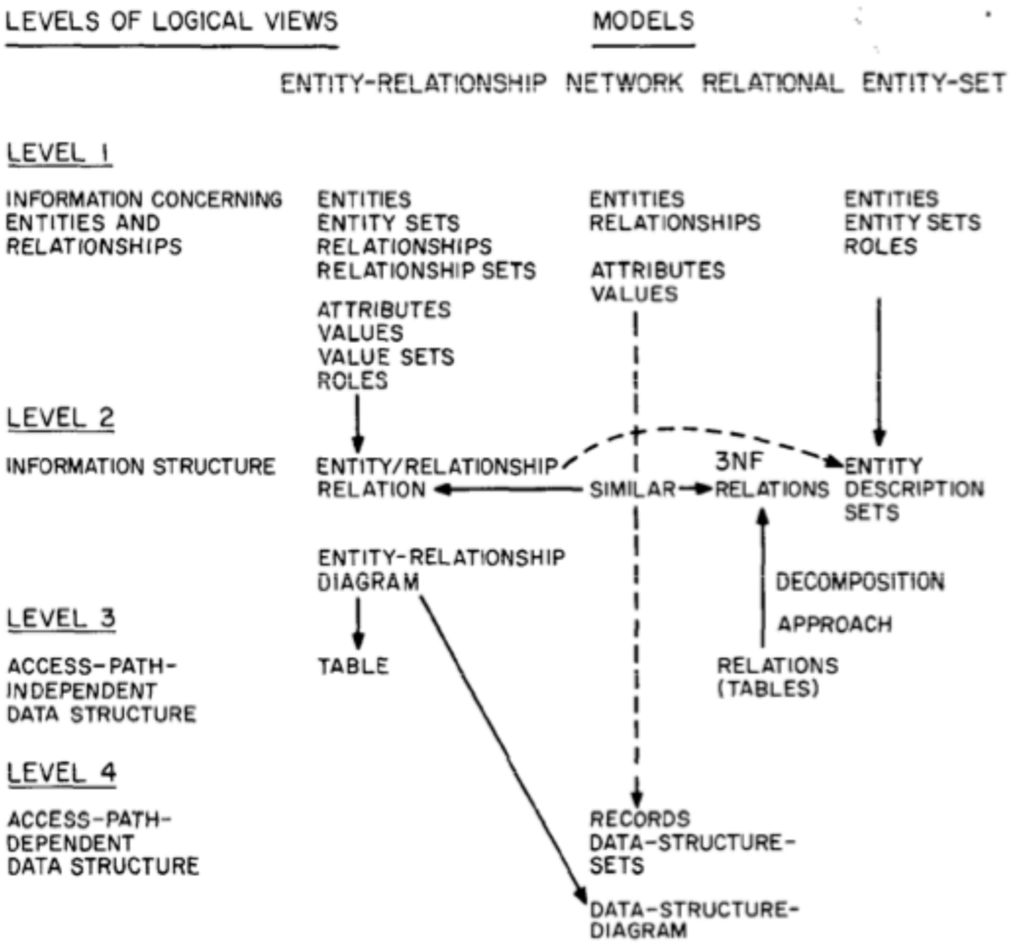

LEVELS OF LOGICAL VIEWS
MODELS
ENTITY-RELATIONSHIP
NETWORK
RELATIONAL
ENTITY-SET
LEVEL 1
INFORMATION CONCERNING ENTITIES AND RELATIONSHIPS
ENTITIES ENTITY SETS RELATIONSHIPS RELATIONSHIP SETS
ATTRIBUTES VALUES VALUE SETS ROLES
ENTITIES RELATIONSHIPS
ATTRIBUTES VALUES
ENTITIES ENTITY SETS ROLES
LEVEL 2
INFORMATION STRUCTURE
ENTITY/RELATIONSHIP RELATION
SIMILAR
3NF
RELATIONS
ENTITY DESCRIPTION SETS
ENTITY-RELATIONSHIP DIAGRAM
DECOMPOSITION
APPROACH
LEVEL 3
ACCESS-PATH-INDEPENDENT DATA STRUCTURE
TABLE
RELATIONS (TABLES)
LEVEL 4
ACCESS-PATH-DEPENDENT DATA STRUCTURE
RECORDS DATA-STRUCTURE-SETS
DATA-STRUCTURE-DIAGRAM

Figure 8: Analysis of Data models [Chen 1976, p. 11]

The ERM languished in the 1970’s, but has then been “wildly successful, namely in database (schema) design“ [Stonebraker & Hellerstein 2005, p. 15], and in the meanwhile a number of commercial tools exist. Chen introduced a simple entity-relationship diagram with entity set and relationship set constructs as presented in Figure 9.

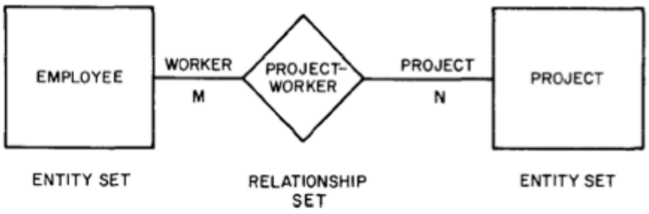


Figure 9: A simple entity relationship diagram [Chen 1976, p. 19, fig. 10].

[Chen 1976] with the ERM suggested extensions to the relational model, to enhance its "semantically impoverished" features (p. 19). "Post relational data were typically called semantic data models" and focused on the notion of classes, "exploiting the concepts of aggregation and generalization" (p. 19), but different from the semantic network model as introduced in the subsequent section.

3.2.2 The Semantic Network Model

In 1969, the Committee on Data Systems Languages (CODASYL)[49] published a specification for a network model database, for which Charles William Bachman later developed logical representation diagrams [Bachmann 1973][50, 51].

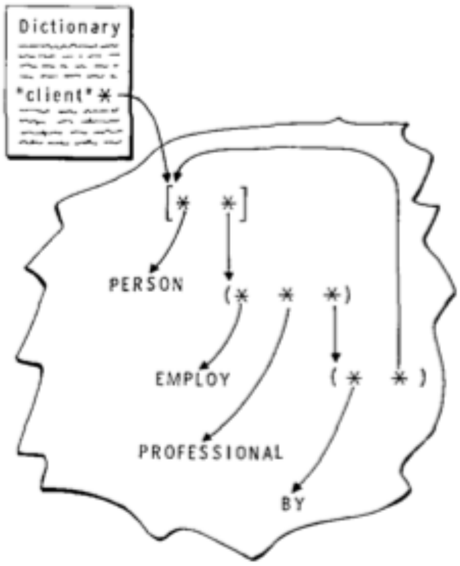


Figure 10: A piece of information in memory [Quillian 1969, p. 462]

[49] The CODASYL Network Model: http://www.remote-dba.net/t_object_codasyl_network.htm (Last visit: January 3rd, 2012)

[50] Bachmann Diagrams: http://en.wikipedia.org/wiki/Bachman_diagram (Last visit: April 8th, 2012).

[51] ACM Turing Award Lecture (ACM is the Association for Computing Machinery)

Ross [Quillian 1969] first introduced *semantic networks* as a mechanism to encode the meanings of words that his program – *the Teachable Language Comprehender* – was capable to comprehend, relating each explicit or implicit assertion to a large memory as with Figure 10. "This memory is a 'semantic network' representing factual assertions about the world" (p. 459). The ambitious goal was then, to "allow representation of everything uniformly enough to be dealt with by specifiable procedures, while being rich enough to allow encoding of natural language without loss of information" (p. 462). Quillian considered factual information encoded to be either as a *unit* or as a *property*: "A unit represents the memory's concept of some object, event, idea, assertion, etc. Thus a unit is used to represent anything which can be represented in English by a single word, a noun phrase, a sentence or some longer body of test. A property, on the other hand, encodes any sort of predication, such as might be stated in English by a verb phrase, a relative clause, or by any sort of adjectival or adverbial modifier" (p. 462). Quillian assumed the best way to explain this is by use of a data-structure diagram, where boxes represent records and directed arrows indicate owner and member records including their cardinality (Figure 10). [Chen 1976, p. 10] refers to his level (4) when explaining the network model as an access-path dependent data structure, but presents it (our Figure 10) at his level (1) view: *Information concerning entities and relationships*, missing the levels (2) *Information Structure* and (3) *Access-Path Independent Data-Structure*. [Bachmann 1973 p. 654] himself introduced it, as "A new basis for understanding is available in the area of information systems. It is achieved by a shift from a computer-centered to the database-centered point of view. This new understanding will lead to new solutions to our database problems and speed our conquest of the n-dimensional data structures which best model the complexities of the real world." (The boxes or units in Figure 10 may refer to what has been discussed earlier – namely, entities referring to identity and converging to objects in information systems.) However, with the definition of new boxes, fomerly-valid arrows may not be continually used and the semantics thus be violated. [Chen 1976, p. 30] therefore correctly asked: "What are the real meanings of the arrows in data-structure diagrams?" [Brachman 1976 pp. 129 ff.] explains: "The basic idea behind the semantic network is a simple one. Information is stored at nodes[52] and 'associations' are represented by links between nodes". His further investigation (pp. 129–130 ff.) revealed, member-relationships typical in class hierarchies were not only used as such "[N]odes for classes are almost universally referred to as 'concept nodes', the implication being that a node should somehow capture *what it means* "to be a member of the corresponding class",

[52] "Note that the only information "stored at" a node is the set of links that impinges on it. We are focusing here on how such a constellation of links can represent those things for which we expect nodes to stand." [Brachman 1976]

and require a logically adequate and consistent definition, which he found to be non-existent. Not only was there no formal definition of *concept* used in semantic networks also the membership relation has been identified problematic. Links as *attributive relationships* were often acknowledged, but usually superficially "[i]t is most often unspecified how to make a concept node actually act as a relation between two other concept nodes – nets are generally not implemented to facilitate such use of concepts" [Brachman 1976, p. 131]. Furthermore, "the fact that the parts fit together in a structured way has been completely ignored" and inheritance of properties to instances "is certainly not apparent from the notation itself" [Brachman 1976, pp. 131–132]. According to [Codd 1970, p. 377], the network or graph model was 'in vogue' for non-inferential systems, but has "spawned a number of confusions, not the least of which is mistaking the derivation of connections for the derivation of relations" (p. 377).

The identified problems with semantic networks needed foundations. In his seminal paper "What's in a concept: structural foundations for semantic networks" [Brachman 1976] examined the fundamentals of the network notation, to understand why it has not been the panacea it was once hoped to be, although it enjoyed such a widespread popularity. He found that "[n]o uniform notation has evolved, no algorithmic procedures for encoding information have been established, and no general assimilation mechanisms exist. Each implementation of a semantic net has adapted the basic node-plus-link idea to its own immediate purpose, creating virtually as many stylized "formalisms" as implementations. [...] implemented networks more often reflect simple concept hierarchies rather than the highly-intertwined knowledge one would expect of human memory" (p. 128); [Brachman 1976] therefore asks for more precision with defining what nodes and links were supposed to represent. That is, to make "explicit several key assumptions that network designers implicitly make about their networks" by concentrating on what is meant by "concept" and how a node and a set of links might represent one; based on philosophical foundations and in the spirit of *What's in a Link* by [Woods 1975], who argues that "if semantic networks are used as a representation for storing human verbal knowledge, then they must include mechanisms for representing propositions without commitment to asserting their truth or belief" (p. 36), and also they must be able "to represent various types of intensional objects without commitment to their existence, in the external world, their external distinctness, or their completeness in covering all of the objects which are presumed to exist" (p. 37), and proposes different mechanisms for handling links, without yet formulating a specification, but rather towards discussing requirements "for an adequate notation and the kind of explicit understanding of what one intends his notations to mean that are required to investigate such questions" [Woods 1975, p.

37]. Semantic networks originated in the field of Compuational Linguistics. However, these fundamental modelling issues are not only relevant and restricted to the modelling and to the understanding of natural language, but also for conceptual modelling with information systems and for the ontological analysis towards descriptions of perceived worlds. *Semantic Networks* as such are not in use today, however, the Semantic Web as a networked system or infrastructure as well as its extension – the Linked Open Data Web – is confronted with similar issues, where now technologies and foundations are available to support the continued ambitions with formal ontologies.

3.2.3 The Relational Model

The relational model by Codd played a major role with system developments since its early introduction in 1969[53] – during a time, where also entity set models were relevant [Chen 1976, p. 10] and network models had their say within database design [Codd 1980, p. 113]. The relational data model by [Codd 1970] "is concerned with the application of elementary relation theory to systems which provide shared access to large banks of formatted data" (p. 377) aimed at future users, whose activities at terminals and application programs "should remain uneffected when the internal representation of data is changed" (p. 377) while assuming that changes will often be needed. Codd saw "no need to specify a physical storage proposal as was required by IMS and CODASYL" [Stonebraker & Hellerstein 2005, p. 10]. [Chen 1976, p. 10] refers to his level (2) and (3) views when pointing to the *relational model* and explains that in the relational model, a relation, is a mathematical relation defined on sets, which are called domains and for means of disambiguation, qualified by roles[54].

[53] According to [Codd 1980] as of 1979, some 40 or more data models were available, where contrary to a then widespread assumption he clarifies, that the relational model (developed in 1969) preceded hierarchical and network models, even though "[h]ierarchical and network systems were developed prior to 1970, but it was not until 1973 that data models for these systems were defined", both of which he himself considered „incomplete" models as of 1979. "Thus, hierarchic and network systems preceded the hierarchic and network models, whereas relational systems came after the relational model and used the model as a foundation." According to [Codd 1980], the widespread use of database systems as such, and their implementation of either the network or the relational model can be regarded an "evidence of the impact of data models on the database field. [...] It is hard to find [a database system] that is not based on either the CODASYL network model or the relational model." He attributes substantial developments in the theory of database structure, research into techniques for optimizing the execution of statements and the separation of semantic from implementation issues or the need to distinguish shared from private variables in programming languages, to the relational model.

[54] "At that time, developers would have to understand the pecularities of each database, as well as how to interact with the underlying hardware. What unified this industry was the widespread adoption of SQL (Structured English Query Language). SQL was an implementation of Edgar F. Codd's relational model, which provided an algebraic basis for modeling databases. The mathematical model assured that all SQL databases would return the same results to the same queries, given the same data. And because most of the database vendors such as IBM adopted the model, programmers could just learn SQL, rather than a new language for each database." (Last visit: December 28th, 2011).

http://www.pcworld.idg.com.au/article/382280/microsoft_researchers_nosql_needs_standardization/

Thus, an attribute name in the relational model is a domain name concatenated with a role name, and basically equivalent to value sets in the ERM, however, the semantics of these terms are different [Chen 1976, p. 26]. Where the *role+attribute* in a relational model is used to distinguish domains with the same name within one relation, an attribute in the entity-relationship model is a function which maps from an entity (or relationship) set into value set(s). [Codd 1970] understands data independence as "independence of application programs and terminal activities from growth in data types and changes in data representation – and certain kinds of *data inconsistency*" (p. 377) and considers the relational view over data „superior in several respects to the graph or network model" as it provides "a means of describing data with its natural structure only – that is, without superimposing any additional structure for machine representation purposes" (p. 377). A little later, [Codd 1980] considers semantic data models an "important contribution to the understanding of the meaning of data in formatted databases" but sees the sore need of some "objective criteria for completeness", as otherwise, semantics is only "a matter of taste" (p. 114).

[Codd 1976]'s relational view over data uses the term relation in its accepted mathematical sense: "Given sets S_1, S_2, ..., S_n (not necessarily distinct), *R* is a relation on these *n* sets if it is a set of *n*-tuples each of which has its first element from $S_{1,}$ its second element from S_2, and so on.[55] We shall refer to S_j as the *jth domain of R*. As defined above, *R* is said to have *degree n*. Relations of degree 1 are often called *unary*, degree 2 *binary*, degree 3 *ternary*, and degree *n n-ary*" (p. 379). The relational model allows for the duplication of domain names, that is, column names. "Whether binary relations (carefully defined with due regard to possible anomalies) are better than relations of higher degree (similarly carefully defined) is of a separate question [...] largely a subjective", however "n-ary relations are unavoidable if one is to support a variety of user views and a variety of queries [...] a data model that does not permit a relationship to be viewed as an entity is clearly inadequate to support these different perceptions" [Codd 1980, pp. 113–114].

3.2.4 The Hierarchical Model

Early hierarchical models (e.g. IMS) lost popularity when relational and network models became standards in database management systems [Stonebraker & Hellerstein 2005, p. 14]. Nowadays, a very popular language to organize hierarchical structures is XML – the

[55] "More concisely, *R* is a subset of the Cartesian product $S_1 \times S_2 \times \cdots \times S_n$." [Codd 1970, p. 379]

Extensible Markup Language. XML and its extensions have regularly been criticized for verbosity and complexity, for lacking efficiency, and for not being the self-describing language it claims to be.[56] Nevertheless, XML is currently the most widely used standard for data exchange. [Stonebraker & Hellerstein 2005, p. 34] predict, that "XML will become an intergalactic data movement standard [but doubt] that native XML DBMSs will become popular." XML provides a formally (but not semantically) declared syntax, which can be validated through so-called Document Type Definitions (DTD's) or XML Schemas, not required in advance. The latest XML 1.0 specification[57] distinguishes between logical and physical structures:

- Logically: "Each XML document contains one or more **elements**, the boundaries of which are either delimited by start-tags and end-tags; or, for empty elements, by an empty-element tag. Each element has a type, identified by name, sometimes called its "generic identifier" (GI), and may have a set of attribute specifications. Each attribute specification has a name and a value."
- Physically: "An XML document may consist of one or many storage units. These are called **entities**; they all have **content** and are all (except for the document entity and the external DTD subset) identified by entity name."

Multiple XML complementing specifications[58] have been published with increased usage. Most common data formats can easily be transformed into XML elements, be it simple text files or Excel spreadsheets, or any proprietary database or file formats. Compared to the syntax of a relational DB schema, XML representations of objects seem very simple and much easier to communicate and understand, especially for people that are not familiar with database technologies. The simplicity with reading XML files at first glance disappears quickly when applied with real-world information management towards integration and with ranging across multiple entity types. The following examples Notation 1, Notation 2, and Notation 3 demonstrate common issues with hierarchical representations.

[56] Major disadvantages of XML: http://about.psyc.eu/Major_disadvantages_of_XML (Last visit: December 28th, 2011) Within the OKKAM project, an XML database has been tested for storage. Although the XML database backend performed well during testing (making use of XQuery), the project experienced scalability issues and finally decided to abandon the XML approach in favour of a relational backend [Bouquet et al. 2006, p. 4]. [Stonebraker & Hellerstein 2005, p. 38] talk of "schema first" and "schema last" systems, where in the latter, instances must be self-describing because no schema gives meaning to records, contrary to the former, where data are always consistent with the pre-existing schema. In their summary "schema last" is considered a niche market.

[57] Extensible Markup Language (XML) 1.0 (Fifth Edition) – W3C Recommendation 26 November 2008: http://www.w3.org/TR/REC-xml/ (Last visit: May 1st, 2012)

[58] XML Specifications: http://www.w3.org/XML/Core/#Publications (Last visit: December 28th, 2011)

As an example, one may consider e.g. person, organisation, or project as entities.

```
<Organisation ORGID=”1”>
     <name>DFKI GmbH</name>
             <hasStaff>
                  <Person PERSID=”2”>
                         <name>Brigitte Jörg</name>
                         <role>Researcher</role>
                         <affiliatedWith>
                               <Organisation>
                                   <name>Deutsches Forschungszentrum für KI</name>
                              </Organisation>
                         </affiliatedWith>
                         <participatedIn>IST World</participatedIn>
                  </Person>
            </hasStaff>
 </Organisation>
```

Notation 1: XML Integration Structure Example

The integration example (Notation 1) represents a full-fledged organisation record at the first level, where relationships *hasStaff* with persons are embedded as full-fledged records. In the given example, a person record realises references to organisation records by its embedding of a partial organisation record, through names. An integration approach according to the way in which information is expressed with (Notation 1) creates ambiguity with organisation names, where many variants occur in different places, which are not physically connected. A solution to this is a reference to the e.g. organisation IDs instead of names. Notation 2 presents an improved storage structure to avoid organisation name duplication from multiple contexts.

Having solved the relationship references leads to the next question. Are the full-fledged person records best embedded under e.g. organisation or better within e.g. project, or e.g. any other entity, or should they be entirely moved to the top level. If it were e.g. at the same level as organisation, and consequently project or any other entity, what need is there for the hierarchy. Where such a top-level *structure* may work fine throughout coverage with explicit attribute-value-only references defined for specified objects, for applications with reified relationship constructs like in CERIF (see 5.1.1) it leads to a *fragmentation* issue, as identified in [Clements & Lockhart 2010, p. 45][59] – see Notation 3.

[59] http://www.st-andrews.ac.uk/crispool/media/crispool%20final%20report%20v2.1%20with%20appendices.pdf (Last visit: May 1st, 2012)

```
<xmlInstances>
     <Organisation ORGID=”1”>
          <name>DFKI GmbH</name>
                  <hasStaff>
                       <Person PERSID=”2”>
                              <name>Brigitte Jörg</name>
                              <role>Researcher</role>
                              <affiliatedWith ORGIDREF=”1”/>
                              <participatedIn>IST World</participatedIn>
                       </Person>
                       <Person PERSID=”3”>
                              <name>Person Name</name>
                              <role>Researcher</role>
                              <affiliatedWith ORGIDREF=”1”/>
                       </Person>
               <hasStaff>
      </Organisation>
<xmlInstances>
```

Notation 2: Improved XML Integration Structure Example

```
<xmlInstances>
     <Organisation ORGID=”1”>
          <name>DFKI GmbH</name>
                  <hasStaff PERSIDREF=”2”/>
                  <hasStaff PERSIDREF=”3”/>
      </Organisation>
      <Person PERSID=”2”>
            <name>Brigitte Jörg</name>
            <role>Researcher</role>
            <affiliatedWith ORGIDREF=”1”/>
            <participatedIn>IST World</participatedIn>
      </Person>
      <Person PERSID=”3”>
            <name>Person Name</name>
            <role>Researcher</role>
            <affiliatedWith ORGIDREF=”1”/>
      </Person>
</xmlInstances>
```

Notation 3: Networked XML Integration Structure Example

A specified hierarchy represents and thus restricts exactly one particular context or situation, e.g. that an organisation explicitly *hasStaff* relationships with persons. However, contextual semantics may be inversely relevant, such as, a person *isHeadOf* an organisation, as with uncountable more cases. This problem has been addressed and e.g. for CERIF XML a solution to streamline the shortcomings was proposed in [Jörg et al. 2012a], introducing an object-centered one-level-only hierarchy, which is underspecified or neutral with respect to particular relationship semantics at first, and allows for a high flexibility with embeddings (see Notation 4). The example in Notation 4 is only an extract, where the entire CERIF 1.4 XML specification[60] allows for multilingual and temporal features.

```
<xmlInstances>
     <Organisation ORGID=”1”>
          <name>DFKI GmbH</name>
          <OrgUnit-Person>
              <Person PERSID=”3”/>
              <RelationshipLabel CLASSREFID=”Staff”/>
          </OrgUnit-Person>
      </Organisation>
      <Person PERSID=”2”>
          <name>Brigitte Jörg</name>
          <OrgUnit-Person>
              <Organisation ORGREFID=”1”/>
              <RelationshipLabel CLASSREFID=”Researcher”/>
          </OrgUnit-Person>
      </Person>
</xmlInstances>
```

Notation 4: Embedded XML Integration Structure Example

The proposed embedded XML structure for the entity organisation in Notation 4 is a flexible and scalable – to a certain extent context-agnostic – approach towards domain conceptualisation, and as such comes close to formal ontological structures, themselves often built on XML. This reified relationship construct is applied in FERON to a certain extend and allows for temporal and functionally open lawful features.

Native XML databases are being further developed and efficiency further improved. Furthermore, a W3C working group[61] aims at improving XML efficiency through EXI – the *Efficient XML Interchange Format*, where however, efficiency comes at the cost of formality.

3.2.5 The Object-oriented Model

Object-oriented modeling languages appeared in the 1970s and late 1980s driven by needs of design support systems [Booch et al. 1998], where applications became increasingly complex

[60] CERIF 1.4 by CERIF Task Group, euroCRIS is licensed under a Creative Commons Attribution-NoDerivs 3.0 Unported License. Permissions beyond the scope of this license may be available at http://www.eurocris.org/CERIF-1.4/.
(Last visit: May 14th, 2012)

[61] Efficient XML Interchange Working Group: http://www.w3.org/XML/EXI/ (Last visit: April 19th, 2012)

and experiments with alternative approaches to system analysis and design began "the number of object-oriented methods increased from fewer than 10 to more than 50 during the period between 1989 and 1994" (p. 10). At that time, the field lacked a common data model; it mostly built on experimental activities, missing a strong theoretical framework, and the semantics of concepts such as types or programs were often ill-defined [Atkinson et al. 1989]. This era was also known for the so-called *impedance mismatch* aiming to bind applications to databases. This, at first failed with transferring research efforts into the commercial marketplace, but was later pushed through start-ups for persistent C++ model system support, and later again extended with a notion for relationships borrowed from the ERM model. "Most of the community decided to address engineering data bases as their target market [...] the market for such engineering applications never got very large [...] Naturally, the OODB vendors focused on meeting these requirements. Hence there was weak support for transactions and queries. Instead, the vendors focused on good performance for manipulating persistent C++ structures." [Stonebraker & Hellerstein 2005, pp. 19–23]

The OODBS Manifesto (Object-Oriented Database System Manifesto) [Atkinson et al. 1989] defined the main characteristics that object-oriented database systems must have, to qualify as such, separating into three categories:

- Mandatory: complex objects, object identity, encapsulation, types or classes, inheritance, overriding combined with late binding, extensibility, computational completeness, persistence, secondary storage management, concurrency, recovery, ad hoc query facility
- Optional: multiple inheritance, type checking and inferencing, distribution, design transactions and versions
- Open: programming paradigm, representation system, type system, uniformity

The Manifesto was not a data model and it obviously missed foundations. From the mid 1990s, a critical mass of ideas has been collected leading to the motivation of developments in UML – the Unified Modeling Language [Booch et al. 1998, p. 11]. Continued efforts resulted in the first releases in June and October 1996 through collaboration by commercial partners contributing to UML 1.0 as a modelling language, being well-defined, expressive, powerful, and applicable to a wide spectrum of problem domains, thus offered to the Open Management Group (OMG) for standardization, and where it has since become widely used (p. 11).

Web tutorials introduce UML as „a standard language for specifying, visualizing, constructing, and documenting artifacts of software systems“, but not as a programming language.[62] There is no generally accepted guideline how to use UML for the modeling of a real world system.

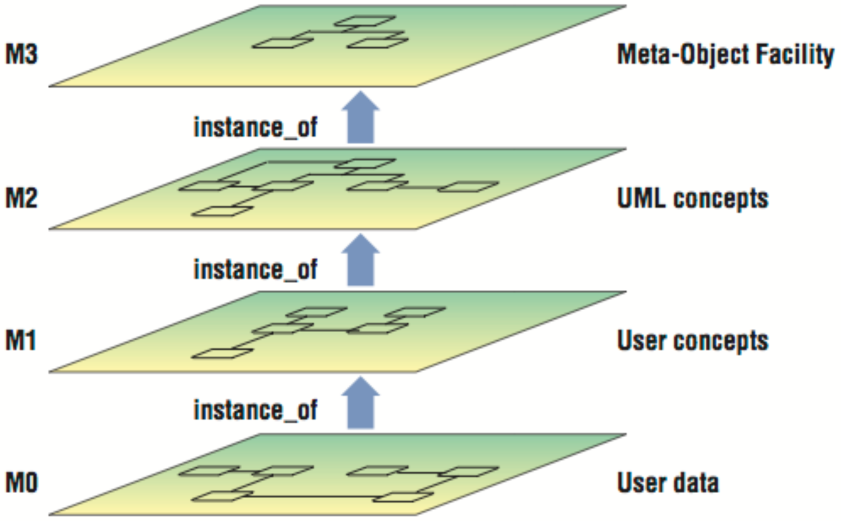


Figure 11: Traditional OMG Modeling infrastructure [Atkinson & Kühne 2003, fig 1]

The traditional OMG modeling infrastructure is depicted in [Atkinson & Kühne 2003, p. 38] (our Figure 11); it refers to the model-driven development (MDD) and thus MDAs, consisting of "a hierarchy of model levels, each (except the top) being characterised as an *instance* of the level above. The bottom level, M0, holds the user data – the actual data objects the software is designed to manipulate. The next level, M1, is said to hold a model of the M0 user data. User models reside at this level. Level M2 holds a model of the information at M1. Because it's a model of a model, it's often referred to as a metamodel. Finally, level M3 holds a model of the information at M2, and therefore is often called the meta-metamodel" (p. 38).

UML is not considered as an object model. It supports object modelling. [Evermann & Wand 2001, p. 354] suggested to extend the use of UML to enable conceptual modeling, mapping UML constructs to a set of real-world concepts 'interpretation mapping'. Their goal is an

[62] Object-oriented concepts hide significant implentation issues and lack transparency down to the bit level; these are crucial in e.g. digital preservation [CCDS Blue Book 2002, p. 26] and equally important in the context of information integration and exchange. Object-oriented concepts represent views, application-specific aspects or customer needs "software that satisfies its intended purpose [...] models to communicate the desired structure and behavior of our system [...] to visualize and control the system's architecture [...] to better understand the system we are building" [Booch et al. 1998, p. 15]. Thus, object-oriented concepts often lack unbiased views upon applications or independence of access and physical storage, which however, are important features for information system integration and towards meaningful information exchange, or re-usability and modularity, and thus, sustainability.

ontology, focused on the ontological meaning of objects and classes, a state of dynamics and of interactions, based on the work by (Bunge 1977, 1979) and applied in a number of subsequent studies, known as the BWW-ontology (Wand and Weber 1989, 1990, 1993)” (p. 356). [Evermann 2009, p. 2] notes: “while UML is not a formal language in the sense that fix-point, model-theoretic or operational semantics are defined for it, it is less ambiguous than a natural language representation.”

3.2.6 Formal Ontology

Formal ontologies “present their own methodological and architectural peculiarities: on the methodological side, their main peculiarity is the adoption of a highly interdisciplinary approach, while on the architectural side the most interesting aspect is the centrality of the role they can play in an information system, leading to the perspective of *ontology-driven information systems*” [Guarino 1998, p. 3]. Within section 2.5 Ontology, the notion of ontology or conceptualisation has been extensively discussed.

> “**<D, R>,** where D is a domain and R is a set or relevant relations on D[63]”
> (Genesereth & Nilson 1987)

While Guarino interpreted *relevant relations* as intensional *conceptual relations*, Gruber understands them as mathematical *extensional relations* “formulated for specific purposes [...] without necessarily operating on a globally shared theory” [Gruber 1993, p. 907], but rather for “a *particular* state of affairs” [Guarino 1998, p. 5].

Guarino proposes as a standard way to represent intensions in a meaningful way independence of a state of affairs; seeing them as “functions from possible worlds into sets” (p. 5) and thus operating on a *domain space* rather than on a certain domain: “We shall define a domain space as a structure <D, W>, where D is a domain and W is a set of maximal states of affairs of such domain (also called *possible worlds*). [...] Given a domain space <D, W>, we shall define a conceptual relation ρ^n of arity *n* on <D, W> as a total function $\rho^n : W \rightarrow 2^{D^n}$ from W into the set of all n-ary (ordinary) relations on D. For a generic conceptual relation ρ, the set $E_\rho = \{\rho(w) \mid w \in W\}$ will contain the *admittable* extensions of ρ. A conceptualization for D can be now defined as an ordered triple C = <D, W, $\mathfrak{R}$>, where $\mathfrak{R}$ is a set of conceptual relations on the domain space <D, W>. We can therefore say that a conceptualization is a set

[63] “In a subsequent paper [Nilsson, N. 1991], Nils Nilsson stresses the importance of the conceptualization for a modeling task.”

of conceptual relations defined on a domain space" [Guarino 1998, p. 5]. With this work, these "conceptual relations defined on a domain space" are considered inline with BWW's *state functions* "representations of the properties" emergent at *thing* and presented in section 3.1 Ontological Foundations, page 36.

As stated in the sections 2.4 Concept and 2.5 Ontology, conceptual modeling and ontology engineering needs a combination of bottom-up (extensions) and top-down (intensions) approaches for the perception of a *reality* (see 3.1 Ontological Foundations, p. 36), and which is compliant with the GDI (see 2.1 Information, p. 25). FERON (chapter 7, p. 210) is modelled in OWL, the Web Ontology Language, an extension of the Resource Description Framework (RDF) and RDF Schema, providing the formal syntax and semantics for machine processing.[64] Common ontology engineering tools such as Protégé[65] implement the entire W3C OWL[66] constructs. These comply with ontological foundations (section 3.1 Ontological Foundations, p. 36), in that they supply *owl:Thing* – Guarino's *Domain Space* and BWW's *Substantial Things* perceived as "the *primary* construct, not the class or natural kinds". The BWW world is made up of "substantial things that possess properties [...] [t]hings change by acquiring or losing properties [...] [a] property can either be *intrinsic* – possessed by the thing itself" and in that correspond with the OWL datatype properties and object properties, or "*mutual* – possessed jointly by two or more things". Any explicit extension (instantiation) of properties is thus understood as a BWW attribute or state function, and refers to Gruber's extensional relation; being an explicit instantiation of Guarino's conceptual relation. *Emergent* BWW properties occur mostly in extensions of yet unknown natural kinds with composite things, and consequently effect BWW's *law*, *lawful* states and therefore *functional schemata* or Guarino's set of conceptual relations. The following extract Notation 5 from the OWL specification gives an overview of the basic constructs constituting the OWL vocabulary, and is worth knowing.

Definition: An ***OWL vocabulary*** V consists of a set of literals V_L and seven sets of URI references, V_C, V_D, V_I, V_{DP}, V_{IP}, V_{AP}, and V_{OP}. In any vocabulary V_C and V_D are disjoint and V_{DP}, V_{IP}, and V_{OP} are pairwise disjoint. V_C, the class names of a vocabulary, contains

[64] The normative formal definition of OWL is provided through the *OWL Semantics and Abstract Syntax* specification http://www.w3.org/TR/owl-semantics/semantics-all.html [W3C 2004], a recommendation by the World Wide Web Consortium (W3C), an international community dedicated to the development of Web standards: http://w3.org/ A newer version OWL 2.0 http://www.w3.org/TR/owl2-overview/ is available since November 2009. (Last visit: January 6th, 2012)

[65] Protégé has been developed at Stanford. It is the most popular tool used in scientific communities with ontology engineering: http://protege.stanford.edu/ (Last visit: July 1st, 2012)

[66] The Web Ontology Language (OWL) is an extension of RDF, providing three sublanguages *OWL Lite*, *OWL DL*, *OWL Full* with increasing expressiveness. http://www.w3.org/TR/owl-features/ (Last visit: January 6th, 2012)

owl:Thing and *owl:Nothing*. V_D, the datatype names of a vocabulary, contains the URI references for the built-in OWL datatypes and *rdfs:Literal*. V_{AP}, the annotation property names of a vocabulary, contains *owl:versionInfo*, *rdfs:label*, *rdfs:comment*, *rdfs:seeAlso*, and *rdfs:isDefinedBy*. V_{IP}, the individual-valued property names of a vocabulary, V_{DP}, the data-valued property names of a vocabulary, and V_I, the individual names of a vocabulary, V_O, the ontology names of a vocabulary, do not have any required members.

Definition: As in RDF, a ***datatype*** *d* is characterized by a lexical space, L(d), which is a set of Unicode strings; a value space, V(d); and a total mapping L2V(d) from the lexical space to the value space.

Defintion: A datatype map D is a partial mapping from URI references to datatypes that maps *xsd:string* and *xsd:integer* to the appropriate XML Schema datatypes.

A datatype may contain datatypes for the other built-in OWL datatypes. It may also contain other datatypes, but there is no provision in the OWL syntax for conveying what these datatypes are.

Definition: Let D be a datatype map. An Abstract OWL Interpretation with respect to D with vocabulary V_L, V_C, V_D, V_I, V_{DP}, V_{AP}, V_O is a tuple of the form: I = <R, EC, ER, L, S, LV> where (with *P* being the power set operator).

- R, the resources of I, is a non-empty set
- LV, the literal values of I, is a subset of R that contains the set of Unicode strings, the set of pairs of Unicode strings and language tags, and the value spaces for each datatype in D
- EC : $V_C \rightarrow P(O)$
- EC : $V_D \rightarrow P(LV)$
- ER : $V_{DP} \rightarrow P(O \times LV)$
- ER : $V_{IP} \rightarrow P(O \times O)$
- ER : $V_{AP} \cup$ { rdf:type } → $P(R \times R)$
- ER : $V_{OP} \rightarrow P(R \times R)$
- L : TL → LV, where TL is the set of typed literals in V_L
- S : $V_I \cup V_C \cup V_D \cup V_{DP} \cup V_{IP} \cup V_{AP} \cup V_O \cup$ { owl:Ontology, owl:DeprecatedClass, owl:DeprecatedProperty } → R
- $S(V_I) \subseteq O$
- EC(*owl:Thing*) = O ⊆ R, where O is nonempty and disjoint from LV
- EC(*owl:Nothing*) = { }
- EC(*rdfs:Literal*) = LV
- If D(d') = d then EC(d') = V(d)
- If D(d') = d then L(″v″^^d') ∈ V(d)
- If D(d') = d and v ∈ L(d) then L(″v″^^d') = L2V(d)(v)
- If D(d') = d and v ∉ L(d) then L(″v″^^d') ∈ R - LV

EC provides meaning for URI references that are used as OWL classes and datatypes. ER provides meaning for URI references that are used as OWL properties. (The property rdf:type is added to the annotation properties so as to provide a meaning for deprecation, see below.)" L provides meaning for typed literals. S provides meaning for URI references that are used to denote OWL individuals, and helps provide meaning for annotations. Note that there are no interpretations that can satisfy all the requirements placed on badly-formed literals, i.e., one whose lexical form is invalid for the datatype, such as *1.5^^xsd:integer*.

S is extended to plain literals in V_L by (essentially) mapping them onto themselves, i.e., S(″l″) = l for l a plain literal without a language tag and S(″l″@t) = <l,t> for l a plain literal with a language tag. S is extended to typed literals by using L, S(l) = L(l) for l a typed literal." [W3C 2004]

Notation 5: Extract from OWL Specification[67]

[67] Direct Model Theoretic Semantics for OWL: http://www.w3.org/TR/owl-semantics/direct.html (Last visit: May 18th, 2012)

3.2.7 Summary

With this work, a formal *Field Extensible Research Ontology* – FERON is modelled, with Protégé, and thus built on a formal syntax, semantics and foundations – as supplied by the tool, which enables OWL modeling inline with the foundations of BWW, Guarino and Gruber. The investigation and presentation of popular modeling grammars was considered important for understanding and thus with respect to taken decisions upon different modeling paradigms and communities, such as e.g. conceptual modeling vs. ontology engineering, closed world vs. open world, relationship and temporal constructs, identity and naming conventions, but also validation mechanisms.

3.3 Modeling Issues

[Sowa 2007, p. 88] recognized, that "[t]he hardest task of knowledge representation is to analyze knowledge about a domain and state it precisely in any language."[68] This work has been guided through foundations where the task was the modeling of "a representation of a real world" – according to BWW "a representation of some perceived reality" [Wand & Weber 1990a, p. 124]. Such an understanding coincides with [Schütte & Rotthowe 1998, p. 243] who consider modeling as "more a definition of structure than a transformation of given structural complexes" (p. 244). Awareness of foundations as well as different modeling techniques incorporated in grammars is thus considered as a critical pre-condition for achieving quality models. This section reflects on some common modeling issues, introduced through different modeling constructs but also viewpoints. Although the goal with this work is not the modeling of an information system as a thing, but a perceived world that information systems ought to be able to model (see section 3.1.1 bullet (1))*,* it is anticipated that FERON will finally be implemented, and is therefore compliant with the architectural paradigms that will be briefly introduced in section 4.2 Information System Architectures.

3.3.1 Open World versus Closed World Assumption

Traditionally, *conceptual* information system models assume a perfect or so-called closed world *CWA*. More and more however, it is noticed, that the real world is not perfect and

[68] [Sowa 2007, p. 88] attributes the shortcomings of approaches since the 1970s to a disjoint logic notation in database design and expert system tools and suggests as first steps "notations that people have used for logic since the middle ages: controlled natural languages supplemented with type hierarchies and related diagrams."

closed but basically open *OWA*, and there exist things that are unknown. [Parsons 1996, p. 355] calls it *imperfect information* and gives a *feeling* for the subject by identifying three different uncertainty types: *uncertainty*, *incompleteness* and *imprecision*. He claims that both, imprecision and incompleteness stem from limitations in the way quantities are measured, but for uncertainty refers to (Bonissone & Tong 1985), being inherently subjective (p. 356).

The differences between *OWA* vs. *CWA* becomes more obvious with transformation or mapping appraoches between e.g. OWL ontologies and frame ontologies, where the former assumes an open world and the latter assumes a closed world in which "everything that is not explicitly said is assumed to be false [...] considerations about large ontologies resulted in the notion of a view as an application-dependent part of an ontology" [Dameron et al. 2005, p. 182]. It is not clear, if a decision for frames should be taken because of the size of the ontology, but it is clear, that within frames[69] ontologies are much easier to read and consequently to understand and communicate. Frames may be, but must not necessarily be application-related – they simply allow to "group together information about each class" [Horrocks et al. 2003, p. 7] and by that, enable to view the entire described *object*. The well-known ontology design tool Protégé supports both modeling paradigms. It namely distinguishes OWL from frame-based and other ontologies (e.g. DAML-OIL, RDF, RDF Schema, etc.). A quick investigation of the ontologies in the *Protégé Ontology Library*[70] reveals that foundational ontologies such as BFO, DOLCE, and OMG, but also domain ontologies such as e.g. a Wine, a Wood, a travel ontology are categorized under OWL, whereas in the frame area, there are e.g. DublinCore, the Science ontology from the former KA^2 project[71], an institutional ontology, a workflow ontology, a Resource-Event-Agent Enterprise (REA) ontology; i.e. there seems a clear and obvious distinction between the selection of one kind towards foundational ontologies being modeled in OWL, whereas domain ontologies are modeled in frames.

RDF and OWL Full are designed for systems in which data may be widely distributed (e.g. via Web). As such, information systems become larger and open, and it is increasingly impractical and finally impossible to know, where all possible data are located. One can

[69] Frames were proposed by Marvin Minsky in his 1974 article *A Framework for Representing Knowledge*. "A frame is a data-structure for representing a stereotyped situation, like being in a certain kind of living room, or going to a child's birthday party. Attached to each frame are several kinds of information. Some of this information is about how to use the frame. Some is about what one can expect to happen next. Some is about what to do if these expectations are not confirmed." [Minsky 1974]. http://web.media.mit.edu/~minsky/papers/Frames/frames.html (Last visit: July 19th, 2011)

[70] Protégé Ontology Library: http://protegewiki.stanford.edu/wiki/Protege_Ontology_Library. The library furthermore suggests google http://www.google.com/search?q=filetype:owl+owl queries or Swoogle http://swoogle.umbc.edu/ queries for more ontologies. (Last visit: July 19th, 2011)

[71] The KA^2 project is now redirected to the http://semanticweb.org portal, collecting science related ontologies. (Last visit: July 19th, 2011)

therefore not generally assume data obtained from a large open system to be complete. If data appear to be missing in one system, one has to assume – in general – that these data *may* exist somewhere else, in another system. This assumption, roughly speaking, is known as the open world assumption [W3C 2009], [Miles & Bechhofer 2009], [Horrocks et al. 2003]. Unlike traditional information systems operating against a fixed set of data sources, the newer architectures such as Linked Open Data applications operate on top of an unbound, global data space, and enable delivery of more complete answers as new data sources appear on the Web [Bizer et al. 2009]. The W3C *Web Ontology Language Guide*[72] explains the implications of *OWA* implementations: “OWL makes an *open world* assumption. That is, descriptions of resources are not confined to a single file or scope. While class `C1` may be defined originally in ontology `O1`, it can be extended in other ontologies. The consequences of these additional propositions about `C1` are *monotonic*. New information cannot retract previous information. New information can be contradictory, but facts and entailments can only be *added*, never *deleted*.”

3.3.2 Conceptual Modeling versus Ontology Engineering

[Gruber 1993a, p. 4] distinguishes between conceptual schemata and ontologies “[w]hile a conceptual schema defines relations on *Data* [see also Chen 1976], an ontology defines terms with which to represent *Knowledge*”. For his translation approach to portable ontology specifications, he thinks of “Data as that expressible in ground atomic facts and Knowledge as that expressible in logical sentences with existentially and universally quantified variables” (p. 4). [Aßmann et al. 2006, p. 249] sees the major distinctive feature between conceptual models and ontologies in that the former are prescriptive while the latter are descriptive. Such a view is also shared by [Calero et al. 2006, p. 64] and [Gruber 1993], in that “[f]ormally, an ontology is the statement of a logical theory” (p. 908). [Spyns et al. 2002, p. 12] point to an important balancing issue between desriptive and prescriptive in this context: “rules, which are important for effective and meaningful interoperation between applications, may limit the genericity of an ontology.” For [Jarrar & Meersmann 2002, p. 1239] “ontology engineering, while similar to data modeling, is substantially more than that, even when the data modeling methodology takes business rules into account”. [Aßmann et al. 2006, p. 264] developed a *meta-pyramid* to explain the role of ontologies within a multi-level modeling approach by

[72] Web Ontology Language Guide: http://www.w3.org/TR/2004/REC-owl-guide-20040210/ W3C Recommendation, February 10th, 2004. (Last visit: March 2nd, 2012)

distinguishing descriptive (*analysis*) from prescriptive (*design*), towards a megamodel of ontology-aware MDE (Figure 12).

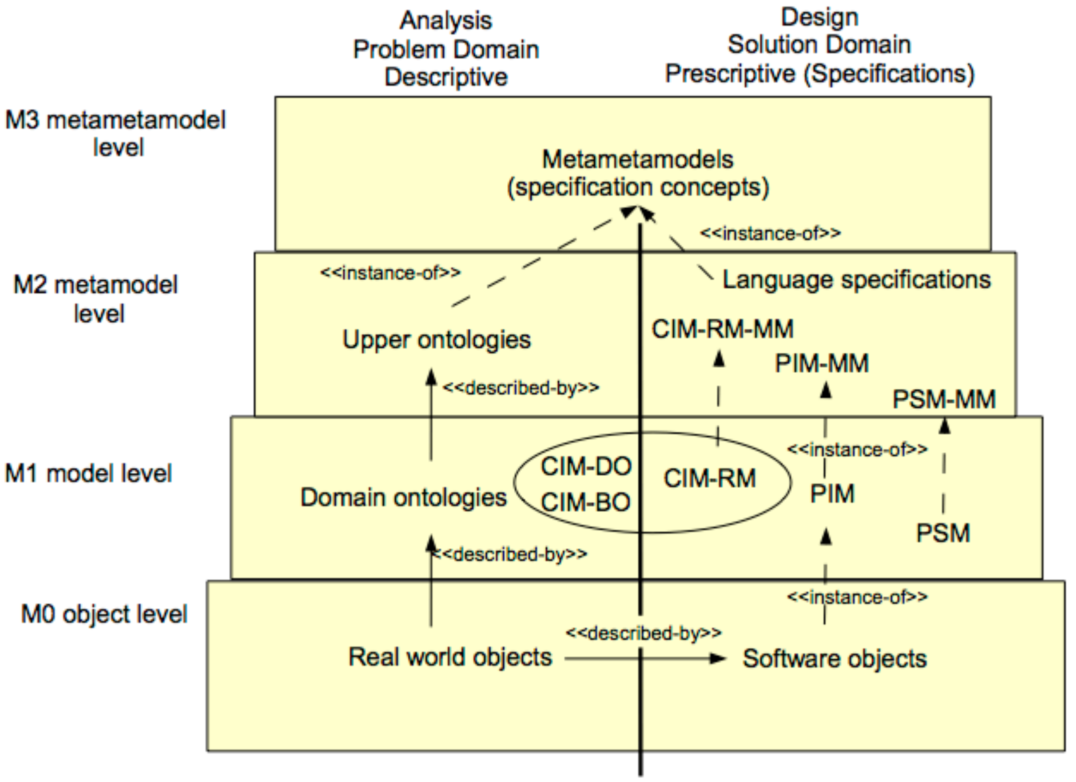


Figure 12: Role of ontologies in a meta-pyramid of MDE [Aßmann et al. 2006, fig. 8]

3.3.3 Naming Conventions or Standards

Naming conventions refer to the applied syntax in a model and often, with vocabularies or formal *ontologies*, there exist recommendations for applying them towards achieving an improved consistency and thus quality of a resulting model in syntax. FOAF, e.g., dedicates a web area to person names[73] refering to the Dublin Core initiative's for representing People's Names in Dublin Core[74], which states, that "it is unlikely that there will be agreement on a single common way of representing names" and upon which FOAF proposes to adopt existing guidelines where possible and roll out own guidelines if necessary. The page

[73] NamesInFOAF: http://wiki.foaf-project.org/w/NamesInFoaf (Last visit: Decmber 26th, 2011)

[74] Reperesenting People's names in Dublin Core: http://dublincore.org/documents/name-representation/ (Last visit: December 26th, 2011).

"NamesInFOAF" notes that "classes for name components do not change across cultures, only the order used for sorting (maybe) and display (sometimes) and concludes, that classes like firstName and lastName no longer make sense."

Usually, metadata descriptions, formats or ontologies have been historically grounded and developed within a particular commmunity addressing particular needs [Gartner 2008, p. 6], [Hirwade 2011, p. 18]. With new technologies, metadata descriptions can often easily be exchanged and re-use is highly recommended. The massive growth of the LOD cloud[75] hosting an increasing number of standards is only one indicator about the numerous metadata standards available; these are not countable anymore (see Figure 13).

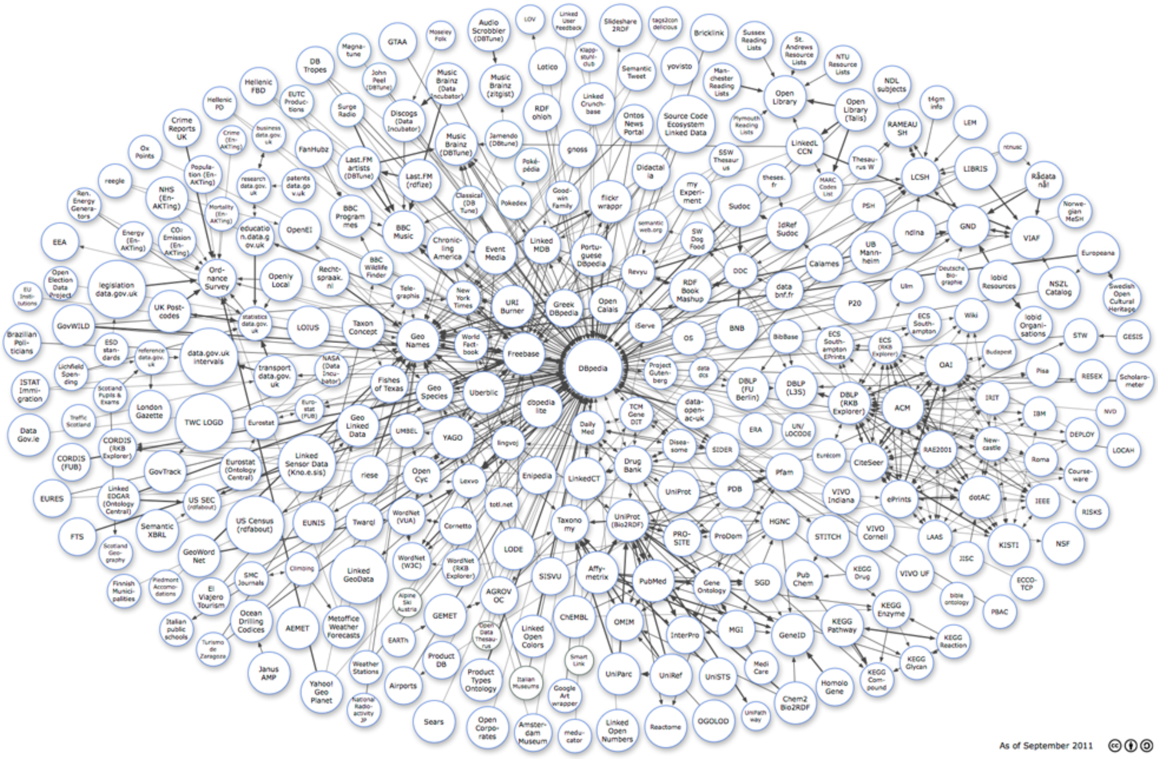


Figure 13: The Linking Open Data Cloud Diagram[76]

To manage access with the LOD cloud's datasets behind these *standards*, a directory has been built – the so-called CKAN[77] directory – aimed at facilitating collaboration, sharing, and finding. [Hirwade 2011, p. 18] identifies "a rapid proliferation of electronic resources" leading to "unpredictability in terms of the availability, accessibility and the authenticity of digital objects" in the wider academic domain; putting *metadata* and even more *metadata*

[75] Linked Data – Connect Distributed Data across the Web: http://linkeddata.org/ (Last visit: December 26th, 2011)

[76] The Linking Open Data cloud diagram: http://richard.cyganiak.de/2007/10/lod/ (Last visit: July 19th, 2011)

[77] The Data Hub – The easy way to get, use and share data: http://thedatahub.org/ (Last visit: July 19th, 2011)

standards in a pole position to manage access, identification and location. Metadata standards do not only grow continuously in numbers, but also in their size and complexity, and where "the task of facilitating metadata in different standards becomes more difficult and tedious" [Nogueras-Iso et al. 2004, p. 611]. Therefore they suggest, for a maximum of usefulness "to use a unique metadata standard in storage labours and provide automated views of metadata in other related standards" (p. 613). Such a suggestion is inline with e.g. the proposed three-level metadata approach for a public sector information and data infrastructure [Houssos et al. 2012].

To qualify "element and attribute names used in the Extensible Markup Language" as defined by a particular community, *Namespaces* "provide a simple method" to identify these by URI references [Bray et al. 2009]. The motivation behind *Namepaces* has been envisioned where applications or XML documents "may contain elements and attributes (here referred to as a "markup vocabulary") that are defined for and used by multiple software modules. One motivation for this is modularity: if such a markup vocabulary exists which is well-understood and for which there is useful software available, it is better to re-use this markup rather than re-invent it. [...] Software modules need to be able to recognize the elements and attributes which they are designed to process, even in the face of 'collisions' occuring when markup intended for some other software package uses the same element name or attribute name." XML [*Bray* et al. 2009] provides a formal declaration of namespaces "A namespace (or more precisely, a namespace binding) is **declared** using a family of reserved attributes. Such an attribute's name must either be **xmlns** or begin **xmlns:**. These attributes, like any other XML attributes, may be provided directly or by default."[78]

3.3.4 Entities and Identification

[Halpin 2006] analyses identity, reference and meaning on the Web, understanding that the "the Semantic Web initiative in particular provoked an 'identity crisis' for the Web due to its use of URIs for both 'things' and web pages and the W3C's proposed solution", arguing by reference to Kripke's *causal theory of reference* as well as to Russel's *direct object theory of reference* and to the Fregean slogan of *priority of meaning over reference* and the notion of logical interpretation, he concludes, that "a full notion of meaning, identity, and reference may be possible, but that it is an open problem on how practical implementations and standards can be created" (p. 1).

[78] Namespaces in XML 1.0 (Third Edition): http://www.w3.org/TR/REC-xml-names/ (Last visit: April 2nd, 2012)

In the *research* domain, the global identifier issue has been recognised as being critical for the quality improvements in information systems and to enable sharing and re-use at large-scale. Science dedicated a comprehensive article to researcher identification [Enserink 2009], the most recent approaches in this respect are reflected in activities around ORCID, whereas e.g. the Virtual International Authority File (VIAF) [79] or CrossRef[80] – the official DOI registration agency – started their operations more than a decade ago to uniquely identify output, i.e. publications in the library domain [e.g. Jörg et al. 2012b, p. 5]. Some countries introduced different concepts for identification of researchers at national level (e.g. the Netherlands with the DAI [van Dijk et al. 2010, pp. 16-17], or Norway by utilizing the Social Security Number). However, these are often unkown beyond national systems and typically not used by researchers.

Where conceptual models support well the descriptions of worlds of interest and therefore the objects within, they have so far not been much concerned with identifications of described objects as such [Evermann & Wand 2001, p. 355]. [Mons et al. 2011, p. 282] propose UUIDs, being unambiguous, non-semantic, stable unique and universal identifiers to which different URIs can be resolved. A proposal for so-called *Cool URIs* has been made [Sauermann & Gyganiak 2008][81]: "A cool URI is one which does not change." – and where they consider it in the duty of a Webmaster "to allocate URIs which will last for a while" and where it is therefore "critical how to design them".

DataCite considers persistent identifiers as the key concept to the service: "A persistent identifier is an association between a character string and an object. Objects can be files, parts of files, persons, organizations, abstractions, etc. DataCite uses Digital Object Identifiers (DOIs)[[82]], at the present time and is investigating the use of other identifier schemes in the future." [DataCite 2011, p. 3]

3.3.5 The Relationship Construct

DBMSs incorporate the referential integrity means to aggregate information contained in distributed tables converging to real world objects. These allow for relationships such as

[79] The world's libraries connected (Virtual Intrnational Authority File) http://www.oclc.org/viaf/ (Last visit: April 2nd, 2012)

[80] http://www.crossref.org/ (Last visit: April 2nd, 2012)

[81] Cool URIs for the Semantic Web: http://www.w3.org/TR/2008/NOTE-cooluris-20081203/ (Last visit: April 2nd, 2012)
Cool URIs don't change: http://www.w3.org/Provider/Style/URI (Last visit: April 2nd, 2012)

[82] "DOIs are administered by the International DOI Foundation, http://www.doi.org/" (Last visit: April 2nd, 2012)

inclusion, aggregation or association. However, “in other disciplines such as linguistics, logic, and cognitive psychology” [Storey 1993, p. 455], but also in computational linguistics, e.g. via automated NLP-driven *Relation Extraction* methods e.g. [Uszkoreit 2007], [Xu 2007], additional semantic relationships have been identified. These are relevant in capturing meaning e.g. [Zimmermann 1993], and are crucial with analysis and design, i.e. knowledge representation. “There are 117 relationships applicable to different knowledge domains in OpenCyc that satisfy the previous [the *parts*] query” [Rodríguez et al. 2009, p. 10]. A clarification over the semantics in LOM identifies ‘references’, ‘is based on’ and ‘requires’ as subtypes of the OpenCyc *parts* relationship, making the fact explicit, that ‘is version of’ relates a LO with another LO and is a refinement of the ‘is based on’ relationship (p. 10).

To represent so-called n-ary relations or reified relations as they are sometimes also ambiguously called, the latest W3C Working Group draft[83] published in 2004 by the Semantic Web Best Practices and Deployment Working Group, a part of the W3C Semantic Web Activity, proposes the use of functional properties with the range of created relation entity types (i.e. Temperature_Relation, Diagnosis_Relation) to host additional attributes or constraints referring to created relation classes. According to the Semantic Web Wiki[84], there are four different kinds of n-ary relationships:

1. additional attributes describing a relation (e.g. high probability)
2. different aspects of the same relation (high, but falling)
3. no distinguished participant, but description of new entity (e.g. purchase)
4. order of relations (lists for arguments)

Whereas some use case examples have been given in the draft document (July 2004) a comprehensive syntax specification has not been defined. At the Semantic Web Wiki page, the approach to n-ary relations is described as follows: “And the RDF/OWL approach to n-ary relations is to map them using binary relations by creating an intermediate entity that serves as the subject for the entire set of relations; this entity is then in turn made the object for a relation in which the main subject is the subject. Since this intermediate entity does not have a real-world name of its own, it is usually given the name of the class to which it

[83] Definint N-ary Relations on the Semantic Web: Use With Individuals: http://www.w3.org/TR/2004/WD-swbp-n-aryRelations-20040721/ (Last visit: April 2nd, 2012)

[84] Semantic Web Wiki: http://semanticweb.org/wiki/N-ary_relations (Last visit: April 2nd, 2012)

belongs, followed by an index number." In more recent activities with respect to property reification W3C drafted a vocabulary[85].

In the traditional hypertext Web, the nature of relationships between two linked documents is implicit, as the data format HTML is not sufficiently expressive to enable typed links. The Semantic Web RDF/OWL syntax allows for typed links, but does not explicitly articulate solutions beyond binary, i.e. n-ary linkages. [Bizer et al. 2009] define Linked Data as "simply about using the Web to create typed links between data from different sources".

[Guizzardi & Wagner 2008, p. 88] recognise "[t]he OMG UML Specification is somehow ambiguous in defining associations. An association is primarily considered to be a 'connection', but, in certain cases (whenever it has 'class-like properties'), an association may be a class: An association class is *"[a] model element that has both association and class properties. An AssociationClass can be seen as an association that also has class properties, or as a class that also has association properties. It not only connects a set of classifiers but also defines a set of features that belong to the relationship itself and not to any of the classifiers.*"[86]

With this work – in FERON – the time-aware relationship construct (section 7.10) is inspired by the investigated activities and by the available formats and descriptions.

3.3.6 Temporal Aspects

Semantic information is not static but meaningful or valid most often only while contextually embedded. That is, semantic information is often dependant on entities in their situational relationships; i.e. related to a particular *time*. [Allen 1983, p. 832] considered the representation of temporal knowledge and temporal reasoning "a core problem of information systems, program verification, artificial intelligence, and other areas involving process modelling" and proposes twelve temporal relations from the perspective of artificial intelligence, knowing "that much of our temporal knowledge is relative, and hence cannot be described by a date (or even a 'fuzzy' date). [...] In particular, the majority of temporal references are implicitly introduced by tense and by the description of how events are related to other events." (p. 836, fig. 4; pp. 834–835). The twelve identified temporal relationships can be and have been condensed to seven, by [Correndo et al. 2010] (Figure 14).

[85] Property Reification RDF Vocabulary: http://www.w3.org/wiki/PropertyReificationVocabulary (Last visit: September 19th, 2012)

[86] http://www.omg.org/spec/UML/20110701/UML.xmi (Last visit: September 19th, 2012)

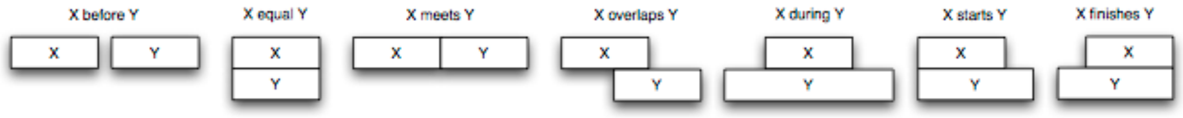


Figure 14: Time Interval Relationships condensed [Correndo et al. 2010, fig. 1]

The relations proposed by Allen are still considered valid by [Correndo et al. 2010, p. 2]. Allen's investigation had been based on previous work that he divided roughly into four categories [Allen 1983, pp. 833–834]:

- *State space approaches:* a state is a description of the world at an instantaneous point in time. Actions are modeled in such systems as functions mapping between states.
- *Datebase systems:* each fact is indexed by a date.
- *Before/After chains:* allows to capture relative temporal information quite directly, however with growth, it suffers from difficult search problems and space problems.
- *Formal models of time:* notable formal models in artificial intelligence; in the situation calculus, knowledge is represented as a series of situations, each being a description of the world at an instantaneous point of time only, without an explicit concept of duration.

"The nature and validity of the information is often related to a time frame and is therefore not universal" [Correndo et al. 2010, p. 2 – *based on Allen 1983*]. It holds for data as well as for information, formal schemas, maps or mappings to undergo 'semantic drifts' [Dunsire et al. 2011, p. 33] that have to be handled. In Research Information Systems a scenario may be the following: *Person X has coordinated a European project Y funded by organization Z in the field B*, it may be relevant for a particular stakeholder, *when* the person coordinated the project. Another stakeholder may be interested in the *duration* of the project, or the *start date* of the project, or when the project will end. A third stakeholder may want to know the areas or units of assessment, in which the project has been classified during a particular time frame.

The usefuleness of time and spatial information has been explored and explained alongside YAGO(2) extensions within [Hoffart et al. 2010, pp. 4 ff.] and there is conviction, that "this would catapult the knowledge bases to a new level of usefulness". E.g. the CERIF ERM data model proposes timestamps with each relationship (link entity), where each link entity is additionally composed of lawful declarable functions. With reference models such as e.g. the FRBR there is no need for time awareness "FRBR does not strive to explicitly account for temporal aspects, such as changes over time" [Le Bœuf 2003, p. 3]. An approach towards temporal representation and management in linked data has been proposed with linked timelines for public sector information[87]. "Some of the data sets have been published already in a Linked Data format, others have been translated within the EnAKTing project[[88]], and many others are waiting to be made available in the Linked Data cloud" [Correndo et al. 2010, p. 1].

3.3.7 Model Validation

[Gruber 1993] explains that for agreements over shared knowledge assumptions and models of the world "ontologies can play a software specification role". This view intersects with the two introduced ontology kinds by [Wand & Weber 1990, p. 1282] (see 3.1.1 bullets (1) and (2)). [Gruber 1993, p. 909] analyses design requirements for shared ontologies and proposes design criteria to guide developments of ontologies for knowledge-sharing based on a *shared conceptualisation*:

1. *Clarity:* An ontology should effectively communicate the intended meaning of defined terms. Definitions should be *objective*. When a definition can be stated in logical axioms, it should be. Where possible, a *complete* definition is preferred. All definitions should be documented with natural language.
2. *Coherence:* An ontology should be coherent; at least, the defining axioms should be logically consistent. Coherence should also apply to the concepts that are defined informally, such as those described in natural language documentation and examples.
3. *Extendibility:* An ontology should be designed to anticipate the uses of the shared vocabulary and be able to integrate new terms in a way that does not require the revision of existing definitions.
4. *Minimal encoding bias:* The conceptualization should be specified at the knowledge level without depending on a particular symbol-level encoding.

[87] "[T]he UK government has launched a public initiative for publishing Public Sector Information (PSI), adopting Linked Data as recommended future best practice. Data sets recently delivered to the public include: government expenses, NHS trusts' performances, public transportation and a whole set of statistics about crime, mortality, census, environment, school and social indicators" [Correndo et al. 2010, p. 1-2].

[88] EnAKTing – Forging the Web of Linked Data: http://www.enakting.org/ (Last visit: April 2nd, 2012)

5. *Minimal ontological commitment:* An ontology should require the minimal ontological commitment sufficient to support the intended knowledge sharing activities. An ontology should make as few claims as possible about the world being modeled, allowing freedom to specialize and instantiate the ontology as needed.

[Gruber 1993, p. 910] consequently identifies himself tradeoffs. I.e., his clarity is mostly about the terms, where ontological commitment is more about conceptualisation; and extendibility may contradict with the minimal ontological commitment requirements: where "a shared ontology need only describe a vocabulary for talking about a domain, a knowledge base may include the knowledge needed to solve a problem or answer arbitrary queries about a domain." What Gruber describes as the minimal encoding bias may contradict with the GDI (see section 2.1, Figure 2), which requires *well-formed-ness* and meaningful *data* for the information *transport*. [Gemino & Wand 2005, pp. 301–302] suggested "clarity within the model may be more important than the apparent complexity of the model when a model is used for developing domain understanding [and where analysts] often strike a balance between simplicity and complexity when communicating information system requirements".

[Shanks et al. 2003, p. 88] refer to theories of ontology for the foundations of representing phenomena in a focal domain, for the modeling of rules and how constructs in modelling grammars should be used. Based on Bunge, they suggest five rules to be adopted by Information System professionals during domain modelling:

(1) Composites and aggregates should be modeled as entities not as relationships.

(2) Relationships should not be modeled with attributes.

(3) Entities should not be modeled with optional attributes.

(4) Conceptual models should clearly distinguish between classes and instances.

(5) Things and their properties should be clearly distinguished in conceptual models.

[Shanks et al. 2003, p. 88] themselves agree, that some of their rules contradict with widely used conceptual modeling practices. Item (2) prevents from assigning time or space to roles, and thus truely conflicts with an understanding endorsed by [Codd 1980, p. 114], in that models not permitting for relationships to be viewed as entities are considered "clearly inadequate to support [...] different perceptions". The above item (3) heavily conflicts with most recent developments towards interconnections and with OWA in systems, such as the LOD Web or the Semantic Web, where completeness cannot be assumed and where everything is basically open or optional and as such perceived as unknown.

The online Enzyclopedia for Information Systems (in German language) provides guiding *Modeling Principles*, so-called *Grundsätze ordnungsgemäßer Modellierung* (GoM) to adress and ensure quality in information models, and where quality goes beyond a syntactic correctness, but takes into account aspects of semantics, representation, organisation and economy "Das hierbei angesetzte Qualitätsverständnis geht über die Betrachtung der syntaktischen Korrekteit der Modelle hinaus und bezieht semantische, repräsentationelle, organisatorische und ökonomische Aspekte mit ein."[89] These principles are as follows (the author's translation from German into English):

- *Principle of Correctness*
  correct representation of facts
- *Principle of Relevance*
  only those facts that support the underlying purpose
- *Principle of Economy*
  adequate proportion between costs and benefit during modeling
- *Principle of Clarity*
  the model has to be understood by stakeholders
- *Principle of Comparability*
  modeling aims at semantic comparability
- *Principle of Modeling Methodology*
  metamodel, consistent multilevel modeling

During the construction or design process, a conceptual model needs to be validated by the stakeholders, whom the resulting system is intended to serve. Such an approach is proposed and inline with MDA's paradigm and ARIS [Scheer 1991], where it is known as *contiuous process improvement* (CPI). If not applied, there is a risk that a model's defects propagate to subsequent design and implementation activities, and, if not discovered until late within the development process, are often costly to correct.

[89] Grundsätze ordnungsmäßiger Modellierung – Enzyklopaedie der Wirtschaftsinformatik: http://www.enzyklopaedie-der-wirtschaftsinformatik.de/wi-enzyklopaedie/lexikon/is-management/Systementwicklung/Hauptaktivitaten-der-Systementwicklung/Problemanalyse-/Grundsatze-ordnungsgemaser-Modellierung (Last visit: April 2nd, 2011)

3.4 Summary

Finally, integration problems are not only of technical nature [Sowa 2007, p. 88], [Hornbostel 2006, p. 31] [Stonebraker & Hellerstein 2005, p. 35], but also subject to fads, trends, politics, standards and co-operation. The need for foundations and a reference methodology seems increasingly accepted and should be applied for improved quality and thus scalability, sustainability and interoperability. Complex domain models are best presented in a graph structure, for which formal ontologies currently seem to be the most appropriate grammar.

4 Information Systems and Architectures

"It is the pervading law of all things organic and inorganic,
Of all things physical and metaphysical,
Of all things human and all things super-human,
Of all true manifestations of the head,
Of the heart, of the soul,
That the life is recognizable in its expression,
***That form ever follows function.** This is the law."*
[Louis Sullivan, born 1856]

The Information Systems (IS) field originates in Information Management. Related research therefore is often called (Management) Information Systems ((M)IS) or Information Systems Science, and has thus mostly been concerned with system development in business or for-profit organizations, focussed on information flow optimization in single enterprises or larger corporations. It has only been recent, that the field recognised the need for cross-organisation information integration [Hevner et al. 2004, p. 29]. A very different evolution can be found behind scientific or research information and communication systems, where a large number of players in a non-coordinated fashion have been involved and nobody owned or controlled the entire system because it has grown organically over decades and considered as "a global, interconnected information system" [Björk 2007, p. 312].

The need for information at a certain time or in a particular situation is usually the first step towards setting up an information system and requires a domain analysis to develop and to design a domain model. System design is therefore often seen as a problem-solving activity [Vessey & Glass 1998, p. 99], [Hevner et al. 2004, p. 1] [Esswein et al. 2010, p. 14], [Siau 2002, p. 107]. To compare the different kinds of information systems according to their capabilities for information management, [Franklin et al. 2005, p. 27] investigate their administrative proximity and semantic integration (see Figure 15). The former indicates how close the various data sources are in terms of administative control and therefore how strong their capabilities for being managed can be accounted to. The latter is a measure of how close the schemas of underlying sources match, i.e. how well their types, names, or units are semantically integrated.

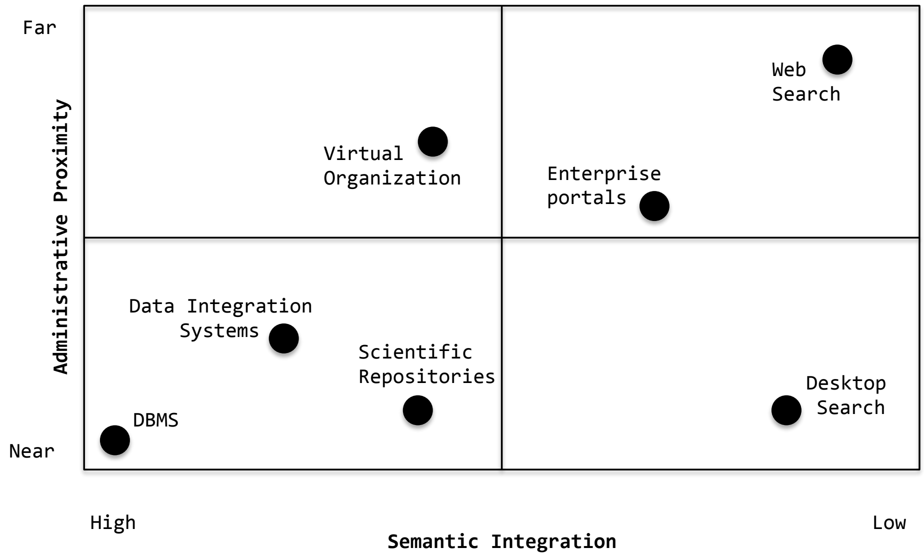


Figure 15: A space of data management solutions [Franklin et al. 2005, p. 27]

Because of their CWA and monolithic character, DBMSs are entirely semantically integrated and thus data are highest matched and closest administrative. On the other end, Figure 15 shows e.g. Web search assuming an OWA that hence encounter low semantic integration, with little (far) administrative proximity, although the supportive semantic (Web) and linking technologies are available. Domain ontologies will certainly be a contribution to narrow the semantic integration gap. On the other end, DBMS, repositories and alike system technologies need to face imperfect worlds and open up their spaces [Krause 2002, p. 25], [Zimmermann 2003]. One may want to recall, that *smaller* information systems in their own sense are aimed at problem solving and have specified needs (prescriptive), whereas large-scale information systems have not been very specific in their problem solving specifications (being descriptive and thus open) except from providing access or answering questions in general; and their lack of semantic integration may therefore be considered a natural consequence; i.e. they are more of an aggregator kind.

[Guarino 1998, p. 3] subsumes under the term ‘information system’ areas such as knowledge engineering, database design and integration, information retrieval and extraction. [Parsons 1996, p. 355] adds imperfect information systems, discussing differences and relationships between idealised models and those dealing with uncertainty. There is a multiplicity of so-called information systems. The focus here is on those that range in the semantic vicinity of this work – namely Research – without a detailed analysis of specific structural features of each system kind, but more towards disambiguation of the information sources that each of them manages. That is, the entities they employ in order to explain their particular conceptual and domain differences or similarities. At this point it is important to recall, that the task of this work is not the modelling of an information system as a thing, but a perceived world that information systems ought to be able to model (see 3.1.1 bullet (1)). Furthermore, this work anticipates implementation and therefore guidance with analysis and design through known architectural frameworks and towards more openness.

4.1 Information System Kinds

Current Research Information Systems (CRIS) have been recognised in the center of the scholarly information interoperability framework. They are surrounded by systems such as ERP (Enterprise Resource Planning), HR (Human Resources), CMS (Content Management Systems), EDMS (Electronic Document Management Systems), and thus intersect with the Finance, Personal and the Enterprise domain. CRISs in particular are most tightly connected with Repositories, E-research or Learning management systems spanning the so-called academic information domain (AID) [van Godtsenhoven et al. 2008, p. 49] as presented in Figure 16.

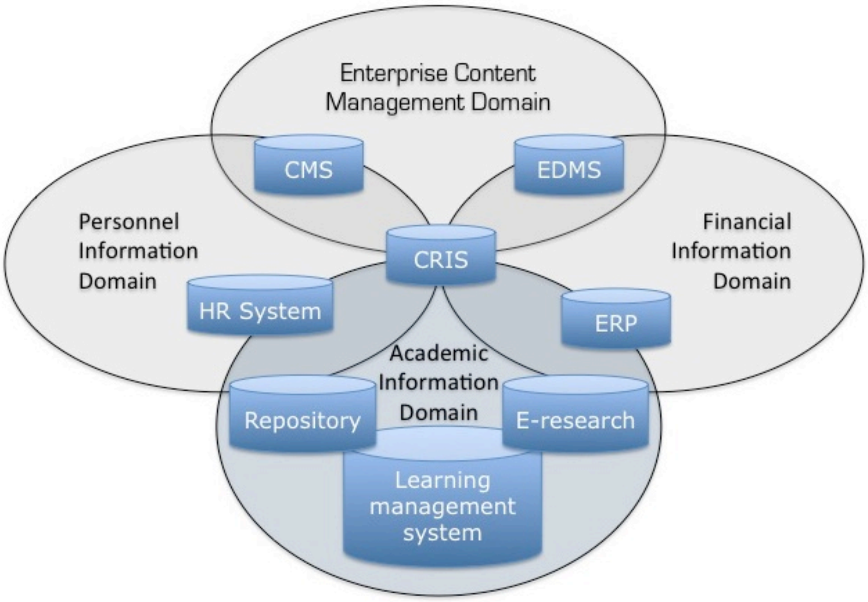

Enterprise Content Management Domain
CMS
EDMS
Personnel Information Domain
Financial Information Domain
CRIS
HR System
ERP
Academic Information Domain
Repository
E-research
Learning management system

Figure 16: The Enhanced AID model [van Godtsenhoven et al. 2008, p. 49]

The strongest current ongoing intersection of CRISs is indeed with scholoarly repositories, and most recently these are utilized[90] and extended towards better coverage of *research data*

[90] Paving the way to an open scientific information space: OpenAIREplus – linking peer-reviewed literature to associated data: http://www.openaire.eu/en/component/content/article/326-openaireplus-press-release (Last visit: December 15th, 2011)

(labelled E-research in Figure 16). The move has happened not least because of imposed mandates from funding organisations e.g. [Ginty et al. 2012, p. 3–4], but certainly also because of expected high benefits from shared *research data* for the researcher, to the University, and to the national research agenda [Wolski et al. 2011, p. 1]. CRISs provide rich means for advanced metadata management, where repositories have been more concerned with deposit for rediscovery, and are most recently reutilized for research data storage. The Open Access (OA) movement has been a strong driver behind the increasing number of repositories[91]. It has often been discussed, if the Education domain belongs to the Research domain (see also chapter 5 Analysis of Research Entities), and the area of learning management systems will therefore be introduced in short to ensure a domain understanding. During history, the scientific domain has been related to (digital) libraries and (digital) archiving or long-term preservation, and these will therefore also be investigated in brief.

4.1.1 Current Research Information Systems

In January 1967 a UNESCO/ICSU Committee was convened to carry out a particular study. The result was published by Unesco's World Scientific Information Programme (UNISIST) as *Study report on the feasibility of a world science information system* [UNISIST 1971]. The report found that national science policies related to allocation of funds and equipment had been familiar, if not resolved, whereas the organisation of informational resources of science has been given less attention, and therefore a joint UNESCO/ICSU project was intended to theme the international deployment of the informational resources of science, that embodies man's knowledge and as such constitutes an essential resource for the work of scientists[92].

This activity reveals that CRIS and CERIF activities are not new, and that the need to develop Current Research Information Systems (CRIS) dates back to the late 70s, where serious efforts for international cooperation among research information systems were made, to survey national scientific and technological potentials for use in the formulation of science

[91] Up-to-date map of Repositories: http://maps.repository66.org/ (Based on data provided by DOAR and OpenDOAR.) (Last visit: April 2nd, 2012)

[92] "[Knowledge] is a cumulative resource; knowledge builds on knowledge as new findings are reported. It is an international resource, built painstakingly by scientists of all countries without regard to race, language, colour, religion or political persuasion. [...] Scientists who are its builders and users ask only that each other's contributions be verifiable; it is, therefore not only a resource; it is a means through which the world's scientists discipline the practice of their professions. It is a medium for the education of future scientists, and a principal reservoir of concepts and data to be drawn on for application to economic and technological development programmes. The recommendations in this report are concerned with the cultivation of this resource, and with the international cooperation to improve its accessibility and use to the end that, as an international resource, it may contribute optimally to the scientific, educational, social, cultural, and economic development of all countries." [UNISIST 1971, p. 1]

policies at national levels. Early efforts towards a *world-science information system* were thus promoted by UNISIST – then simply called "registers of current research" (UNISIST 1975)[93]. Their main objectives were "to enhance communication among scientists concerning ongoing projects" and "to provide an effective information base to managers of the national R&D program" (UNISIST 1975). A Reference Manual for machine-readable bibliographic descriptions has then been published in three editions [UNISIST 1974, 1981, 1986][94], aimed at documentation databases, demonstrating "The flow of scientific and technical information" (in [Fjordback Søndergaard et al. 2003, p. 281, fig. 1]). The developed model is *user*-driven and addressed at *Producers* of *Information Sources*. *Information Sources* were separated into three types, *informal*, *formal* and *tabular* sources.

The first are talks and lectures, the second are distinguished into *published* and *unpublished* kinds, where the former refers to *Publishers* and *Editors*, with *Libraries*, *Abstracting* and *Indexing Services* supplying *Abstracts*, *Indexes*, *Catalogs*, *Guides*, *Referral Services*, the latter refers to *Clearing Houses* with e.g. *Thesis Reports*. Finally, all converge to so-called *Information Centers* supplying e.g. *Special Bibliograpies*, *Reviews*, or *Syntheses* to *Users*. The third kind of information sources (tabular) are provided by so-called *Data Centers* that host *Primary Sources* (selection, production, distribution) and *Secondary Sources* (analysis and storage, dissemination) as well as *Tertiary Services* (evolution, compression, consolidation) to supply *Quantified Surveys*. The UNISIST model as replicated in [Fjordback Søndergaard et al. 2003, p. 281, fig. 1] did already identify *Information Centers* and *Data Centers* as two distinct entities. The *Information Centers* certainly match well with the concept of repositories, (see next section 4.1.2 Scholarly Repositories). They contain *Publications* or *Unpublished* grey literature, and *Data Centers* match especially with the concept of research data, which has become of high interest recently, and where repositories are increasingly being utilized for storage (see therefore a separate section on 4.1.3 Data Repositories (E-Research)). The motivation behind CRISs goes beyond descriptions of pure 'Information Source' (content) entities, in that they allow for descriptions of entities such as person, organisation, service, and many more (see Figure 17).

[93] CORDIS Archive: CERIF – CERIF 2000 fundaments: http://cordis.europa.eu/cerif/src/fundaments.htm (Last visit: May 16th, 2012)

[94] "The model was a product of four years of co-operation between the United Nations Educational, Scientific and Cultural Organization (UNESCO) and the International Council of Scientific Unions (ICSU)" [Fjordback Søndergaard et al. 2003, p. 278]. The model was updated [Fjordback Søndergaard et al. 2003, p. 303, figure 5] towards being more adequate with new technologies and to support a better understanding within other disciplines, especially the Social Sciences and Humanities However, the update entirely misses out the very important concept of the "Data Center" in assuming that tabular data are entirely "present in printed books and journals and unpublished documents" (p. 282).

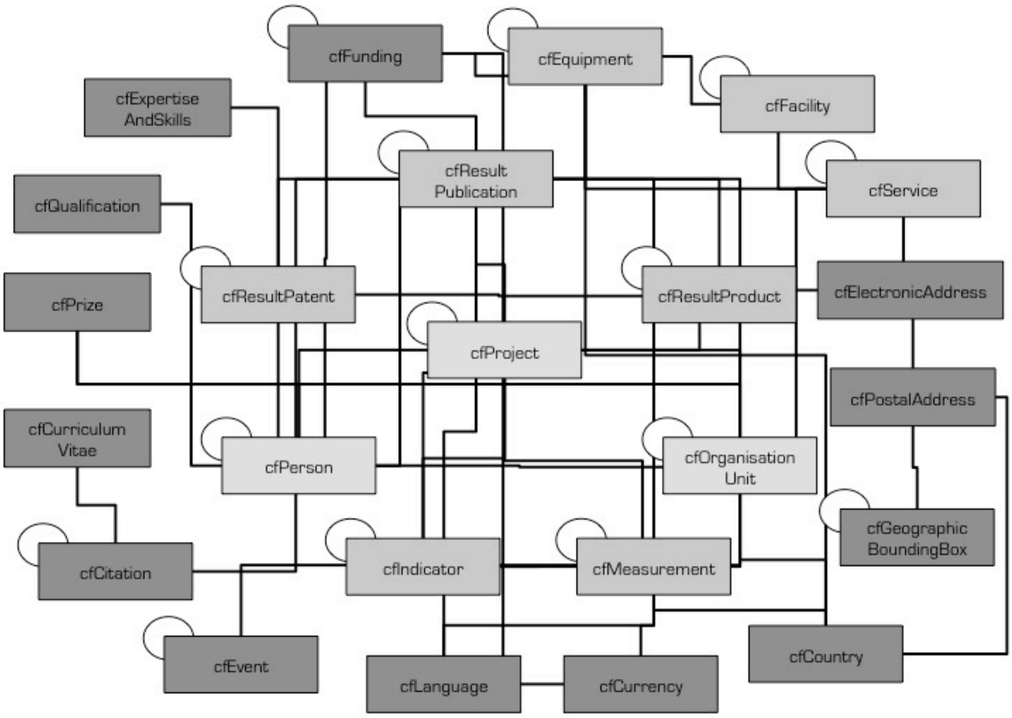


Figure 17: CERIF entities and relationships [Jörg et al. 2012, p. 9, fig 1]

The UNISIST *Publishers* or *Editors* concepts in this sense, match well with the CRIS concept of organisation and the UNISIST concepts of *Libraries*, *Abstracting* or *Indexing* match well with the CRIS concept of a service, under which also the two kinds of repositories are subsumed. In Europe, activities around CRISs have always been tightly related to CERIF. The European Commission published CERIF 1991 and its successor CERIF 2000 as a “EU recommendation to Member States”[95], before in 2002 the responsibility of the activities have been entrusted to euroCRIS, a non-profit organization registered in the Netherlands. CRISs are gaining more interest because they allow for representations of complex research worlds [van Godtsenhoven et al. 2008, p. 49], [Hornbostel 2006, p. 29], [Jeffery & Asserson 2006, p. 3], [Asserson & Jeffey 2004, p. 31], [Asserson et al. 2002, pp. 8–9], [Zimmermann 2002, p. 1]. As indicated in sections 2.1 Information, and 2.2 Data and Metadata of this work, it is of little value to consider data, artifacts or information objects in isolation “their meaning is derived from their relationships to each other” [Pepe et al. 2010, p. 1].

An advanced relationship construct is the strength of CRISs through their underlying extended CERIF ER-model, maintained in so-called link entities connecting research entities. CERIF and thus CRISs identify and maintain the important entities in the Research domain,

[95] CORDIS comprehensive information about CERIF, CRISs and their history: http://cordis.europa.eu/cerif/ (Last visit: July 2nd, 2012)

and their relationships. The perceived Research world of CRISs is thus composed of entities such as person, organisation, project, publication, patent, research product, funding, facility, equipment, service, expertise and skill, qualification, prize, cv, citation, event, language, currency, country, geographic binding, postal and electronic address as indicated in Figure 17.

Figure 17, shows the CERIF[96] Research entities and indicates their relationships as simple lines for a better readability. However, in ER-models and thus, finally with running CRIS implementations, these relationships are maintained as entities (so-called Link Entities) and in fact, are underspecified either unary or binary relationships, where their linking names; i.e. types or roles are not explicit constructs of the formal model, but maintained as vocabulary terms within the CERIF Semantic Layer, each having its own identifier [Jörg et al. 2012b, p. 32]. The CERIF entities, including the link entities and the Semantic Layer will be investigated in more details within chapter 5 Analysis of Research Entities, where each will be analysed towards its convergence into a Research object and its adoption in FERON.

4.1.2 Scholarly Repositories

Digital repositories may be considered as containers for scholarly output, hosting information sources, such as pre-print publictions, theses, technical reports, working papers (also known as Literature (see section 5.2.3.2), digitised text, but also image collections, and increasingly research data. The EC-funded DRIVER[97] projects and successors OpenAIRE and OpenAIREplus aim at the creation of “a cohesive, robust and flexible, pan-European infrastructure for digital repositories, offering sophisticated services and functionalities for researchers, administrators and the general public inspired by the vision to build a Europe and worldwide digital repository infrastructure, which follows the principle of linking users to knowledge” [van der Graaf & van Eijndhoven 2008, p. 11]. [Swan 2012, p. 21] identified

[96] The CERIF model is defined at three different levels (inline with common architectural frameworks in the next section). It maintains loosely coupled conceptual, logical and physical descriptions. Here, the conceptual level is presented, i.e. MDA’s *CIM* view (see 4.2.3). For education purposed, the CERIF entities have conceptually often been grouped into base entities (cfPerson, cfProject, cfOrganisationUnit), result entities (cfResultPublication, cfResultPatent, cfResultProduct) and so-called 2nd Level entities. This conceptual structure is however not reflected at logical or physical level, where all entities are underlying consistent constructs.

[97] Digital Repositories Infrastructure Vision for European Research (DRIVER): http://search.driver.research-infrastructures.eu/ With DRIVER II there was a move to a production-quality infrastructure and efforts led to the launch of a new international organization (COAR) the Confederation of Open Access Repositories: http://www.coar-repositories.org/ (Last visit: April 20th, 2012) http://www.driver-repository.eu/DRIVER-Objectives.html The OpenAIRE project is a successor of DRIVER and DRIVER II - implementing the EC Open Access pilot. It is succeded by the OpenAIREPlus project, which kicked-off in December 2011 towards scientific data integration with open access publications: http://www.openaire.eu/en/component/content/article/76-highlights/326-openaireplus-press-release (Last visit: April 20th, 2012).

four major types of repositories based on OpenDOAR[98]: institutional (83%), subject-specific (11%), specialised (4%), government (2%). Developments with repositories started in the 1990s [Scholze & Maier 2012, p. 205]. Where CRISs are mainly concerned with quality in metadata (context) maintained via multiple entities, the repository community is more interested in (content) for global open access. [van Godtsenhoven et al. 2008, p. 52] surveyed the European repository landscape and explain its main focus: "The global Open Access repository community is more concerned with providing access to the full-text than with precision and consistency in the metadata. Metadata is also important, but the primary goal is providing access" to push the Open Access (OA) movement.

The Berlin declaration[99] on OA to knowledge in the Sciences and Humanities supports activities that are inline with the Budapest OA Initiative, the ECHO Charter and the Bethesda statement on OA Publishing defining an OA contribution as: "Establishing open access as a worthwhile procedure ideally requires the active committment of each and every individual producer of scientific knowledge and holder of cultural heritage. Open access contributions include original scientific research results, raw data and metadata, source materials, digital representations of pictorial and graphical materials and scholarly multimedia material". Where the initial goal with repositories was to push back the publishers and their vast fees, the state as of today is still unsatisfying and the publishers' budgets are as big as ten years ago[100]. New policy guidelines for the development and promotion of OA by UNESCO have been published [Swan 2012][101]. Basically there are two known approaches to achieve OA – and these have been thoroughly reflected in [Swan 2012, pp. 20–22], including business models, which are still being discussed and thus, as hybrid models still maintained.

- Green: repositories collect and host all research output
- Gold: OA journals[102] publish research output
- Hybrid OA: publishers offer OA if authors pay

[98] OpenDOAR – Directory of Open Access Repositories (Last visit: April 20th, 2012): http://www.opendoar.org/

[99] Berlin Declaration on OA to Knowledge in the Sciences and Humanities (October 22nd, 2003) supports activities inline with the Budapest OA Initiative, the ECHO Charter and the Bethesda Statement on OA Publishing from June 20th, 2003 (http://www.earlham.edu/~peters/fos/bethesda.htm): http://www.zim.mpg.de/openaccess-berlin/berlin_declaration.pdf (Last visit: April 20th, 2012)

[100] Academic publishers make Murdoch look like a socialist (Last visit: August 29th, 2011): http://www.guardian.co.uk/commentisfree/2011/aug/29/academic-publishers-murdoch-socialist

[101] Policy Guidelines for the development and Promotion of Open Access (Last visit: April 2nd, 2011): http://unesdoc.unesco.org/images/0021/002158/215863e.pdf

[102] Directory of Open Access Journals (Last visit: April 2nd, 2011): http://www.doaj.org/

Most repositories manage their bibliographic information through descriptive Dublin Core (DC) elements (investigated in chapter 5 Analysis of Research Entities). These are also specified in the Open Archives Initiative – Protocol for Metadata Harvesting (OAI-PMH)[103] [Swan 2012, p. 20], [van Godtsenhoven et al. 2008, p. 52].

Increasingly, the repository community is becoming aware of the need for contextual and quality metadata and repositories are slowly transforming into CRISs. At the same time, CRISs are incrementally equipped with storage mechanisms: “Distinction between CRIS / repository starts to blur”[104]. According to [Scholze & Maier 2012, pp. 205–206] “there is now a wider uptake of the question as to how and to what purpose CRIS and repositories should integrate or interact”. The two communities are in communication and officially started their collaboration by announcing joint interest in their ‘Rome Declaration’[105]. The EC-funded Open Access Pilot OpenAIREplus will align the OpenAIREplus data model to CERIF[106].

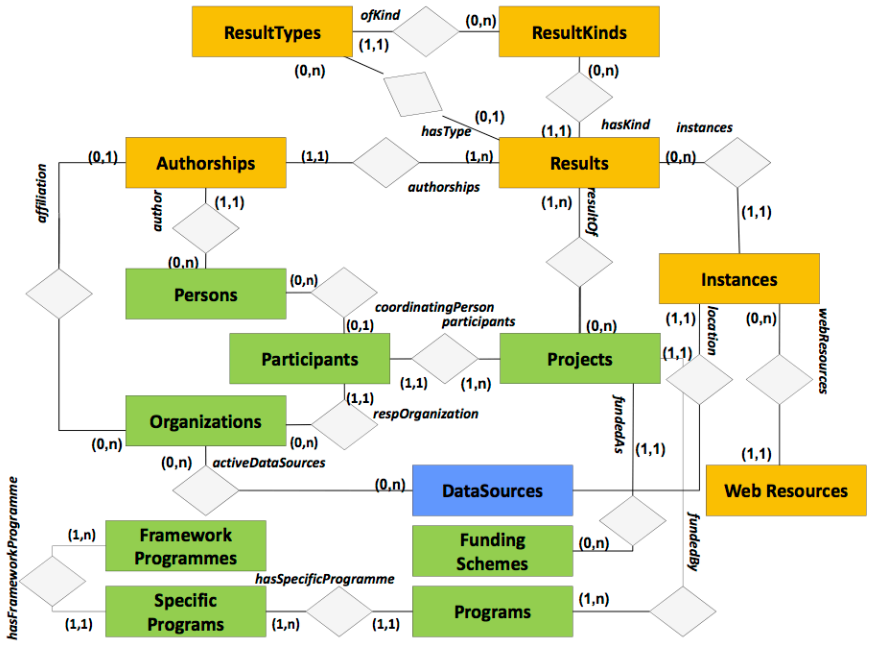


Figure 18: OpenAIRE (Open Access Infrastructure for Research in Europe)[107]

[103] OAI PMH – Open Archives Initiative Protocol for Metadata Harvesting (Last visit April 2nd, 2011): http://www.openarchives.org/pmh/

[104] EPrints: A Hybrid CRIS/Repository (slides by Leslie Carr, University of Southampton) (Last visit: April 2nd, 2011): http://eprints.soton.ac.uk/271048/1/hybridCRISIR.pdf

[105] ROME Declaration by the CRIS and OAR community published in October 2011): http://www.openaire.eu/en/about-openaire/publications-presentations/public-project-documents/doc_details/308-rome-declaration-on-cris-and-oar.

[106] EuroCRIS NewsFlash Issue 49, December 2011 http://www.eurocris.org/Uploads/Web%20pages/newsflash/Newsflash%2049.pdf

[107] OpenAIRE Data Model (slide 22 by Paolo Manghi, CNR, Instituto di Scienza e Tecnologie dell’Informazione „A.Faedo“) (Last visit: April 20th, 2012): http://www.eurocris.org/Uploads/Web%20pages/seminars/Seminar_2011/Session%203%20-%20Paolo%20Manghi.pdf

Figure 18, shows the entities underlying the OpenAIRE repository. These are research output called *Results*, distinguishing types and kinds. Furthermore, agents such as person and organisations, but also *Projects*, *Data Sources* and Funding entities such as *Programs*, and *Funding Schemes*. The model or its physical implementation will not be further investigated, because with OpenAIREplus it will converge to CERIF. The snapshot provides insight into the repository world through the underlying presented entities and reveals the intersections with CRISs.

4.1.3 Data Repositories (E-Research)

The sharing of data is not a new topic in research and policy circles and dates at least back to the 1980s. "Data sharing" for [Borgman 2011, p. 3] "is the release of research data for use by others". It has become a critical element in getting research grants with many funding organisations e.g. [Wolski et al. 2011, p. 1], [Ginty et al. 2012, p. 3–4], where often the "ability to articulate what her or his data are, how they will be managed, how they will be shared, and if not shared why, will influence whether or not a project is funded" [Borgman 2011, p. 3–5]. A European high level expert group on Scientific Data submitted a report, *Riding the wave – How Europe can gain from the rising tide of scientific data* to the European Commission [EC-SDI Report 2010] to provide a vision and an action plan. "Like Australia, the United States, the United Kingdom, Canada, the European Union, and many countries in Asia are investing heavily in the development of national infrastructures" [PMSEIC 2006, p. 9].

"The data deluge has arrived" [Borgman 2011, p. 2] in the wider scientific, i.e. as well in the CRIS community[108] and particularly with repositories – these are increasingly utilised for storage of research data, and a registry of data repositories has just started to being populated, namely DataCite[109], set up as a non-profit organisation towards "building on the approach developed by the German National Library of Science and Technology (TIB) and promote the use of Digital Objects Identifiers (DOI) for datasets". Like publication repositories, data repositories also aim to harvest via OAI-PMH, based on DublinCore (OpenAIREplus – the EC Open Access pilot – is being streamlined towards research data as previously indicated).

[108] CERIF for Datasets (C4D) project: http://cerif4datasets.wordpress.com/about/ and the DMP project: http://datamanagementplanning.wordpress.com/ (both JISC funded projects in the UK) to investigate research data, and where with C4D, a proposal for a Dataset ontology based on CERIF has been developed in a first draft [Bokma & Garfield 2012, p. 11, fig. 3], indicating storage in the repository, recognising subject classificatons and identifying relevant contextual entities such as project, organization, person and publication. (Last visit: April 20th, 2012)

[109] DataCite – Helping you to find, access, and reuse research data: http://datacite.org/ (Last visit: April 20th, 2012)

Managing and sharing of research data requires thorough analysis; and "the largely ad hoc approach to managing such data [...] is now beginning to be understood as inadequate" [Uhlir and Schröder 2007, p. 36]. Thus, modeling research data is considered "an increasingly challenging task that warrants a rethinking of its design" [Li et al. 2010, p. 137].

The DataCite list of repositories provides in-sight in the multiplicity of the different types of subject areas contributing *data*; spanning from technical sciences, climate, fluid dynamics, earth sciences, social, historical, political, arecheological, economic and ecological, marine, oceanography, environmental, genomics, bioinformatics, astronomy, chemistry, child care, education, chrystallography, atmospheric, geo sciences, to humanities, conservation, demographic, population, psychological health and medical care, drug addiction, illnesses, crimonology, forestry, geophysics, glaciological or earth sciences, earth quakes, civil engineering, neuroscience, biochemistry, literature and linguistics, visual arts, timeseries data, and many more. [Borgman et al. 2011, pp. 26 ff.] explain: "Data take many forms, both physical and digital. They are much more than numbers in a spreadsheet: data can be samples, software, field notes, code books, instrument calibrations, archival records, or a myriad of other information objects, none of which may stand alone", and presents four rationales behind their collection:

1. to reproduce or to verify research
2. to make the results of publicly funded research available to the public
3. to enable others to ask new questions of extant data
4. to advance the state of research and innovation

To indicate the complexity behind data collections, an analysis of data practices by [Borgman et al. 2011, p. 9, fig. 1] is presented. It illustrates multi-dimensionally the underlying purposes for collections (Figure 19). The dimensions are "neither exhaustive nor mutually exclusive", but reflect clearly the multiple fields or disciplines and thus the multiple approaches with data collections.

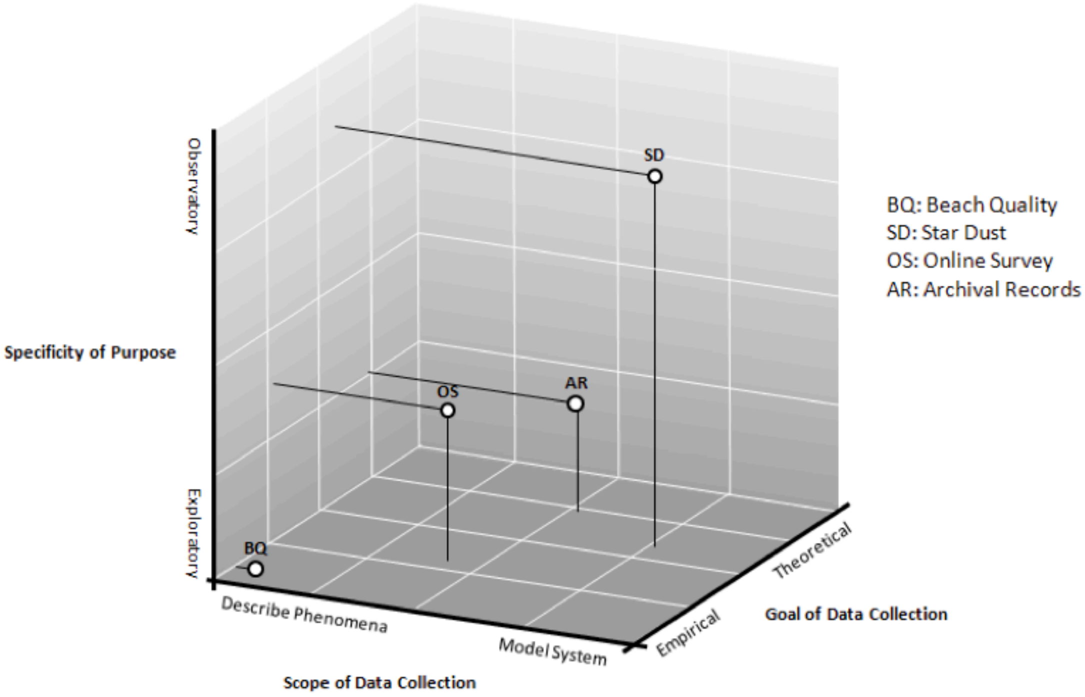


Figure 19: Purposes for collecting data. By J. C. Wallis [Borgman et al. 2003, p. 9, fig 1]

Besides scholarly repositories addressing the storage of research data in the scientific world, there is the ever growing Linked Open Data Cloud "using the Web to connect related data that wasn't previously linked, or using the Web to lower the barriers to linking data currently linked using other methods. More specifically, Wikipedia defines Linked Data as "a term used to describe a recommended best practice for exposing, sharing, and connecting pieces of data, information and knowledge on the Semantic Web using URIs and RDF."[110]

4.1.4 Digital Libraries

The Digital Library Manifesto as a component of the final deliverable of the DELOS model[111], subsumes under the term *Digital Library* "systems [that] range from digital object and metadata repositories, reference-linking systems, archives and content administration systems (mainly developed by industry) to complex systems that integrate advanced digital library services (mainly developed in research environments) [...] as yet there is no agreement on what Digital Libraries are and what functionality is associated with them" resulting in a

[110] Linked Data – Connect Distributed Data across the Web: http://linkeddata.org/ (Last visit: December 28th, 2011)

[111] The Digital Library Reference Model (DELOS) work was funded by the European Commission and had been continued in the DL.org project – Digital Library Interoperability, Best Practices and Modeling Foundations: http://www.dlorg.eu/ until December 2010. (Last visit: July 24th, 2011)

lack of interoperability and reuse of content and technologies [Candela et al. 2011, p. 13]. The so-called "Digital Library Universe" builds on main concepts such as: content, functionality, user, quality, architecture, policy, and is organized according to a tree-tier framework [Candela et al. 2011, pp. 19–22, figs. I.3-1; I.4-1]. For each of the main concepts an individual concept map has been designed.

Figure 20 shows the so-called DELOS "Content Domain Concept Map" from where it is clear, that *Information Objects* are in the center of interest with *Digital Libraries*. Each belongs to a *Collection*, with extensions in *Resource Sets* and implied by *Intensions* through queries expressed upon the Information Object, which is additionally expressed through an ontology, and where the entire *Content Domain* itself consists of these *Information Objects* (managed and defined by the underlying system), which have *Manifestations*, *Views*, and *Editions*, and are understood as basically being *Resources*. For *Resource*, there exists a separate DELOS Resource Domain Concept Map, which is indicated in the upper background of Figure 20 – taking into account the purpose associated with a *Resource*.

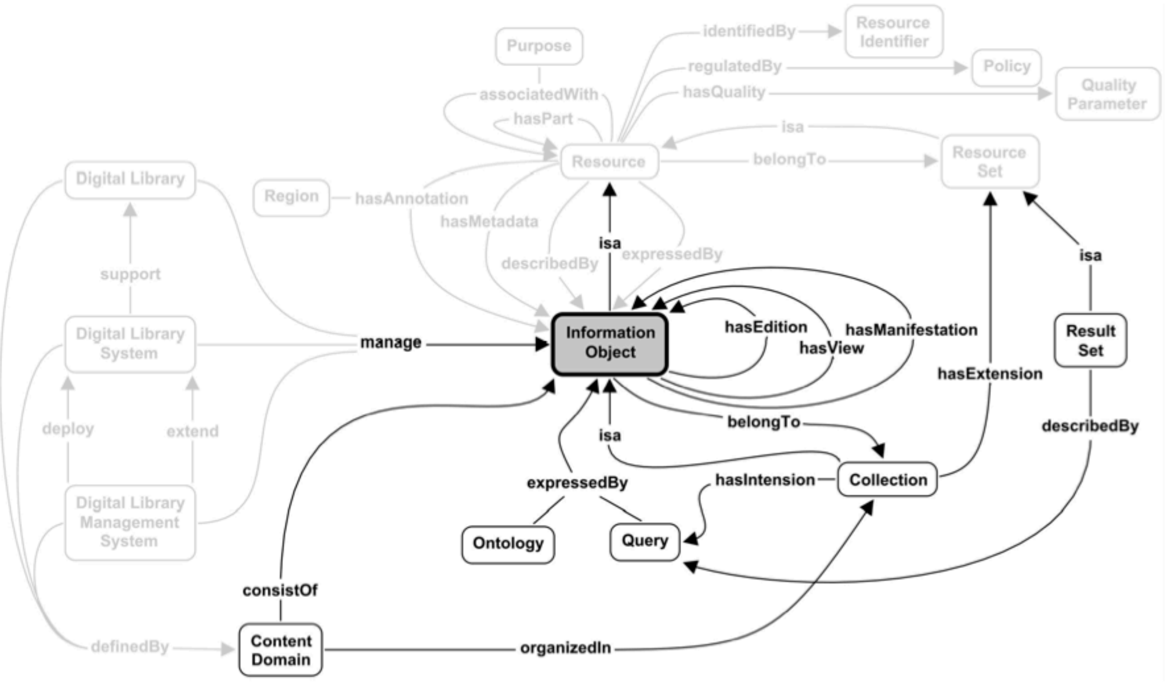


Figure 20: Content Domain Concept Map [Candala et al. 201, p. 37, fig II.2-3]

Figure 20, reveals obvious differences between reference models and domain models. Where the former aim at taking into account functionalities, user views, qualities, architectures, or policies, the latter concentrate on describing the domain as a perceived world of entities, entirely agnostic as to possible applications, functionalities, etc. but focused on describing the entities and their relationships at a meta (domain) level. The *Information Object* in Digital

Libraries as a *Resource* – resembles *Information Resources* in the UNISIST model and thus intersects with the repository concept, whereas CRISs are more dedicated to contextual views with e.g. Person in the role of Author, or Editor e.g. associated with an information object or other research objects. The DELOS model does not seem to maintain such additional entities in the system, but concentrates on the management of *Information Objects* in the manner of repositories. The reference model is not aimed at being used for implementation, but – as its name indicates – for reference with communication and stakeholders representing and applying systems and views; it is thus more of a foundational nature [Thomas 2006, p. 7].

A domain model for digital libraries is available with Europeana, a very popular EC-funded project in the Digital Humanities to manage Cultural Heritage objects. It builds on latest SW and LOD technologies, i.e. assumes an open world. "[The Europeana Data Model] EDM is an attempt to transcend the respective information perspective of the various communities constituting Europeana, such as museums, archives, audio-visual collections and libraries. EDM is not built on any particular community standard but rather adopts an open, cross-domain Semantic Web-based framework that can accomodate particular community standards such as LIDO, EAD or METS" [EDM Primer 2010, p. 5][112]. The EDM class hierarchy in Figure 21 shows the employed concepts and Web standards.

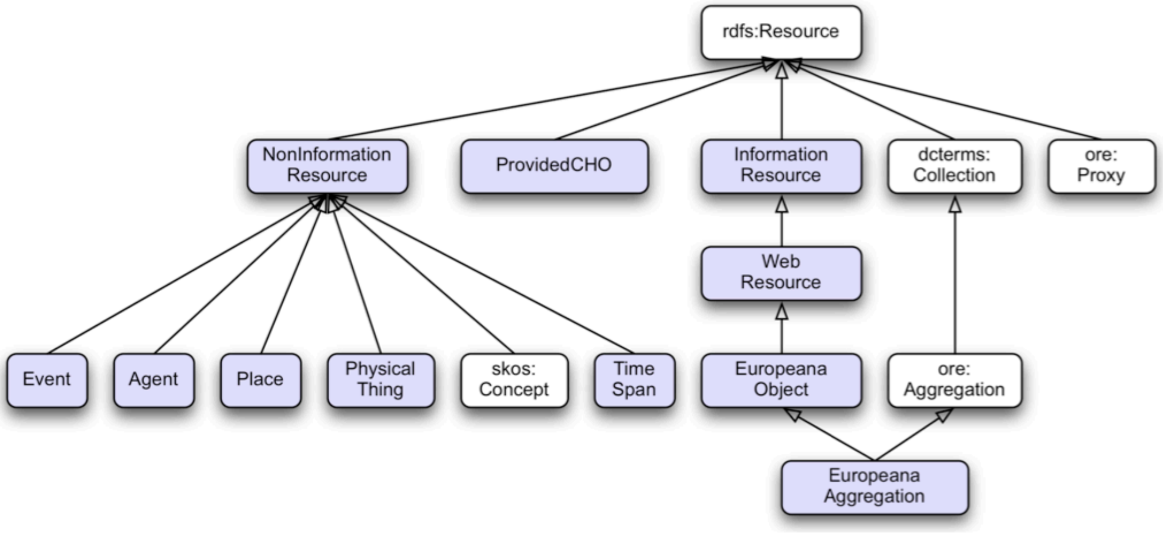


Figure 21: The EDM Class hierarchy [EDM 2012, p. 6, fig. 1][113]

[112] Europeana Data Model Primer: http://pro.europeana.eu/documents/900548/770bdb58-c60e-4beb-a687-874639312ba5 (Last visit: December 28th, 2011)

[113] Europeana Data Model v. 5.2.3: http://pro.europeana.eu/documents/900548/bb6b51df-ad11-4a78-8d8a-44cc41810f22 (Last visit: December 2nd, 2012)

The classes introduced by EDM are shown in light blue rectangles. The classes in white rectangles are re-used from other schemas: the schema is indicated before the colon [EDM 2012, p. 6, fig. 1]

Figure 21, shows the *rdfs:Resource* concept at the top, holding an *rdfs* namespace, thus refering to the RDF standard (which will be elaborated a bit within section 4.3 Architectural Styles – 4.3.2 The Semantic Web) from which the defined concept is borrowed. The EDM distinguishes between *InformationResource*, *NonInformation Resource* (inline with the W3C's TAG group kinds), *ProvidedCHO*, *dcterms:Collection* and *ore:Proxy*. It considers *WebResource* an *EuropeanaObject* and an *EuropeanaAggregation*, that is, *ore:Aggregation* upon *dc:termsCollection*, whereas *NonInformationResource* covers *rdf:Resources* different from *InformationResources* or *WebResources*, such as *Event*, *Agent*, *Place*, *PhysicalThing*, *skos:Concept*, *TimeSpan*. The EDM employs ontological constructs such as *skos:Concept*, *Time Span*, *Physical Thing*, *Place, dcterm:collection*, *ore:Proxy* and design-wise resembles FERON more than any other presented approach. This is surely due to the underlying technology or modeling grammar and constructs – being a formal ontology (as indicated by the top concept through the rdfs namespace), but also, because the EDM mostly describes a perceived world and not a particular system. However, the EDM is not entirely consistent with respect to domain descriptions and functional descriptions, e.g. it employs functional *technology* such as *ore:Aggregation* in parallel to domain description classes such as *Information Resource*.

4.1.5 Digital Archives

The Open Archives Initiative has developed and promoted interoperability standards for more than 10 years, the most popular is the OAI-PMH – the Open Archives Initiative Protocol for Metadata Harvesting. Through the Object Reuse and Exchange solution (OAI-ORE) they provide a technical foundation for the handling of aggregations of Web resources towards exchange by URIs, in order to overcome the problem of aggregation identity. That is, the union of aggregation elements that "describe the constituents or boundary of an aggregation" are described by a document as to which resources are part of the aggregation and which are merely related to it [Lagoze & Van de Sompel 2008][114]. The OAI-ORE approach is built on RDF and therefore follows an open world assumption.

The Consultative Committee for Space Data Systems (CCSDS) recommends OAIS – an ISO Reference Model for an Open Archival Information System. [CCSDS Blue Book 2002, p. 1-1] understand *archives* as a "wide variety of storage and preservation functions and systems

[114] Open Archives Initiative – Object Reuse and Exchange: http://www.openarchives.org/ore/1.0/primer
(Last visit: December 2nd, 2011)

[...] preserve records, originally generated by or for a government organization, institution, or corporation, for access by public or private communities [...] in such forms as books, papers, maps, photographs, and film [...] The major focus for preserving this information has been to ensure that they are on media with long term stability and that access to this media is carefully controlled" [p. 2-1]. The model resembles the DELOS reference model greatly, in that it models functions such as e.g. *Data Management*, *Archival Storage*, *Administration*, *Preservation Planning*, *Access* upon *Data* and *Information Objects*. It understands the *Information Object* yielded from *Representation Information* being interpreted from a *Data Object* keeping in mind the *Producer* and *Consumer* view, thus, additionally resembling the UNISIST model; these are all reference models and in that of a foundational nature [Thomas 2006, p. 7].

The OAIS model classifies the *Information Object* in a taxonomy (Figure 22). As a reference model it was not meant for a particular implementation design, discipline or organisation, but "[S]tandard[s] developers are expected to use this model as a basis for further standardization in this area" [CCSDS Blue Book 2002, p. 1–2].

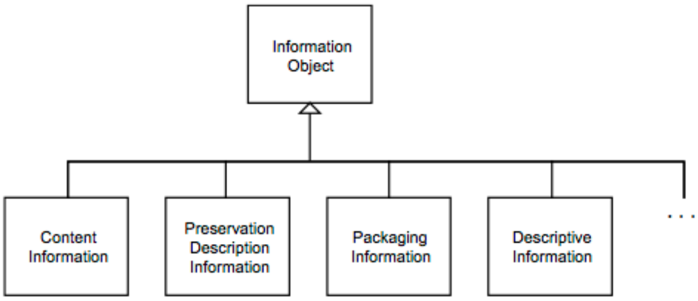


Figure 22: OAIS Information Object Taxonomy [CCSDS Blue Book 2002, p. 4-24, fig. 4-12]

4.1.6 Learning Management Systems

[van Godtsenhoven et al. 2008] introduce IMS-CP (version 1.1.4) as „the de facto standard for packaging educational or learning content for transport across Learning Management Systems (LMSs) and Virtual Learning Environments (VLEs)" (p.131)[115]. The latest public IMS-GLC HTML version[116] defines Learning Information Services (LIS) as to "how systems

[115] [van Godtsenhoven et al. 2008, p. 132] explain that "it is usually implemented with the metadata-set defined in IMS-Learning Resource Meta-Data (IMSMD) specification v1.2.1 or IEEE 1484.12.3 standard for XML Schema binding for Learning Object Metadata (LOM) defined in IEEE 1484.12.1", and that discrepancies "between IMSMD and IEEE LOM [...] have been realigned" (p. 133).

[116] IMS GLC Learning Information Service Specification Version 2.0 – Final Release Version 1.0 (Date Issued: 30 June 2011): http://www.imsglobal.org/lis/lisv2p0/CMSv1p0InfoModelv1p0.html (Last visit: May 22nd, 2012)

manage the exchange of information that describes people, groups, memberships, courses and outcomes within the context of learning [...] There is no such thing as a Course object. Instead Courses are reflected in four types of object[s] each of which has its own SourceId." These four types of course objects are defined [as in Figure 2.2 Structure of a Course][117]:

- **Course Template:** identification of the basic and definition of the course e.g., Biology 101. A course template will, in general have one or more course offerings associated with it.
- **Course Offering:** the allocation of the course to an academic session e.g., Maths 101 Semester 1. A course offering will, in general, have one or more course sections associated with it.
- **Course Section:** the assignment of teaching resources to the scheduled activities that constitute the course e.g. English 101 Semester 2 Seminars.
- **Association:** the association of two or more course sections for some educational purpose.

The "unique official formal standardization body for e-Learning at international level" is ISO/IEC JTCI SC36 [Stracke 2010, pp. 359 ff.], defined as "Standardization in the field of information technologies for learning, education and training to support individuals, groups, or organizations, and to enable interoperability and reusability of resources and tools", and subsequent, there is ISO/IEC 19788 – the MLR (Metadata for Learning Resources) standard – 'a multi-part' standard and its first part provides the General Framework for Metadata and Application Profiles that is completely interoperable and compatible with Dublin Core".

[117] IMS GLC Course Management Service Information Model Version 1.0: http://www.imsglobal.org/lis/lisv2p0/CMSv1p0InfoModelv1p0.html#_Toc297634773 (Last visit: December 2nd, 2011)

The COLIS (Collaborative Online Learning and Information Services) activities back in 2003 pictured systems as in Figure 23, where relationships to LOM repositories are given, and where also the library domain is visible in that it provides the catalogues through which the learner gets hold of the harvested content within the Learning Management System. The entire learning environment thus overlaps with UNISIST, DELOS, Europeana and the OAIS models, in that it is also concerned with managing of information resources.

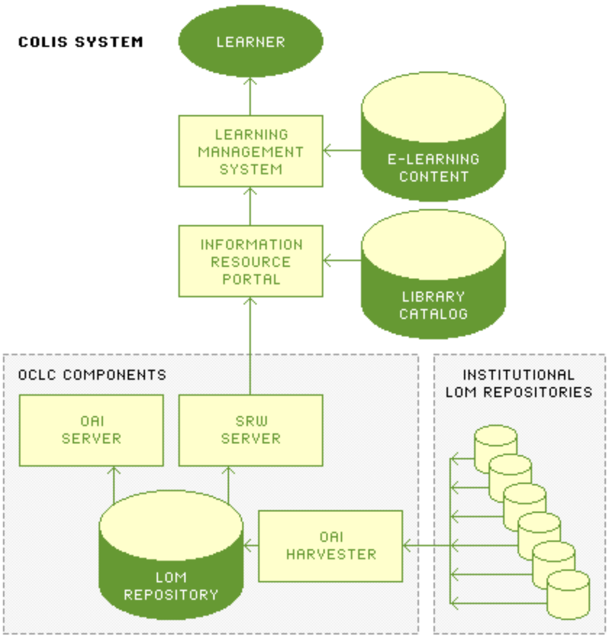


Figure 23: COLIS System Architecture[118]

4.2 Information System Architectures

In Information Systems Science, systems are usually planned and built in a top-down fashion [Björk 2007, p. 312], in their earlier days they mostly concentrated on *information processing* strictly for particular problem solving activities. However, it has been recognised that beyond a strict information processing view an advanced knowledge understanding is crucial for continuous innovation and i.e. the success of a company. “In an economy where the only certainty is uncertainty, the one sure source of lasting competitive advantage is knowledge [...] the holistic approach to knowledge at many Japanese companies is also founded on another fundamental insight. A company is not a machine but a living organism. Much like an individual, it can have a collective sense of identity and fundamental purpose” [Nonaka 1991, p. 96]. Two very well-known names with (management) information system frameworks and architectures are John A. Zachman, through his *Framework for information systems architectures*[119] [Zachman 1987] and August Wilhelm Scheer behind ARIS –

[118] COLIS was a project of the Australian national government focusing on the development of collaborative online learning and information services, where OCLC – The world’s libraries connects was a member. The goal of the project was a Metadata Switch, i.e. metadata mappings: http://www.oclc.org/research/activities/past/orprojects/mswitch/5_colis.htm (Last visit: May 28th, 2012).

[119] Zachman Framework: http://en.wikipedia.org/wiki/Zachman_Framework (Last visit: January 4th, 2012)

Architecture of Integrated Information Systems[120] [Scheer 1991] for business process optimisation, supplying starting points or guidance for information system development, as well as MDAs (see 4.2.3). This work has been inspired[121] and guided by the three-layer ARIS concept, in that conceptualisation guides formalisation before implementation – continuously and vertically repeated. It is considered inline with the paradigm of MDAs (see 4.2.3), where formal models are core for interoperation or networked systems with loosely coupled vertical levels. This also highly intersects with Zachman's framework. As indicated, there is a clear distinction between a perceived world that is modeled in a domain ontology, and a prescriptive information system *specification* model. The task of this work is a description of two perceived (integrated) domains, but not the design of a system to manage the information in the two domains, i.e. user needs, access control, archiving, interfaces. FERON is a work at the conceptual level – and the introduced architectural frameworks are intended for guidance and awareness towards subsequent system implementation; which is considered crucial with conceptual modeling.

4.2.1 Zachman Framework for Information Systems Architecture

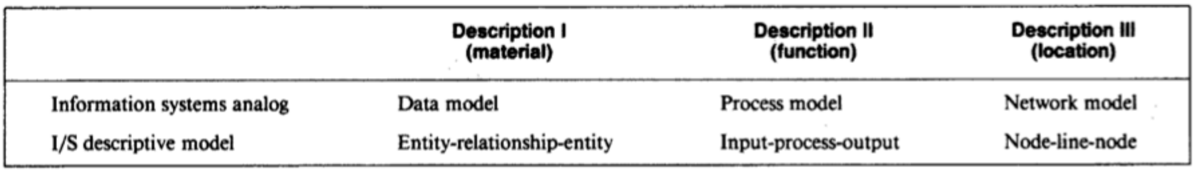
To manage the increasing size and complexity with information system implementations, [Zachman 1987] proposes to use "some logical constructs (or architecture) for defining and controlling the interfaces and the integration of all of the components of the system", but, which he presents as discipline agnostic; independent of "strategic planning methodology" (p 276). He developed "an objective, independent basis" by consulting the field of architecture itself, starting from *Bubble Charts* – equal to concepts (p. 280) – in the sense of "I'd like to build a building" – which are then transcribed into *Requirements* resulting in a work breakdown structure that equals an *architect's drawings* replicating an *owner's perspective* which is then further continued in an *architect's plan* which equals the *engineering design*, that is – the *designer's view*. Zachman's framework (1987) continues and is, though explained within a manufacturing context, developed generically. He identifies as important "three fundamental architectural representations, one for each "player in the game", that is, the owner, the designer and the builder": "The owner has in mind a product that will serve some purpose. The architect transcribes this perception of a product into the owner's perspective. Next the architect translates this representation into a physical product, the

[120] Architecture of Integrated Information Systems (Last visit: January 4th, 2012): http://en.wikipedia.org/wiki/Architecture_of_Integrated_Information_Systems

[121] The author studied Information Systems with August Wilhelm Scheer at the University of Saarbrücken.

designer's perspective. The builder then applies the constraints of the laws of nature and available technology to make the product producible, which is the builder's perspective" (p. 281).

[Zachman 1987, p. 281] recognises as a significant observation of these architectural representations "that each has a different *nature* from the others. They are not merely a set of representations, each of which displays a level of detail greater than the previous one. [...] each of the architectural representations differs from the others in *essence*, not merely in level of detail" the first being described as to *WHAT* the thing is made of, the second as to *HOW* the thing works, and the third as to *WHERE* the flows or connections exist and applies them to the world of information systems (Figure 24), where *material* is "stuff the things is made of" and function "would likely be called a process" and location "would likely be called the network model, in which the focus is on the flows (connections)" (p. 283).



	Description I (material)	**Description II (function)**	**Description III (location)**
Information systems analog	Data model	Process model	Network model
I/S descriptive model	Entity-relationship-entity	Input-process-output	Node-line-node

Figure 24: Information systems analogs [Zachman 1987, p. 283, table 4]

This work aims at modeling a perceived world of Research in general and LT in particular, agnostic of any business model or business strategy. In that, it follows Zachman's material description, describing *WHAT* Research in general and LT in particular is made of.

4.2.2 Architecture of Integrated Information System (ARIS)

Similarly to Zachman, Scheer with ARIS stresses the importance of different vertical levels and horizontal views towards domain understanding and finally for design and system implementation. These levels are called *Requirement Definition*, *Design Specification*, and *Implementation Description* and are often reflected as pictured in Figure 25, through *Conceptual*, *Technical* and *Physical* levels. Horizontally, ARIS employs views upon *Organization*, *Data*, *Controll* or *Process*, and *Functions*, towards system integration within organisational boundaries. Each view repeats the three introduced levels in that a vertical relationship of loosely coupling (backwards and forwards) exists, much in the spirit of Model-driven Architectures (MDAs) and Zachman's representation levels. ARIS therefore provides a comprehensive and holistic view over organisation-relevant information system components for integration, but at the same time views single items towards their contribution to the whole, and enables thus an understanding of the whole by introducing the single items (throughout aware of the three levels of integration) but started conceptually from the top; that is, with an idea or the need and purpose for the system setup. In the spirit of *WHAT* is the problem to be solved, which has to be logically or technologically described, and finally prescriptive enough for a physical implementation that supports the initial concept or idea. During the whole process, the architect must be aware, that changes in the implementation and logic may or may not imply changes in the concept.

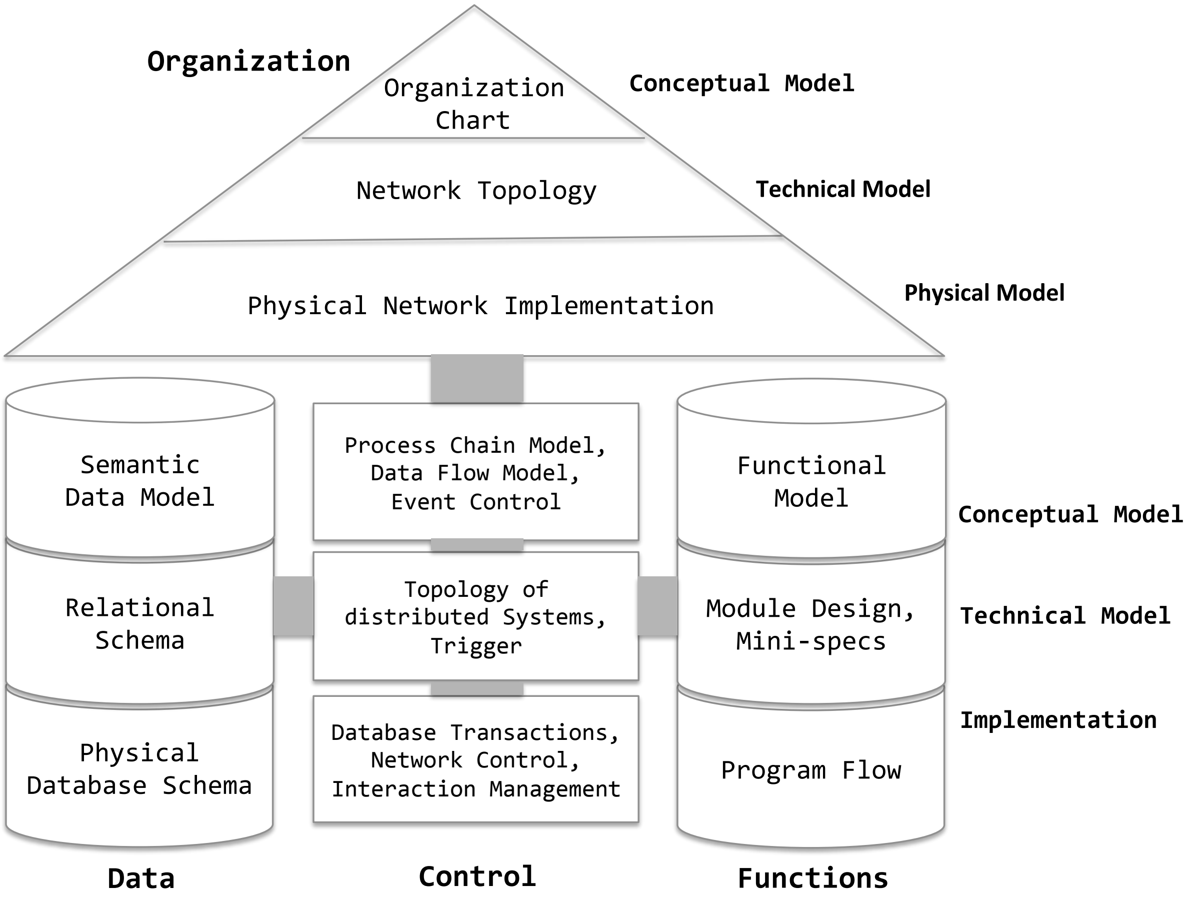


Figure 25: Architecture of Integrated Information Systems (ARIS) [Scheer 1991]

This work aims at modeling a perceived world of Research in general and LT in particular, agnostic of any business model or business strategy, and in that it follows the ARIS Conceptual Model.

4.2.3 Model-driven Architectures

Model-driven Engineering (MDE) is a variant of refinement-based[122] software development in which models are loosely coupled, but connected in a systematic way. The *refinements* may be what [Fielding 2000, p. 7] calls design documentation "architectural design and source code structural design, though closely related, are separate design activities" referring to [Perry & Wolf 1992], who call it rationale what influences the evolution of an architecture, but doing so in replacing lower-levels contrary to [Perry & Wolf 1992] where it is part of the architecture itself. MDA as a specific incarnation of MDE is an approach to application modelling. The most significant feature is the independence of system specifications from implementation technologies or platforms. "The system definition exists independently of any implementation model and has formal mappings to many possible platform infrastructures (e.g., Java, XML, SOAP)" [Poole 2001, p. 2]. The OMG[123] MDA [Brown 2004, p. 318] follows four principles:

- Models expressed in a well-defined notation are a cornerstone to understanding of systems for enterprise-scale solutions.
- The building of systems can be organized around a set of models by imposing a series of transformations between models, organized into an architectural framework of layers and transformations.
- A formal underpinning for describing models in a set of metamodels facilitates meaningful integration and transformation among models, and is the basis for automation through tools.
- Acceptance and broad adoption of this model-based approach requires industry standards to provide openness to consumers, and foster competition among vendors.

"To support these principles, the OMG has defined a specific set of layers and transformations that provide a conceptual framework and a vocabulary for MDA. Notably, OMG identifies four types of models: Computation Independent Model (CIM), Platform Independent Model (PIM), Platform Specific Model (PSM), and an Implementation Specific

[122] "However, since MDA discerns platform-specific information as the main criterion for refinement, the entire process is much more structured than the „free-style" refinement of the 1970s. Also, in MDA, all models are graph-based, while standard refinement worked mainly for syntax trees." [Aßmann et al. 2006, p. 252]

[123] "Founded in 1989, the Object Management Group, Inc. (OMG) is an open membership, not-for-profit computer industry standards consortium that produces and maintains computer industry specifications for interoperable, portable and reusable enterprise applications in distributed, heterogeneous environments. [...] OMG specifications address middleware, modeling and vertical domain frameworks." [OMG 2008]

Model (ISM)" [Brown 2005, p. 318] and by that, goes one step further towards applications than ARIS (Figure 25). Because "MDA discerns platform-specific information as the main criterion for refinement" [Aßmann et al. 2006, p. 253], it is therefore clearly distinguished from just refinement-based software development.

The MDA guided the approach of this work to formally describe two complex domains of interest – anticipating subsequent specific implementations (MDA's CIM level).

4.3 Architectural Styles

Networked information systems have to deal with high complexity and a robust setup requires a thorough analysis of well-established architectural methods to perform the desired tasks, and the guideline that "form follows function" should never be ignored [Fielding 2000, p. 1]. Multiple so-called architecture definition languages (ADLs) have been developed. [Di Nitto and Rosenblum 1999] exploited them for analysis and design processes towards networked systems – "*middleware-induced architectural styles*" – where "[a] style defines a set of general rules that describe or constrain the structure of architectures and the way their components interact". Their unexpected conclusion was "that the top-down approach adopted by the software architecture community in the development of languages and tools seems in many ways to ignore the results that practitioners have achieved (in a bottom up way) in the definition of middlewares" [Di Nitto & Rosenblum 1999, p. 9]. Motivated by that, [Fielding 2000] started by investigating the existing architectural styles to demonstrate with his doctroral thesis how one particular style (the Representational State Transfer (REST) architectural style) has been used to guide the design and development of the architecture for the Web, and describes the lessons learned from applying REST. This work refers to Fielding's seminal and famous work addressing architectural issues, but will not further investigate them: "Software architecture research investigates methods for determining how best to partition a system, how components identify and communicate with each other, how information is communicated, how elements of a system can evolve independently, and how all of the above can be described using formal and informal notations." Fielding's work addresses architectural issues in a Web environment inspired by Berners-Lee's (1996)[124] idea about the Web, being "a shared information space through which people and machines could communicate".

[124] The World Wide Web: Past, Present and Future: http://www.w3.org/People/Berners-Lee/1996/ppf.html (Last visit: July 1st, 2012)

It is anticipated that conceptual models are preserved in architectural networks, i.e. in the components and connectors referring to the processing of their elements. A formal domain ontology will therefore be of high value for semantic integration with information system architectures such as the Web, the Semantic Web and the Linked Open Data Web (see Figure 15), and furthermore, e.g. support orchestration of services with SOAs, direct responses in Question-answering (QA) systems, guide Emergent Software Architectures and distributed (harvesting) systems, but also Community-driven and thus Crowd-sourcing systems, Social Networks and e-Infrastructures; these are subsumed under the wider concept of architectural styles and not all investigated in detail. At this point, a brief explanation of the WWW is provided because it is the basis of the Semantic Web and its extension the Linked Open Data Web (both are technologically-driven by formal ontologies – the modeling method applied in this work).

4.3.1 The World Wide Web

The World Wide Web (WWW) known as *the Web* "is a system of hyperlinked documents accessed via the Internet" that emerged from a proposal by Tim Berners-Lee, then working with CERN[125]. A Web document or so-called hypertext document is structured through underlying markup defined in the Hypertext Markup Language (HTML), a W3C standard[126] and recommendation for a document structure to organise the content of Web documents. HTML provides a syntax for document descriptions through embedded markup elements such as <head> <title> <body> or <p> for paragraph and e.g. <a> (anchor) achieving linkage. However, it does not provide a specification for semantics. Therefore, HTML documents aka Web pages have often been extended by Dublin Core elements to semantically enhance the documents beyond just structural features. Dublin Core provides metadata elements such as subject or creator – these can be embedded in the HTML document head for recognition by search engines. The first W3C HTML recommendations[127] date back to 1997. These have become incrementally stricter in formality and with validation towards well-formedness, increasingly employing XML and integrating formal semantics – thus, slowly transforming the World Wide Web into a Semantic Web (SW) and into a Web of Linked Open Data (LOD).

[125] WWW: http://en.wikipedia.org/wiki/World_Wide_Web (Last visit: April 2ne, 2011)

[126] HTML: http://www.w3.org/TR/#tr_HTML (Last visit: April 2nd, 2011)

[127] HTML Recommendation: http://www.w3.org/TR/tr-technology-stds#tr_HTML (Last visit: April 2nd, 2011)

4.3.2 The Semantic Web

The Semantic Web (SW) is aimed at adding machine-readable meaning to the current Web. The vision behind is reflected in the definition by [Hendler et al. 2002]: “The Semantic Web is an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation. It is based on the idea of having data on the Web defined and linked such that it can be used for more effective discovery, automation, integration and reuse across various applications.”[128] The SW is often explained to incorporate the transformation of the *Web of documents* towards the *Web of data*, and refers to the vision of the Web of Linked Data (LOD). According to [Wilks 2008, p. 41], “the concept of the Semantic Web has two distinct origins, and this bifurcation persists in two differing lines of SW research: one closely allied to notions of documents and natural language processing (NLP) and one not”[129]. This strongly overlaps with the view of [Marshall & Shipman 2003, p. 1] who attribute SW developments to “the anxiety over the apparent disorder of this new world of digital documents [...] A second comes from the field of Artificial Intelligence, with its maturing sense of the kinds of computation that can take place given formal representations”. For their analysis [Marshal & Shipman 2003, p. 57] structure the Semantic Web into three different kinds (Figure 26): (1) an information access scenario in which retrieval is supported by semantic metadata; (2) a globally distributed knowledge base in the sense of Berners-Lee; (3) an infrastructure for the coordinated sharing of data and knowledge.

Figure 26 takes a perspecive from a human and machine user distinguishing particular and universal knowledge, the former being “limited to the author’s original motivation for publishing something on the Web [...] the universal, useful in any context” [Marshal & Shipman 2003, p. 58].

[128] Since its inception, which is often attributed to the famous paper by [Berners-Lee et al. 2001] in the Scientific American, the Semantic Web has become a serious research subject e.g. [Wilks 2008, p. 41] with a growing number of conferences and journals, but also dedicated organisations and assocations. Although a killer application is still missing (maybe IBM’s Watson) will bring the breakthrough, the Semantic Web gets slowly down to business. “Semantic Web gets down to business” Computerworld: http://www.computerworld.com/s/article/9209118/The_semantic_Web_gets_down_to_business (Date issued February 22nd, 2011)

[129] Wilk’s assumption refers to the philosophy of Ludwig Wittgenstein, that natural langauge is human’s primary method of conveying meaning and that other methods of conveying meaning (formalisms, science, mathematics, codes, and so on) are parasitic upon it. His point is that the Semantic Web inevitably rests upon some technology within the scope of IE to annotate raw texts to derive entity types, then fact databases, and later ontologies; thus blurring the distinction between language and KR. According to Karen Spärck Jones in “What’s new about the Semantic Web” (2004) „words stand for themselves“ and not for anything else and therefore cannot be recorded in a general way, especially if it is content in a specific domain. According to Wilks, Spärck Jones put it mischieveously, IR has gained from “decreasing ontological expressiveness”. “We still ‘dial’ numbers when we make a phone call. Even though telephones no longer have dials; so not even number-associated concepts are safe from time; according to Wilks predicates don’t mean this year what coders meant by them 20 years earlier. The Semantic Web present offers no solution to this problem” [Wilks 2008].

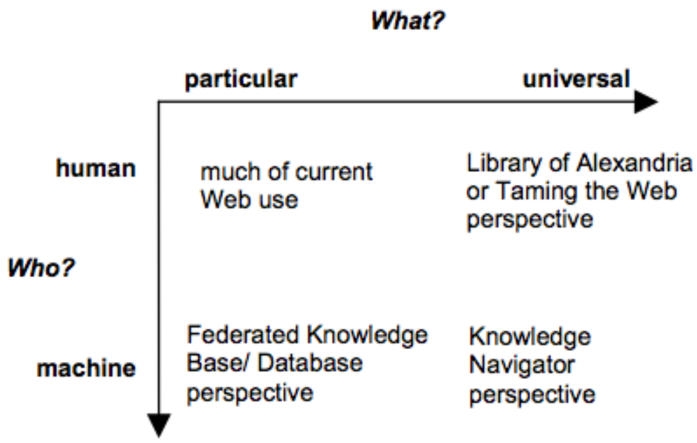

What?
particular
universal
human
much of current Web use
Library of Alexandria or Taming the Web perspective
Who?
machine
Federated Knowledge Base/ Database perspective
Knowledge Navigator perspective

Figure 26: Three perspectives on the Semantic Web [Marshal & Shipman 2003, p. 2, fig 1]

Most often the SW has been explained in technical terms, by the famous layer cake, which underwent substantial changes and variations during recent years. The SW architecture as shown in Figure 27 is from the perspective of software applications and components [Harth et al. 2010, p. 11].

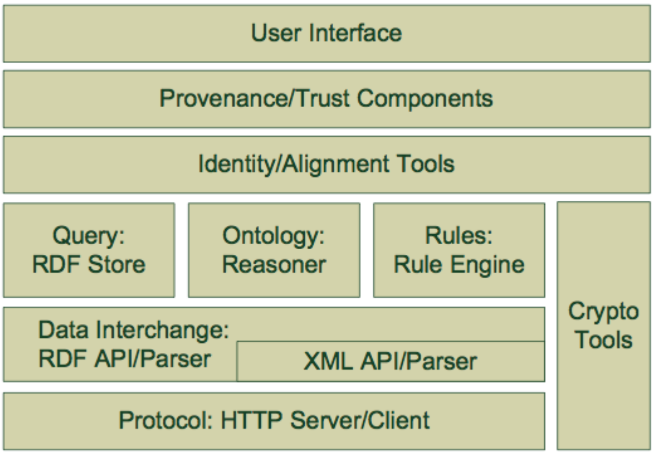


Figure 27: Semantic Web Components Architecture [Harth et al. 2010, p. 11, fig. 1.2]

Within this section, the Semantic Web is introduced – including its structure, that is, the syntactic representation of information – it builds on RDF – the Resource Description Framework, often implemented in XML, as indicated in Figure 27. In RDF, descriptions are formally realised in triples, originating from knowledge representation, artificial intelligence, data management, conceptual graphs, frames, and relational databases [Manola & Miller 2004][130].

[130] RDF Primer: http://www.w3.org/TR/rdf-syntax/ (Last visit: April 2nd, 2011)

“RDF is based on the idea of identifying things using Web identifiers (called *Uniform Resource Identifiers*, or *URIs*), and describing resources in terms of simple properties and property values. This enables to represent simple statements about resources as a graph of nodes and arcs representing the resources, and their properties and values [Figure 28]”.

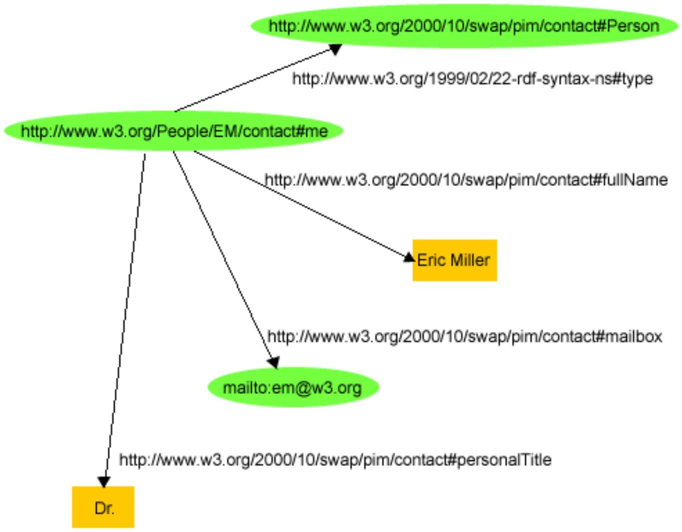

http://www.w3.org/2000/10/swap/pim/contact#Person
http://www.w3.org/1999/02/22-rdf-syntax-ns#type
http://www.w3.org/People/EM/contact#me
http://www.w3.org/2000/10/swap/pim/contact#fullName
Eric Miller
http://www.w3.org/2000/10/swap/pim/contact#mailbox
mailto:em@w3.org
http://www.w3.org/2000/10/swap/pim/contact#personalTitle
Dr.

Figure 28: An RDF Graph describing Eric Miller [Manola & Miller 2004]

The corresponding XML representation of Figure 28 is given with Notation 6.

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
             xmlns:contact="http://www.w3.org/2000/10/swap/pim/contact#">
  <contact:Person rdf:about="http://www.w3.org/People/EM/contact#me">
    <contact:fullName>Eric Miller</contact:fullName>
    <contact:mailbox rdf:resource="mailto:em@w3.org"/>
    <contact:personalTitle>Dr.</contact:personalTitle>
  </contact:Person>
</rdf:RDF>
```

Notation 6: RDF/XML describing ERIC Miller [Manola & Miller 2004]

The RDF syntax has been applied and extended to OWL, the Web Ontology Language – a W3C recommendation in its second generation *OWL 2*[131] for the Semantic Web with formally defined meaning. Where OWL 1 uses URIs (Uniform Resource Identifiers), OWL 2 uses IRIs (Internationalized Resource Identifiers); these must be absolute, i.e. not relative (section 2.4) [Motik et al. 2009]. OWL 2 is defined to use the datatypes from XML Schema

[131] OWL 2 Web Ontology Language http://www.w3.org/2004/OWL/ (Last visit: April 2nd, 2011)

Definition Language (XSD) (assuming version 1.1 progresses towards recommendation[132]). The application of either "OWL 2 Direct Semantics" or "OWL 2 RDF-based Semantics" are acknowledged as to being two alternative ways of assigning meaning to OWL 2 ontologies, where the latter is fully compatible with the RDF Semantics, and where the "backwards compatibility with OWL 1 is, to all intents and purposes, complete: all OWL 1 Ontologies remain valid OWL 2 Ontologies, with identical inferences in all practical cases". OWL 2 adds 'syntactic sugar' [Motik et al. 2009] and offers new expressivity, including:

- keys
- property chains
- richer datatypes
- data ranges
- qualified cardinality restrictions
- asymmetric, reflexive and disjoint properties
- enhanced annotation capabilities

"OWL 2 also defines three new profiles and a new syntax [OWL 2 Manchester Syntax]. In addition, some of the restrictions applicable to OWL DL have been relaxed; as a result, the set of RDF Graphs that can be handled by Description Logics reasoners is slightly larger in OWL 2" [Motik et al. 2009].

4.3.3 The Linked Data Initiative

The Linked Data initiative aims at sharing structured data on the Web, where the Web is rapidly transforming from a Web of documents into a Web of data (interconnected by typed links) and where applications have to deal with an unbound, global data space, contrary to a fixed set of data sources [Bizer et al. 2009, p. 1–2]. This unbound data space is often called 'open world' and builds on RDF graphs, where anything can be connected to any thing, but where efficient and sustainable 'linking', 'grouping', or 'recording' is yet to be efficiently solved, and functions such as, how sets of data elements should be assembled into 'packages' or 'records' for particular applications towards community needs [Dunsire et al. 2011, p. 35] is yet to be specified.

(Berners-Lee 2006)[133] recommends a set of 'principles' for publishing data on the Web:

[132] The latest XML Schema Definition Language Recommendation has been published as a Candidate Recommendation on July 21st, 2011: http://www.w3.org/TR/xmlschema11-1/ (Last visit: April 2nd, 2011)

[133] Linked Data Initiative: http://www.w3.org/DesignIssues/LinkedData.html (Tim Berners Lee 2006). (Last visit: April 2nd, 2011)

- use URIs as names for things
- use HTTP URIs so that people can look up those names
- when someone looks up a URI, provide useful information,
  using the standards (RDF, SPARQL)
- include links to other URIs, so that they can discover more things

Linked data is closely related to the general Web architecture in terms of representing and describing *Resources*, as it builds on RDF. However, the particular resource type is *Data*, and consequently, contrary to the traditional understanding of library records, linked data are "more focused on statements rather than records" [Dunsire et al. 2011, p. 29].

4.4 Summary

Several kinds of information systems have been introduced to provide insight into different modeling approaches as well as into the domains that are content-wise related to the domain of interest, namely Research. This short excursion demonstrated the substantial overlap with managing information resources and other sources across the introduced domains. Each of them can certainly profit from cross-communication, CRISs and OA repositories have already started communication with joint projects and a recent public declaration. Beyond a problem solving approach from where information system analyses starts, architecturural frameworks support with managing the complexity that systems ought to be able to handle, and for which a three-layer architecture concept matured, which starts from informal or conceptual ideas (business-driven or problem-solving situations) towards implementation, with increased formal prescriptions and application commitments. The frameworks were not only presented to demonstrate the complexity, but also, to explain, at which of the levels of formality this work is to be found – namely at the highest conceptual level – with a minimal formality, but awareness of subsequent levels. This work is not aimed at modelling of an information system as a thing, but a perceived world that information systems ought to be able to model (see 3.1.1 bullet (1)). The difference is explained in some more detail and becomes more obvious from within section 6.1 where systems of the two kinds are discussed. Modelling FERON profitted from tools that supplied adequate modelling constructs, i.e. enable formal, founded modeling according to latest technological standards. FERON is clearly addressed at the human reader to enable the understanding of the two domains – namely Research in general and LT in particular – anticipating information system implementations and integration. FERON is introduced formally and faces the inherent complexity in a structured way. It enables a required efficiency, quality and hence sustainability as well as re-usability with dedicated setups and interfaces. FERON is thus applicable for further machine processing and implementations.

5 Analysis of Research Entities

"No entity without identity."

[Quine 1969 in Ontological Relativity]

Research is a global activity. The involved entities and related processes overlap between nations and regions. A research project in one country is likely to be based on previous research in several other countries; and many research projects are transnational. Knowledge about research activities in one country may influence the scientific strategy including the priorities and resources provided, in another country. There is the obvious need to share research information across countries or between organisations within one country. Research Information (RI) is information about entities spanning the Research domain, i.e. information about scientists, funding opportunities, ongoing and completed projects, events, facilities, equipments, products, patents, results or outcome, services, resources, and many more (see Figure 29). RI is of interest to a multiplicity of stakeholders; researchers, research managers, research strategists, publication editors, intermediaries or brokers, the media, and thus the general public, and it is as such a significant factor for decision makers and strategists in the scientific environment, and beyond the academy for society as such. Having such an impact, RI has to be collected thoughtfully and carefully and to be preserved systematically, in order to support society and the individuals within most effectively [Zimmermann 2002], [Hornbostel 2006], [EUROHORCS-ESF 2008]. Scientific activity aims at wealth creation and improvements in the quality of life. Science in Society has been identified as an important research area. At European level, this has been reflected in the *Capacities* modul of the 7th Framework Programme[134]. This important aspect is taken into account by embedding Research into a societal context. The intersections with society are manifold and Figure 29 is not meant to be exhaustive; in other contexts there may exist more or less of the identified entities. There is a reference to the 'blurring' boundaries and increased interdisciplinarity of science, i.e. where the commercial sector enters into the scientific arena. It is therefore a consequence, that the concepts representing relevant research entities, intersect with societal properties, and which is demonstrated with Figure 29. To prevent from too much complexity in FERON, the societal intersections will not be further discussed; they are implicitly available with Research entities. The current work concentrates on analysing the Research

[134] Science in Society Home, EC – CORDIS – FP7: http://cordis.europa.eu/fp7/sis/ (Last visit: April 2nd, 2011)

domain in general and Language Technology in particular, and is guided by ontological foundations towards a field-agnostic formal Research ontology to enable field extensions – FERON (chapter 7).

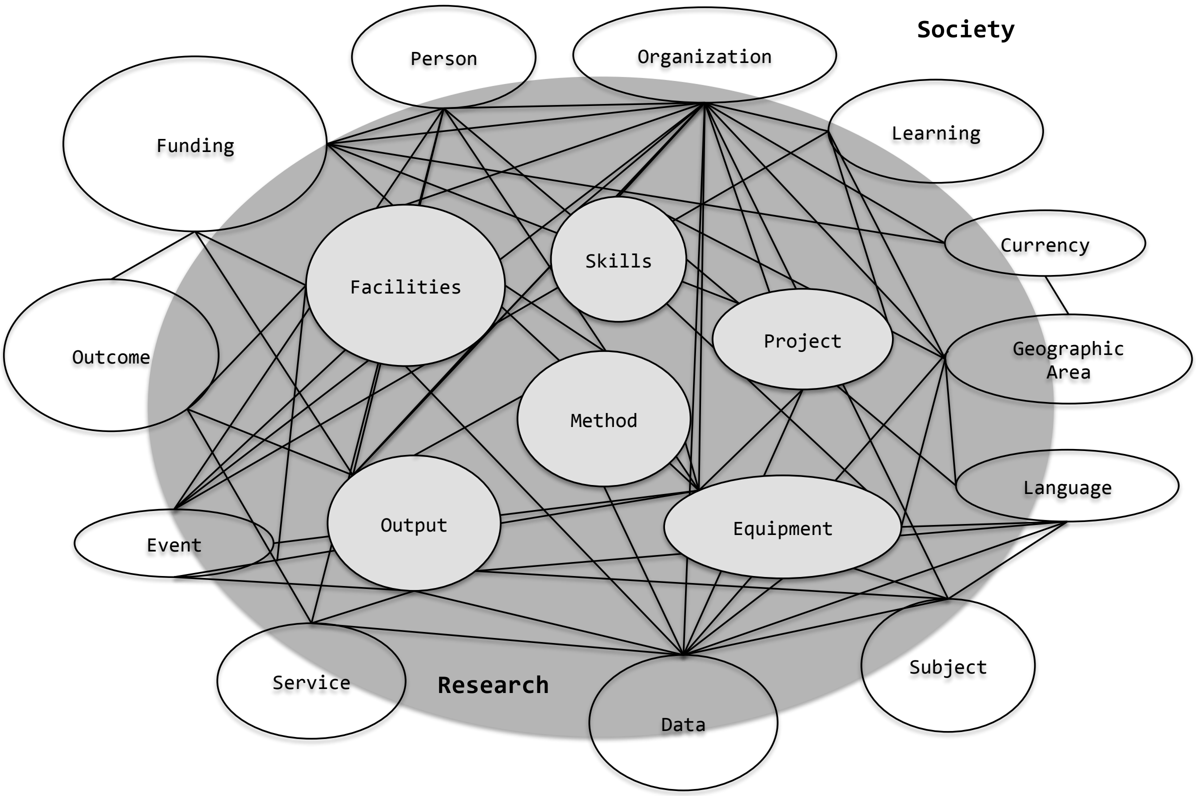


Figure 29: Research Entities embedded in Society viewable from various and selected contexts

Figure 29 can be read from any direction; it starts e.g. with Funding. Research is dependent on funding and increasingly outcomes are investigated by figures of measurable output and other indicators; also known as impact. Funding is often related to projects, organisations, persons, facilities, equipment, or services, where the direction is clear (another terminology may well label funding as income). There are uncountable numbers of relationships between entities in research, e.g. manager, coordinator, participant, contributor relationships between person and project or organisation. Organisations are in relationships with projects and many other entities. Person and organisation have been subsumed under the concept of an agent. Persons have skills that may be associated with e.g. equipment (e.g. responsible), subjects (e.g. information systems; computational linguistics, medicine, economics, etc.), or language (e.g. native speaker), and they live in a country (e.g. as a resident, or are visitors). Output is in the optimum case related to equipment, facility and subject, and additionally associated with person, project or organisation. Events need facilities, and equipment, an organisation that hosts, and persons that organise, attend and present. Organisations manage their accounts in a currency assigned to geographic regions. An international project may require for accounting a currency in US Dollars, to which the partners need to comply. An organisation may define the corporate language as Chinese, and because language is the means for communication and intrinsic in information, it may be essential with handling equipment, output or subject. Technologies or methods are associated with data and these are related to knowledge organisation systems (e.g. subjects), geographic areas, equipment, facilities, and output. With FERON, LT is modeled as a scientific field utilising natural language data, descriptions and tools or services; i.e. LT Resource, LT KOS, LT Infrastructure. Learning is an entity interfacing research and society through people (e.g. professors, assistants or students). Research is emerging within communities represented by people and organisations related to subjects (mostly known as disciplines or fields). Currency, language and geographical regions are important, e.g. in a relationship to funding or skills. The scientific paper as output is still the most important means to measure impact through its citations. The work performed and published in a research paper may result from a project, and the project – before funding – had to be proposed to a funding body (organisation) for evaluation and review towards decision upon funding, inline with the funding programme. Proposals are often submitted under specific themes or terms, at a local, national, European or at international level, and they involve multiple people employed in multiple organisations. These perform the work proposed once the proposal is accepted for funding. The given example processes and functions are far from being complete and are only provided to give an idea of the complexity inherent in research-related activities, and to rise awareness of the manifold relationships – many yet to be identified – that need to be maintained within the research ecosystem: “Managing research is complicated.” [RIM Report 2010, p. 28]

Complexity is reduced by structure. Conceptual modeling requires at first the identification of the main entities and the relationships these maintain. This analysis will therefore identify and characterise the entities considered most relevant in a Research context while applying a neutral relationship construct introduced in detail with section 7.6 under FERON. To construct the perceived world of FERON, publicly available formats, standards and descriptions have been collected and analysed – these may not always be formal (see section 4.1 – Information System Kinds). While automated approaches aim at identifying similarities in concepts by statistical or rule-based methods, this approach is truely a human investigation and aggregation, i.e. a collection of concepts, and where consequently the granularity with descriptions differs due to differences in the investigated resources, or in some cases knowledge simply does not yet exist, e.g. with describing research methods. It is anticipated, that this is owed to the fact that entities had enjoyed differences in priority during the course of modeling history and technological influences, but also due to their intrinsic heterogeneous nature.

The analysis here anticipates the structure of FERON (Figure 67), which, by employing ontological commitments identifies the substantial things modeled as classes. With each class, FERON distinguishes between properties of two kinds, namely intrinsic properties “possessed by the thing itself” and mutual properties “possessed jointly by two or more things”. These ontological principles are technologically reflected within OWL, where the former is called a “datatype property”, and the latter is called an “object property”; and both of the constructs are available through the ontology modeling tool Protégé. Furthermore, mutual properties are understood as to being relationships and must therefore reflect temporal aspects of attributions (according to BWW these are called state functions). To account for the fact, that BWW attributes – and thus FERON attributes – are “representations of the properties of a thing as perceived by an observer”, this work is consequently aware of their occurences in multiple worlds (which BWW calls functional schemata, each being “a set of attributes used to describe a set of things”), and for which technologically, FERON utilises namespaces. Namespaces are the means foreseen with schema assignments, originating from XML (see introduction in 3.3.3 Naming Conventions or Standards); where each namespace is identified through an individual URI (to ensure the required lawfulness according to 3.1.1 Bunge-Wand-Weber Ontology (BWW)), as indicated in section 7.4 Namespaces. To manage lawful state functions (e.g. by controlled vocabularies, thesauri, classification systems), FERON applies KOSs and LT KOSs (see section 7.11). These are themselves grounded, i.e. employ ontologically declared constructs such as class or concept (see 5.2.4 Knowledge Organisation Systems (KOS)).

5.1 Research

The Organisation for Economic Co-operation and Development (OECD) defines: “Research and experimental development (R&D) comprise creative work undertaken on a systematic basis in order to increase the stock of knowledge, including knowledge of man, culture and society, and the use of this stock of knowledge to devise new applications.” This definition cites the latest (sixth) edition of the Frascati Manual [OECD 2002] for a comprehensive definition of R&D and related activities. The first Frascati Manual issued in 1963 was exclusively concerned with the measurement of human and financial resources devoted to R&D, often referred to as R&D ‘input’ data. However, in a knowledge-based economy “it has become increasingly clear that such data need to be examined within a conceptual framework that relates them both to other types of resources and to the desired outcomes of given R&D activities" [OECD 2002, p. 14]. The Manual is consistent with UNESCO recommendations for all scientific and technological activities (UNESCO 1978), but specific to needs of OECD member countries maintaining similar economic and scientific systems. The 2002 Manual aims to enable productions of statistics for use in various *models* of the S&T system.

For survey purposes, the Frascati Manual distinguishes between R&D and

- Education and Training
- Other related scientific and technological activities
- Other industrial activities
- Administration and other supporting activities

which has been proposed as to being excluded from R&D measurements. R&D and these related activities may be considered under two headings: the family of scientific and technological activities (STA) and the process of scientific and technological innovation[135].

UNESCO had developed a broader concept of the STA, whereas Frascati [OECD 2002, p. 18] deals only with the measurement of R&D; i.e. basic research, applied research and experimental development.

Boundaries between basic and applied research are increasingly blurring [EC Report 2010, pp. 24 ff.], and the transfer of findings into products or services is increasingly seen as an integral part of the research process, i.e. innovation being science in society. Therefore, the UNESCO descriptions are briefly presented as included in the *Recommendation concerning the International Standardisation of Statistics on Science and Technology* (UNESCO 1978).

[135] Wilhelm von Humboldt, considered the unity of Research and Education important for a productive relationship between the teaching staff and students. His well-known concept of "Unity of Research and Education" (in German *Einheit von Forschung und Lehre*) guided universities world-wide: "Universitäten in aller Welt orientieren sich an dem von Humboldt geprägten Ideal der Einheit von Forschung und Lehre. Hierzu gehören die Weitergabe von Wissen aus dem Geist der Forschung und die Idee der forschenden Lehre. Studierende und Lehrende sind durch die kritische Auseinandersetzung mit den Wissensbeständen sowie in der aktiven Mitarbeit an der Erweiterung des Wissens vereint. Deshalb fördert die Humboldt-Universität die sozialen und kommunikativen Kompetenzen ihrer Mitglieder und unterstützt deren eigene Initiativen." [Humboldt University Website Extract] (Last visit: March 1st, 2011) Humboldt's highest aim with the "unity of research and teaching" was the cultivation of character belonging to the entirety of humanity, not the ascertainment of truth, or a steady stream of technological breakthroughs making us healthier, wealthier and (possibly) wiser [McNeely 2002, pp. 32 ff.]. Since universities had been teaching institutions before Humboldt, his concept aimed at bringing research into universities without throwing out teaching. Humboldts idea became reality in the less radical version of the Humboldtian pattern that made its way into many European university systems. More recently, the Humboldtian and pre-Humboldtian pattern of the relationship between teaching and research have been considered inadequate [Schimank & Winnes 2000, p. 407]. Curiosity-driven research seems to have a less prominent role with research performed in the new mode, what [Schimank & Winnes 2000, p. 407] lament to be "the greatest loss to research from a move towards the post-Humboldtian pattern". The European Union established a programme, where individual resarchers can apply for individual funding. The 'FET'-type projects within the EC Framework Programme still offers some options.

It says "scientific and technological activities (STA) comprise scientific and technical education and training (STET), and scientific and technological services (STS). The latter services include for example, S&T activities of libraries and museums, translation and editing of S&T literature, surveying and prospecting, data collection on socio-economic phenomena, testing, standardisation and quality control, client counseling and advisory services, patent and licensing activities by public bodies" [OECD 2002, p. 18].

The R&D concept of OECD is defined similarly by UNESCO, but "is thus to be distinguished from STET and STS" (p. 18). "Unfortunately, while indicators of R&D output are clearly needed to complement input statistics, they are far more difficult to define and produce" (p. 17). The so-called "Frascati family" suggests *Innovation*, *Technology balance of payments*, *Patents*, *Human Resources*, be possible measurements for R&D output, but, also considers *Bibliometrics*, *High-technology*, and *Globalisation* methodological S&T frameworks, as well as *Education classification*, *Education statistics*, and *Training statistics* to which further details have been provided in their Annex [OECD 2002, p. 16, table 1.1].

It is recalled, the R&D concept specified in the Frascati Manual [OECD 2002][136] was written by, and for national experts in member countries who collect and issue national R&D data, and submit responses to R&D surveys. It was intended as a set of guidelines on measurements of scientific and technological activities. The Frascati Manual increases the understanding of the R&D environment by providing "internationally accepted definitions of R&D and classifications of its component activities" aimed at being reused, with an attempt to make R&D surveys consistent [OECD 2002, p. 3, p.15, p. 17]. Here, only the list of headings from the Manual's chapters are provided:

- Basic Definitions and Conventions
- Institutional Classification
- Functional Distribution
- Measurement of R&D Personnel
- Measurement of Expenditures devoted to R&D

[136] "The Frascati Manual, which was developed by the Organisation for Economic Co-operation and Development (OECD), is the global standard for collecting R&D statistics. However, the manual was specifically designed for industrialized countries. As a result, statisticians from diverse regions often face difficulties adapting this standard to produce cross-nationally comparable statistics that accurately reflect the specific contexts and policy issues of their countries. [...] UIS Technical Paper No.5 (available in English) examines the Frascati Manual from the perspective of developing countries. It provides detailed information on how to interpret Frascati concepts and methodologies. The publication also presents a series of recommendations concerning data collection issues that are not addressed in the Frascati framework. It will eventually serve as the basis for an annex to the manual."

http://evaluation.zunia.org/post/measuring-rd-challenges-faced-by-developing-countries/ (Last visit: April 2nd, 2012).

- Survey Methodology and Procedures
- Government Budget Appropriations or Outlays for R&D by Socio-economic Objectives (GBOARD)

Multiple stakeholders have analyzed the Research domain from different perspectives. This work aims at a generic formal model to describe the Research domain scaling towards field extensions. The Research domain is analysed by identifiying and describing its main entities (*substantial things* according to BWW) and their relationships. These are considered a prerequisite for domain understanding, i.e. the base for modeling FERON – *F*ield-*e*xtensible *R*esearch *ON*tology in chapter 7. The subsequent investigation starts from the first top-down view as presented in Figure 29 (an anticipation of FERON from the long-term activity of the author (see section 1.1)). It is further refined through the analysis of selected publicly available definitions, standards and formats; they were applied where perceived relevant. The task is considered a modeling, design or analysis approach rather than a mapping and therefore, employed concepts are not duplicated; concept origins are furthermore indicated by namespaces.

5.1.1 CERIF

The Common European Research Information Format (CERIF) is a relational-structured ERM representation of the research domain, it origins from European activities (4.1.1 Current Research Information Systems). CERIF is furthermore, an EC Recommendation to Member States; the latest release (1.3) [Jörg et al. 2012] identifies the main entities (Figure 17, presents a conceptual *CIM* view) and a multiplicity of inter-relationships between them. CERIF specifies (inline with introduced architectural frameworks in section 4.2) at three different levels: conceptually, logically and physically. Where specification documents are truely conceptual *CIM*, the ERM is more formal (logic) but still platform independent *PIM*, and the same holds for CERIF XML. The automatically generated CERIF SQL scripts for particular databases are considered truely platform specific *PSM*; *ISMs* employ particular vocabularies. Following the intensional structure of FERON, within section 5.2 Research Entities, the individual CERIF entities are investigated in even more detail; i.e. their intrinsic and mutual – their datatype and object properties. With this section insight is provided into the overal CERIF ERM structure and constructs by presentation of a small model extract with Figure 30, because an overview of CERIF entities has already been given within section 4.1.1 Current Research Information Systems. The CERIF entities and relationships are organised consistently across the model, and the constructs and mechanisms presented are therefore valid and applicable with the entire model[137].

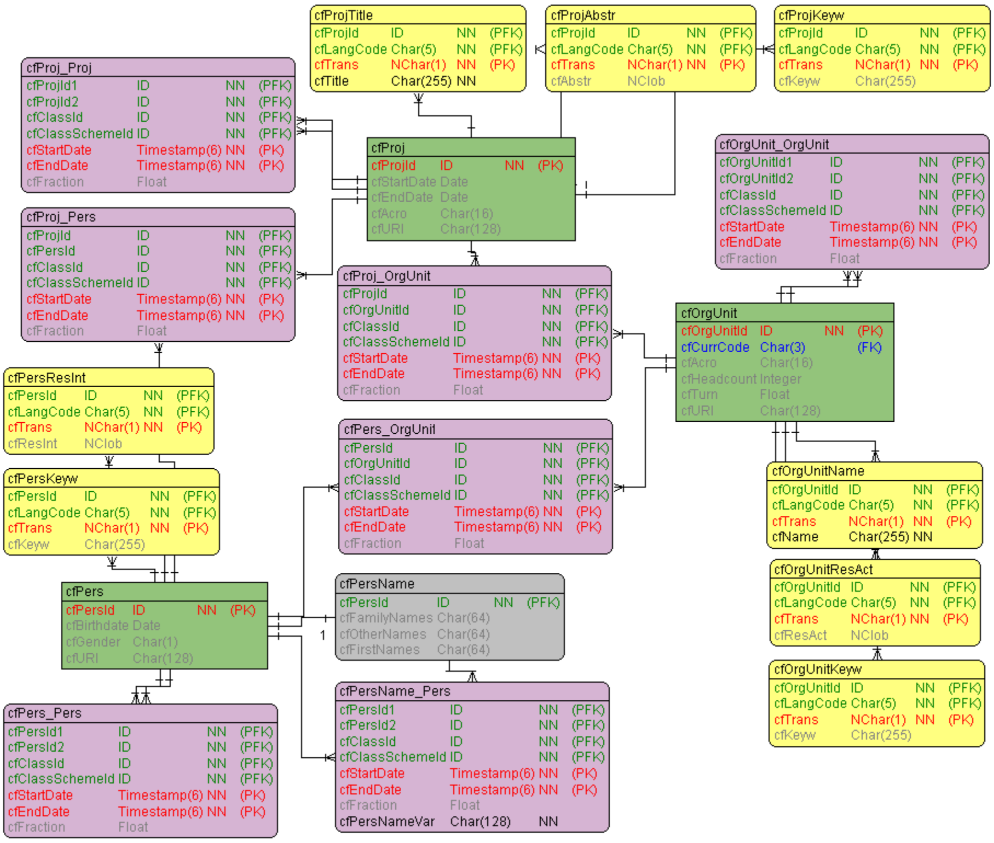


Figure 30: CERIF Base entities and link entities [Jörg et al. 2012, p. 11, fig. 3]

In the CERIF ERM, each entity, e.g. *cfPerson*, *cfProject*, or *cfOrganisationUnit* is described by intrinsic attributes. For example a *cfPerson* is represented by an identifier *cfIdentifier*, by *cfFirstNames*, *cfFamilyNames*, *cfGender*, *cfBirthdate* and a *cfURI*. In the relational (ERM) universe, there are entities and not concepts or classes (these are usually ontologically-inspired), The same is true for CERIF with attributes rather than properties. In addition to intrinsic attributes, there are mutual attributes – these are in fact relationships[138]. In CERIF relationships are celled *link entities*; they are modeled semantical-neutral. A CERIF link entity can either be of a binary or unary kind. Examples for underspecified binary CERIF link entities are *cfPerson_Person*, *cfPerson_OrganisationUnit*, unary link entities are e.g. *cfPerson_Classification*, or *cfOrganisationUnit_Classification*. The CERIF relationship or *link entity* construct is agnostic to particular contexts. It supplies for semantical-neutral

[137] CERIF Releases published at the euroCRIS website: http://www.eurocris.org/Index.php?page=CERIFreleases&t=1 http://cordis.europa.eu/cerif/ (Last visit: July 1st, 2012)

[138] In ontology-driven systems, the intrinsic and the mutual properties finally converge to objects in a perceived world.

relationship constructs a consistent syntax. The capturing of explicit or situational semantics happens through its so-called Semantic Layer – a conceptual construct to define classes, their relationships and multiple classification scheme assignments upon CERIF syntax and structure. The CERIF Semantic Layer as such may best be viewed as a declared knowledge organization system (KOS). It allows to maintain the vocabulary terms that are organised in e.g. flat lists, through taxonomies, terminologies, or ontologies upon semantical-declared CERIF syntax (The CERIF Semantic Layer will be introduced in more detail in section 5.2.4 Knowledge Organisation Systems (KOS)). It is mentioned here, because of its relevance with mutual attributes, i.e. relationships. Every CERIF *link entity* is by nature semantically neutral, it does not have a reading direction *cfPers_OrgUnit*, *cfProj_Pers*, but requires a semantic label, i.e. the function, with each record (e.g. manager, co-ordinator), and each function is assigned to a classification scheme, i.e. the namespace. The label or function in CERIF, is physically a reference to a class *cfClass* entity through its identifier as indicated in Figure 30, where link entities such as *cfProj_Proj*, *cfProj_Pers*, *cfProj_OrgUnit*, *cfPers_OrgUnit* are presented in abbreviated physical syntax, with class references through the class identifier *cfClassId* to single class records, and where the *cfClass* entity and thus its records are physically embedded and semantically valid by means defined through the CERIF Semantic Layer. In addition, each CERIF *link entity* requires a classification scheme *cfClassSchemeId* reference and time values *cfStartDate* and *cfEndDate* with each function. Furthermore, each CERIF link entity allows for an optional *cfFraction* value (e.g. percentage), as indicated in Figure 30. With CERIF, mutliple constructs (entities, their intrinsic attributes and mutual attributes (relationships), i.e. *link entities* including their semantics (functions by *cfClassId* references)) converge to an object or a concept (*substantial thing* according to BWW) in the sense of an ontological universe by aggregation of entities through internal identifiers, and where CERIF is open or scalable towards any available vocabulary, and thus for functional and schema (namespace) assignment.

Figure 30, shows the CERIF base entities (cfPers, cfProj, cfOrgUnit) and their relationships *link entities* with *cfClassId* and *cfClassSchemeId* references to the Semantic Layer. Furthermore, it indicates multilingual properties of CERIF entities in e.g. project titles *cfProjTitle* or organisation names *cfOrgUnitName*. The abbreviated syntax is inline with physical implementations, i.e. databases through SQL and applied in CERIF XML. The entire CERIF ERM follows this structure and employs the constructs as presented. CERIF provides a stable but scalable model syntax – neutral in semantics but scalable through a triple-like relationship construct *link entity* with reference to the so-called Semantic Layer. The *link entities* provide a powerful means to maintain the dynamics in science and allow for either binary, and unary relationships with time-aware fractional features. More of the particular intrinsic and mutual CERIF features will be further investigated under particular sections within 5.2 Research Entities. CERIF is maintained as an ERM and in parallel first with the CERIF 2006–1.1 release [Jörg et al. 2007a], an XML interchange format had been developed [Jörg et al. 2007b]. With the latest release CERIF 1.4, the CERIF XML format has been substantially improved [Jörg et al. 2012a] allowing for object-centered representations. A person object valid in CERIF 1.4 XML could be represented as in Notation 7.

```
<CERIF>
    <cfPers>
        <cfPersId>person-brigitte-joerg-internal-id</cfPersId>
        <cfBirthdate>****-**-**</cfBirthdate>
        <cfGender>f</cfGender>
        <cfPersName>
            <cfFamilyNames>Jörg</cfFamilyNames>
            <cfFirstNames>Brigitte</cfFirstNames>
        </cfPersName>
        <cfKeyw cfLangCode="EN" cfTrans="o">CERIF; CRIS; Information Systems</cfKeyw>
        <cfPers_EAddr>
            <cfEAddrId>brigitte.joerg@eurocris.org</cfEAddrId>
            <cfClassId>any-vocabulary-email-identifier-uuid</cfClassId>
            <cfClassSchemeId>any-vocabulary-scheme-identifier-uuid</cfClassSchemeId>
            <cfStartDate>2000-08-15T00:00:00</cfStartDate>
            <cfEndDate>2012-03-31T00:00:00</cfEndDate>
        </cfPers_EAddr>
        <Proj_Pers/>
        <Pers_ResPubl/>
        <Pers_Prize/>
        <Pers_OrgUnit/>

        <!-- … -->
    </cfPers>
</CERIF>
```

Notation 7: CERIF Person object representation inline with CERIF 1.4 XML Schema[139]

Where CERIF 1.4 XML allows for object-centered structures ERMs are relational in their nature; objects are only *created* with e.g. queries, in applications or at interface interaction

[139] CERIF 1.4 XML by the CERIF task group, euroCRIS, is licensed under a Creative Commons Attribution-NoDerivs 3.0 Unported License. Permissions beyond the scope of this license may be available at http://www.eurocris.org/CERIF-1.4/ (Last visit: May 24th, 2012).

time. A list of all CERIF entities and link entities is available with [Jörg et al. 2012]. In Table 3, the CERIF entities are compared with VIVO classes and CASRAI concepts[140].

5.1.2 VIVO

The VIVO project aims at enabling a "National Networking of Scientists" [Krafft et al. 2010] [Corson-Rikert et al. 2012][141]. The VIVO ontology is introduced as to having been developed from an „entity-relationship ontology model" behind the Cornell University Library portal to provide a single point of access for a virtual Life Sciences community by organising and presenting information about people, research, and education activities in an integrated view – transcending campus, college and department structure – aimed at usage by the Cornell faculty, students, administrative and service officials, prospective faculty and students, external sponsors, and the public [Lowe et al. 2007, p. 1]. Where the public library portal[142] reflects the ontology structure from a user's perspective with entry points such as People, Organizations, Research and Events, this work investigated the formal VIVO 1.4[143] ontology with Protégé (visualized with OntoGraf in Figure 31). Here, individual VIVO structure and constructs are investigated – a more detailed analysis of Research Entities will be presented under section 5.2.

VIVO 1.4 imports concepts from multiple sources (indicated by namespaces in Table 3). Its underlying syntax is OWL (RDF) with namespaces being part of the URIs, e.g., *DCMI Metadata Terms* for the object property *publisher*, and e.g. *vitro*, for the annotation property *vitro:descriptionAnnot* in Notation 8. Finally, *rdf:type* is employed to e.g. define the publisher property as an OWL object property *owl#ObjectProperty*, or the Person class as an OWL class *owl#Class*. The *Person* class is imported from the *FOAF* ontology (Notation 9) whereas the VIVO class *Project* is not imported, but self-declared as indicated in Notation 10.

[140] A CERIF ontology does not yet exist. With euroCRIS, a new task group has been initiated for Linked Open Data (LOD) and a collaboration has started with CASRAI and VIVO by signing a Memorandum of Understanding to progress with work in this respect.

[141] NATURE News: *Networking in VIVO*. Nature, Vol. 462, Number 5, November 2009.

[142] Cornell University Library Portal: http://vivo.library.cornell.edu/ (Last visit: March 29th, 2011)

[143] VIVO ontology downloaded from http://vivoweb.org/ontology/core/ (Issued Date, November 2011) via SourceForgeNet: http://sourceforge.net/projects/vivo/files/Ontology/vivo-core-public-1.4.owl/download. VIVO is described as „a semantic web project built on the Jena semantic web framework, and is an application to facilitate the discovery of researchers and collaborators across the country and internationally" (http://sourceforge.net/projects/vivo/).

```
<rdf:Description rdf:about="http://purl.org/dc/terms/publisher">
    <vitro:descriptionAnnot rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Public definition
          source: http://dublincore.org/2008/01/14/dcterms.rdf#. Examples of a Publisher
          include a person, an organization, or a service. Typically, the name of a
          Publisher should be used to indicate the entity.
    </vitro:descriptionAnnot>
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#ObjectProperty"/>
    <skos:scopeNote xml:lang="en">Used to link a bibliographic item to ist
          publisher.</skos:scopeNote>
    <rdfs:label xml:lang="en-US">publisher</rdfs:label>
</rdf:Description>
```

Notation 8: VIVO OWL description of an object property (publisher)

```
<rdf:Description rdf:about="http://xmlns.com/foaf/0.1/Person">
    <rdfs:subClassOf rdf:nodeID="A20"/>
    <rdfs:subClassOf rdf:nodeID="A21"/>
    <rdfs:subClassOf rdf:nodeID="A22"/>
    <rdfs:label xml:lang="en-US">Person</rdfs:label>
    <rdfs:subClassOf rdf:nodeID="A23"/>
    <vitro:shortDef rdf:datatype="http://www.w3.org/2001/XMLSchema#string">The most general
         classification of a person</vitro:shortDef>
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
    <rdfs:subClassOf rdf:nodeID="A24"/>
    <rdfs:subClassOf rdf:nodeID="A25"/>
    <rdfs:subClassOf rdf:nodeID="A26"/>
    <rdfs:subClassOf rdf:resource="http://xmlns.com/foaf/0.1/Agent"/>
    <rdfs:subClassOf rdf:nodeID="A27"/>
</rdf:Description>
```

Notation 9: VIVO OWL description of an imported FOAF concept (Person)

```
<rdf:Description rdf:about="http://vivoweb.org/ontology/core#Project">
    <rdfs:subClassOf rdf:nodeID="A33"/>
    <vitro:shortDef rdf:datatype="http://www.w3.org/2001/XMLSchema#string">An endeavor, frequently
         collaborative, that occurs over a finite period of time and is intended to achieve a
         particular aim.</vitro:shortDef>
    <rdfs:label xml:lang="en-US">Project</rdfs:label>
    <rdfs:subClassOf rdf:nodeID="A155"/>
    <rdfs:subClassOf rdf:nodeID="A117"/>
    <vitro:descriptionAnnot rdf:datatype="http://www.w3.org/2001/XMLSchema#string">An endeavor,
         frequently collaborative, that occurs over a finite period of time and is intended to
         achieve a particular aim.
    </vitro:descriptionAnnot>
    <rdfs:subClassOf rdf:nodeID="A135"/>
    <rdfs:subClassOf rdf:nodeID="A136"/>
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
```

```
     <rdfs:subClassOf rdf:nodeID="A116"/>
     <rdfs:subClassOf rdf:nodeID="A95"/>
     <rdfs:subClassOf rdf:nodeID="A82"/>
     <rdfs:subClassOf rdf:nodeID="A181"/>
     <rdfs:subClassOf rdf:nodeID="A179"/>
</rdf:Description>
```

Notation 10: VIVO OWL description of the VIVO concept (Project)

The namespaces available through XML syntax underlying the RDF and OWL in Notation 8, Notation 9, Notation 10 are not visible in Figure 31, with the visualisation, but are available with the Table 3: Crosswalk between CERIF – VIVO – CASRAI

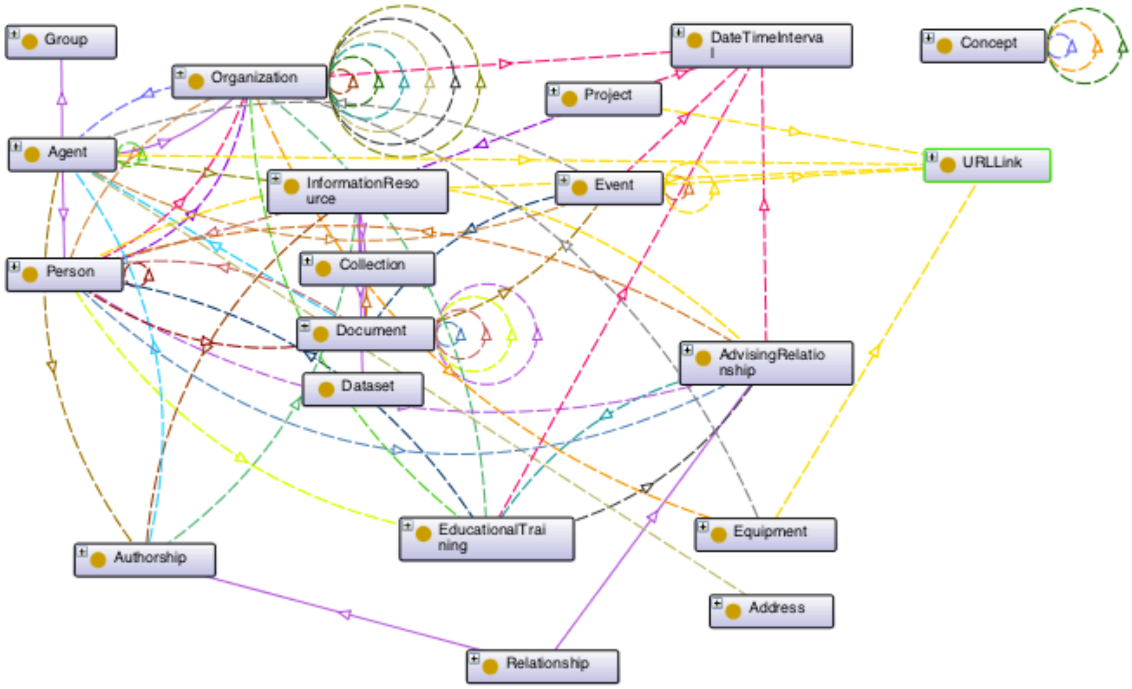


Figure 31: VIVO 1.4 core ontology visualized with OntoGraf

VIVO subsumes *Person, Group* and *Organization* classes under *Agent*. The *Agent* class in VIVO is imported from FOAF and maintains e.g. relationships or mutual properties with classes such as *InformationSource*, *Event*, *Address*, and *Authorship*. *Authorship* is a class subsumed under the *Relationship* class. *Relationship* is an explicit VIVO class with a short definition „a reified relationship“ subsuming the two classes *Authorship* and *AdvisingRelationship* to which the class *DateTimeInterval* is a subclass – where domain and range definitions are open. With the class *DateTimeInterval* VIVO defines *start* and *end* properties as subproperties of a *dateTimeValue* property with *DateTimeValue* ranges. VIVO maintains a *Role* class with super-classes to inherit from *dateTimeInterval* and *Description* and where *Role* subsumes e.g. *ResearcherRole*, *EditorRole*, *TeacherRole*, *MemberRole*. An *InformationResource* in VIVO is as class to subsume the imported BIBO classes such as *Collection*, or *Document*, but also VIVO classes such as *Dataset*. Table 3 provides a crosswalk betwen CERIF entities, VIVO classes and CASRAI concepts. More particular intrinsic and mutual VIVO properties will be further investigated under particular FERON concepts in section 5.2 Research Entities.

5.1.3 CASRAI

The *Consortia Advancing Standards in Research Administration* (CASRAI) is a not-for-profit standards development organisation to provide "a forum and the mechanisms required to standardize the data that researchers, their institutions and their funders must produce, store, exchange and process throughout the life-cycle of research activity"[144]. In the Program section of the CASRAI website under *Dictionary*, CASRAI provides a set of profiles and templates, and further defines included subsets[145]:

- *Research Activity Profile v0.9 (draft)*: Information about, and unique identification of a specific research activity (program or project).
- *Research Personnel Profile v1.1*: Information that fully describes a person conducting research activity.
- *Academic Funding CV*: A standard template for a CV to be attached by a researcher to a funding application.
- *Non-academic Funding CV*: A CV intended to by attached funding to a funding application but from a non-academic participant in the proposed research.
- *Student CV*: A CV maintained by a student and used for submission to applicable funding applications.
- *Abridged CV*: A CV used when only a basic amount of data about an individual is required.

The CASRAI Research Activity Profile v0.9 (draft) and v1.1.0 include descriptions for e.g. *Identification*, *Details*, *Team*, *Partners*, *Funding Requests*. Where the *profile info* indicates, that a *Research Activity Profile* is included in documents such as *Statement of Intent (SOI)*, and *Research Activity Full Profile*. A *Statement of Intent (SOI)* is thus itself profiled and defined as "The minimum information about a research activity that may be required by a

[144] CASRAI http://casrai.org/ (Last visit: May 27th, 2012)

[145] CASRAI Dictionary – Search Interface: http://dictionary.casrai.org/ (Last visit: May 27th, 2012)

prospective research funder in advance of submission of a full funding application", and it inherits the descriptive features from the *Research Activity Profile* (namely *Identification*, *Details*, *Team*, *Partners*, *Funding Requests*) and by inclusion in Statement of Intent (SOI) is indicated as to being recursive. Like the *Research Activity Profile* as such, it is included in the *Research Activity Full Profile*. The *Research Activity Full Profile* is defined as "An output of the RA profile" and the root concept; and is recursive because of its "inclusion of documents" *Research Activity Full Profile* statement.

The investigation of the included documents will now proceed to the included parts of the Research Activity Profile and Full Profile. It starts with *Identification*, which is defined as "Information that captures basic information about, and unique identification of, a specific research activity". *Identification* belongs to the element class *Grouping* and is part of the *Research Activity Full Profile*. It includes as parts *Activity Info* and *Research Location*, and is included in documents *Statement of Intent (SOI)*, and *Research Activity Full Profile*. An *Activity Info* belongs to the element class of *Record Type* and is defined as "Information that allows unique identification and classification of the research activity". It inhertits the "inclusion in document" *Statement of Intent (SOI)* and *Research Activity Full Profile*. The *Activity Info* itself includes parts (properties); *Activity ID*, *Activity Type*, *Activity Parent ID*, *Activity Short Title*, *Activity Long Title*, *Activity Description*, *Keywords*, *Temporal Classification*, *Research Classification*. Each of them has declared field types and is defined. The same holds for *Research Locations*, which is a sibling of *Activity Info* from the Element Class *Grouping*. It is defined as "The specific locations where the activity will be actively conducting research", and it belongs to the Element Class *Record Type* being part of *Identification*. It includes parts such as *Location Geo Tag*, *Location Municipality*, *Location Percent Effort*. For a better readibility, a visualisation of the CASRAI Research Activity Profile concepts down to the level of *Grouping* in Figure 32 is presented.

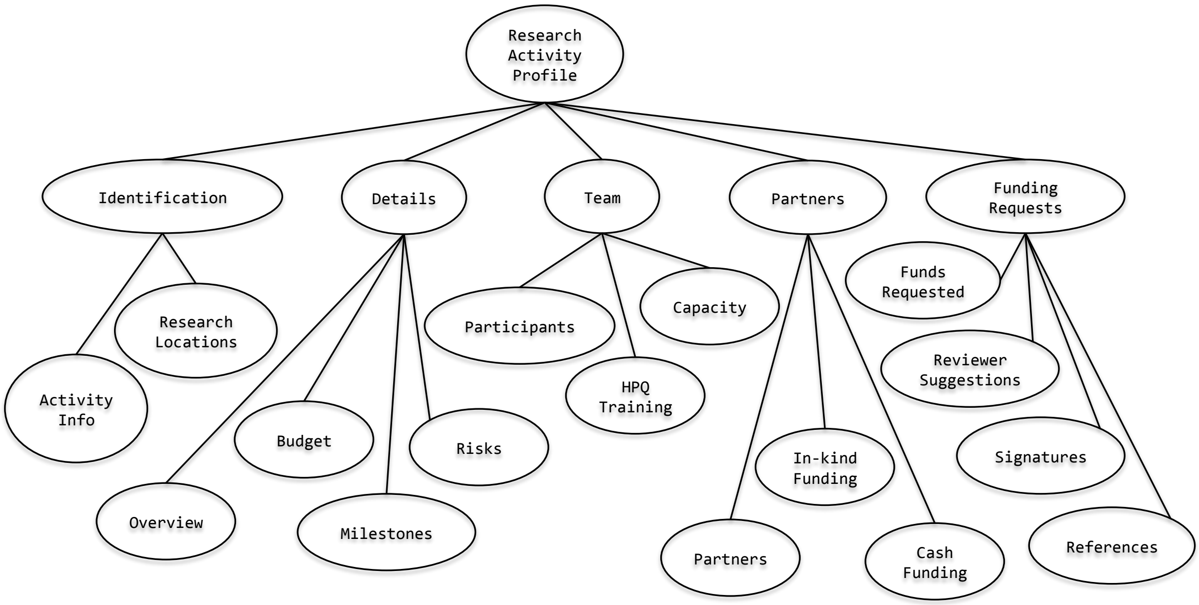


Figure 32: CASRAI Research Activity Profile concepts (CASRAI level of Grouping)

An investigation of the CASRAI *Research Personnel Profile v1.1* reveals, like the previous one, it belongs to the CASRAI element class of *Profile*. It includes as parts, *Identification*, *Contact*, *Education*, *Employment*, *Distinctions*, *Funding*, *Contribution*. These again belong to the CASRAI element class of *Grouping*. It is noted, that *Identification* as part of the *Research Personnel Profile v1.1* is different in parts from the *Identification* concept under *Research Activity Profile*. Here, it contains record types such as *Person Info*, *Language Competencies*, *Citizenships*, *Career Status*, *Research Classification*.

The CASRAI *Academic Funding CV template* belongs to the CASRAI element class of *Profile*. It includes as parts, *Identification*, *Contact*, *Education*, *Employment*, *Distinctions*, *Funding*, *Contributions*. These again belong to the CASRAI element class of *Grouping*. The *Identification* as part of the *Funding CV profile* is the same as in the Research Personnel Profile (see also Figure 33). This is indicated in the template in that the Academic Funding CV is included in the documents *Academic Funding CV*, *Non-academic Funding CV*, *Student CV*, *Abridged CV*, *Research Personnel Profile*. The same holds for *Contact*, which has been introduced within *Research Personnel Profile*. The *Contact* being thus included in documents such as *Academic Funding CV*, *Non-academic Funding CV*, *Student CV*, *Abridged CV*, and *Research Personnel Profile* as presented in Figure 33. The same applies for *Education*, *Employment*, *Distinctions*, *Funding* and *Contributions*, which have been introduced in the context of Figure 33.

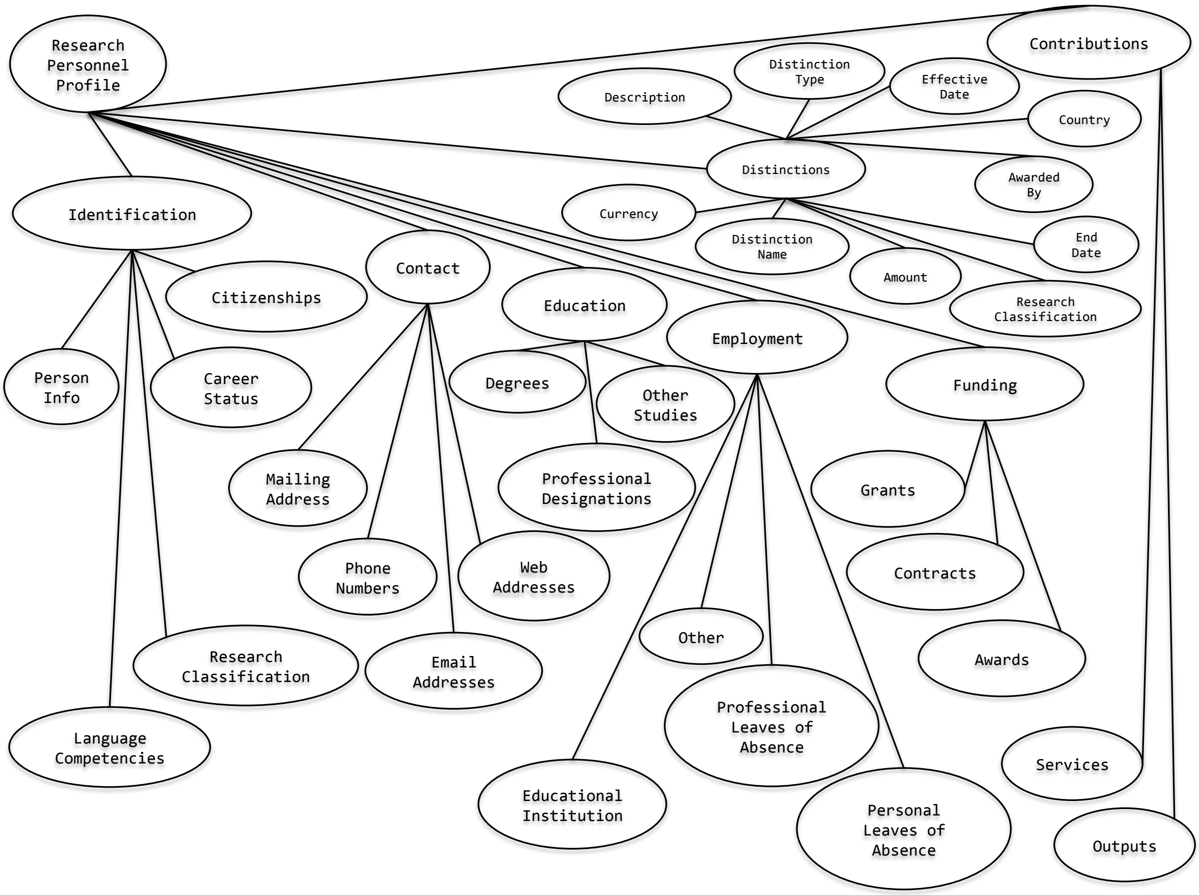


Figure 33: CASRAI Research Personnel Profile concepts (CASRAI level of Grouping)

Continuing in the order of the above list leads to an investigation of the CASRAI *Non-Academic Funding CV template*. It also belongs to the CASRAI element class of *Profile*, and includes as parts, *Identification*, *Contact*, *Education*, *Employment*, *Distinctions*, *Funding*, *Contributions*. These again belong to the CASRAI element class of *Grouping* and have been entirely investigated within the *Research Personnel Profile* (see also Figure 33). The same applies for *Student CV* and *Abridged CV*, which will therefore not be further investigated. Each of them includes the same parts. The CASRAI *Record Types* below the just introduced *Groupings* will be further investigated within 5.2 Research Entities in the course of intrinsic or mutual properties with the FERON ontology concepts. Table 3 provides a crosswalk between the CERIF entities, VIVO classes and CASRAI concepts.

5.1.4 Overview and Summary

With this overview, a crosswalk between the different model constructs is presented, namely CERIF ERM entities, VIVO Ontology classes and CASRAI Dictionary groupings. It is not called a mapping, because the conceptual constructs behind the three *grammars* are very different, and within FERON they will be organised according to the perceived world. Table 3 is therefore meant for guidance with subsequent sections, where individual concepts will be discussed and compared in detail, by anticipating the FERON model.

Table 3: Crosswalk between CERIF – VIVO – CASRAI

CERIF **ERM Entities**	**VIVO** **Ontology Classes**	**CASRAI** **Dictionary Groupings**	**field-specific**[146]
cfFunding	vivo:Agreement	Funding	
cfPerson	foaf:Person	Research Personnel Profile	
cfOrganisationUnit	foaf:Organization foaf:Group	Employment Funding Organization Educational Institution Partners Team	
cfProject	vivo:Project	Research Activity Profile	▨
cfResultPublication	bibo:Document bibo:Collection	Outputs	
cfResultProduct	vivo:Dataset vivo:Software	Outputs	▨
cfResultPatent	bibo:Patent	Outputs	
cfFacility	vivo:Facility	--	▨
cfEquipment	vivo:Equipment	--	▨
cfService	vivo:Service	Service	▨
cfMedium	bibo:Image	Outputs	
cfPostal Address	vivo:Address	Contact	
cfElectronic Address	vivo:URLLink	Contact	
cfGeographic BoundingBox	vivo:GeographicalEntity	--	
cfCountry	vivo:Country	Country	
cfEvent	event:Event	--	

[146] Those entities considered field specific have been indicated in Table 3 by grey background cells.

cfCurriculumVitae	--	Academic Funding CV Non-academic Funding CV Student CV Abridged CV	
cfLanguage	--	Language Competencies	
cfExpertiseAndSkills	vivo:Position vivo:AcademicDegree	Career Status Professional Designations	
cfPrizeAward	vivo:Award	Awarded By	
cfQualification	vivo:AcademicDegree	Degrees	
	vivo:EducationalTraining	Education	
cfCurrency	--	Currency	
cfCitation	--	--	
cfMetrics	--	--	
cfMeasurement	--	--	
cfIndicator	--	--	
	vivo:IssuedCredential	--	
cfClassification	skos:Concept	Research Classification	
cfClassificationScheme	--	--	
cfEntity1_Entity2	vivo:Relationship vivo:Role	--	
cfInternalIdentifiers	vivo:URLLink	Identification	
cfPostalAddress; cfAddressline	vivo:Location	Research Locations	
cfStartDate/cfEndDate *(in Link entities)*	vivo:TimeInterval vivo:TimeValue vivo:Time Precision	Effective Date End Date	
--	--	Risks	
--	--	Budget	
Additional: Research Method			

5.2 Research Entities

With FERON, a *F*ield *E*xtensible *R*esearch *On*tology is modelled. Within the following sections, while introducing each entity (each being thus identified as a *substantial thing* (3.1.1 Bunge-Wand-Weber Ontology (BWW) in the perceived world), those FERON classes and sub-classes that require field-specific attention are investigated in more details – namely *Information Resources* and *Non-Information Resources* such as *Activity*, *Measurement* and *Infrastructure*; *Product* and *KOS* as indicated in grey color with Figure 34.

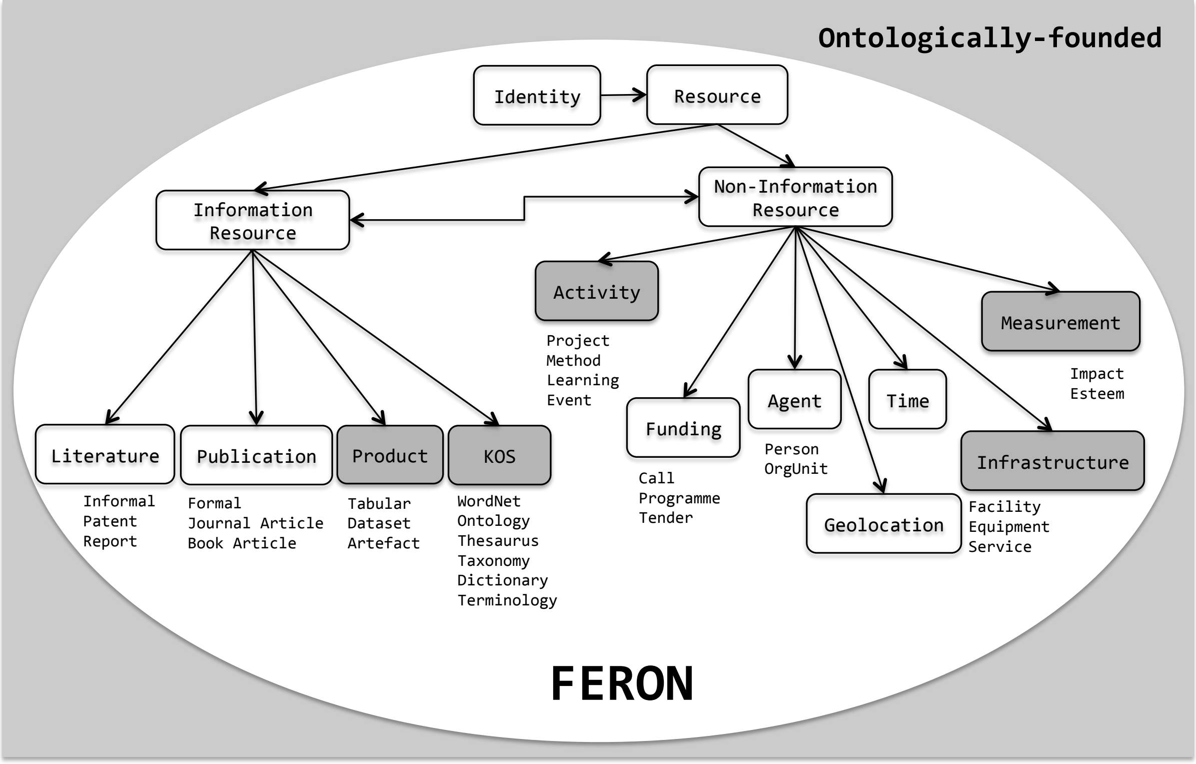


Figure 34: FERON – abstract view, indicating field extension concepts in grey

With this chapter, the single entities below abstract classed from Figure 34 are analysed, to demonstrate a generic Research domain perception level. In chapter 6 Analysis of Language Technology Entities, field specific entities are investigated, i.e. LT entities. In chapter 7 FERON is presented formally with reference to subsections for more details with respect to time-aware relationships, identities, namespaces, geographic location and KOSs.

5.2.1 Resource

The most distributed application of the *Resource* concept on the Web is realised through the Dublin Core Metadata Intitiative (DCMI) by specifications for its *Metadata Element Set*[147] managing a vocabulary based on fifteen properties "for use in resource description"; where the term *core* has been explained to be chosen "because its elements are broad and generic, usable for describing a wide range of resources" [Dublin Core 2010]. While the Dublin Core Metadata Element Set, Version 1.1 suggests a description of resources by the 15 elements such as *contributor*, *coverage*, *creator*, *date*, *description*, *format*, *identifier*, *language*, *publisher*, *relation*, *rights*, *source*, *subject*, *title*, *type*, it focuses on the terms in the DCMI vocabularies that are "intended to be used in combination with terms from other, compatible vocabularies in the context of application profiles and on the basis of the DCMI Abstract Model (DCAM)" (see Figure 35).

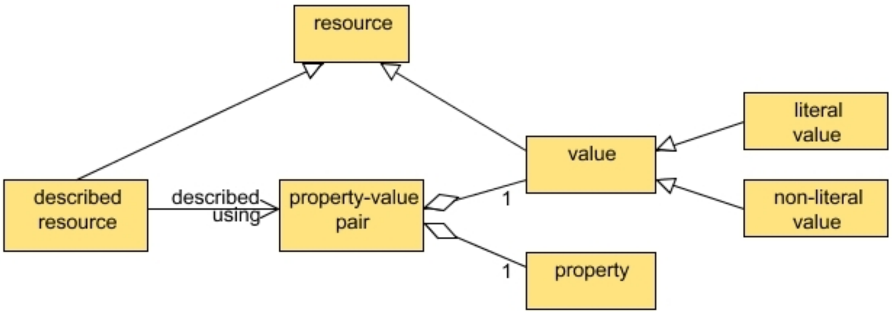


Figure 35: Dublin Core Abstract Model[148]

The DCAM in Figure 35, presents an *abstract* view of a resource being described by one or more property-value pairs, each made up of one property and one value, where each value is itself a resource "the physical, digital or conceptual entity or literal that is associated with a property when a property-value pair is used to describe a resource, and therefore, each value is either a literal value or a non-literal value". A *Resource* as proposed may thus be applied for any entity – and for this work that implies the entire Research domain, i.e. all Research entities. DCAM is therefore perceived as a meta-model of a *Resource*. In MDE terms, the model in Figure 35 is a mixture of the *CIM* and *PIM* view; it defines datatypes at an abstract level distinguishing two kinds – namely literal and non-literal values. However, it misses a

[147] The DC set belongs to a larger set of metadata vocabularies and technical specifications, where a full set (DCMI-Terms) includes sets of resource classes (including the DCMI-TYPE vocabulary), vocabulary encoding schemes and syntax encoding schemes. Dublin Core Metadata Element Set, Version 1.1 – a DCMI Recommendation: http://dublincore.org/documents/2010/10/11/dces/ (Last visit: July 20th, 2011)

[148] DCMI Abstract Model: http://dublincore.org/documents/2007/04/02/abstract-model/ (Last visit: July 20th, 2011)

comprehensive definition of the property with implications for relationships. It defines a property as a “specific aspect, characteristic, attribute or relation used to describe resources”, where however relation is rather to be perceived as the mathematical concept of a relation, and not to a relationship according to the concept of a graph. In a note it is indicated, that DCAM semantics “does not explicitly define a formal semantics for the Abstract Model. The intention is that the formal semantics can be defined by RDF and RDF Schema semantics” as specified by W3C, and to which they therefore refer. The definition is as indicated in Notation 11[149]. RDF is inherent in common ontology modeling tools and thus in OWL.

```
<rdf:Property rdf:about="http://www.w3.org/1999/02/22-rdf-syntax-ns#type">
    <rdfs:isDefinedBy rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#"/>
    <rdfs:label>type</rdfs:label>
    <rdfs:comment>The subject is an instance of a class.</rdfs:comment>
    <rdfs:range rdf:resource="http://www.w3.org/2000/01/rdf-schema#Class"/>
    <rdfs:domain rdf:resource="http://www.w3.org/2000/01/rdf-schema#Resource"/>
</rdf:Property>
```

Notation 11: Formal rdf:Property definition with domain (Resource) and range (Class)

There is still debate about the *Resource* concept and its kinds within the Semantic Web Community. In the terminology section of the introductory document of the Hypertext Transfer Protocol HTTP/1.1 (an often cited document referring to RFC 2616 [Fielding et al. 1999] [150]) a *Resource* is defined as “A network data object or service, identified by a URI” and what BWW may call *composite thing* (see 3.1.1 Bunge-Wand-Weber Ontology (BWW)). According to [Halpin et al. 2009, p. 123] in *An Ontology of Resources: Solving the Identity Crisis*, the resource definition of [Fielding et al. 1999] was broadened by Berners-Lee in his RFC 2396[151], defining a URI’s syntax as stating, that “A resource can be anything that has identity”. This definition has been partly adopted in the DCMI Glossary; where the use of the

[149] RDF provides the syntax to represent properties as relationships given its domain and range and with Dublin Core, the issue that formal relationships had not been sufficiently reflected has finally been recognised. “[Dublin Core] has been criticized for it use of unqualified DC and all the problems this has resulted in due to bad quality metadata” [Technology Watch Report 2008, p. 41], [Jeffery 1999], [Asserson & Jeffery 2004, p. 5]. Where the DCAM could easily be extended through a recursive relationship at the *Resource* node, the change implications for running systems would be massive. http://www.w3.org/1999/02/22-rdf-syntax-ns#Property (Last visit: April 2nd, 2011)

[150] HTTP/1.1 Introduction: http://www.w3.org/Protocols/rfc2616/rfc2616-sec1.html (Last visit: April 2nd, 2011)

[151] [Halpin et al. 2009] accidentially refer to RFC 2398 where the correct number should read RFC 2396: Uniform Resource Identifiers (URI): Generic Syntax: http://tools.ietf.org/html/rfc2396 (Last visit: April 2nd, 2011)

term *Resource* is discussed by reference to both, *non-Web accessible things* and *Web-accessible things* and continued with RFC 3986[152], the current IETF[153] RFC, which states:

“This specification does not limit the scope of what might be a resource; rather, the term ‘resource’ is used in a general sense for whatever might be identified by a URI. Familiar examples include an electronic document, an image, a source of information with a consistent purpose (e.g., “today's weather report for Los Angeles”), a service (e.g. an HTTP-to-SMS gateway), and a collection of other resources. A resource is not necessarily accessible via the Internet; e.g., human beings, corporations, and bound books in a library can also be resources. Likewise, abstract concepts can be resources, such as the operators and operands of a mathematical equation, the types of a relationship (e.g., “parent” or “employee”), or numeric values (e.g., zero, one, and infinity)” [Berners-Lee et al. 2005].

W3C’s Technical Architecture Group (TAG) distinguishes between *Information Resources* and *Non-Information Resources* (also called *Other Resources*)[154]. The distinction is applied with FERON. Accordingly, information resource is something whose „essential characteristics can be conveyed in a message“ [Jacobs & Walsh 2004], and in their *Architecture of the Web* (2001) they state: “By design a URI identifies one resource. We do not limit the scope of what might be a resource. The term ‘resource’ is used in a general sense for whatever might be identified by a URI. It is a convention on the hypertext Web to describe Web pages, images, products, catalogs, etc. as “resources” [...] Information resources associated with a non-information resource need to have their own URIs. They are themselves distinct resources and provide representations. They may have uses other than providing additional information about the non-information resource. However, the fact that they are associated with a non-information resource is important”. The Open Archives Initiative with the Object Reuse and Exchange (OAI-ORE) specification additionally describes aggregation boundaries through URIs [Lagoze & Van de Sompel 2008] [155].

[Bouquet et al. 2007, p. 2] provide another view over the resource concept. Beyond identifiers, they suggests to distinguish between resources as *things* and *abstract objects* and

[152] RFC 3986: Uniform Resource Identifier (URI): Generic Syntax: http://www.ietf.org/rfc/rfc3986.txt
[Berners-Lee et al. 2005]

[153] Internet Engineering Task Force (IETF): http://www.ietf.org/ (Last visit: April 2nd, 2011)

[154] Derefercing HTTP URIs: http://www.w3.org/2001/tag/doc/httpRange-14/2007-05-31/HttpRange-14#sec-information-resources (Last visit: May 24th, 2012).

[155] Both, the ore:Aggregation concept and the W3C’s TAG concept distinguishing information resources from non-information-reources has been applied with the EDM (see section 4.1.4).

to label the former *entities* and the latter *logical resources* and to claim, that any attempt of *forcing* the use of the same URI for logical resources is in principle likely to fail, whereas the use of the same URI for entities should be enforced: “the claim is that there are compelling *theoretical reasons* why the Semantic Web (and any other semantically driven information system) should not force people to use shared URIs for logical resources, but only (or mostly) *practical reasons* why people do not use shared URIs for entities.” They consider it crucial to distinguish between the two types in order to achieve semantic interoperability and efficient knowledge integration, not least due to the fact that 99% of research effort refers to the problem of designing shared ontologies, and, designing methods for aligning and integrating heterogeneous (T-Box) ontologies. [Haase 2004, pp. 205–206] considers it important to distinguish broad-coverage from general-purpose representation; and categorises the Dublin Core metadata standard as broad-coverage but special purpose representation designed to serve very general, typical-bibliographic purposes, generally applicable to a broad range of media items.

FERON follows [Haase 2004, pp. 205–206]’s view applying the Dublin Core (DC) Resource concept as a top-level *dcterms:Resource*, i.e. root concept, under which it subsumes all other concepts. Formally, the DCAM-proposed RDF/OWL construct supported by Protégé is implemented. However, the DCMI Elements in FERON are not considered *simple* properties but modeled through a neutral relationship construct, i.e. enable formally declared and lawful functions in the spirit of Bunge (see 7.10 Time-aware Relationships). That is, FERON – inspired by CERIF – does not explicitly model relationship properties (e.g. *is-manager-of*; *is-participant-in*), but a generic *dcterms:relation* property upon the range of a *Relation* class. This work aims at modeling of a generic field-extensible Reseach ontology and perceives explicitly modeled properties to be meaningful most often only within specific application contexts – very often not enough scalable and easily outdated. The time-aware relationship class (7.10) is semantically neutral at first but extensible towards contextual or situational functions. It enables to capture the dynamics in Science, which has to be taken care of when anticipating system design.

Dublin Core was aimed at resource description, and has been extensively applied with scholarly repositories (see 4.1.2 Scholarly Repositories) for storage of research *output*, i.e. publications in the wider sense. FERON aims at describing the Research domain in general to allow for field extensions; i.e. LT in particular, and is therefore in need of a time-aware relationship construct and of federate identifiers anticipating an open world.

Since the Web has evolved, in 2008, the DCMI introduced formal domain and range specifications of its properties instead of its initial natural language definitions. In order to not affect the conformance of already existing implementations of the "simple Dublin Core" in RDF, fifteen new properties identical to those of the DCMI Element Set have been created, where qualified properties are defined as subproperties of the simple element properties. The set of elements is part of a larger set of vocabularies and technical specifications maintained. The full set of vocabularies is identified by the namespace *dcterms* DCMI-TERMS[156]. It includes a set of resource classes, also the DCMI type vocabulary DCMI-TYPE, vocabulary encoding schemes, and syntax encoding schemes. The DCMI vocabularies are intended for use in combination with terms from other, compatible vocabularies in the context of application profiles and on the basis of the DCAM. The DCMI namespace policy describes how DCMI terms are assigned with Uniform Resource Identifiers (URIs) and sets limits on the range of editorial changes that may allowably be made to the labels, definitions, and usage comments associated with existing DCMI Terms. The initial 2008 *DCMI Metadata Terms* have been superseded in 2010. The index of DCMI's terms however is considered a helpuful insight into the range of coverage, and therefore presented here with Table 4. It is considered as a reference to "all metadata terms maintained by the Dublin Core Metadata Initiative", where on the website each term is specified with a minimal set of attributes – namely: *Name*, *Label*, *URI*, *Definition*, *Type of Term*, and an applicable set of attributes for additional information – namely: *Comment*, *See*, *References*, *Refines*, *Broader Than*, *Narrower Than*, *Has Domain*, *Has Range*, *Member Of*, *Instance Of*, *Version*, *Equivalent Property*.

Table 4: Index of DCMI Metadata Terms

Properties in the */terms/* namespace	abstract, accessRights, accrualMethod, accrualPeriodicity, accrualPolicy, alternative, audience, available, bibliographicCitation, conformsTo, contributor, coverage, created, creator, date, dateAccepted, dateCopyrighted, dateSubmitted, description, educationLevel, extent, format, hasFormat, hasPart, hasVersion, identifier, instructionalMethod, isFormatOf, isPartOf, isReferencedBy, isReplacedBy, isRequiredBy, issued, isVersionOf, language, license, mediator, medium, modified, provenance, publisher, references, relation, replaces, requires, rights, rightsHolder, source, spatial, subject, tableOfContents, temporal, title, type, valid
Properties in the legacy */elements/1.1/* namespace	contributor, coverage, creator, date, description, format, identifier, language, publisher, relation, rights, source, subject, title, type
Vocabulary Encoding Schemes	DCMIType, DDC, IMT, LCC, LCSH, MESH, NLM, TGN, UDC

[156] DCMI Metadata Terms: http://dublincore.org/documents/dcmi-terms/ (Issued date: January 14th, 2008) has been replaced by http://dublincore.org/documents/2010/10/11/dcmi-terms/. (Last visit: May 1st, 2012)

Syntax Encoding Schemes	Box, ISO3166, ISO639-2, ISO639-3, Period, Point, RFC1766, RFC3066, RFC4646, URI, W3CDTF
Classes	Agent, AgentClass, BibliographicResource, FileFormat, Frequency, Jurisdiction, LicenseDocument, LinguisticSystem, Location, LocationPeriodOrJurisdiction, MediaType, MediaTypeOrExtent, MethodOfAccrual, MethodOfInstruction, PeriodOfTime, PhysicalMedium, PhysicalResource, Policy, ProvenanceStatement, RightsStatement, SizeOrDuration, Standard

In FERON, the *dcterms:Resource* class employs the *dcterms:identifier* data-type property and the *dcterms:relation* object property upon the range of a *Relation* class, and features a *federated-identifier* property upon the range of a *FederatedIdentification* class as intrinsic properties. FERON labels the *dcterms:relation* property as a *relationship*, because it is an explicitly-named object property pointing to the URI of a resource without additional functional features – whereas relation in the mathematical sense as in FERON follows the *Relation* concept and is modeled as a class (7.10 Time-aware Relationships).

5.2.2 Non-Information Resource

The W3C’s Technical Architecture Group (TAG) distinguishes between *information resources* and *non-information resources* (also called *other resources*). This categorisation is applied with FERON as indicated in Figure 34. A comparable distinction is also implemented with the Europeana Data Model (see Figure 21). Now the introduced abstract class *w3c-tag:NonInformationResource* will be unfolded, and the sub-classes analysed. The *w3c-tag:NonInformationResource* as well as *w3c-tag:InformationResource* inherit their intrinsic properties from *dcterms:Resource* – namely the data-type property *dcterms:identifier* and the two object properties *federated-identifier* and *dcterms:relation* upon the ranges of *FederateIdentification* and *Relation*, both subsumed under *Time*, a sub-class of *w3c-tag:Non-InformationResource*, which features its own intrinsic properties, namely the *cerif:keywords* and *cerif:description*; these are propagated to sub-classes. At this point a difference is perceived compared to the *w3c-tag:InformationResource* class, where a title is intrinsic and where language as a property emerges. With the *w3c-tag:NonInformation Resource* and some sub-classes, however, a *name* instead of a *title* holds, namely with the *cerif:Infrastructure*, *cerif:Measurement*, and *Geolocation* classes, and especially with classes subsumed under *foaf:Agent*, where *cerif:title* semantically matches more with the *bibo:suffixName* but is not perceived as the title of an agent (resource) itself.

5.2.2.1 Agent

The *Agent* class is known from the popular FOAF[157] vocabulary, where it is defined as "things that do stuff. A well known sub-class is Person, representing people. Other kinds of agents include Organization and Group". The *foaf:Agent* class has been considered useful "where Person would have been overly specific"[158]. The *foaf:Agent* class anticipates the concept of identities which "sometimes belong to software bots", or "physical artifacts". It has subclasses *foaf:Organization*, *foaf:Group*, and *foaf:Person*, these have been entirely imported into VIVO with a subset of FOAF properties. The list of the *foaf:Agent* properties according to the specification are: *weblog*, *icqChatID*, *msnChatID*, *account*, *age*, *mbox*, *yahooChatID*, *tipjar*, *jabberID*, *status*, *openid*, *gender*, *interest*, *holdsAccount*, *topic_interest*, *aimChatId*, *birthday*, *made*, *skypeID*, *mbox-sha1sum*. These will not be further investigated. The list shows, that most of the *Agent*'s properties are indeed means for identification with tools or services and in FERON perceived as federated identifiers. FERON subsumes *foaf:Agent* under *w3c-tag:NonInformationResource*, from which it inherits properties. The *foaf:Agent* class is thus an abstract class in FERON without intrinsic properties and subsumes the *foaf:Person* and *cerif:OrganizationUnit* classes.

5.2.2.1.1 Person

FERON's *foaf:Person* class is a subclass of *foaf:Agent*. FERON does not employ the *foaf:Person* properties, but imports properties from authoritative Person formats originating in the Research domain. FERON's *foaf:Person* inherits the *dcterms:identifier*, and a *federated-identifier* property as well as a *dcterm:relation* property from *dcterms:Resource*. In FERON, the multiple FOAF identifier (see above) properties with *foaf:Agent* are not

[157] Since its creation in mid-2000, the Friend of a Friend (FOAF) project is evolving towards creating a Web of machine-readable pages describing people, the links between them and the things they create and do; it is a contribution to the Web. According to the latest Vocabulary Specification 0.98 (August 9th, 2010), FOAF is defined as a dictionary of named properties and classes using W3C's RDF technology. "FOAF integrates three kinds of networks: *social networks* of human collaboration, friendship and association; *representational networks* to describe a simplified view of a cartoon universe in factual terms; *information networks* that use Web-based linking to share independently published descriptions of this inter-connected world [...] In FOAF descriptions, there are only various kinds of things and links, which we call *properties*. The types of the things we talk about in FOAF are called *classes*. FOAF explicitly allows other vocabulary or local extensions to be mixed in with FOAF terms; it has been designed to be extended. The FOAF vocabulary is maintained in the FOAF Wiki. According to the FOAF specification document, the main FOAF terms are grouped in three broad categories: - Core: where related work is with: DublinCore; SKOS; DOAP; SIOC; Org vocabulary; Bio vocabulary; - Social Web: where related work is with: Portable Contacts; W3Cs Social Web group; - Linked Data utilities: where it began as the 'RDFWeb' project; FOAF "is not a standard in the sense of ISO" or that associated with W3C, but rather considered an "Open Source or Free Software project".

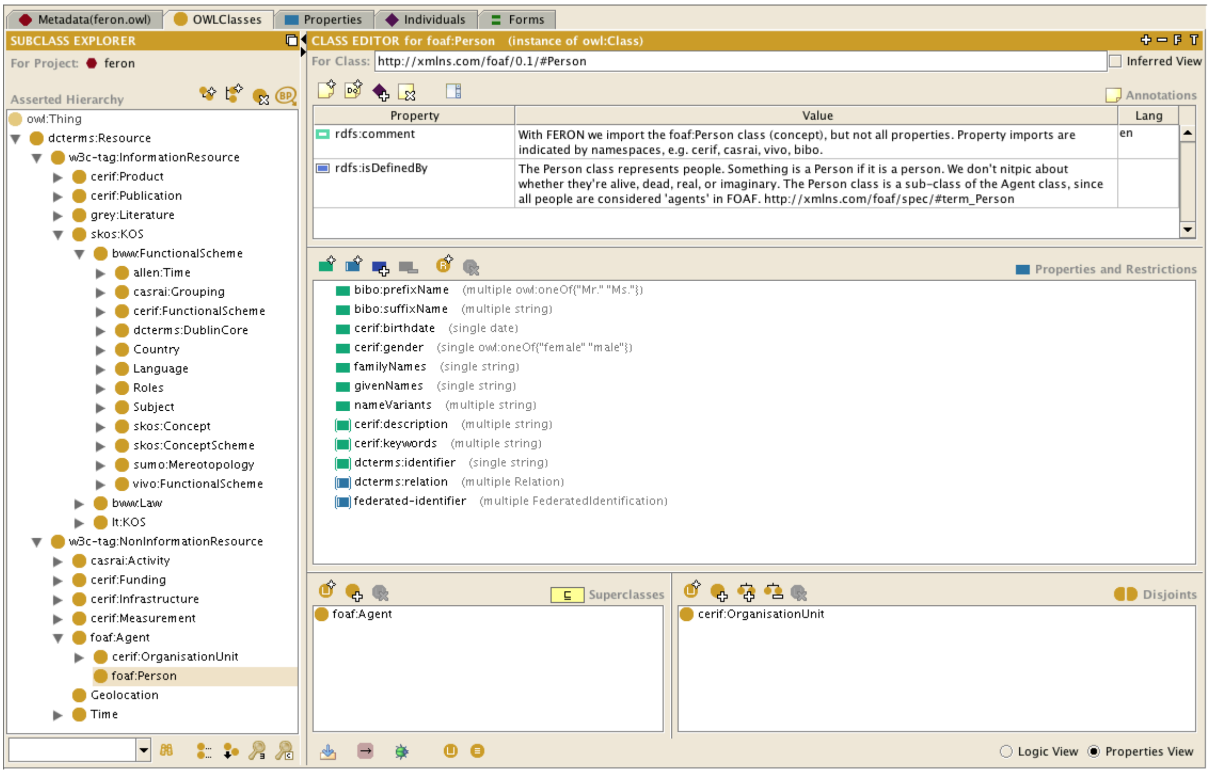
[158] Friend of a Friend Vocabulary specification of term Agent: http://xmlns.com/foaf/spec/#term_Agent (Last visit: May 1st, 2012).

imported, but identifier types are subsumed under a *FederateIdentification* class, i.e. *ORCID*, *ResearcherID*, *SkypeID*, *OpenID, URI*. VIVO does also not import the identifiers from FOAF, but employs *researcherId*, *orcidId* and *scopusId* as subproperties of an *identifier* datatype property. The investigation of CASRAI's dictionary reveals *Identification* as a *Grouping*. In the *Research Personnel Profile* it subsumes *Person Info*, *Language Competencies*, *Research Classification*, *Career Status*, *Citizenships* – and is thus a composition of multiple features, i.e. *grouping*, and not only a single attribute. CERIF attributes a *cfPersonId* for system-internal record identification, and the CERIF task group announced the development of a *cfFederatedIdentifier* entity with the upcoming release[159].

In FERON, *FederatedIdentification* is modelled as a class under which identifier instances are recorded and typed as sub-classes. In addition to FERON's *dcterms:identifier* or *federated-identifier* properties there are name attributes for person record disambiguations. E.g. VIVO applies two attributes from FOAF, namely *foaf:firstName* and *foaf:lastName* and introduces a *vivo:middleName* datatype property. The CERIF format assigns *cfFirstNames*, *cfOtherNames*, *cfFamilyNames* via a separate *cfPersonName* entity. In CASRAI, *Person Info* corresponds to *First Name*, *Middle Name*, *Family Name*, *Previous Family Name*. Often related to a person name is a title, which VIVO imports from BIBO as *bibo:prefixName* (e.g. Mr. or Ms.), and *bibo:suffixName* (e.g. M.A. or PhD)[160]. CERIF does not explicitly model a person title, it applies the *cfPerson_Class* linking mechanism to assign titles to a person. CASRAI reflects titles via *Salutation* "The title that forms a part of a person's full name", or *Presented Name* "The preferred presentation of the full name of the person when printed". A *foaf:Person* allows for multiple choices with *familyName*; it also employs *lastName*, *family_name*, and *surname* (this may cause confusion and ambiguity) and *firstName*. FERON employs the *foaf:Person* class *givenNames*, *familyNames* and *nameVariants*. Furthermore, the *cerif:birthdate* and the *cerif:gender* attributes; as *foaf:birthday* is perceived only as a day. The CASRAI *Date of Birth* concept suits the CERIF attribute and provides a definition that is imported in FERON.

[159] Impressive Turnout at CERIF Tutorial and UK Data Surgery in Bath: http://isc.ukoln.ac.uk/2012/02/13/impressive-turnout-at-cerif-tutorial-and-uk-data-surgery-in-bath/ (Last visit: April 2nd, 2012)

[160] BIBO ontology: http://purl.org/ontology/bibo/ (Last visit: April 2nd, 2012)



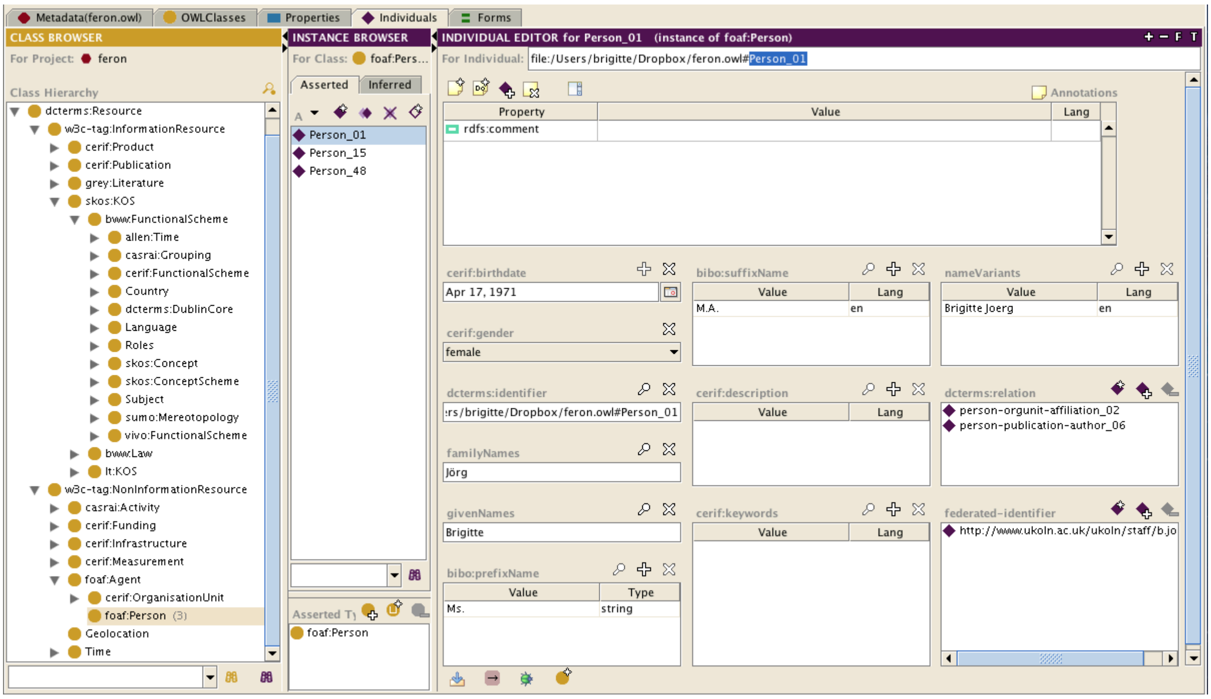
For Class: http://xmlns.com/foaf/0.1/#Person

Property	Value	Lang
rdfs:comment	With FERON we import the foaf:Person class (concept), but not all properties. Property imports are indicated by namespaces, e.g. cerif, casrai, vivo, bibo.	en
rdfs:isDefinedBy	The Person class represents people. Something is a Person if it is a person. We don't nitpic about whether they're alive, dead, real, or imaginary. The Person class is a sub-class of the Agent class, since all people are considered 'agents' in FOAF. http://xmlns.com/foaf/spec/#term_Person	

bibo:prefixName	(multiple owl:oneOf{"Mr." "Ms."})
bibo:suffixName	(multiple string)
cerif:birthdate	(single date)
cerif:gender	(single owl:oneOf{"female" "male"})
familyNames	(single string)
givenNames	(single string)
nameVariants	(multiple string)
cerif:description	(multiple string)
cerif:keywords	(multiple string)
dcterms:identifier	(single string)
dcterms:relation	(multiple Relation)
federated-identifier	(multiple FederatedIdentification)

Figure 36: FERON imported foaf:Person class with additional imported properties

Person is perceived a valid class and with FERON it is thus modelled as a subclass of *foaf:Agent*. With (Figure 36) the intrinsic properties of FERON's *foaf:Person* concept have been investigated, but mutual properties (i.e. relationships) have yet to be analysed,. With FERON, a *Relation* is modelled as a class in the range of a *dcterms:relation* object property within the *dcterms:Resource* domain. FERON's *foaf:Agent* as well as its subclass *foaf:Person* inherit this *dcterms:relation* property to allow for multiple relationships upon the *Relation* class. A person relationship in the role of e.g. a Researcher with e.g. an organisation will thus be recorded as a *Relation* instance. Each relationship record or instance has thus its own URI (e.g. relationship#person-project-manager) to which e.g. the person record – via the FERON *dcterms:relation* property – refers, and of which multiple are allowed. Because there is an uncountable number of relationships that e.g. a person maintains within the Research ecosystem, they will not be further investigated in more detail here, but explained in a more dedicated section 7.10 Time-aware Relationships. Figure 37 shows two recorded person relationship instances (reference to their Protégé URIs from within the selected *Person_01* (highlighted) instance; one in the function of affiliation with orgunit *person-orgunit-affiliation*, the second in the function of author with publication *person-publication-author*).



For Project: feron

For Class: foaf:Pers...

For Individual: file:/Users/brigitte/Dropbox/feron.owl#Person_01

Property	Value	Lang
rdfs:comment		

cerif:birthdate: Apr 17, 1971

cerif:gender: female

dcterms:identifier: :rs/brigitte/Dropbox/feron.owl#Person_01

familyNames: Jörg

givenNames: Brigitte

bibo:prefixName

Value	Type
Ms.	string

bibo:suffixName

Value	Lang
M.A.	en

cerif:description

Value	Lang

cerif:keywords

Value	Lang

nameVariants

Value	Lang
Brigitte Joerg	en

dcterms:relation: person-orgunit-affiliation_02; person-publication-author_06

federated-identifier: http://www.ukoln.ac.uk/ukoln/staff/b.jo

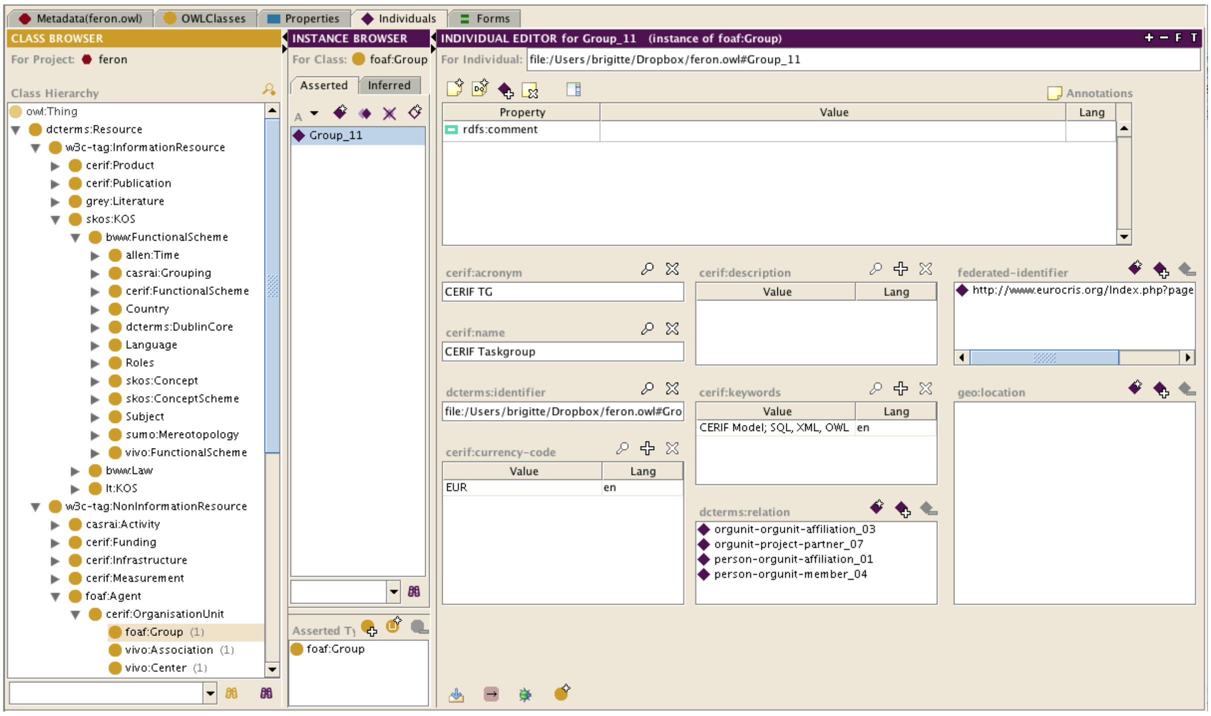
Figure 37: FERON Person instance viewed from within Protégé

5.2.2.1.2 Organisation

The *foaf:Organization* class is defined as a sub-class of *foaf:Agent*, where it is not further described through properties. In FERON, the CERIF entity *cfOrganisationUnit* is considered more appropriate for the Research ecosystem, and therefore subsumed under *foaf:Agent* instead of *foaf:Organization* which is not employed. The *foaf:Group* class is hence already covered by the CERIF entity *cfOrganisationUnit*; where *Group* is perceived as a functional type; i.e. a subclass thereof. VIVO employs an *abbreviation* datatype *Literal* and various labelled object-properties with *foaf:Organization* (e.g. *hasCurrentMember*, *contributingRole*, *dateTimeInterval*, *hasGeographicLocation*, *hasSubOrganization*, *primaryEmail*, *featuredIn*, *webpage*, *email*, *mailingAddress*). CASRAI does not explicitly consider the organisation concept, but employs kinds such as *Funding Organisation* or *Educational Institution* or *Team*, and roles such as *Partners*. CERIF with *cfOrganisationUnit* employs attributes such as *cfAcronym* and *cfCurrencyCode*; and multilingual attributes *cfName*, *cfResearchActivity*, *cfKeywords* are applied in FERON as string fields. The *cerif:name* is defined functional; allowing for one value only (relying on the language construct of RDF/OWL as inherent in Protégé which allows multiple translations). The Science ontology[161] downloaded from the

[161] The Ontology of SCIENCE is a „slightly improved version of the KA^2 ontology developed by Knowledge Annotation Initiative of the Knowledge Acquisition Community. The original ontology is available at the KA^2 [the link is redirected to http://semanticweb.org/wiki/Main_Page] portal, coded in F-Logic, DAML and OIL and on the Spanish mirror of

public Protégé Ontology Library includes *Organization* as a concrete class describing slots such as *name: required String*, *Location-Place: required Country or State or City*, *Head: requires Person*, *employs: requires multiple Employee*, *develops: multiple Product*, *carries-Out: multiple Project*. FERON follows VIVO and employs for *cerif:OrganisationUnit* a *geo:location* object property upon the range of a *Geolocation* class (for more information see section 7.11 Geographic Location). The *Science* ontology represents a very specific context, with relationships explicitly modeled upon explicit ranges, e.g. *Organization employs Employee*, or *Organization finances Project*, or *Organization publishes Scientific Document*.



For Individual: file:/Users/brigitte/Dropbox/feron.owl#Group_11

Property	Value	Lang
rdfs:comment		

cerif:acronym: CERIF TG

cerif:name: CERIF Taskgroup

dcterms:identifier: file:/Users/brigitte/Dropbox/feron.owl#Gro

cerif:currency-code

Value	Lang
EUR	en

cerif:description

Value	Lang

cerif:keywords

Value	Lang
CERIF Model; SQL, XML, OWL	en

dcterms:relation

- orgunit-orgunit-affiliation_03
- orgunit-project-partner_07
- person-orgunit-affiliation_01
- person-orgunit-member_04

federated-identifier

- http://www.eurocris.org/Index.php?page

geo:location

Figure 38: FERON organization unit (foaf:Group) record

In FERON, relationships or mutual properties are managed through the *dcterms:relation* property with reference to the *Relation* class. An organisation relationship in the role of e.g. *Affiliation* with e.g. another organisation will thus be recorded as an instance. An organisation can maintain multiple relationships, each is an instance of *dcterms:Resource* under the class *Relation*, with its own URI (e.g. *relationship#orgunit-person-employer*) to which e.g. the organisation instance refers, and of which multiple are allowed. Because there are uncountable numbers of relationships that an organisation can maintain within the *Research* ecoystem (see Figure 38), these will not be further investigated in detail here, but a

Ontoloingua“. http://protege.stanford.edu/ontologies/ontologyOfScience/ontology_of_science.htm (Last visit: June 7th, 2012) dedicated section 7.10 Time-aware Relationships will explain the formal FERON relationship construct.

5.2.2.2 Activity

Activity has been recognised as a broader concept in the Research domain, which is often further specified e.g. as a Project (e.g. in CERIF), but sometimes also as a Scientific Event. With FERON, the *casrai:Activity* concept and subsume *Method* and *Learning* is employed. Where *scientific Method* obviously *is a* scientific activity, *Learning* is more *about* scientific activity, i.e. the transfer of scientific results or knowledge to the education system and in fact a science to society activity. The *Learning* class in FERON is considered relevant but only for linkage and will therefore not elaborated in more details. CASRAI dedicated a profile to *Research Activity*, which was introduced in section 5.1.3, composing parts such as *Identification*, *Team*, *Partners*, *Funding*, *Details*, and *Funding Requests*. In FERON, *casrai:Activity* is subsumed under *w3c-tag:NonInformationResource*, from where it inherits the *dcterms:identifier*, the *federated-identifier*, and the *dcterms:relation* property. In FERON, the *casrai:Activity* is a grouping concept such as is *foaf:Agent*; it is anticipated that it will not contain instances, but it introduces intrinsic properties such as *cerif:acronym*, *cerif:title*, *cerif:description* and *cerif:keywords* which are propagated to subclasses. With some ontology tools or versions, classes that are not populated are often called *abstract* – contrary to *concrete* classes containing instances.

5.2.2.2.1 Project

Project is also defined in FOAF, where however, it is still in the status of *testing*. The FOAF specification additionally refers to DOAP – the *D*escription *O*f *A* *P*roject. DOAP is “an RDF schema and XML vocabulary to describe software projects, and in particular open-source”[162] “Open Source Projects”[163], and is thus not aimed to describe research projects. In CERIF, *Project* emerged historically as the first entity where it was the only concept when activities started in the early 1970s. Back then, it has been the conviction, that Research Activity can be tracked entirely through Project records – the record back then being a one-dimensional

[162] Description of a Project: http://en.wikipedia.org/wiki/Description_of_a_Project (Last visit: June 7th, 2012)

[163] XML Watch: Describe open source projects with XML, Part 1: http://www.ibm.com/developerworks/xml/library/x-osproj.html (Last visit: June 7th, 2012)

(library-card-like) record, and not the multi-dimensional record first proposed in the CERIF 2000 recommendation. The Science ontology downloaded from the Protégé library includes *Project* as a concrete class (with subclasses such as *Development-Project* and *Research-Project*) and assigns *slots* for description: *Carried-out-By: required multiple Organizations*; *is-Financed-By: multiple Organization*, *Members: multiple Employee*; *name: required String*; *produces: multiple Product*; *Project-Head: required multiple Academic-Staff*; *Project-Publication: multiple Instance of Scientific-Document*; *Topics: required multiple Research Topic*; these are in fact – except from an intrinsic *name* string property all relationships, i.e. mutual properties. [Luzi et al. 2004, pp. 14–18] categorise the main steps of a project live-cycle into two categories, *Activities* and *Documents*. Where the former is further categorised as *Project Proposal Elaboration*, *Project Assignment*, and *Project Execution*, the latter is considered project output such as *Research Report*, *Research Results*, *Description*, and input such as *Project Forms*, *Call for Proposal*, *Activity Plan* or *Administrative documentation*. In FERON, most project *output* documents would be subsumed under the *w3c-tag:InformationResource* class, in the classes *cerif:Publication*, *cerif:Patent*, *cerif:Product* (inspired by the *unisist:tabular concept* thus subsuming Data – the *(vivo:)Dataset(s))*, and e.g. Luzi's *Call for Proposal* would be recorded as a type, i.e. subclass of *Funding*. Both, the *output* and the *funding* in FERON are referred to a project by relationships and therefore not explicitly modeled.

[Luzi et al. 2004, pp. 14–18, fig. 5] model a project life-cycle by assigning intrinsic and mutual properties to each step. They thus reveal nicely the dynamics in scientific activities and the change in activity-relevant properties upon contextual or situational circumstances (see Figure 39). The process of document production from [Luzi et al. 2004, fig. 5] is presented to support the understanding of *Project* as a Research *activity*, where relationships take a prominent role (Figure 39), e.g. *Department*, *Programme*, *Reviser* are in fact relationships.

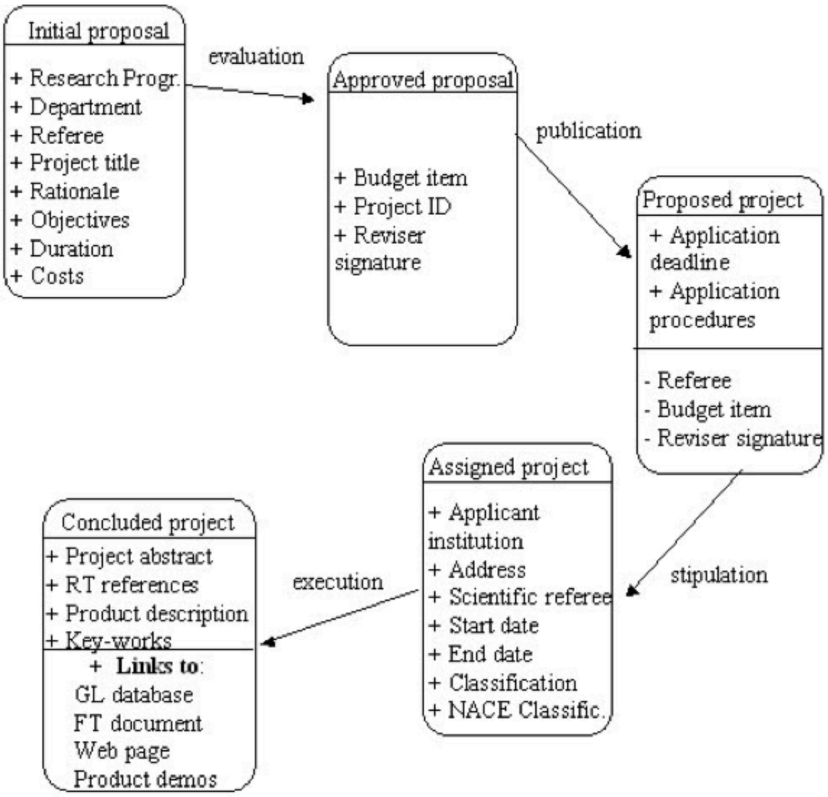
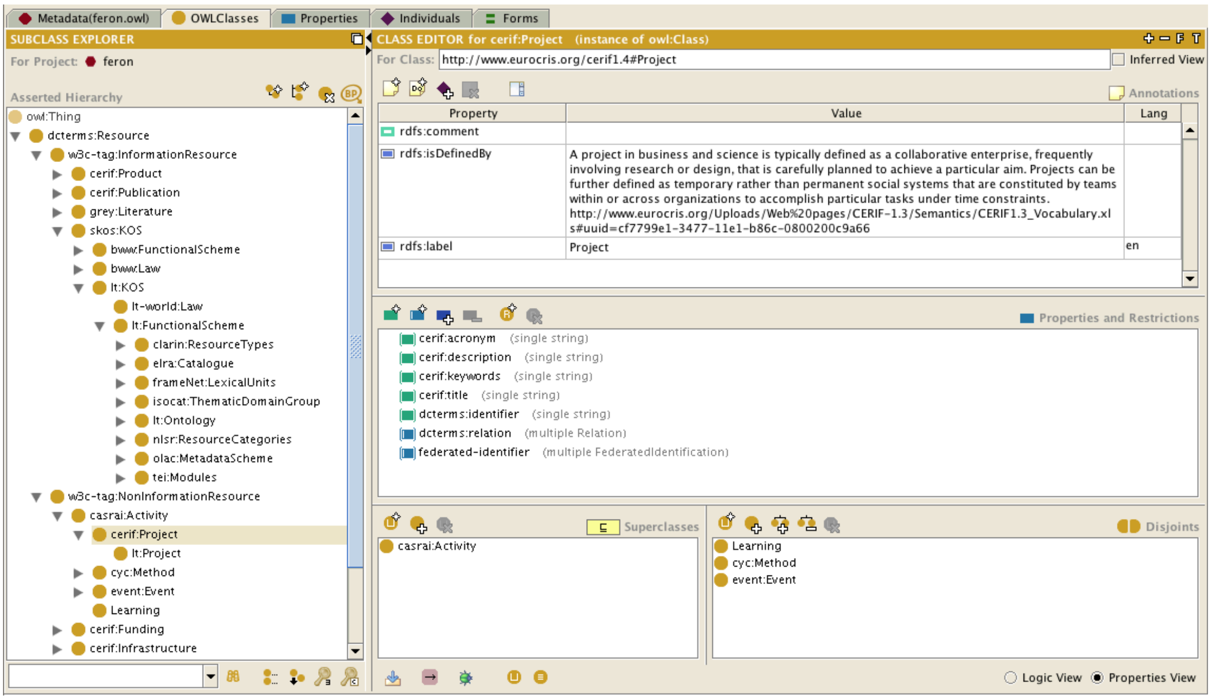


Figure 39: Production of GL documents and content update process
[Luzi et al. 2004, pp. 14–18, fig. 5]

The described live-cycle may be similar across organisations – however, life-cycle modeling is a step ahead of FERON. FERON provides the underlying means for process modeling, i.e. a generic Research ontology identifying the main entities and thus, the range of possible relationships and rules. With FERON workflow, i.e. logical constraints are not modeled. These are tightly contextual and would mean an additional step in formality towards PSM and even ISM, i.e. a concrete application or system, which is beyond the scope of this work.

In FERON, relationships or mutual properties are managed through a *dcterms:relation* property with reference to a *Relation* class (see section 7.10 Time-aware Relationships). A Project relationship in the role of e.g. a *Partner* with e.g. an *Organisation* will thus be recorded as a *Relation* instance. A project can maintain multiple relationships, each is an instance of *dcterms:Resource* under the class *Relation*. Each Relation instance record has its own URI (e.g. *relationship#orgunit-project-partner*) to which e.g. the organisation record refers via the FERON *dcterms:relation* property, and of which multiple are allowed.



For Class: http://www.eurocris.org/cerif1.4#Project

Property	Value	Lang
rdfs:comment		
rdfs:isDefinedBy	A project in business and science is typically defined as a collaborative enterprise, frequently involving research or design, that is carefully planned to achieve a particular aim. Projects can be further defined as temporary rather than permanent social systems that are constituted by teams within or across organizations to accomplish particular tasks under time constraints. http://www.eurocris.org/Uploads/Web%20pages/CERIF-1.3/Semantics/CERIF1.3_Vocabulary.xls#uuid=cf7799e1-3477-11e1-b86c-0800200c9a66	
rdfs:label	Project	en

Figure 40: FERON Project indicating field extension viewed in Protégé

FERON's *cerif:Project* class is truely a candidate for field extension (Figure 40), which is indicated by a FERON subclass *lt:Project*. This represents a class for projects in the field of Language Technology, and for which a new field-specific mutual, i.e. object property is introduced – namely *lt:relation* to allow for LT-specific project relationship recordings. The *lt:relation* property ranges upon the *lt:Relation* class, which is a sub-class of the generic *Relation* class and therefore inherits the *Relation* class properties.

5.2.2.2.2 Event

The Event Ontology[164] as imported by VIVO defines *Event* as "An arbitrary classification of a space/time region, by a cognitive agent. An event may have actively participating agents, passive factors, products, and a location in space/time." The overview of terms [Raimond & Abdallah 2007] also define classes such as *Factor:* "Everything used as a factor in an event", *Product* „Everything produced by an event", and properties: *agent*, *agent_in*, *factor*, *factor_of*, *hasAgent*, *hasFactor*, *hasLiteralFactor*, *hasProduct*, *hasSubEvent*, *isAgentIn*, *isFactorOf*, *literal_factor*, *place*, *producedIn*, *product*, *sub_event*, *time* (Figure 41).

[164] The Event Ontology Version 1.0: http://motools.sourceforge.net/event/event.html [Raimond & Abdallah 2007] (Last visit: June 4th, 2012).

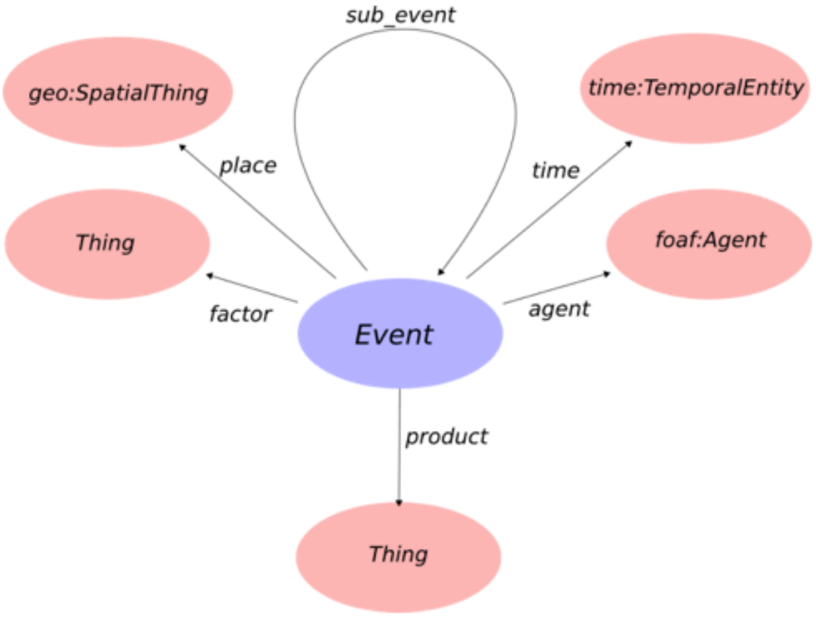
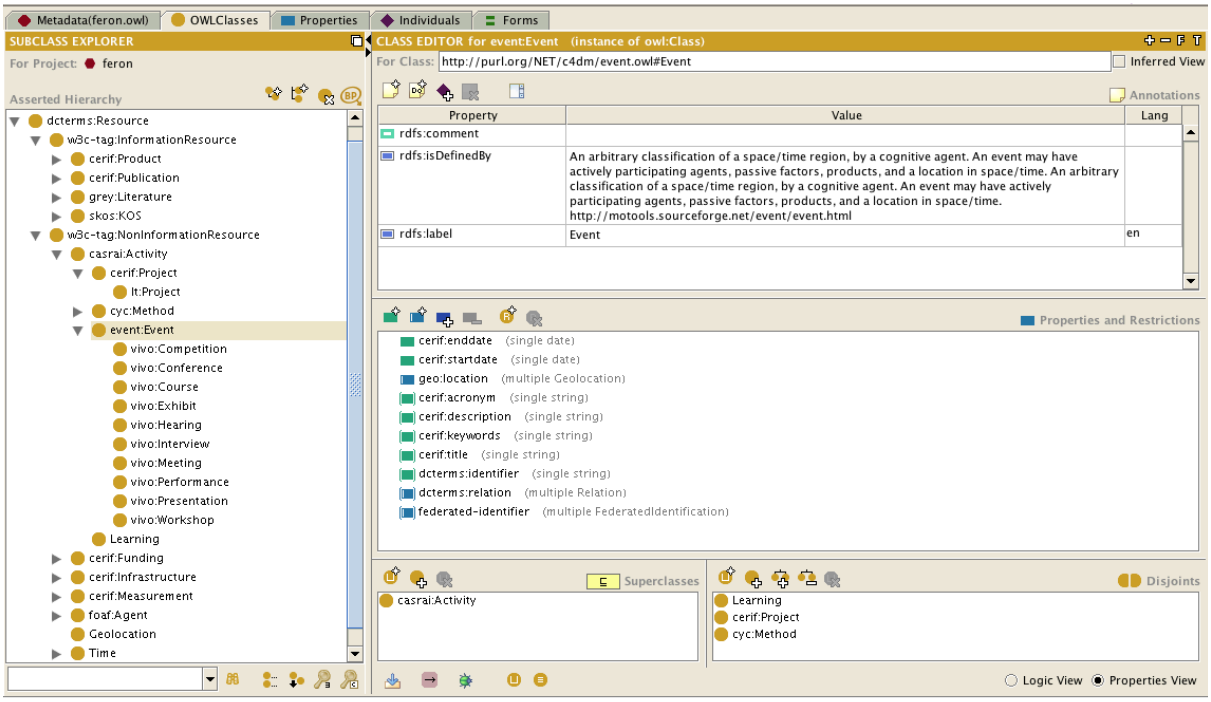


Figure 41: The Event Ontology [Raimond & Abdallah 2007]

In FERON, the Event ontology's [Raimond & Abdallah 2007] *Event* concept is applied, which is thus *event:Event*. VIVO subsumes under the class *event:Event Competition*, *Conference*, *Course*, *Exhibit*, *Hearing*, *Interview*, *Meeting*, *Performance*, *Presentation*, *Workshop*. These are applied as types, i.e. subclasses with the FERON *event:Event* class. VIVO employs with *event:Event* literals (i.e. intrinsic properties) such as the *contactInformation*, *description*, and *DateTimeIntervals*, but also geographic information, such as *domesticGeographicFocus*, *geographicFocus*, *hasGeographicLocation*, *international GeographicFocus* ranging over *GeographicRegion* and *GeographicLocation* imported from the Geo ontology (7.11 Geographic Location). FERON refers to *Geolocation* through the property *geographic-location* (slightly adopted from VIVO) – to indicate geographic features – but does not further specify them. It employs the *cerif:startdate* and the *cerif:enddate* properties at the *event:Event* class, where VIVO defines *DateTimeIntervals* as being "a specific period or duration, defined by (optional) start and end date/times". Furthermore, VIVO employs a mutual property *Role* upon the range of a *Role* class, and a *webpage* property upon the range of a *URLLink* class. It is perceived similar to FERON's inherited properties *dcterms:relation* and *federated-identifier* and therefore not imported.

The Science ontology downloaded from the Protégé library includes *Scientific Event* as a concrete class under an abstract *Event* class, and *Scientific Event* is further specified by two sub-classes *Live-Scientific-Event* and *Scientific-Publication-Event*, and inherits from Event the slots *name: required String*, *Initial-Date: required Date*, *Home-Page*, *Final-Date*, *Event-Publication: Scientific Document*. The *Scientific Event* class adds additional slots for description, such as: *Code*, *Acceptance-Date*, *Deadline*, *Final-Paper-Due*, *Home-Page*, *Organizing-Chair*, *Organizing-Committee*, *Program-Committee*, *Scientific-Chair*, *Subject-Areas*. CERIF describes scientific *Event* with attributes such as *cfEventId*, *cfCountryCode*, *cfCityTown*, *cfFeeOrFree*, *cfStartDate*, *cfEndDate*, *cfURI*, *cfName*, *cfDescription*, *cfKeywords* and by multiple relationships with *cfFunding*, *cfEquipment*, *cfOrganisationUnit*, *cfPerson*, *cfProject*, *cfResultPublication*.



Property	Value	Lang
rdfs:comment		
rdfs:isDefinedBy	An arbitrary classification of a space/time region, by a cognitive agent. An event may have actively participating agents, passive factors, products, and a location in space/time. An arbitrary classification of a space/time region, by a cognitive agent. An event may have actively participating agents, passive factors, products, and a location in space/time. http://motools.sourceforge.net/event/event.html	
rdfs:label	Event	en

cerif:enddate (single date)
cerif:startdate (single date)
geo:location (multiple Geolocation)
cerif:acronym (single string)
cerif:description (single string)
cerif:keywords (single string)
cerif:title (single string)
dcterms:identifier (single string)
dcterms:relation (multiple Relation)
federated-identifier (multiple FederatedIdentification)

Superclasses: casrai:Activity

Disjoints: Learning, cerif:Project, cyc:Method

Figure 42: FERON event:Event class

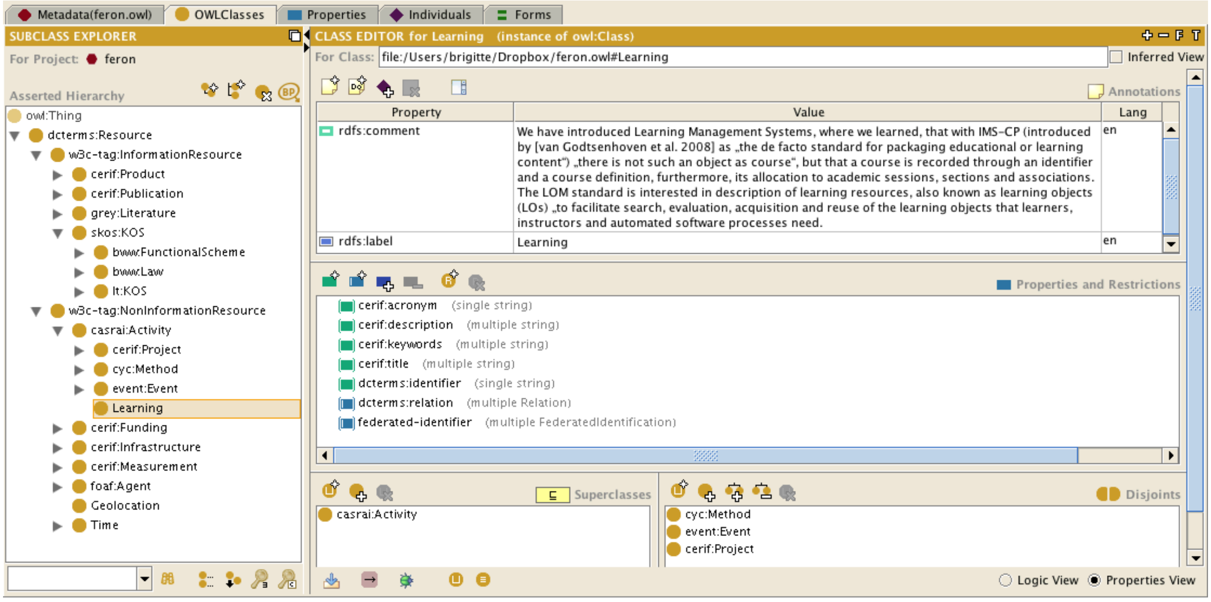
With FERON, relationships or mutual properties are managed by a *dcterms:relation* property upon a *Relation* class (see section 7.10 Time-aware Relationships). To give an idea of the range the Event Ontology in Figure 41 was presented. The Event Ontology deals with the notion of reified events. It defines one main **Event** concept, and an event may have a location, a time, active agents, factors and products”. The FERON *Event* class is not necessarily a candidate for field extension. Although events are field specific, field specific properties are not considered immediately intrinsic but more mutual and are thus maintained through the functional relationship construct. Relationships with *SubjectAreas* as proposed by the Science ontology are reflected in FERON through *dcterms:relation* in the range of records upon subclasses of the *skos:SKOS* class – which at implementation level may employ logical constraints or rules as to only allow for lt-related functions. The FERON *event:Event* class imports sub-classes from *event:Event* like VIVO, namely *vivo:Competition*, *vivo:Conference*, *vivo:Course*, *vivo:Exhibit*, *vivo:Hearing*, *vivo:Interview*, *vivo:Meeting*, *vivo:Performance*, *vivo:Presentation*, *vivo:Workshop* – as indicated in Figure 42.

5.2.2.2.3 Learning

In section 4.1.6 Learning Management Systems were introcuded, to indicate, that IMS-CP (as introduced by [van Godtsenhoven et al. 2008] is "the de facto standard for packaging educational or learning content"), and where "there is not such an object as course", but that a course is recorded through an identifier and a course definition, furthermore, its allocation to academic sessions, sections and associations. The LOM standard is interested in describing learning resources, also known as learning objects (LOs) "to facilitate search, evaluation, acquisition and reuse of the learning objects that learners, instructors and automated software processes need. LOM is well suited for LO cataloguing and string-based searches, but it lacks of the semantic expressiveness to enable semantic searches. For this reason applications using LOs metadata are evolving their metadata representations by adding semantic structures (Al-Khalifa and Davis 2006)" [cited in Rodriguez et al. 2009, p. 1][165]. "LOM also provides several kinds of relationships between LOs organized as a recommended list of appropriate values (vocabulary, in terms of LOM). These relationships are based on the relationships proposed by Dublin Core" [cited in Rodriguez et al. 2009, p. 2] and provided by the IEEE LOM standard: *is part of* (has part); *is version of* (has version), *is format of* (has format), *is referenced by* (references), *is based on* (is basis for), *is required by* (requires). LOM does not explicitly include the notion of LO class [Rodriguez et al. 2009, p. 2] and therefore does not allow for establishing relationships at class level. The internal organisational structures in the IEEE LOM standard define: *Atomic*, *Collection*, *Networked*, *Hierarchical*, *Linear* [Rodriguez et al. 2009, p. 3, table 2]. VIVO subsumes *vivo:Course* under *event:Event*, with the properties presented in section 5.2.2.2.2 Event, but additionally models *EducationalTraining* as a class, being a subclass of *dateTimeIntervall* and *supplementalInformation*, and with subclasses such as *Internship*, *MedicalResidency*, and *PostdoctoralTraining*. VIVO describes the class *vivo:EducationalTraining* with properties such as *vivo:trainingAtOrganization*, or *vivo:departmentOrSchool* ranging upon *foaf:Organization*, and *vivo:advisingContributionTo*,

[165] Online Version [Rodriguez et al. 2009, p. 1] http://personal.uoc.edu/personalonto/files/MTSR09_rodriguez.pdf (Last visit: June 6th, 2012).

vivo:educationalTraining with the domain *foaf:Person*, and a *vivo:EducationalTraining* class in the range of properties such as *vivo:degreeOutcomeOf*. In FERON, the Learning class will not be elaborated – it is too big a subject on its own and more suited for the Science-to-Society area – only the inherited properties are presented, namely *cerif:acronym, cerif:title*, *cerif:keywords, cerif:description*, *dcterms:identifier*, *federated-identifier*, *dcterms:relation*, which allows for linkage with the imported *vivo:Course* class and thus for a basic description of Learning Objects (Figure 43). No subclasses are modeled as e.g. types of Learning.



owl:Thing
- dcterms:Resource
  - w3c-tag:InformationResource
    - cerif:Product
    - cerif:Publication
    - grey:Literature
    - skos:KOS
      - bww:FunctionalScheme
      - bww:Law
      - lt:KOS
  - w3c-tag:NonInformationResource
    - casrai:Activity
      - cerif:Project
      - cyc:Method
      - event:Event
      - Learning
    - cerif:Funding
    - cerif:Infrastructure
    - cerif:Measurement
    - foaf:Agent
    - Geolocation
    - Time

Property	Value	Lang
rdfs:comment	We have introduced Learning Management Systems, where we learned, that with IMS-CP (introduced by [van Godtsenhoven et al. 2008] as „the de facto standard for packaging educational or learning content") „there is not such an object as course", but that a course is recorded through an identifier and a course definition, furthermore, its allocation to academic sessions, sections and associations. The LOM standard is interested in description of learning resources, also known as learning objects (LOs) „to facilitate search, evaluation, acquisition and reuse of the learning objects that learners, instructors and automated software processes need.	en
rdfs:label	Learning	en

cerif:acronym (single string)
cerif:description (multiple string)
cerif:keywords (multiple string)
cerif:title (multiple string)
dcterms:identifier (single string)
dcterms:relation (multiple Relation)
federated-identifier (multiple FederatedIdentification)

Superclasses: casrai:Activity

Disjoints: cyc:Method, event:Event, cerif:Project

○ Logic View ● Properties View

Figure 43: FERON Learning class description modeled with Protégé

5.2.2.2.4 Method

Traditional information sources about *Research Methods* collected terminology together with methods and techniques, e.g. (Lewis-Beck, Bryman and Liao, 2004 cited in [Sicilia 2010, p. 249]). "However, these are not prepared for use in computer-based systems, but only provide the main definitions from which models or ontologies can be devised. Elements of the scientific method can be found in existing thesauri. For example, in the UNESCO Thesaurus, the microthesaurus 2.05 is devoted to the "Scientific approach" and provides terms for research work (e.g. *Design*, *Experiments*, *Case studies*) and research methods (e.g. *Qualitative analysis*, *Sampling*, *Forecasting*). Also, there are some "scientific equipments" defined in other microthesaurus (with terms as *Laboratory equipments*, *Microscopes* or

Plankton recorders). It can be used as a point of departure for an ontology of methods, however it is rather generic and incomplete. [...] There are only some scattered reports concerning research methods terminologies or ontologies, however, to our knowledge there have been no attempts to systematically describe research outcomes using method-aware descriptions." [Sicilia 2010, p. 249].[166]

In Cyc, a query for *method* results in a collection of technique with an English ID of a so-called *TechniqueType*, and aliases such as *method*, *methods*, and *techniques*. A *TechniqueType* is defined as „the collection of types of practical actions that require specific skills and are used for specific purposes or tasks. Although *TechniqueType* is a specialization of *SkilledActivityType*, techniques are usually not main, general or primary activities, but rather sub-activities that accomplish more specific elements of more general *SkilledActivityTypes*. For example, *DoingMath* would not be a instance of *TechniqueType*, but a spec of *SolvingAMathematicalEquation* could be as long as it represents a particular approach towards solving a math problem. The Cyc *TechniqueType* incorporates *scientific technique*, *teaching method* and *verification method* as subtypes. The *scientific technique* is defined as "the collection of types of actions which are specific to the pursuit of scientific fields of study. The collection includes (at least) laboratory techniques, techniques used to gather data in the field, and those involved in the use of any *ScientificInstrument* for any scientific purpose"[167]. CASRAI considers *Research Technique* as output and defines it: "A practical methods or skills applied to particular tasks identified as part of the research." In FERON, the Cyc alias-concept of method is applied, being *cyc:Method*, while not further sub-typed with *techniqueType*, *teaching* or *verification*, but rather with field-specific subclasses e.g. *lt:Method* to anticipate scientific field methods. FERON's *cyc:Method* class is truely a candidate for field extension with sub-classes such as *lt:Method*, *csMethod* or *mathMethod*. Within the *lt:Method* class, the *lt:relation* property emerges inline with, and as explained under section 7.6. A method in FERON is thus (being at the same level as Learning) described through inherited intrinsic properties such as *cerif:acronym*, *cerif:description*, *cerif:keywords*, *cerif:title*, and the emerging properties with *dcterms:Resource* – namely *dcterms:identifier*, *dcterms:relation*, *federated-identifier*.

[166] EXPO (http://expo.sourceforge.net/) – the Ontology of scientific experiments, defines over 200 concepts for creating semantic markup about scientific experiments, using OWL. It is in used by the Robot Scientist: http://www.aber.ac.uk/en/cs/research/cb/projects/robotscientist/ (Last visit: June 11th, 2012)

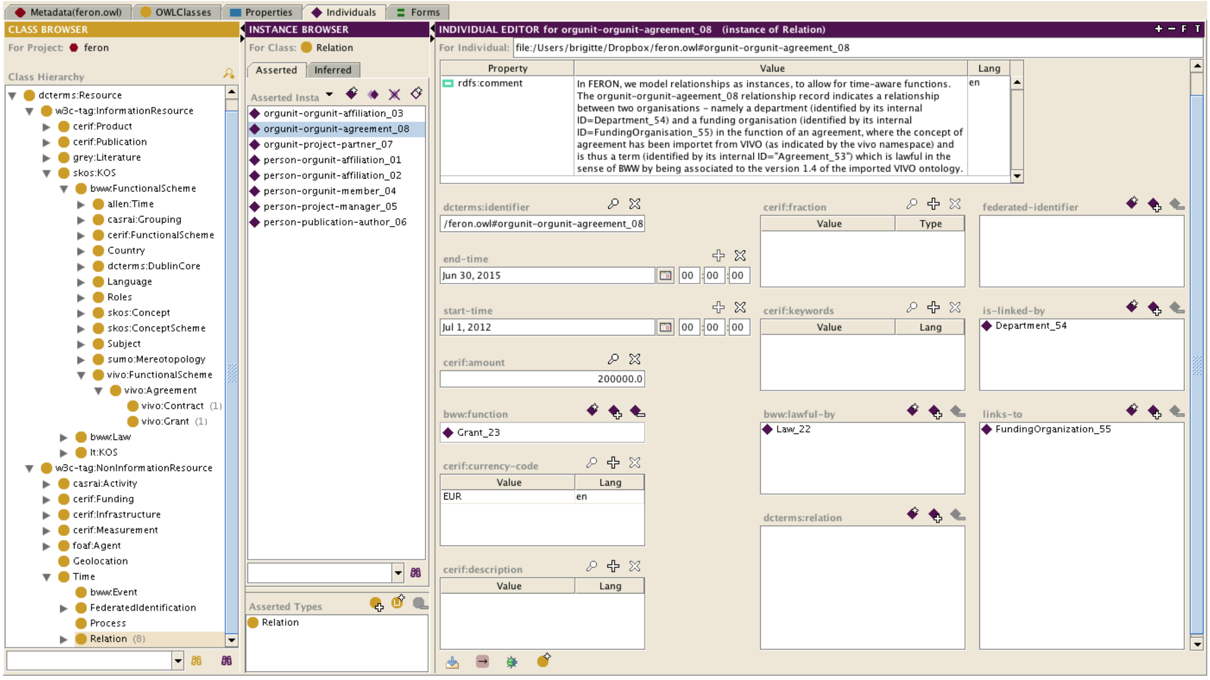
[Li et al. 2010, p. 139] explain: "EXPO was developed in a top-down manner by extending concepts in the Suggested Upper Merged Ontology (SUMO). Although very comprehensive, these models are fairly verbose and not very suitable as models for developing data management systems."

[167] OpenCyc Collection: scientific technique: http://sw.opencyc.org/concept/Mx4rvyI005wpEbGdrcN5Y29ycA (Last visit: June 12th, 2012)

FERON features a contextually-neutral relationship construct (section 7.10 Time-aware Relationships) upon the range of a *Relation* class which is also consistently applied for lt-specific relationships, i.e. *lt:relation*.

5.2.2.3 Funding

The FERON class *Funding* is inspired by the CERIF entity *cfFunding*. It is as sister-class of the *casrai:Activity* under *w3c-tag:NonInformationResource* and employs intrinsic properties such as *cerif:title*, *cerif:description*, *cerif:keywords*, *cerif:acronym* and *cerif:amount*. In FERON types of *Funding* such as *cerif:Programme*, *cerif:Call*, *cerif:Tender* are sub-classed and follow the CERIF 1.3 Vocabulary[168]. FERON manages mutual properties through relationships. Investigating VIVO reveals *vivo:FundingOrganization* as a subclass of *vivo:Organization* and vice versa, implying they are considered same-as classes in logical terms.



For Individual: file:/Users/brigitte/Dropbox/feron.owl#orgunit-orgunit-agreement_08

Property	Value	Lang
rdfs:comment	In FERON, we model relationships as instances, to allow for time-aware functions. The orgunit-orgunit-ageement_08 relationship record indicates a relationship between two organisations - namely a department (identified by its internal ID=Department_54) and a funding organisation (identified by its internal ID=FundingOrganisation_55) in the function of an agreement, where the concept of agreement has been importet from VIVO (as indicated by the vivo namespace) and is thus a term (identified by its internal ID="Agreement_53") which is lawful in the sense of BWW by being associated to the version 1.4 of the imported VIVO ontology.	en

dcterms:identifier: /feron.owl#orgunit-orgunit-agreement_08

end-time: Jun 30, 2015 00:00:00

start-time: Jul 1, 2012 00:00:00

cerif:amount: 200000.0

bww:function: Grant_23

cerif:currency-code

Value	Lang
EUR	en

cerif:description

Value	Lang

cerif:fraction

Value	Type

cerif:keywords

Value	Lang

bww:lawful-by: Law_22

dcterms:relation

federated-identifier

is-linked-by: Department_54

links-to: FundingOrganization_55

Figure 44: FERON Agreement relationship record view from within Protégé

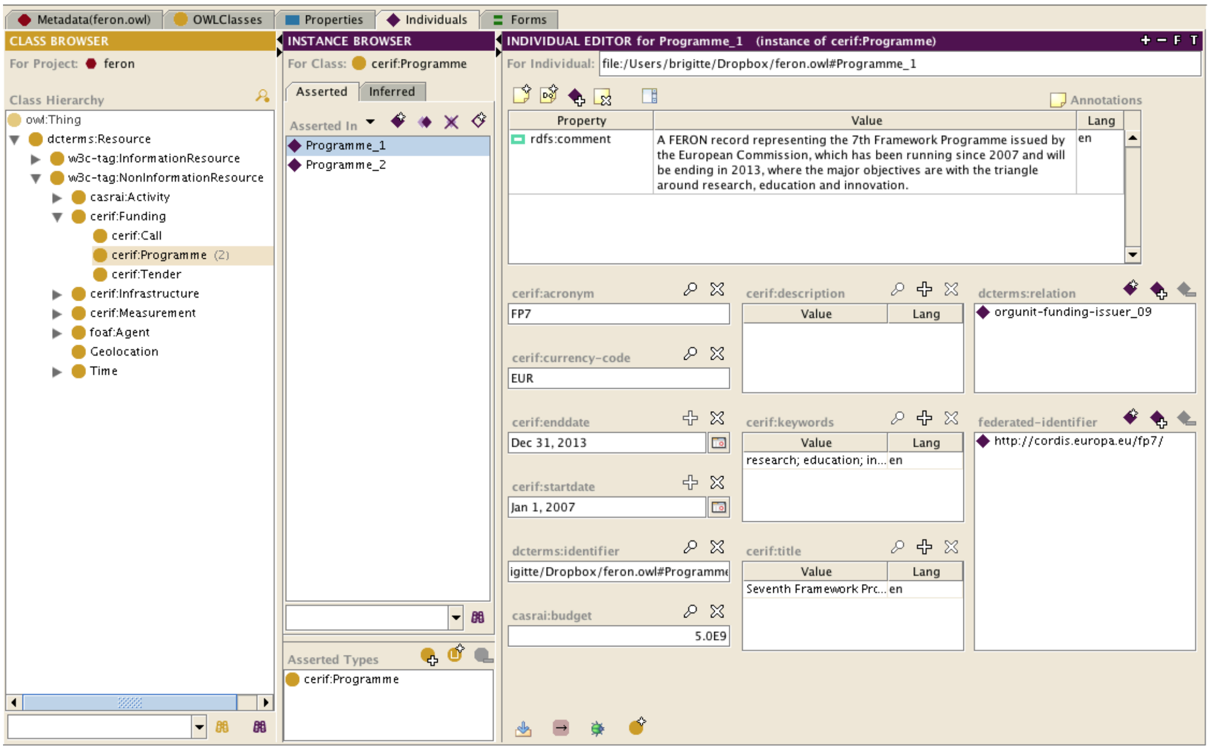
[168] The CERIF 1.3 Vocabulary is provided as an Excel sheet following the structure of the CERIF ERM. Each defined term has an assigned uuid for unique identification: http://www.eurocris.org/Uploads/Web%20pages/CERIF-1.3/Specifications/CERIF1.3_Semantics.pdf (Last visit: April 2nd, 2012)

The VIVO ontology does not explicitly contain a *Funding* class but employs an *Agreement* class with sub-classes *Contract* and *Grant*. FERON models them like in VIVO, by importing them as functional scheme *vivo:FunctionalScheme vivo:Agreement* for reference in relationships such as between organisations, e.g. between *vivo:FundingOrganization* and *vivo:Department* as depicted in Figure 44. CASRAI subsumes *Budget* as *Details* of an Activity profile, and employs FundingRequest. These concepts are considered as belonging to pre-award process descriptions from the perspective of e.g. a University. Figure 44 shows an explicit record of a FERON agreement relationship as an instance indicating a time-aware lawful function, where in the shown record a valid time-span starts with *July 1st, 2012* and ends in *June 30th, 2015*. The function is an agreement as defined by VIVO, hence a reference to the contained instance *Grant_23* in FERON, which is inline and lawful according to *Law_22* and thus inline with BWW. The *orgunit-orgunit-agreement_08* record (which has an additional counting number in the URI for uniqueness) describes a relationship between two instantiated organisation-units – one a department identfied as *Department_54*, the other a funding organisation identified as a *FundingOrganisation_55*. The FERON relationship record allows for features such as *cerif:amount* and *cerif:currency-code* imported from CERIF, which are especially relevant in a funding context. Because the introduced relationship records are instances of the *Relation* concept, which itself is a *dcterms:Resource (*sub-class of *w3c-tag:NonInformationResource)*, the *federated-identifier*, *dcterms:relation*, *dcterms:identifier* properties are inherited. For more details about the relationship construct see section 7.10 Time-aware Relationships.

VIVO provides a rich set of organisation types with its ontology, as subclasses of *foaf:Organization*, and which FERON employs except from *vivo:Program*, subsumed under Funding – thus perceived as a Funding type which FERON already covers with the *cerif:Programme* from the CERIF 1.3 Vocabulary. If *Grant* is considered as a funding type as in CASRAI, then in FERON, it would indeed be model as a subclass of *Funding*. The same holds for *Contract* and *Award* if understood as types of *Funding*. However, if *Contract* or *Grant* like in VIVO are considered an *Agreement* anticipating a relationship, then *Agreement* as modeled in FERON is considered a function of inter-organisational relationships and therefore subsumed under functional scheme. In the case of FERON, the *vivo:FunctionalScheme* anticipates lawfulness according to the *bww:FunctionalScheme* as explained above with investigation of the recorded instance and as depicted in Figure 44.

5.2.2.3.1 Programme

Funding Programmes are types of *Funding* (where in VIVO, Program is a subclass of Organisation) and in FERON comparable to *Call* and *Tender*. A FERON Funding Programme (the concept *cerif:Funding*), is thus featured by intrinsic properties such as *cerif:acronym*, *cerif:title*, *cerif:description*, *cerif:keywords*. FERON imported *casrai:budget* as a datatype property, to distinguish it clearly from a *cerif:amount* in relationship records as shown in the previous section. Budget is clearly allocated to *Funding* and amount is more in relationships and anticipated e.g. with contracts or grant agreements. Finally, *Funding Programme* is considered a non-information resource *w3c-tag:NonInformationResource* and a sub-class of *dcterms:Resource*. It inherits the *federated-identifier* and *dcterms:relation* and the *dcterms:identifier* properties. A Funding Programme record in FERON is described with Protégé as presented in Figure 45.



owl:Thing
- dcterms:Resource
  - w3c-tag:InformationResource
  - w3c-tag:NonInformationResource
    - casrai:Activity
    - cerif:Funding
      - cerif:Call
      - cerif:Programme (2)
      - cerif:Tender
    - cerif:Infrastructure
    - cerif:Measurement
    - foaf:Agent
    - Geolocation
    - Time

Programme_1
Programme_2

Asserted Types: cerif:Programme

For Individual: file:/Users/brigitte/Dropbox/feron.owl#Programme_1

Property	Value	Lang
rdfs:comment	A FERON record representing the 7th Framework Programme issued by the European Commission, which has been running since 2007 and will be ending in 2013, where the major objectives are with the triangle around research, education and innovation.	en

cerif:acronym: FP7

cerif:currency-code: EUR

cerif:enddate: Dec 31, 2013

cerif:startdate: Jan 1, 2007

dcterms:identifier: igitte/Dropbox/feron.owl#Programme

casrai:budget: 5.0E9

cerif:description

Value	Lang

cerif:keywords

Value	Lang
research; education; in...	en

cerif:title

Value	Lang
Seventh Framework Prc...	en

dcterms:relation: orgunit-funding-issuer_09

federated-identifier: http://cordis.europa.eu/fp7/

Figure 45: FERON Funding Programme record describing FP7

5.2.2.3.2 Income

Where in FERON, the Funding Programme is indicated as being a sub-class and thus type of *Funding*, the concept of *Income*, which CASRAI defines as budget "Information detailing the projected revenues and expenditures for the Research Activity" is truely a more complex one and inherently perceived *financial*, and thus considered a *Measurement* in FERON. In Figure 16, a view of the Enhanced Academic Domain (AID) was presented, where CRISs are seen in the center – mediating between *satellite systems* such as ERP systems (essentially rooted in the financial domain). In FERON, a *casrai:budget* property emerges at *Funding*; to anticipate budgets assigned with *Funding* e.g. *programme*, *call*, or *tender*, and where details or particular amounts *cerif:amounts* of income or expenditures (costs) are recognised as especially occuring within relationship recordings as indicated and explained in the previous section alongside Figure 44. FERON aims at analysing and thus explaining the Research domain in general and LT in particular, and does not further elaborate funding-related income or expenditure calculations, which are considered outside of the scope of this work due to essentially roots in financial accounting practices and further to be managed and maintained with ERP systems to ensure compliance with existing laws.

5.2.2.4 Measurement

Measurements have become increasingly important internationally and as well at European level "Universities rankings are increasingly popular" [EC Report 2010, *foreword*]. In Germany, there is e.g. the so-called 'Exzellenzinitiative' which started in 2005, where in the UK, there is currently a change from RAE (Research Assessment Exercise) to REF (Research Excellence Framework)[169] to become effective in 2014, and where the UK universities will have to report their outcomes and inform funding bodies towards their selective allocation and decision over future funding. The first initiatives towards formalizing in this respect started in Australia "Measuring the impact of Research"[170], which have guided further developments in the UK – in particular the findings in the JISC-funded MICE[171] project from which a conceptual model (Figure 46) resulted, that was proposed for uptake in the CERIF standard. The *MICE* model was taylored to impact indicators and measurements, and the CERIF implementation has been further influenced by another UK project in this respect,

[169] Research Excellence Framework is the new system for assessing the quality of research in UK higher education institutions (HEIs). It will replace the RAE 2008: http://www.ref.ac.uk/ A rough overview of assessment activities in Europe and beyond has been provided in [Jörg 2012, table 2]

[170] Measuring the Impact of Research: http://www.atn.edu.au/docs/Research%20Global%20-%20Measuring%20the%20impact%20of%20research.pdf (Last visit: April 2nd, 2011)

[171] Measuring Impact under CERIF (MICE) – a JISC-funded project: http://mice.cerch.kcl.ac.uk/ (Last visit: April 2nd, 2011)

namely *CERIFy*[172] modeling business cases (amongst them Esteem which was perceived as to being inverse to impact). The MICE proposal as well as the CERIFy results have finally influenced the CERIF 1.3 model employing generic Measurment *cfMeasurement* and Indicator *cfIndicator* entities.

Measurement has become a critical factor in Research activities and especially with respect to outputs and outcomes, i.e. impact. However, the concepts over *Measurement* and *Indicator* are not particularly rooted in Research but rather inherited from a higher level. *Measure* is e.g. a concept in the SUMO ontology (see Figure 7), where it is subsumed under *Numeric*. Because of its importance, in FERON the *Measurement* concept is a class under *w3c-tag:NonInformationResource*. However, will not be further elaborated as a class and possible sub-classes such as *Impact* or *Innovation*, because *Measurement* is essentially not a Research entity – but a means about (to measure) Research (see also section 5.1 discussion about the range of Research and involved entities). [Sicilia 2010] in this context cited Bunge (1967) who considered "science a style of thinking, and as with any human outcome, there is a need to distinguish between its final outcomes – knowledge – and its work process – research work" [Sicilia 2010, p. 247].

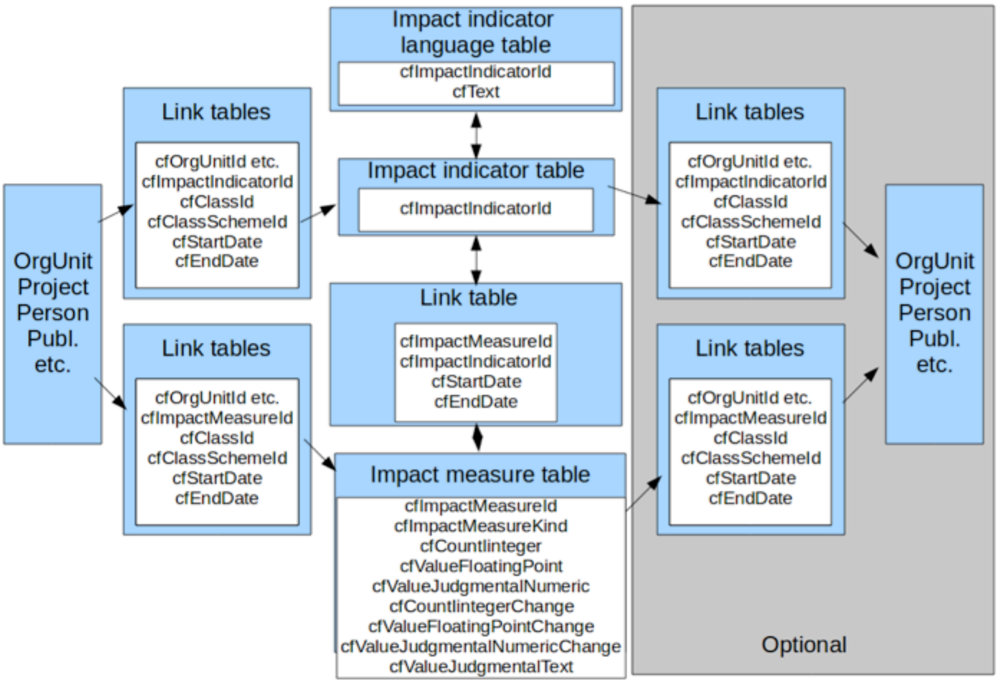


Figure 46: MICE (Measuring Impact under CERIF) conceptual model [Cox et al. 2011]

[172] CERIFy – a JISC-funded project to increase the engagement with CERIF in the UK Higher Education sector: http://cerify.ukoln.ac.uk/ (Last visit: June 7th, 2012)

With the *Beyond Impact* project[173], the Open Science Foundation indicated interest in this respect. Furthermore, CASRAI presented results with first approaches towards measuring impact by drafting a catalogue of impact indicators[174] and by organising a first CASRAI conference under the theme "Occupy Impact"[175] in October 2012: "Measuring the impact of a highly diverse research community is a hard problem but one that we need to 'occupy' together as a community in order to make it something we own collectively."

5.2.2.5 Infrastructure

The EC-funded MERIL[176] project defines: "European Research Infrastructure is a facility or (virtual) platform that provides the scientific community with resources and services to conduct top-level research in their respective fields. These research infrastructures can be single-sited or distributed or an e-infrastructure, and can be part of a national or international network of facilities, or of interconnected scientific instrument networks." With its definition it follows definitions from the European Commission and the European Strategy Forum on Research Infrastructures (ESFRI)[177]. For this work, another definition of *Cyberinfrastructure* provided by the National Science Foundation Blue-Ribbon Advisory Panel [Atkins et al. 2003] is considered particularly pregnant, because it includes a view of involved entities "the opportunity is here to create cyberinfrastructure that enables more ubiquitous, comprehensive knowledge environments that become functionally complete for specific research communities in terms of people, data, information, tools and instruments and that include unprecedented capacity for computational, storage and communication. Such environments enable teams to share and collaborate over time and over geographic, organizational, and disciplinary distance. They enable individuals working alone to have access to more and better information and facilities for discovery and learning. They can serve individuals, teams and organizations in ways that revolutionize what they can do, how they do it, and *who*

[173] Beyond Impact (an Open Science Foundation funded project): http://beyond-impact.org/ (Last visit: May 2nd, 2012)

[174] A Framework for Research Impact Data Standards: http://www.casrai.org/sites/casrai.org/files/draft_catalogue_of_impact_indicators.pdf (Last visit: May 2nd, 2012)

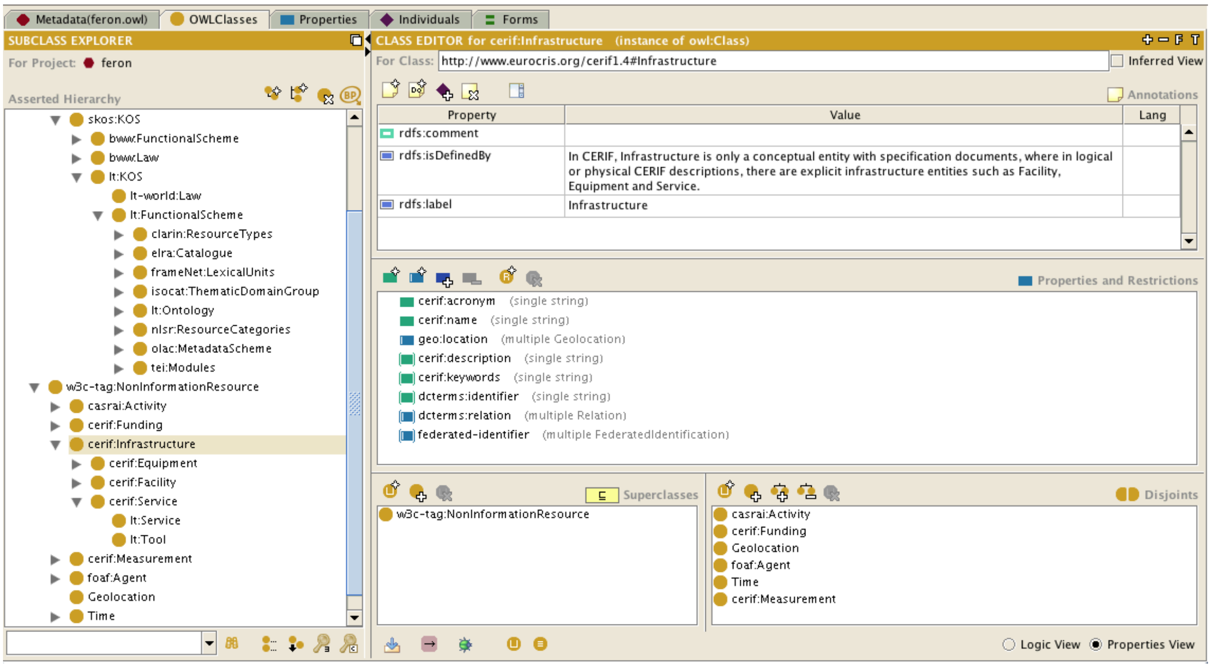
[175] "The theme of reconnect12 is: **Occupy Impact**. We feel the 'occupy' meme fits the subject well. In our case occupy is not about protest or revolution. It is about getting inside a difficult issue and tackling it as a community." http://reconnect.casrai.org/ (Last visit: May 2nd, 2012)

[176] Mapping of the European Research Infrastructures (MERIL): http://www.esf.org/activities/science-policy/research-infrastructures/meril-mapping-of-the-european-research-infrastructure-landscape/what-is-meant-by-research-infrastructures.html (Last visit: June 7th, 2012)

[177] European Strategy Forum on Research Infrastructure (ESFRI): http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=esfri (Last visit: April 2nd, 2011)

participates” (pp. 12–13). In CASRAI, *Infrastructure* is one type out of a list of Funding Target Types such as *Equipment*, *Establishment*, *Infrastructure*, *Operating*.

The structure of this section follows the FERON ontology and is inspired by the CERIF concept of *Research Infrastructure* composed of the classes *cerif:Facility*, *cerif:Equipment*, and *cerif:Service*. Infrastructure is thus a concept that is particularly important with FERON, because it is highly field-specific. With chapter 6 Analysis of Language Technology Entities, the field specific requirements will be further discussed, where now, the three FERON *Infrastructure* entities will be presented in some detail through Figure 47. More specific features such as e.g. time-aware relationships, identities, namespaces, geographic location or KOSs, will be investigated in chapter 7 FERON – *F*ield-*e*xtensible *R*esearch *ON*tology.



For Project: feron

- ▼ skos:KOS
  - ▶ bww:FunctionalScheme
  - ▶ bww:Law
  - ▼ lt:KOS
    - lt-world:Law
    - ▼ lt:FunctionalScheme
      - ▶ clarin:ResourceTypes
      - ▶ elra:Catalogue
      - ▶ frameNet:LexicalUnits
      - ▶ isocat:ThematicDomainGroup
      - ▶ lt:Ontology
      - ▶ nlsr:ResourceCategories
      - ▶ olac:MetadataScheme
      - ▶ tei:Modules
- ▼ w3c-tag:NonInformationResource
  - ▶ casrai:Activity
  - ▶ cerif:Funding
  - ▼ cerif:Infrastructure
    - ▶ cerif:Equipment
    - ▶ cerif:Facility
    - ▼ cerif:Service
      - lt:Service
      - lt:Tool
  - ▶ cerif:Measurement
  - ▶ foaf:Agent
  - Geolocation
  - ▶ Time

CLASS EDITOR for cerif:Infrastructure (instance of owl:Class)

For Class: http://www.eurocris.org/cerif1.4#Infrastructure ☐ Inferred View

Property	Value	Lang
rdfs:comment		
rdfs:isDefinedBy	In CERIF, Infrastructure is only a conceptual entity with specification documents, where in logical or physical CERIF descriptions, there are explicit infrastructure entities such as Facility, Equipment and Service.	
rdfs:label	Infrastructure	

cerif:acronym (single string)
cerif:name (single string)
geo:location (multiple Geolocation)
cerif:description (single string)
cerif:keywords (single string)
dcterms:identifier (single string)
dcterms:relation (multiple Relation)
federated-identifier (multiple FederatedIdentification)

Superclasses: w3c-tag:NonInformationResource

Disjoints: casrai:Activity, cerif:Funding, Geolocation, foaf:Agent, Time, cerif:Measurement

Figure 47: FERON’s cerif:Infrastructure class with subclasses

5.2.2.5.1 Facility

In the CERIF 1.3 Vocabulary, *Facility* is defined as “a space or equipment necessary for conducting research”. With FERON, *cerif:Facility* is modeled as a type and thus a sub-class of *cerif:Infrastructure*, which inherits FERON’s *federated-identifier*, *dcterms:dentifier* and the *dcterms:relation* properties from *dcterms:Resource*. The *cerif:Facility* in FERON is essentially subsumed under *w3c-tag:NonInformationResource*, its *cerif:Infrastructure* sub-class features a functional *cerif:acronym*, and *cerif:name*, *cerif:description*, *cerif:keywords* datatype properties. Furthermore, it maintains an object property *geo:location* upon the range of *Geolocation*. VIVO employs *vivo:Facility* as a sub-class of *vivo:GeographicLocation*, itself a sub-class of *vivo:Location* further classified by *Building* and *Room*. VIVO does not explicitly model properties with *vivo:Facility*, but employs a definition available from the Free Online Dictionary describing *Facility* as being "[d]istinct from the organization that runs it; e.g., a laboratory may be an organization but may be run by another organization and only consist of facilities housing equipment or services. Can be a building or place that provides a particular service or is used for a particular activity. Use the specific Building or Room whenever possible." The Science ontology does not consider a *Facility* class.

FERON subsumes *lt:Facility* under *cerif:Facility* (alike it subsumes the *lt:Service* below the *cerif:Service*) to indicate openness with respect to field extensions (Figure 47).

5.2.2.5.2 Equipment

The CERIF 1.3 Vocabulary employs WordNet for the definition of *Equipment*, that is, "an instrumentality needed for undertaking or to perform a service"[178]. In FERON, the class *cerif:Equipment* is a type of infrastructure and therefore a subclass of *cerif:Infrastructure*. It inherits *federated-identifier*, *dcterms:identifier* and *dcterms:relation* from *dcterms:Resource* because *cerif:Facility* and *cerif:Service* are essentially a *w3c-tag:NonInformationResource*. The *cerif:Infrastructure* class introduces a functional *cerif:acronym* datatype property, and *cerif:name*, *cerif:description*, *cerif:keywords* datatype properties. Furthermore, it features the object property *geo:location* upon the range of the *Geolocation* class. VIVO models the *vivo:Equipment* as a subclass of *vivo:Thing* without further classification. VIVO provides a *vivo:freetextKeyword* with datatype *Literal* and a *vivo:webpage* property upon the range of the *vivo:URLLink* class for *vivo:Equipment*. The VIVO ontology employs as short definition of *vivo:Equipment*: "A physical object provided for specific purpose, task or occupation." The Science ontology downloaded from the public Protégé library does not consider *Equipment* as a class. In CASRAI, *Equipment* is one type out of a list of *Funding Target Types* such as *Equipment*, *Establishment*, *Infrastructure*, *Operating*.

FERON subsumes *lt:Equipment* under *cerif:Equipment* (alike it subsumes the *lt:Service* below the *cerif:Service*) to indicate openness with respect to field extensions (Figure 47).

[178] WordNet Search 3.1 "equipment": http://wordnetweb.princeton.edu/perl/webwn?s=equipment (Last visit: June 7th, 2012)

5.2.2.5.3 Service

In the CERIF 1.3 Vocabulary, *Service* is defined as “an exchange for money or other commodities where an enduser receives support from a supplier”. FERON models the *cerif:Service* as a type and thus subclass of the *cerif:Infrastructure* class, to inherit *federated-identifier*, *dcterms:identifier* and *dcterms:relation* properties from *dcterms:Resource*. The *cerif:Service* is a *w3c-tag:NonInformationResource* subsumed under *cerif:Infrastructure*. The *cerif:Infrastructure* class features a functional *cerif:acronym* property, and *cerif:name*, *cerif:description*, *cerif:keywords* datatype properties. Furthermore, it maintains an object property *geo:location* upon the range of the *Geolocation* class. Taking a look at VIVO reveals *vivo:Service* as a subclass of *vivo:Thing*. VIVO models an object property *vivo:contributingRole* upon the range of *vivo:Role* and as with *vivo:Equipment*, a webpage property upon the range of a *URLLink* class. VIVO employs a short definition for the *vivo:Service* class: “A regularly offered service in support of an academic, research, or administrative function (not personal or professional service by an individual).” The Science ontology downloaded from the public Protégé library does not consider *Service* as a class.

FERON subsumes *lt:Service* under *cerif:Service* to indicate openness with respect to field extensions (see Figure 47).

5.2.3 Information Resource

During the last few decades, research information has mostly been interpreted as output in the format of publications. This is obvious through common means of measurement such as the journal impact factor[179], citations, or the h-index[180], which are based on publications only. Also patents have obviously played a role, being e.g. a single entity in CERIF, a class in VIVO (subclass of *bibo:Document*), and a kind of output in CASRAI. Growing significance is assigned to research data, which has become an “increasingly important re-usable product of research“ [Wolski et al. 2011, p. 1], and which most recently is often directly linked to funding or monetary income streams: “The Engineering and Physical Sciences Research Council (EPSRC) [a funding body in the UK] currently identifies research data metadata as a key part of the outputs from its funded activities” [Ginty et al. 2012, p. 3–4], and requests all institutions in receipt of funding to “have developed a clear roadmap to align their policies

[179] The Thomson Reuters Impact Factor: http://thomsonreuters.com/products_services/science/free/essays/impact_factor/ (Last visit: May 2nd, 2012)

[180] H-Index: http://en.wikipedia.org/wiki/H-index (Last visit: May 2nd, 2012)

and processes with EPSRC's expectations"[181]. The European Commission (EC) within its Open Access (OA) strategy fosters enhanced access to research data and results – at both European and national levels[182]. Another indicator for the importance of data (not only research-induced data) in general is the growing number of public websites hosting so-called *public sector information*, such as *data.eu*[183], *Data.gov*[184], *Opening up government*[185], *Australian National Data Service*[186]. Non-profit organisations such as DataCite[187] are dedicated to establishing easier access to research data and to increase their acceptance as legitimate citable contributions as well as to support archiving.

Outputs in the format of publications are often stored in institutional repositories[188], where the underlying format – in most cases has been Dublin Core. Dublin Core (DC) is entirely based on the concept of a *Resource* and because of its widespread use in most of todays repositories, it has been investigated and discussed in depth with section 5.2.1 Resource. The *dcterms:Resource* is perceived as an overarching concept and thus is a class in FERON subsuming all other classes. Besides the introduced concept of Non-Information Resource in section 5.2.2, the *Information Resource* concept will now be elaborated and thus unfolded. It has been perceived and will hence be reflected in FERON, that there are three basic kinds of information resources – this approach follows thus the UNISIST model (1971) where these kinds have first been identified and distinguished – *formal*, *informal* and *tabular*. Such a categorisation is still considered very appropriate, especially with the emerging significance of data. However, with FERON – representing a perceived world of Research – in subsequent sections, *formal* is named *Publication*, *informal* is named *Literature* and *tabular* is named *Product*.

FERON's *w3c-tag:InformationResource* class inherits intrinsic properties from its superclass *dcterms:Resource* namely *dcterms:identifier*, *federated-identifier*, *dcterms:relation*, and these

[181] EPSRC Policy Framework on Research Data - Impact, Timescales and Support: http://www.epsrc.ac.uk/about/standards/researchdata/Pages/impact.aspx (Last visit: May 2nd, 2012)

[182] Science and Society Home page Research "The first results of the Open Access Pilot in FP7 will represent important inputs into the Commission's deliberations on the next steps needed to enhance access to research data and results at both the European and national levels." http://ec.europa.eu/research/science-society/index.cfm?fuseaction=public.topic&id=1300 (Last visit: May 2nd, 2012)

[183] No content available at the moment with: http://www.data.eu/ (Last visit: May 2nd, 2012)

[184] Data.gov An Official Website of the United States Goernment: http://www.data.gov/ (Last visit: May 2nd, 2012)

[185] Data.Gov.UK^{Beta} Opening Up Government: http://data.gov.uk/ (Last visit: May 2nd, 2012)

[186] Australian National Data Service: http://www.ands.org.au/ (Last visit: May 2nd, 2012)

[187] DataCite – Helping you to find, access, reuse research data: http://datacite.org/ (Last visit: May 2nd, 2012)

[188] Depending on the output type, the institutional repository may only store a pre-print of the published version, where the final version is with the publisher holding the copyright.

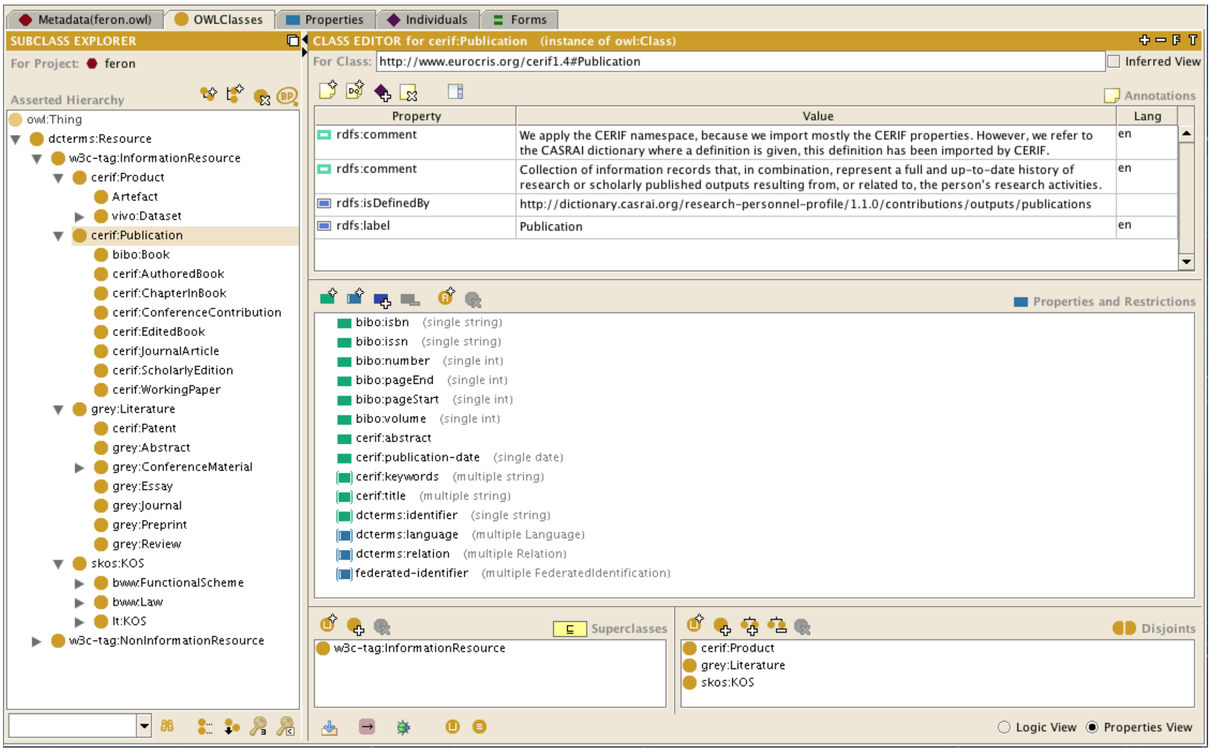
are further propagated to sub-classes. In addition, FERON features own intrinsic properties with *w3c-tag:InformationResource*, namely *cerif:keywords* and *cerif:title* and – very important – a *dcterms:language* object-type property upon the range of a *Language* class under *skos:KOS* and thus subsumed under *bww:FunctionalScheme*.

Contrary to the *w3c-tag:Non-InformationResource* class, there is no *cerif:description* property at this level, because with both classes *cerif:Publication* and *grey:Literature* one commonly talks of an abstract, and therefore a *cerif:abstract* property is featured with the classes *cerif:Publication* and *grey:Literature*, whereas a *cerif:description* property emerges with *cerif:Product* (e.g. *vivo:Dataset*) and with the *skos:KOS* and thus its sub-classes. VIVO models *InformationResource* as a sub-class of *Thing*, with *Document* as a sub-class thereof, subsuming types such as, *Article*, *AudioDocument*, *Catalog*, *Manuscript*, *Image*, and *Software*, *Thesis* etc. CASRAI labels similar types *Outputs* revealing a result-driven; e.g. a funder's view.

5.2.3.1 Publication

A indicated, the *cerif:Publication* class is perceived upon the *formal* as defined in the concept borrowed from the UNISIST (1971) model. This implies an understanding of *formal* in the sense of peer-reviewed rather than *formal* in a technical understanding or in knowledge representation speech. FERON's class *cerif:Publication* thus reflects the fact of a required 'accountability' (i.e. peer-review) as applied in assessment exercises, distinguishing a *formal* publication (e.g. journal article or book chapter) from *informally* published work which in fact FERON perceives as *grey:Literature*, being thus a class with its own sub-typed classes such as e.g. *Preprint*, *Conference Material*, *Essay*, *Review* or *Patent*. FERON aims at the representation of *formal* and *informal* information material in research information systems (where there is not the library or cataloguing ambition of collecting and granting access to all worldwide published resources [ISBD 2007] and collocated information, but) to record contextual, high-quality metadata understood as significant components of the research ecosystem towards more integration and exchange by improved (founded) semantics and thus towards better human understanding.

A FERON *cerif:Publication* is thus a *dcterms:Resource*; from which it it inherits intrinsic properties *federated-identifier*, *dcterms:identifier* and *dcterms:relation*. In addition, it inherits intrinsic properties from *w3c-tag:InformationResource*, namely *cerif:keywords*, *cerif:title* and *dcterms:language* (here, a difference to *w3c-tag:NonInformationResource* is perceived, where the *cerif:title* is not featured at this level, but only at lower levels because some non-information sub-classes feature a *cerif:name* property instead). In addition, *cerif:Publication* imports publication-inherent properties *cerif:publication-date*, *cerif:keywords*, *cerif:abstract*, and *bibo:pageStart*, *bibo:pageEnd*; *bibo:volume*; *bibo:number*, *bibo:isbn*, *bibo:issn* as indicated in Figure 48. The *cerif:Publication* class is typed through sub-classes such as *cerif:AuthoredBook* or *cerif:JournalArticle*[189], *bibo:Book*. It is obvious, that given types overlap with known *grey:Literature* types (depending on the definition of formal (e.g. only peer-reviewed)), and thus – in the course of time and scientific development, these subclasses may be interchanged between existing descriptions. With FERON it must therefore be ensured, that intrinsic properties in between these two classes are identical, so that at record level (in information systems), there is no loss of information if a type changes, in case that the sub-types change their super-class (such flexibility is important for implementations).



For Class: http://www.eurocris.org/cerif1.4#Publication

owl:Thing
dcterms:Resource
w3c-tag:InformationResource
cerif:Product
Artefact
vivo:Dataset
cerif:Publication
bibo:Book
cerif:AuthoredBook
cerif:ChapterInBook
cerif:ConferenceContribution
cerif:EditedBook
cerif:JournalArticle
cerif:ScholarlyEdition
cerif:WorkingPaper
grey:Literature
cerif:Patent
grey:Abstract
grey:ConferenceMaterial
grey:Essay
grey:Journal
grey:Preprint
grey:Review
skos:KOS
bww:FunctionalScheme
bww:Law
lt:KOS
w3c-tag:NonInformationResource

Property	Value	Lang
rdfs:comment	We apply the CERIF namespace, because we import mostly the CERIF properties. However, we refer to the CASRAI dictionary where a definition is given, this definition has been imported by CERIF.	en
rdfs:comment	Collection of information records that, in combination, represent a full and up-to-date history of research or scholarly published outputs resulting from, or related to, the person's research activities.	en
rdfs:isDefinedBy	http://dictionary.casrai.org/research-personnel-profile/1.1.0/contributions/outputs/publications	
rdfs:label	Publication	en

bibo:isbn (single string)
bibo:issn (single string)
bibo:number (single int)
bibo:pageEnd (single int)
bibo:pageStart (single int)
bibo:volume (single int)
cerif:abstract
cerif:publication-date (single date)
cerif:keywords (multiple string)
cerif:title (multiple string)
dcterms:identifier (single string)
dcterms:language (multiple Language)
dcterms:relation (multiple Relation)
federated-identifier (multiple FederatedIdentification)

Superclasses: w3c-tag:InformationResource

Disjoints: cerif:Product, grey:Literature, skos:KOS

Figure 48: FERON's cerif:Publication class definition view

FERON manages relationships or mutual properties through a *dcterms:relation* (labelled relationship) property upon the range of a *Relation* class. A *cerif:Publication* relationship in the role of e.g. an *Author* with e.g. a *foaf:Person* is thus recorded as an instance. As with all FERON entities, publication instances can be extended through multiple relationships, each

[189] E.g. in the UK assessment framework or in the Norwegian CRIStin system there is an official list of countable output types. These are required with submissions. There is awareness, that the current types or subtypes may well change between the current categories of *cerif:Publication* and *grey:Literature* – or may be perceived formally different.

being itself an instance of *dcterms:Resource* under the class *Relation* (a subclass of *Time*). A *Relation* record has its own URI (e.g. *relationship#person-publication-author_00*) to which e.g. the publication and person records refer, and of which multiple are allowed. Because there are uncountable numbers of relationships that a publication can maintain within the *Research* ecoystem (see Figure 38), it will not be further investigated here, but within section 7.10 Time-aware Relationships.

The International Federation of Library Associations and Institutions (IFLA) is an established and recognised organisation; the digital library community has a massive experience with managing knowledge, i.e. information resources (e.g. FRBR[190], ICP, ISBD, RDA). This is acknowledged, and therefore insight into the FRBR reference model (it could essentially be perceived as an upper ontology) is provided with this work. The cataloguing community with digital libraries and open access systems is increasingly moving to the same information space, and their work undertaken is therefore highly relevant also for the research ecosystem. The International Cataloguing Principles (ICP) with an obvious user-driven approach is investigated, but also the FRBR model is considered very relevant as a theoretic or upper model describing the bibliographic domain. The statement of principles which is commonly known as the ‘Paris Principles’ as approved by the International Conference on Cataloguing Principles in 1961, achieved to serve as a basis for international standardisation in cataloguing, where “most of the cataloging that were developed worldwide since that time followed the Principles strictly or at least to a high degree” [ICP 2009, p. 1]. With its latest statement, IFLA explains an effort to adapt the Principles to online catalogues and beyond; the first principle is to serve users and the scope is broadened from textual works to all types of materials, and furthermore, to all aspects of bibliographic and authority data [ICP 2009, p. 1]. The document contains a glossary explaining the relevant ICP terms with references to FRBR and IME ICC (IFLA Meetings of Experts for an International Cataloguing Code), such as *Agent*, *Concept*, *Content Type*, *Collection*, *Access Point*, *Creator*, *Entity*, *Event*, *Identifier*, *Relationship*, *Object*, etc.

Not only do Research-related activities and processes undergo dramatic changes, but also libraries are confronted with enormous change, new roles and positions. [Bianchini & Guerrini 2009, pp. 106 ff.] consider the change as a “switch from the functions of the catalog” – finding (specific search) and collocating (search for like material) – “to the needs of users” as reflected in the recommended models [FRBR 2009], [ICP 2009] particularly

[190] Functional Requirements for Bibliographic Records (FRBR): http://www.ifla.org/functional-requirements-for-bibliographic-records (Last visit: July 1st, 2012)

difficult to manage, because interaction is required with simultaneously ongoing change processes that inherit their own complexity[191]. [Bianchini & Guerrini 2009, p. 107] miss a coordinating body to guide the changes and the relationships between current models and national cataloging codes, and are convinced "[t]he Bibliographic universe can be managed only through unceasing interaction between theory and practice", and believe "there must be a fundamental break with past practice, in order to make room for completely new models and tools". They consider it particularly urgent "to reach agreement on a definition of the correct relationships between FRBR, ISBD, and national, multinational, and international codes – chiefly RDA [Resource Description and Access]" (pp. 106–107). [192]

The FRBR approach started with identification of relevant entities, their attributes and the types of relationships between the entities to produce a conceptual model that supports the mapping of attributes and relationships to various user tasks mainly from a user's perspective (including not only library clients and staff, publishers, distributors, retailers, providers and users of information services outside traditional library settings), taking into account a wide range of applications (the context of purchasing or acquisitions, cataloguing, inventory management, circulation and interlibrary loan, and preservation, for reference, information retrieval), and cover a comprehensive range of materials, media and formats (pertaining textual, cartographic, audio-visual, graphic, three-dimensional, paper, film, magnetic tapes, optical media, accoustic, electric, digital and optical recording modes), without assumptions about structure or content of the bibliographic record, or an intension to design bibliographic databases. FRBR associated data with bibliographic entities are only considered to an extent where they function as headings or index entries for the records. As a consequence, the FRBR model is considered comprehensive in scope but not exhaustive in terms of the entities, attributes and relationships, and is therefore not considered a fully developed data model. Its focus is on „as far as possible, a ‚generalized' view" of the bibliographic universe. The study group recognised the need to extend the bibliographic model towards authority

[191] The International Federation of Library Associations and Institutions (IFLA) initiated a set of cataloging principles (1961) and established international standards for the form and content of bibliographic descriptions (1971) which have served as the bibliographic foundation for a variety of new and revised national and international cataloging codes. This period was strongly influenced by huge technological changes, by economic pressure to reduce cataloging costs, and by an enormous growth of published output, and by a growing need to support the user, that formed the backdrop of the 1990 Stockholm Seminar [FRBR 1997].

[192] The changing environment brought together the participants of the Stockholm Seminar to develop the terms of reference for a commonly shared understanding of the bibliographic record that addresses users' needs and covers the broad range of requirements associated with various types of material and contexts by stating its purpose and scope as follows [FRBR 1997, p. 2]:*"The purpose of this study is to delineate in clearly defined terms the functions performed by the bibliographic record with respect to various media, various applications, and various user needs. The study is to cover the full range of functions for the bibliographic record in its widest sense- i.e., a record that encompasses not only descriptive elements, but access points (name, title, subject, etc.), other "organizing" elements (classification, etc.), and annotations."*

data and further analysis is needed of the entities in the focus for subject authorities, thesauri, and classification schemes, and relationships between those [FRBR 1997][193].

With FRBR, the key objects of interest to users of bibliographic data have been identified and divided into groups. The FRBR model entities and descriptions as subsequently presented are extracted from the report [FRBR 1997]. The entities in the first group (see Figure 49) describe different aspects of user interest in the products of intellectual or artistic endeavour.

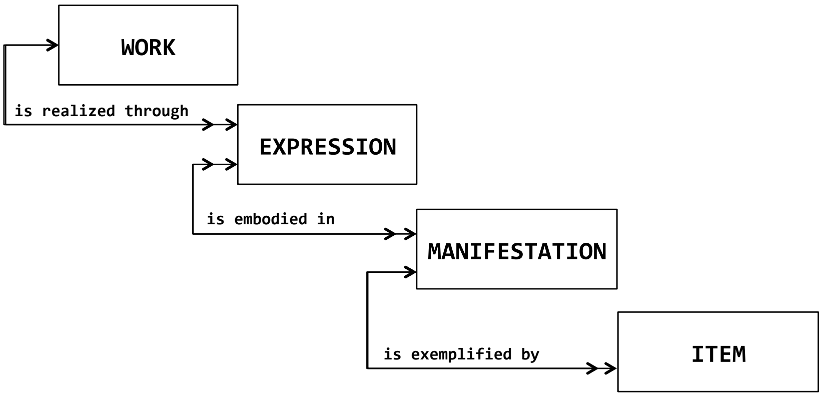


Figure 49: Group 1 Entities and Primary Relationships [FRBR 1997, fig. 3.1]

The entities defined as *Work* (a distinct intellectual or artistic creation) and *Expression* (the intellectual or artistic realization of a *work*) reflect the intellectual or artistic content. The entities defined as *Manifestation* (the physical embodiment of an *Expression* of a *Work*) and *item* (a single exemplar of a *Manifesta tion*) reflect the physical form. Relationships depicted in Figure 49 indicate that a *Work* may be realised through one or more than one *expression* (hence the double arrow on the line that links *Work* to *Expression*). An *Expression*, on the other hand, is the realization of one and only one *work* (hence the single arrow on the reverse direction of that line linking *Expression* to *Work*). An *Expression* may be embodied in one or more than one *Manifestation*; likewise a *Manifestation* may embody one or more than one *Expression*. A *Manifestation*, in turn, may be exemplified by one or more than one *Item*; but an *Item* may exemplify one and only one *Manifestation*.

The entities in the second group (outlined in bold with Figure 50) represent the responsibles for the intellectual or artistic content, the physical production and dissemination, or the

[193] "The basic elements of the model developed for the study--the entities, attributes, and relationships--were derived from a logical analysis of the data that are typically reflected in bibliographic records. The principal sources used in the analysis included the *International Standard Bibliographic Descriptions* (ISBDs), the *Guidelines for Authority and Reference Entries* (GARE), the *Guidelines for Subject Authority and Reference Entries* (GSARE), and the *UNIMARC Manual*. Additional data were culled from other sources such as the *AITF Categories for the Description of Works of Art*, from input provided by experts who were consulted as drafts of the report were being prepared, from an extensive review of published user studies, and from comments received as part of the world-wide review of the draft report." [FRBR 1997]

custodianship of the entities in the first group. The entities in the second group include *person* (an individual) and *corporate body* (an organisation or group of individuals and/or organisations). The diagram Figure 50 depicts the type of *responsibility* relationships that exist between entities in the second group and the entities in the first group. The diagram indicates that a *work* may be created by one or more than one *person* and/or one or more than one *corporate body*. Conversely, a *person* or a *corporate body* may create one or more than one *work*. An *expression* may be realised by one or more than one *person* and/or *corporate body*, and a *person* or *corporate body* may realise one or more than one *expression*. A *manifestation* may be produced by one or more than one *person* or *corporate body*; a *person* or *corporate body* may produce one or more than one *manifestation*. An *item* may be owned by one or more than one *person* and/or *corporate body*; a *person* or *corporate body* may own one or more than one *item*.

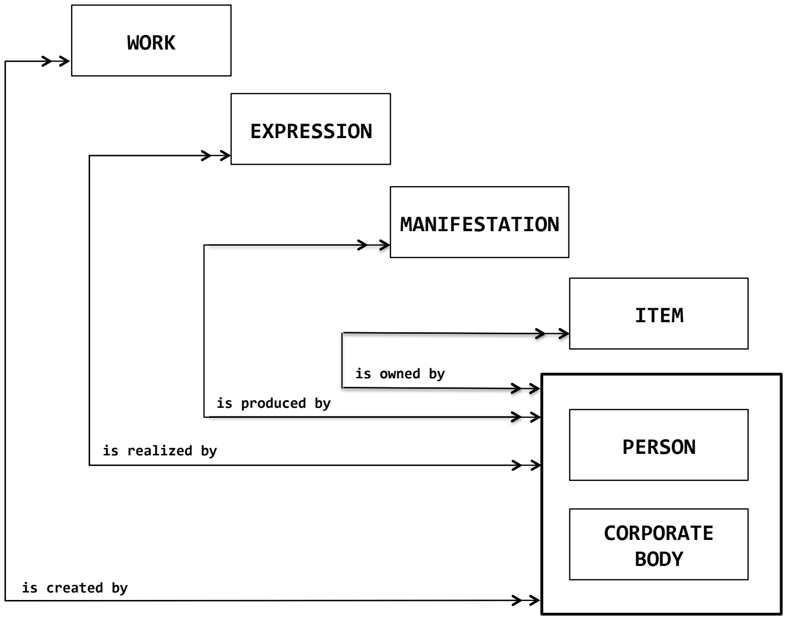


Figure 50: Group 2 Entities and "Responsibility" Relationships [FRBR 1997, fig. 3.2]

The entities in the third group (outlined in bold in Figure 51) represent an additional set of entities that serve as the subjects of *works*. The group includes *concept* (an abstract notion or idea), *object* (a material thing), *event* (an action or occurrence), and *place* (a location).

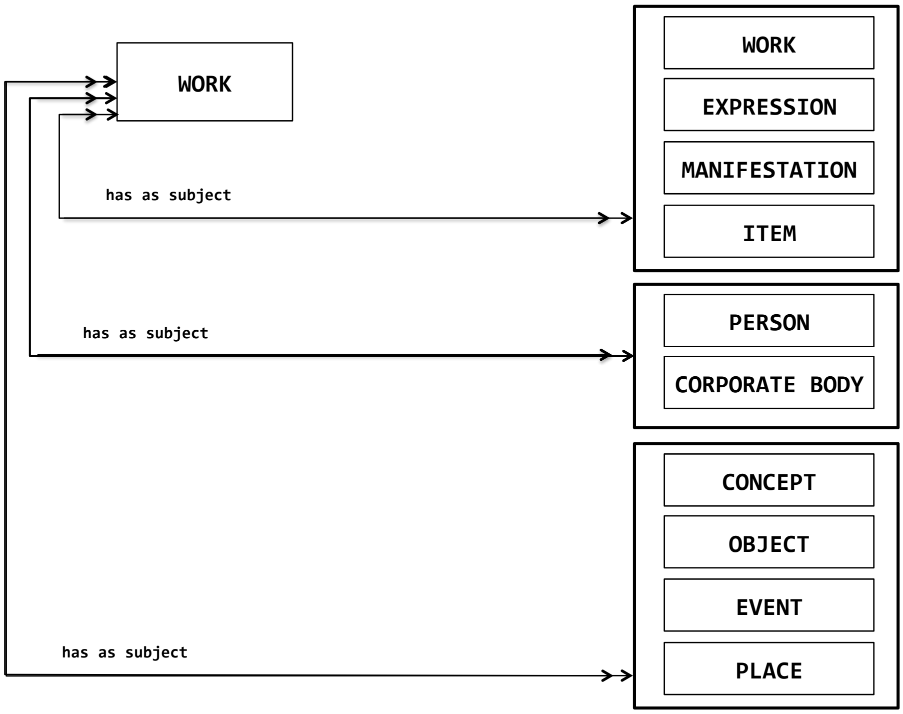


Figure 51: Group 3 Entities and "Subject" Relationships [FRBR fig. 3.3]

Figure 51 depicts the ‘subject’ relationships between entities in the third group and the *work* entity in the first group Figure 49. The diagram indicates that a *work* may have as its subject one or more than one *concept, object, event,* and/or *place*. Conversely, a *concept, object, event,* and/or *place* may be the subject of one or more than one *work*. The diagram also depicts the *subject* relationships between *work* and the entities in the first and second groups. The diagram indicates that a work may have as its subject one or more than one *work*, *expression*, *manifestation*, *item*, *person*, and/or *corporate body.*

[Bianchini & Guerrini 2009] present FRBR as the current accepted theoretical model for cataloguing, developed at a very high level of logic and founded on well-defined ideas about the objects that constitute the bibliographic universe (works, documents, authors, publishers, etc.), that places those objects into groups with special attributes and relationships. FRBR is introduced as a conceptual model of entities and relationships – never dealing with data descriptions and presentation, or how data must be communicated – and focusses on the function of data and on entities; it “does not cover the extended range of attributes and relationships that are normally reflected in authority records” [ISBD 2007], but has two objectives: “to provide a clearly defined, structured framework for relating the data that are recorded in bibliographic records to the needs of the users of those records” and “to recommend a basic level of functionality for records created by national bibliographic agencies"[194]. [Bianchini & Guerrini 2009] consider the 'FRBR catalog' a non-adequate term and the model not useful for a set of cataloging rules, where essential descriptive attributes are considered absent. The FRBR model concepts and relationships have been transformed to RDF (Davis & Newsman 2005)[195] allowing for relationships with external vocabularies – but no owned entity attributes. A recent RDF version has been published by the FRBR Review Group in the open metadata registry[196], covering attributes as well as entities and relationships – but with no reference to external vocabularies, and not employing the group classes due to reasons of non-shared characteristics of containing entities, but only in support of the ERM model simplification towards usage inline with the RDA namespace [Dunsire et al. 2011, p. 33].

International Standard Bibliographic Description – "The ISBD's main goal is, and has been since the very beginning, to offer consistency when sharing bibliographic information." [ISBD 2007][197] The ISBD Review Group agreed to avoid using FRBR Terminology in the ISBD, but nevertheless introduced some changes in terminology, among them the use of the term "resource" rather than "item" or "publication" [ISBD 2007, p. 7]. They finally believed the development of a table to detail the relationship of each of the elements specified would satisfy the need to make clear that the ISBDs and FRBR enjoy a harmonious relationship. [Le Bœuf 2003] nicely demonstrated that a disambiguation in the different semantic meanings is not that easy: "— when we say "book", what we have in mind may be a distinct, merely physical object that consists of paper and a binding (and can occasionally serve to wedge a table leg); FRBR calls it: "Item"; — when we say "book", we also may mean "publication", as when we go to our bookseller's and ask for a publication identified by a given ISBN: the particular copy does not matter to us, provided it belongs to the general class of copies we require and pages are not missing; FRBR calls it: "Manifestation"; — when we say "book", as in "Who wrote that book?", we may have a specific text in mind, the intellectual *content* of a publication; FRBR calls it: "Expression"; — when we say "book",

[194] citing Elaine Svenonius, *The Intellectual Foundation of Information Organization* in [Bianchini & Guerrini 2009]

[195] The FRBR model in RDF format: http://vocab.org/frbr/core.html (Davis & Nesman 2005) (Last visit: June 7th, 2012)

[196] FRBR in the Open Metadata Registry: http://vocab.org/frbr/core.html (Last visit: May 2nd, 2012)

[197] "The **International Standard Bibliographic Description (ISBD)** dates back to 1969, when the IFLA Committee on Cataloguing (subsequently renamed the Standing Committee of the IFLA Section on Cataloguing, now known as the Standing Committee of the IFLA Cataloguing Section) sponsored an International Meeting of Cataloguing Experts. This meeting produced a resolution that proposed creation of standards to regularize the form and content of bibliographic descriptions. As a result, the Committee on Cataloguing put into motion work that ultimately would provide the means for a considerable increase in the sharing and exchange of bibliographic data. This work resulted in the concept of the International Standard Bibliographic Description (ISBD), which has now endured for more than 30 years. The individual formats to which the ISBD concept has been applied are now used by bibliographic agencies, national and multinational cataloguing codes, and cataloguers in a wide variety of libraries throughout the world, because of their potential for promoting record sharing." http://www.ifla.org/en/about-the-isbd-review-group

we eventually may mean an even higher level of abstraction, the conceptual content that underlies all of its linguistic versions, either the original or a translation; the "thing" that an author may recognise as his/her own, even in, say, a Japanese translation and even though he/she cannot speak Japanese and cannot therefore be held as responsible for the Japanese text; FRBR calls it: "Work".

5.2.3.2 Literature

In FERON, Grey Literature[198] is perceived as an adequate representation of the *informal* concept as introduced with the UNISIST (1971) model, to imply non-peer-reviewed and thus *informally* published work. Therefore the class *grey:Literature* is subsumed under FERON's *w3c-tag:InformationResource* class. Its intrinsic and mutual properties to describe the underlying records themselves overlap with the *cerif:Publication* class although grey literature has not undergone a *formal* review process. This work did not want to introduce a new abstract class to subsume *cerif:Publication* and *grey:Literature* although a thought has been given the often used concept of *Output*. Output was considered biased towards a particular view, because e.g. a publication may as well function as input for another publication. A FERON *grey:Literature* class is identical in features with the *cerif:Publication* class, i.e. a *dcterms:Resource* and inherits the intrinsic properties *federated-identifier*, *dcterms:identifier* and *dcterms:relation*. In addition, it inherits intrinsic properties from the class *w3c-tag:InformationResource*, namely *cerif:keywords*, *cerif:title* and *dcterms:language*. In addition, the *grey:Literature* class features datatype-properties *cerif:abstract*, *cerif:title*, *cerif:publication-date*, *bibo:pageStart*, *bibo:pageEnd*; *bibo:volume*; *bibo:number*, *bibo:isbn*, *bibo:issn*.

The *cerif:Patent* class is subsumed under *grey:Literature*; patents are basically field-agnostic. That is, their intrinsic properties do not reflect field specific features and in fact, patents are published independent of the domain to which they may be assigned, i.e. classified. In FERON, fields or areas or domains are usually reflected by functional references to authorised classification systems, such as the International (IPC) or the European Patent Classification (EPC)[199]. The IPC and the EPC as such are KOSs in FERON and functional

[198] Grey Literature Network Service (GreyNet): http://www.greynet.org/ (Last visit, January 4th, 2012). Grey Literature is a field in library and information science that deals with the production, distribution, and access to multiple document types produced in all levels of government, academics, business, and organization in electronic and print formats not controlled by commercial publishing i.e. where publishing is not the primary activity of the producing body. Grey Literature Typology: http://code.google.com/p/grey-literature-typology/ (Last visit: January 4th, 2012).

[199] International Patent Classification: http://www.wipo.int/classifications/ipc/ (Last visit: May 2nd, 2012)

linkage is managed through the relationship construct that is elaborated in section 7.10 Time-aware Relationships.

The *grey:Literature* class and the *cerif:Publication* class are not considered field specific; i.e. there are no LT sub-classes below them. A patent in CERIF is described by properties such as *cerif:registration-date*, *cerif:approval-date*, *cerif:patent-number*, a *country* reference and multiple relationships, e.g. with publication, person, project, organisation, etc. In VIVO, a *bibo:Patent* is a sub-class of *bibo:LegalDocument*, which is a sub-class of *bibo:Document*, which is a sub-class of *vivo:InformationResource*, maintaining intrinsic properties such as *date-issued* and *publisher* upon the range of *foaf:Organization*.

5.2.3.3 Data

Whereas the concept of a publication is fairly good understood and defined through existing standards evolving through history, being more or less discipline-agnostic, the descriptions of research data are highly related to disciplines, i.e. fields. Furthermore, teams "who may be widely distributed, have to agree upon what data will be collected, by what techniques and instruments, and who has the rights and responsibilities to analyze, publish, and release those data" [Borgman 2011, p 13]. [Uhlir & Schröder 2007, p. 36] define "public research data" as "data that are generated through research within government organizations, or by academic or other not-for-profit entities, as well as public data used for research purposes, but not necessarily produced primarily for research (e.g. geographic or meteorological data, or socioeconomic statistics produced by or for government organizations)."

DataCite[200] has been setup as a registry for research data. A registration requires a minimum set of core metadata with dataset registration; the key concept to the service is a domain agnostic "*persistent* approach to access, identification, sharing, and re-use of datasets [...] to serve scholars in a range of dicsciplines, from the sciences, social sciences and humanities" – namely, a persistent identifier. For accurate and consistent identification with citations and retrieval a minimum set of mandatory metadata has to be provided at the time of identifier registration, and data centers and submitters may also choose optional properties for increased clarity with identification. The version 2.2 of the DataCite Schema [DataCite 2011, p. 4] gives the following mandatory attributes:

[200] DataCite Metadata Schema: http://schema.datacite.org/meta/kernel-2.2/doc/DataCite-MetadataKernel_v2.2.pdf (Last visit: May 2nd, 2012)

Table 5: Mandatory DataCite Properties

ID	Property
1	Identifier (with type attribute)
2	Creator (with name identifier)
3	Title (with optional type attributes)
4	Publisher
5	PublicationYear

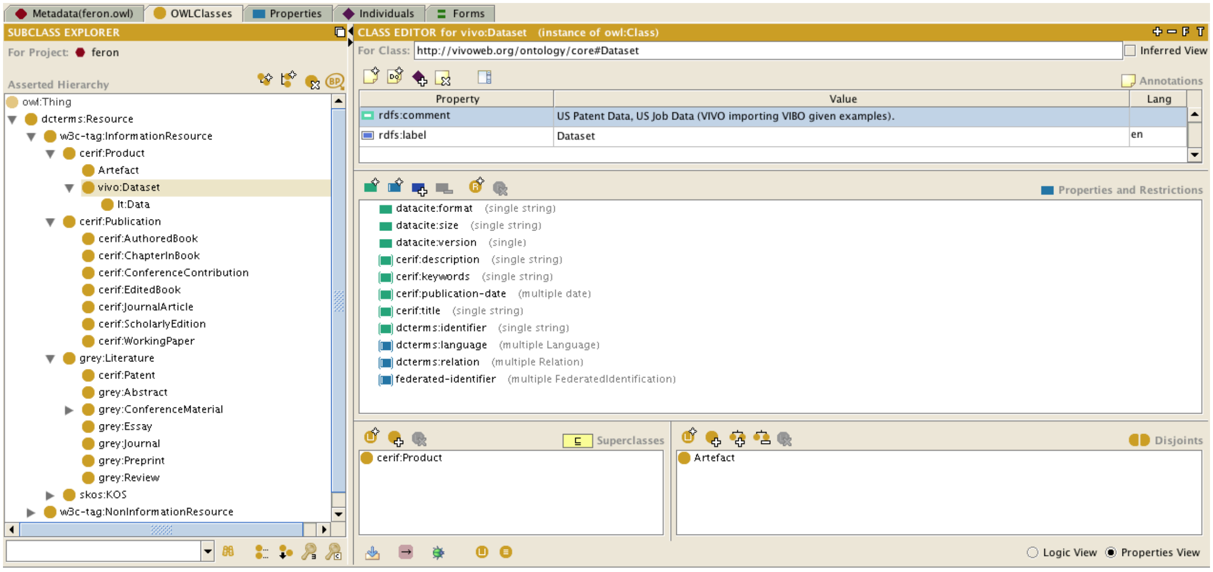
Table 6: Optional DataCite Properties

ID	Property
6	Subject (with schema attribute)
7	Contributor (with type and name identifier attributes)
8	Date (with type attribute)
9	Language
10	Resource Type (with description attribute)
11	AlternateIdentifier (with type attribute)
12	RelatedIdentifier (with type and relation type attributes)
13	Size
14	Format
15	Version
16	Rights
17	Description (with type attribute)

The attributes in Table 5 and Table 6 reveal the known Dublin Core elements. However, these are not ontological-driven metadata descriptions, where a creator or publisher would be rather the role in a relationship between e.g. a person and a dataset, and not an attribute of the dataset itself (this issue has been elaborated within section 5.2.1 Resource).

The Science ontology downloaded from the public Protégé library includes *Product* as a concrete class with *Software-Component* as a sub-class. A *Product* is then described through slots such as *name: required String*, *Developed-By: required multiple Organization*, *Under-*

Project: *Project*. In CERIF, the *cfResultProduct* entity subsumes datasets, where a further concept elaboration is on the agenda. The latest version CERIF 1.3 is limited with product and dataset-related features. It currently employs the *cerif:Title*, *cerif:Description*, *cerif:Keywords* and *identifiers* for the *cfResultProduct* entity. In VIVO, the dataset class *vivo:Dataset*, which is employed with FERON, is a subclass of *vivo:InformationResource*, from where it inherits all properties: *vivo:dateTimeValue*; *vivo:domesticGeographicFocus*, *vivo:freetextKeyword*, *vivo:geographicFocus*, *vivo:hasSubjectArea*, *vivo:internationalGeographicFocus*, *vivo:webpage*, *bibo:editor*, *bibo:translator*, *vivo:features*, *vivo:informationProductOf*, *vivo:InformationResourceInAuthorship*, *vivo:informationResourceSupportedBy*, as indicated in Figure 52.



```
owl:Thing
▼ dcterms:Resource
    ▼ w3c-tag:InformationResource
        ▼ cerif:Product
              Artefact
            ▼ vivo:Dataset
                  lt:Data
        ▼ cerif:Publication
              cerif:AuthoredBook
              cerif:ChapterInBook
              cerif:ConferenceContribution
              cerif:EditedBook
              cerif:JournalArticle
              cerif:ScholarlyEdition
              cerif:WorkingPaper
        ▼ grey:Literature
              cerif:Patent
              grey:Abstract
            ▶ grey:ConferenceMaterial
              grey:Essay
              grey:Journal
              grey:Preprint
              grey:Review
        ▶ skos:KOS
    ▶ w3c-tag:NonInformationResource
```

For Class: http://vivoweb.org/ontology/core#Dataset

Property	Value	Lang
rdfs:comment	US Patent Data, US Job Data (VIVO importing VIBO given examples).	
rdfs:label	Dataset	en

datacite:format (single string)
datacite:size (single string)
datacite:version (single)
cerif:description (single string)
cerif:keywords (single string)
cerif:publication-date (multiple date)
cerif:title (single string)
dcterms:identifier (single string)
dcterms:language (multiple Language)
dcterms:relation (multiple Relation)
federated-identifier (multiple FederatedIdentification)

Superclasses: cerif:Product

Disjoints: Artefact

○ Logic View ● Properties View

Figure 52: FERON's imported vivo:Dataset class subsuming the lt:Data class in Protégé

With FERON, most of the given properties are considered mutual properties – and in FERON these are not explicitly modeled, but, employ the relationship class construct (7.10 Time-aware Relationships) to apply the functions defined under the *skos:KOS* classes (e.g. *dcterms:Creator*, *bibo:Translator*, *bibo:Editor*, etc.). From the DataCite Kernel (2.2) as presented with Table 5 and Table 6, FERON employs the properties *datacite:format*, *datacite:size*, *datacite:version* with the *vivo:Dataset* class, but with this work, will not further elaborate on the more generic imported concept *cerif:Product*.

Data i.e. *vivo:Dataset* is considered the most field-specific concept or class where LT Data extensions are subsumed as sub-classes. More details will be elaborated within section 6.3 LT Resource.

5.2.4 Knowledge Organisation Systems (KOS)

Knowledge organisation systems (KOS) are systems such as thesauri, classification schemes, subject headings, taxonomies, topic maps, folksonomies and similar types of controlled vocabularies in support of organising knowledge. A clear distinction or similarity between controlled vocabularies or *terminologies* and ontologies is still discussed. A Special Issue of Applied Ontology – *Ontologies and Terminologies: Continuum or Dichotomy* – defines terminology as "a set of terms, which represents the system of concepts for an area and for an application. These terms remain linguistic entities and linguistic information may be associated with them. Term organisation is usually not constrained by any formal logics or description, which may lead to problems like cyclicity and redundancy with a terminology. As for ontologies, they are built upon formal specification and constraints and describe also a system of concepts and associated properties for a specific area. They are intended to be used by computers and automatic applications" [201].

The history of knowledge organisation systems clearly refers to the Library and Information Science (LIS) field, where various systems have been developed, and some of them have been in use for more than a century[202] and still are. With the Web and enabling technologies in networked information systems the production of KOSs proliferated beyond the traditional library environment into research, and furthermore into markets and society. However, their wider use and re-use on the Web requires formalization and standardization inline with emerging technologies. Alongside recent developments, and especially with the Semantic Web and with Linked Open Data[203], the deployment and growth of vocabularies including so-called micro-formats exploded through community efforts. There are multiple operating standardization bodies[204], and with the Web increasingly *open* standards[205] are becoming popular.

[201] Special Issue of Applied Ontology: http://natalia.grabar.perso.sfr.fr/AO-CALL/ (planned publication in Summer 2012, Last visit: July 24th, 2011)

[202] The Dewey Decimal Classification was developed in 1876. E.g. In 1898 the Library of Congress Subject Headings (LCSH) „converted from an author- plus a classed-catalog to a dictionary catalog, which incorporated author, title, and subject entries into a single file. In: A brief history of the Library of Congress Subject Headings, and introduction to the centennial essays. http://catalogingandclassificationquarterly.com/ccq29nr1-2ed.htm (Last visit: January 8th, 2012) In May 2009, the Library of Congress announced the launch of Linked Data Subject Headings: http://blogs.talis.com/panlibus/archives/2009/05/library-of-congress-launch-linked-data-subject-headings.php (Last visit: January 8th, 2012)

[203] Linked Data: http://linkeddata.org/ Wikipedia defines Linked Data as „a term used to describe a recommended best practice for exposin, sharing, and connecting pieces of data, informaiton, and knowledge on the Semantic Web using URIs and RDF.“ (Last visit: Janauary 8th, 2012)

[204] The World Standards Services Network (WSSN) (http://www.wssn.net/WSSN/listings/ links_international.html) provides an overview of the internationally recognised standard bodies. Those relevant in the context of this work are the International Organization for Standardization (ISO) as a body for all fields except electrical and electronic engineering, whereas the International Electrotechnical Commission (IEC) is exactly recognised for that. Furthermore, there is IETF – the

All the analysed formats were openly available on the Web. With the next sections known syntactically and semantically declared KOSs to manage vocabularies will be investigated – namely: SKOS, the CERIF Semantic Layer, and in very brief SBVR.

5.2.4.1 Simple Knowledge Organisation System (SKOS)

SKOS, the Simple Knowledge Organization System [Miles & Bechhofer 2009][206] is a highly popular system and the W3C Recommendation defining "a common data model for sharing and linking knowledge organization systems via the Web [...] The fundamental element of the SKOS vocabulary is the *concept*". The Willpower[207] glossary as referred to in the SKOS Primer, defines concepts as „the units of thought – ideas, meanings, or (categories of) objects and events – which underly many knowledge organisation systems, concepts exist in the mind as abstract entities which are independent of the terms used to label them" (see also chapter 2.4). The Recommendation informs about SKOS's background and motivation, which is data sharing for "bridging several different fields of knowledge, technology and practice" in machine-readable form. The need resulted from the activities and accumulated experience in the library and information sciences, where the important point for SKOS is "that, in addition to their unique features, each of these families shares much in common, and can often be used in similar ways [...], there is currently no widely deployed standard for

Internet Engineering Task Force towards internet architecture and operation, the International Federation of Library Associations and Institutions (IFLA) for bibliographic control and other aspects of library matters, UNESCO – the United Nations Educational, Scientific and Cultural Organization with an interest in scientific and technological information and documentation, libraries and archives. With the foundation of the Web, in 1994, the World Wide Web (W3C) Consortium was founded at the Massachusets Institute of Technology, Laboratory of Computer Science (MIT/LCS), with support by the European Commission (EC) and the Advanced Research Projects Agency (DARPA), to develop Web standards and to lead the Web to its full potential. Furthermore, there is OASIS – for Advancing Open Standards for the Information Society and the Open Management Group (OMG) to develop enterprise integration standards, addressing middleware, modeling and vertical domain frameworks. Officially the release of a "standard" refers to ISO, W3C publishes "recommendations", and OMG provides "specifications".

[205] Free Software Foundation Europe (FSFE) – Open Standards – Definition: http://www.fsfe.org/projects/os/def.en.html "An Open Standard refers to a format or protocol that is (i) subject to full public assessment and use without constraints in a manner equally available to all parties; (ii) without any components or extensions that have dependencies on formats or protocols that do not meet the definition of an Open Standard themselves; (iii) free from legal or technical clauses that limit its utilisation by any party or in any business model; (iv) managed and further developed independently of any single vendor in a process open to the equal participation of competitors and third parties; (v) available in multiple complete implementations by competing vendors, or as a complete implementation equally available to all parties." (Last visit: January 8th, 2012)

[206] The W3C Recommendation for SKOS has been developed by the Semantic Web Deployment Working Group, which is part of the W3C Semantic Web Activity. "The elements of the SKOS data model are classes and properties, and the structure and integrity of the data model is defined by the logical characteristics of and interdependencies between those classes and properties. [...] However, SKOS is not a formal knowledge representation language." http://www.w3.org/TR/2009/REC-skos-reference-20090818/ [Miles & Bechhofer 2009]. [Isaac & Summers 2008] is referred to in [Miles & Bechhofer 2009] as an informative guide.

[207] Willpower Glossary of terms relating to thesauri and other forms of structured vocabulary from information retrieval: http://www.willpowerinfo.co.uk/glossary.htm (Last visit: January 8th, 2012)

representing these knowledge organisation systems as data and exchanging them between computer systems.”

SKOS is built on RDF where concepts are identified by URIs (4.3.2 The Semantic Web) labeled with lexical strings in one or more natural languages, and these can refer to many schemes and thus be assigned to one or more notation via lexical codes, or documented with notes of various types. The data model provides a basic set of properties. In addition, SKOS concepts can be interlinked via “semantic relation properties”[208], where again the data model provides support for hierarchical and associative links, and, can be extended by third parties towards more specific needs.

```
<rdf:RDF
     xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
     xmlns:skos="http://www.w3.org/2004/02/skos/core#">
     <skos:Concept rdf:about="http:/example.com/concept/0001">
          <skos:inScheme rdf:resource="http:/example.com/thesaurus"/>
     </skos:Concept>
</rdf:RDF>
```

Notation 12: SKOS Concept Definition in RDF [Miles & Bechhofer 2009]

A simple RDF statement to describe a *skos:Concept* and its scheme *skos:inScheme* assignment, is shown in Notation 12. SKOS allows for three kinds of labels “Preferred Lexical Labels”, “Alternative Lexical Labels” and “Hidden Lexical Labels” – indicated in Notation 13, but which will not be further explored here. In KOSs semantic relationships play a crucial role for defining concepts, because meaning is not just defined by concept names or labels, but also in relationships and even in relationships maintained with other vocabulary concepts. SKOS mirrors fundamental categories of relationships as used in thesauri such as: *skos:broader*, *skos:narrower*, and *skos:related*, *skos:closeMatch*, *skos:exactMatch*, *skos:hasTopConcept*. In addition, it provides documentary properties: *skos:note*, *skos:scopeNote*, *skos:definition*, *skos:example*, *skos:historyNote*, *skos:editorialNote* and a *skos:changeNote*. FERON features the *skos:definition* and the *skos:example* datatype properties with the *skos:KOS* class.

[208] “The W3C’s specification to generic schema for thesauri was initially produced by the DESIRE project [Cross et al. 2000] and further developed in the Limber project Matthews et al 01. This work formed the basis of SKOS Core. The SKOS Primer and the SKOS Reference replace the former Core and Core Vocabulary specifications. The SKOS Core was developed as draft of an RDF Schema for thesauri compatible with relevant ISO standards. Further work extended it to multilingual thesauri, and mappings between thesauri, and developed some pilot tools; see SWAD Reports for the deliverables on Thesauri.” (*ids.snu.ac.kr/w/images/f/f1/SC18.pdf*)

```
ex:pineapples rdf:type skos:Concept;
skos:prefLabel "pineapples"@en;
skos:prefLabel "ananas"@fr;
skos:definition "The fruit of plants of the family
     Bromeliaceae"@en;
skos:definition "Le fruit d'une plante herbacée de la famille des
broméliacées"@fr.
```

Notation 13: SKOS example record in N3 notation [Miles & Bechhofer 2009]

Being based on RDF, SKOS concepts can be easily created and used as stand-alone entities. However, for improved quality and towards interoperability, the re-use of defined vocabularies is highly recommended "[o]n the Semantic Web the true potential of data is unleashed when it is interlinked" and „concepts usually come in carefully compiled vocabularies, such as thesauri or classification schemes. SKOS offers the means of representing such KOSs using the skos:ConceptScheme class" [Miles & Bechofer 2009]. If a concept scheme has been created, it can be linked through the *skos:inScheme* property as indicated in Notation 14. To provide efficient access to the entry points of broader/narrower concept hierarchies, SKOS offers a *skos:hasTopConcept* property.

```
ex:mammals rdf:type skos:Concept;
   skos:inScheme ex:animalThesaurus.


ex:cows rdf:type skos:Concept;
   skos:broader ex:mammals;
   skos:inScheme ex:animalThesaurus.

ex:fish rdf:type skos:Concept;
   skos:inScheme ex:animalThesaurus.

ex:animalThesaurus rdf:type skos:ConceptScheme;
   skos:hasTopConcept ex:mammals;
   skos:hasTopConcept ex:fish.
```

Notation 14: SKOS examples in N3 notation [Miles & Bechhofer 2009]

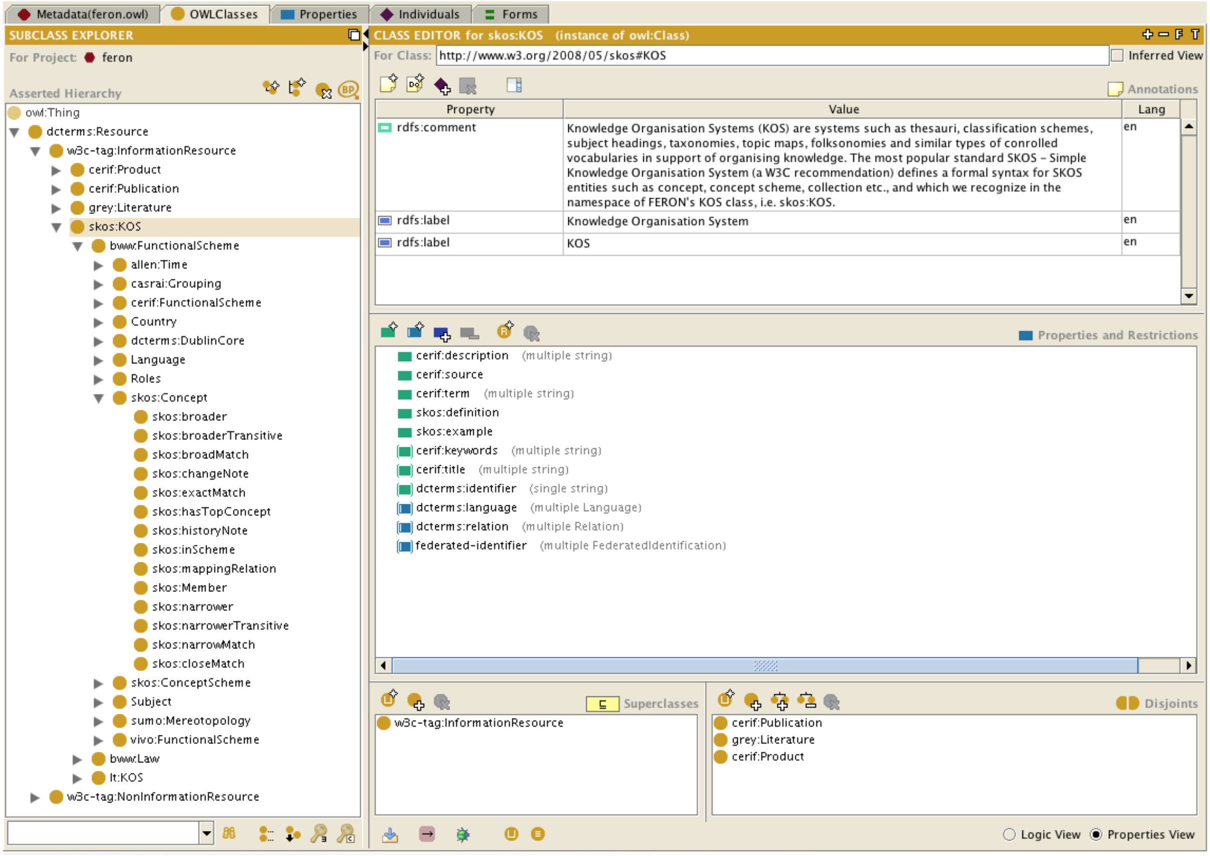
SKOS suggests to map concepts by proposing properties such as *skos:exactMatch*, *skos:closeMatch* or the introduced semantic relations *skos:broadMatch*, *skos:narrowMatch* and *skos:relatedMatch* and resolves extensions or imports from other schemes with OWL

owl:imports and by means of the Semantic Web identifiers – namely URIs distinguished through namespaces. SKOS refers to Dublin Core for *subject* i.e. field references and employs e.g. *creator* as a documenting feature. For advanced features "[w]hen KOSs are not Simple Anymore" [Miles & Bechhofer 2009] proposes:

- Grouping of concepts based on specific criteria
- Advanced documentation by means of complex resources
- Establishing relationships between labels of concepts
- Creation of complex concepts from simple ones (coordination)
- Assessing transitive hierarchical relationships
- Representing notations for concepts

This work will not further investigate the advanced SKOS features – because it is not aimed at a formally exhaustive model of e.g. an entire thesaurus or ontology representation – but more interested in understanding the conceptual level. An overview is provided in Table 8 based on [Miles & Bechhofer 2009][209], where a comparison of the simple SKOS entities with the CERIF Semantic Layer has been investigated at CIM level [is table 2 in Jörg et al. 2011]. FERON models the *skos:KOS* class under the *w3c-tag:InformationResource* class as a container for all kinds of knowledge organisation systems (KOSs), and employs the *skos* namespace to indicate the SKOS *Concept* approach with subsumed (imported) KOSs. Protégé provides the *Class* concept as a formal-declared semantic construct with each modeled *thing (i.e. concept in terms of ontology)* – supported in RDF and OWL (section 3.2.6 Formal Ontology).

[209] SKOS Namespace Document [Miles & Bechhofer 2009] http://www.w3.org/2009/08/skos-reference/skos.html



For Class: http://www.w3.org/2008/05/skos#KOS

Property	Value	Lang
rdfs:comment	Knowledge Organisation Systems (KOS) are systems such as thesauri, classification schemes, subject headings, taxonomies, topic maps, folksonomies and similar types of conrolled vocabularies in support of organising knowledge. The most popular standard SKOS – Simple Knowledge Organisation System (a W3C recommendation) defines a formal syntax for SKOS entities such as concept, concept scheme, collection etc., and which we recognize in the namespace of FERON's KOS class, i.e. skos:KOS.	en
rdfs:label	Knowledge Organisation System	en
rdfs:label	KOS	en

Figure 53: FERON's SKOS inspired KOS classes

All FERON relationships, are recorded as instances upon functions refering to *skos:KOS* sub-classes (each corresponding to a particular functional scheme in the spirit of Bunge to ensure lawfulness). Where SKOS relationships between concepts such as with *skos:broader* or *skos:narrower* are modeled (following RDF and OWL) as object-type properties, with FERON these are featured functional classes, i.e. sub-classes of the *skos:KOS* class for reference from within recorded relationship instances. FERON’s *skos:KOS* subsumes *bww:FunctionalScheme*, which subsumes the *skos:ConceptScheme* with sub-classes such as *skos:Collection*, *skos:OrderedCollection* and *skos:MemberList (see* Figure 53*)*.

VIVO employs the *skos:Concept* class without property specifications – giving a short definition of Concept – as “[a]n idea or notion; a unit of thought.”

5.2.4.2 CERIF Semantic Layer

The CERIF Semantic Layer is a conceptual construct embedded in – and syntactically inline with – the CERIF ER-Model as introduced in section 5.1.1. The central entity in the Semantic Layer is a *cfClass*, to which all related entities are assigned to by an internal *cfClassId* identifier attribute, see Figure 54, where *cfClassId* is defined as primary key (PK) and thus inherited in entities such as classification description *cfClassDescr*, classification example *cfClassEx*, classification definition *cfClassDef*, classification term *cfClassTerm*, the recursive classification-classification link entity *cfClass_Class*, and the recursive classification-scheme *cfClassScheme_ClassScheme* link entity.

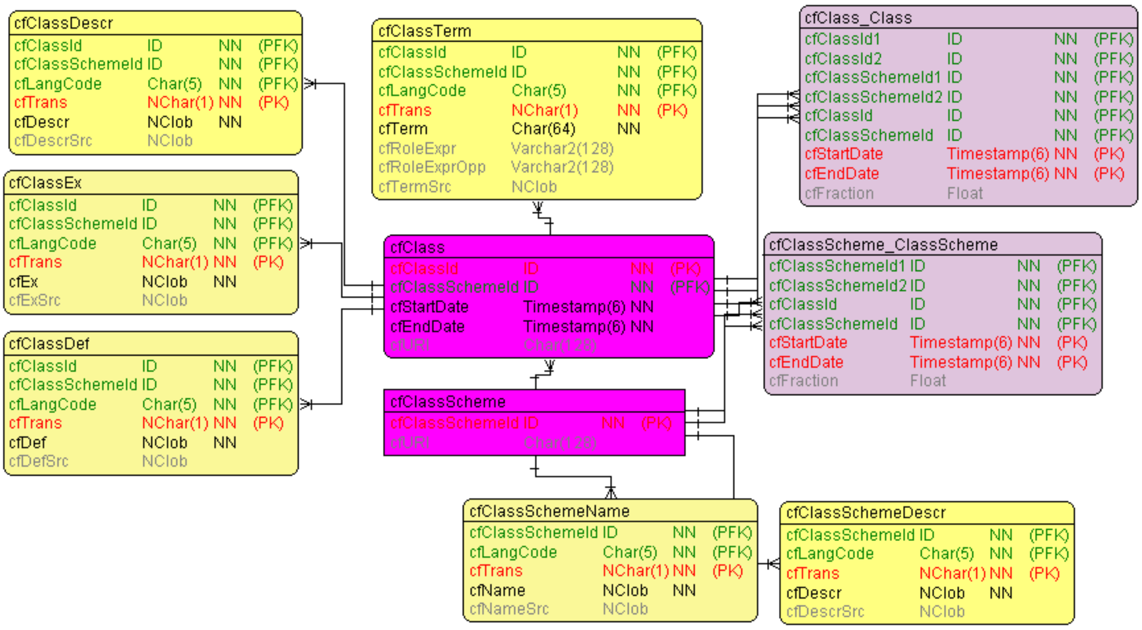

cfClassDescr
cfClassId ID NN (PFK)
cfClassSchemeId ID NN (PFK)
cfLangCode Char(5) NN (PFK)
cfTrans NChar(1) NN (PK)
cfDescr NClob NN
cfDescrSrc NClob
cfClassEx
cfClassId ID NN (PFK)
cfClassSchemeId ID NN (PFK)
cfLangCode Char(5) NN (PFK)
cfTrans NChar(1) NN (PK)
cfEx NClob NN
cfExSrc NClob
cfClassDef
cfClassId ID NN (PFK)
cfClassSchemeId ID NN (PFK)
cfLangCode Char(5) NN (PFK)
cfTrans NChar(1) NN (PK)
cfDef NClob NN
cfDefSrc NClob
cfClassTerm
cfClassId ID NN (PFK)
cfClassSchemeId ID NN (PFK)
cfLangCode Char(5) NN (PFK)
cfTrans NChar(1) NN (PK)
cfTerm Char(64) NN
cfRoleExpr Varchar2(128)
cfRoleExprOpp Varchar2(128)
cfTermSrc NClob
cfClass
cfClassId ID NN (PK)
cfClassSchemeId ID NN (PFK)
cfStartDate Timestamp(6) NN
cfEndDate Timestamp(6) NN
cfURI Char(128)
cfClassScheme
cfClassSchemeId ID NN (PK)
cfURI Char(128)
cfClass_Class
cfClassId1 ID NN (PFK)
cfClassId2 ID NN (PFK)
cfClassSchemeId1 ID NN (PFK)
cfClassSchemeId2 ID NN (PFK)
cfClassId ID NN (PFK)
cfClassSchemeId ID NN (PFK)
cfStartDate Timestamp(6) NN (PK)
cfEndDate Timestamp(6) NN (PK)
cfFraction Float
cfClassScheme_ClassScheme
cfClassSchemeId1 ID NN (PFK)
cfClassSchemeId2 ID NN (PFK)
cfClassId ID NN (PFK)
cfClassSchemeId ID NN (PFK)
cfStartDate Timestamp(6) NN (PK)
cfEndDate Timestamp(6) NN (PK)
cfFraction Float
cfClassSchemeName
cfClassSchemeId ID NN (PFK)
cfLangCode Char(5) NN (PFK)
cfTrans NChar(1) NN (PK)
cfName NClob NN
cfNameSrc NClob
cfClassSchemeDescr
cfClassSchemeId ID NN (PFK)
cfLangCode Char(5) NN (PFK)
cfTrans NChar(1) NN (PK)
cfDescr NClob NN
cfDescrSrc NClob

Figure 54: CERIF Semantic Layer construct - CERIF version 1.4 [Jörg et al. 2011]

Figure 54 shows the CERIF Semantic Layer entities at PIM, i.e. ERM level, where the heart is the classification entity *cfClass* tied to a scheme *cfClassScheme* via a foreign key (PFK) to preserve the system-internal unique identification. It additionally requires dates and allows for a URI *cfURI*. The *cfClassId* identifies a class or concept to which multiple terms, descriptions, definitions, or examples can refer in multiple languages: *cfClassTerm*, *cfClassDescr*, *cfClassDef*, *cfClassEx*. A recursive classification entity, the *cfClass_Class* link entity, allows for conceptual mappings or relationship kind variations within and across multiple classification systems or schemes, such as e.g. synonym or broader term. A class scheme is identified by its own *cfClassSchemeId*, has a name and allows for a description, *cfClassSchemeDescr*. Each related multilingual entity also employs a *cfSrc* attribute to inform about the source of the term, description, definition, example or name. A term describes the function behind a relationship (in CERIF called Link Entity) as indicated in Table 7 [Jörg et al. 2011, table 1], where e.g. in a person-publication relationship the role ‘authoring’ leads to a role expression *cfRoleExpr* ‘is author of’, and inversely *cfRoleExprOpp* 'is authored by', where the property in the person entity is 'author', and in the publication entity it is 'authored'.

Table 7: CERIF Term, role and expression examples [Jörg et al. 2011, table 1]

CERIF Link Entity	Term	Role	Relationship 1	Relationship 2
cfPers_ResPubl	**Author**	authoring	is author of	is authored by
cfPers_OrgUnit	**Commissioner**	commissioning	is commissioner at	is commissioned by
cfPers_OrgUnit	**Manager**	managing	is manager of	is managed by
cfProj_Pers	**Manager**	managing	is manager of	is managed by
cfProj_Fund	**Pending**	pending	is pending	is pending for
cfOrgUnit_OrgUnit	**Part**	(is-a)	is part of	has part
cfOrgUnit_OrgUnit	**Funder**	funding	is funded by	is funder of
cfOrgUnit_OrgUnit	**Member**	membership	is member of	has member
cfOrgUnit_OrgUnit	**Acquisition**	acquiring	was acquired by	has acquired

In CERIF, roles are always inaugurated through link entities, as indicated in column 1 of Table 7, and as presented in Figure 54 with the ERM extract of some CERIF entities. The CERIF link entity construct has very much influenced the FERON relationship construct. (euroCRIS) CERIF started to develop a Research domain vocabulary[210] and has recently formed strategic partnership with CASRAI and VIVO towards a standardized canonical vocabulary (ontology) for the Research domain.

5.2.4.3 Semantics of Business Vocabulary and Rules (SBVR)

The Open Management Group [OMG 2008] specified (v1.0) as "vocabulary and rules for documenting the semantics of business vocabularies, business facts, and business rules, as

[210] CERIF 1.3 Vocabuary (Last visit: April 8th, 2012): http://www.eurocris.org/Uploads/Web%20pages/CERIF-1.3/Semantics/CERIF1.3_Vocabulary.xls

well as an XMI schema for the interchange of business vocabularies and business rules among organizations and between software tools“ [OMG 2008]. The Business Semantics Methodology has been implemented by Collibra[211] and is essentially constituted of a set of complementary cycles aimed at creating a set of consolidated language neutral semantic patterns for application in a variety of semantic environments. The BSM allows for a strong community driven development of a shared conceptual model [see for an example in Jörg et al. 2011, p. 7].

SBVR aims at documenting vocabularies, business facts and rules, supporting linguistic analysis of the text behind vocabularies, conceptualized for business people rather than for automated processing. The SBVR 'Business Vocabulary' distinguishes between so-called semantic communities (similar to CIM) by their shared understanding of concepts, while speech communities use language expressions in vocabularies stored as a term dictionary (similar to PIM) inline with the concepts of the semantic community. In SBVR terms, Research is a semantic community (conceptually perceived, i.e. CIM); the CERIF entities, declared types and roles are speech communities (i.e. perceived as PIM), where each maintains the vocabularies to be structured inline with the CERIF model, and where a conceptual term (i.e. perceived as PIM) corresponds to the global dictionary (i.e. being semantically-declared and thus PIM).

5.2.4.4 Summary

Where SKOS provides the formal syntax to represent and interchange a controlled vocabulary, SBVR extends the understanding of concepts towards business facts and rules upon e.g. SKOS, and is therefore more conceptual than formal. SKOS aims to overcome legacy structures by defining a common data model for sharing and linking knowledge organization systems via the Web, based on RDF at PIM level. CERIF provides a formal and semantically declared contextual-neutral datamodel towards interoperability of research information systems. It incorporates the Semantic Layer as a conceptual construct in support of knowledge organisation. [Jörg et al. 2011, table 2] investigated the classes and properties under the SKOS namespace and map them to the CERIF Semantic layer entities (PIM); see Table 8.

[211] Collibra: http://www.collibra.com/ (Last visit: January 8th, 2012)

Table 8: Basic SKOS to CERIF Mapping [Jörg et al. 2011, table 2]

SKOS Types	SKOS Entity	SKOS-CERIF-Mapping (CIM)	CERIF Entity
Class	Collection	is type of	cfClassScheme
Class	Concept	is a	cfClass
Class	ConceptScheme	is a	cfClassScheme
Class	OrderedCollection	is type of	cfClassScheme
lexical	altLabel	is term of class in lexical scheme	cfTerm
mapping	broadMatch	is term of class in mapping scheme	cfTerm
semantic rel.	broader	is term of class in semantic rel. scheme	cfTerm
semantic rel.	broaderTransitive	is term of class in semantic rel. scheme	cfTerm
docu	changeNote	is time-stamped descr. in new/old class	cfClassDescr
mapping	closeMatch	is term of class in mapping scheme	cfTerm
docu	definition	is a	cfDef
map prop	editorialNote	is time-stamped descr. in new/old class is descr. in mapping scheme is cerif publication record reference	cfClassDescr cfClSchDescr cfResPubl
docu	example	is a	cfEx
conc schemes	hasTopConcept	is term of class in concept scheme	cfTerm
lex label	hiddenLabel	is term of class in lexical scheme	cfTerm
docu	historyNote	is time-stamped descr. in new/old class is descr. in docu scheme is cerif publication record reference	cfClassDescr cfClSchDescr cfResPubl

conc schemes	inScheme	inherent cerif linkage (mandatory)	cfClSchID
map prop	mappingRelation	is term of class in mapping scheme	cfTerm
conc coll.	member	is term of class in conc coll. scheme	cfTerm
conc coll.	memberList	is term of class in conc coll. scheme	cfTerm
map prop	narrowMatch	is term of class in mapping scheme	cfTerm
semantic rel	narrower	is term of class in semantic rel. scheme	cfTerm
semantic rel	narrowerTransitive	is term of class in semantic rel. scheme	cfTerm
notations	notation	is	cfClassScheme
docu	note	is time-stamped descr. in new/old class is descr. in docu scheme is cerif publication record reference	cfClassDescr cfClSchDescr cfResPubl
lex label	prefLabel	is tem of class in lexical scheme	cfTerm
semantic rel	related	is term of class in semantic rel. scheme	cfTerm
map prop	relatedMatch	is term of class in mapping scheme	cfTerm
docu	scopeNote	is time-stamped descr. in new/old class is descr. in docu scheme is cerif publication record reference	cfClassDescr cfClSchDescr cfResPubl
semantic rel	semanticRelation	is term of class in semantic rel. scheme	cfTerm
conc schemes	topConceptOf	is term of class in concept schemes	cfTerm

Because of its formal syntax and declared semantics (PIM), CERIF can store – within the Semantic Layer – terms representing structural relationships and by that is able to emulate dictionaries, lexicons, thesauri and domain ontologies. Moreover, since it exhibits at entity instance level a triple structure, all the usual logical processing operations are sustainable. CERIF thus becomes a superset over semantic stores such as dictionaries, thesauri, or ontologies and a mapping to CERIF allows resolution of conflicts in term representation and meaning [Jörg et al. 2011] within the world of Relational databases for which it is designed, i.e. underlying a closed-world assumption. Table 8 presents a mapping from SKOS (PIM) to CERIF (PIM), but will not go into details of mapping additional constraints like ranges and domains, or inverse property definitions as available with advanced SKOS. It is very important at this point, to recognise the two distinct roles KOSs play. On the one hand they support modeling by supplying declared means to describe the modeling constructs (syntax) semantically, on the other hand, they apply these defined constructs to describe the things themselves.

5.3 Conclusion

Analysing the various available descriptions revealed the many-fold approaches to manage the access and maintenance of domain knowledge and where certainly history and available methods should be utilised with guiding of future steps. Increased collaboration, cross-field fertilisation and openness is required for achievements in interoperability.

6 Analysis of Language Technology Entities

In the previous chapter Analysis of Research Entities the entities constituting our Research ontology have been introduced. To demonstrate its field-extensibility, now those entities considered most relevant in Language Technology[212] are being analysed. First, therefore a brief overview of the LT field is given, before the LT entities are investigated in more detail through publicly available descriptions. These are not always equally formal and in most cases differently structured because of diverse underlying technologies. The approach here is thus again a human investigation but also an integration of *substantial field things* into the anticipated structure of FERON – *F*ield-*e*xtensible *R*esearch *ON*tology (chapter 7) guided by ontological commitments.

6.1 Language Technology

Human Language Technology (HLT) roots in Artificial Intelligence and has thus a history of more than 50 years. The European Commission funded HLT for some 40 years with an emphasis on Machine Translation (MT) throughout 1980-1990, resulting in some pioneering MT and Translation Memory technologies. After a period of low visibility, the EC support for HLT has been revived due to new political commitments following the enlargement of the EU and where challenges emerge from global markets towards the overcoming of still significant language barriers. In recognition of the ”importance of languages in the digital age” the current EC work programme includes a specific challenge “Technologies for Digital Content and Languages” with particular support for Small and Medium-sized Enterprises (SMEs)[213].

In an overview [Uszkoreit 2006][214] describes the field of Language Technology at the intersection of *multimedia & multimodality technologies*, *speech technologies*, *text technologies* and *knowledge technologies* (see Figure 55 as in [Uszkoreit 2006, p. 1]).

[212] Intro to CL “What is Computational Linguistics”: http://www.coli.uni-saarland.de/~hansu/what_is_cl.html (Uszkoreit 1996 and 2000) Understanding of Contents through Understanding of Languages (Last visit: June 4th, 2012): http://ec.europa.eu/information_society/events/cf/ict2010/document.cfm?doc_id=14787

[213] European Commission Website about Language Technologies: http://cordis.europa.eu/fp7/ict/language-technologies/ (Last visit: April 1st, 2012)

[214] Cited in a paper https://helda.helsinki.fi/bitstream/handle/10138/29375/sprakvisreport.pdf?sequence=2 as to have been “accessed in 2006” hence „[Uszkoreit 2006]“.

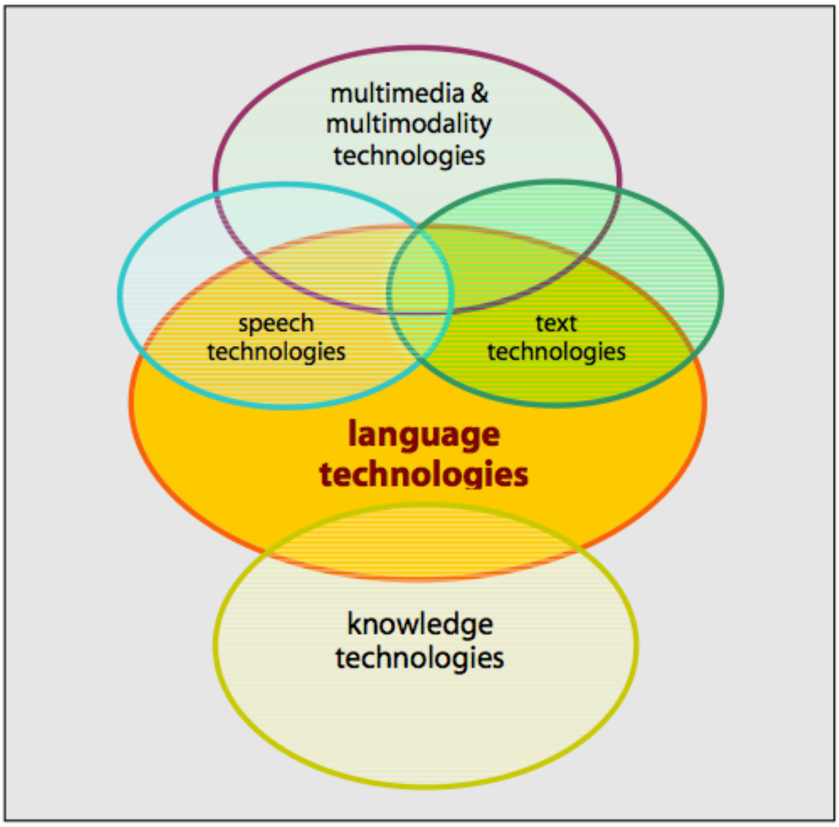


Figure 55: "Language Technology - A First Overview" [Uszkoreit 2006, p. 1]

[Uszkoreit 2006, p. 4] explains: "As the investigation and modelling of human language is a truly interdisciplinary endeavour, the methods and language technology come from several disciplines: computer science, computational and theoretical linguistics, mathematics, electrical engineering and psychology." These different dimensions and influences to the field have been identified and analyzed in [Uszkoreit et al. 2003], and were designed towards an implementation with the LT World portal [Jörg & Uszkoreit 2005, fig. 1], an ontology-driven research information system – often also called *Virtual Information Center* in the field of Language Technology [Jörg et al. 2010][215].

Figure 56 [Jörg & Uszkoreit 2005, fig. 1] shows that *Language Technology* inherits properties from generic concepts such as *Technology* and *Languages*, these are propagated to *LT World* and its subsumed concepts.

[215] LT World: http://www.lt-world.org/ (Last visit: May 2nd, 2012)

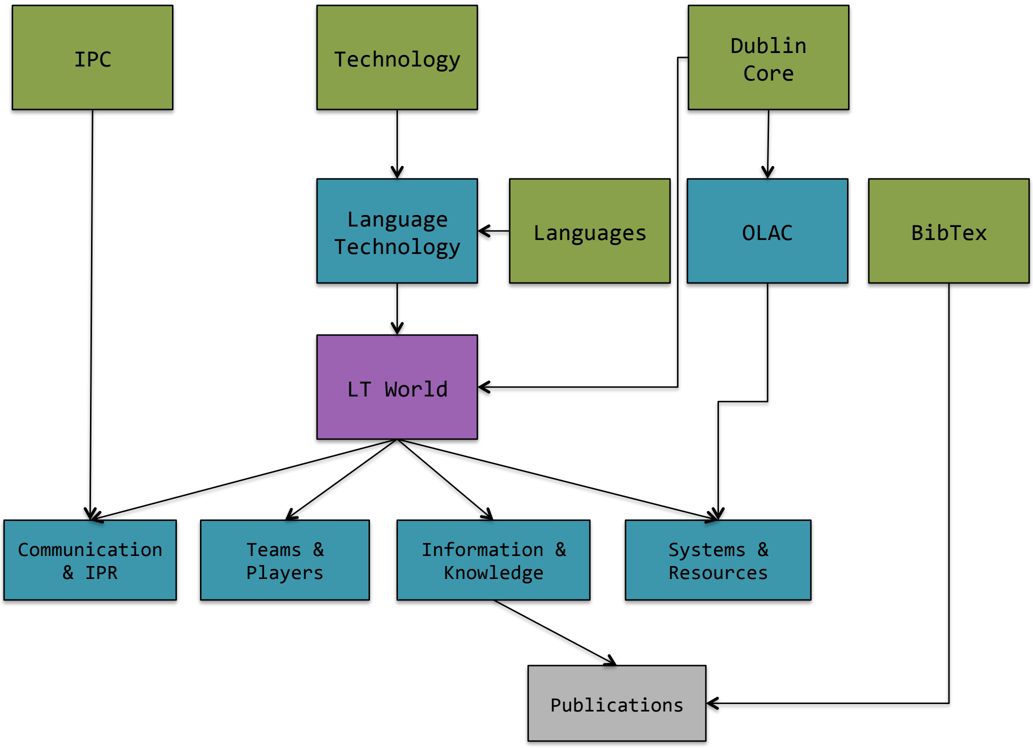


Figure 56: Conceptual LT World structure indicating multiple inheritance [Jörg & Uszkoreit 2005, fig. 1]

In FERON, language technology *lt:technology* is perceived as a method, and *language* as an intrinsic property emergent in information resources; the *w3c-tag:InformationResource* class features an object-property *dcterms:language* upon the range of a *Language* class (a sub-class of *skos:KOS* under *bww:FunctionalScheme*). At first, this seems inline with the conceptual *LT World* structure presented in Figure 56, where *Languages* are propagated to the concept of *Language Technology* and thus inherent in *LT World*[216] and its subsumed concepts, i.e. also in *Information & Knowledge* resources and sister-concepts. However, where FERON is ontologically founded and aimed at being domain agnostic but field-extensible, the LT World Ontology was developed pragmatically and for usage with setting up a field-specific public portal recognising *Languages* to be inherent in *Technology*, i.e. *Language Technology* and therefore propagated to underlying resources; with constraints implemented mostly through e.g. user-interfaces at "presentational structure" level. FERON aims at describing a perceived world that information systems ought to be able to model (see 3.1.1 bullet (1)), where the LT World Ontology followed an "information system as a thing" approach (see 3.1.1 bullet (2)).

[216] The *LT World* concept as presented in Figure 56 is thus to be perceived as a class to feature portal or system specific properties and to subsume abstract classes (access views). FERON (as indicated in 3.1.1 bullet (1)) is not a model of an information system, but a perceived world that information systems ought to be able to model.

The *LT World* structure (Figure 56) subsumes abstract classes such as *Communication & IPR*, *Players & Teams*, *Information & Knowledge*, *Systems & Resources* – these mirror the navigation structure. With *LT World*, field-specific features emerge from the LT Ontology (explained in section 6.3.3.1 and introduced in [Uszkoreit et al. 2003]). These LT features are represented by the *Language Technology* concept in Figure 56 and particularly applied under *Systems & Resources*, but also under *Players & Teams* with *Projects* and *Organizations* and furthermore, under *Communication & IPR* with *Patents* and *News*, and later *Events*, and under *Information & Knowledge* with *Technologies*. LT World never recorded publications because the community maintained a publication collection through the ACL Anthology[217] [Bird et al. 2008], and the ACL Anthology Network offered high-value measurements and services for the community. A linkage with ACL Anthology's publication resources has been incorporated in the LT World portal (see left-column navigation in screenshot Figure 58) through the *ACL Anthology Searchbench* [Schäfer 2012] developed within the TAKE project, funded by the German Federal Ministry of Education and Research[218]. It is worth mentioning, that in FERON *Publications* (5.2.3.1) are not perceived as field-specific; they neither have field-specific intrinsic nor field-specific mutual properties[219]. But, FERON perceives *information resources* such as e.g. *Publications* with LT markup or LT annotation field-specific products *cerif:Product* and as such *lt:Data* under the *vivo:Dataset* class. The class describing *w3c-tag:InformationResource* features and propagates a generic object-property *dcterms:relation* to sub-classes, and the *lt:relation* object-property is equally propagated from *lt:Classes* to instances of *lt:Project*, *lt:Method*, *lt:Facility*, *lt:Equipment*, *lt:Tool*, *lt:Data*, *lt:Service* and *lt:Relation.*

In FERON (Figure 57), language technology *lt:Method* is perceived as a kind of *non-information resource* whereas language resources *lt:Data* and *language description lt:KOS* are kinds of *information resources*.

[217] ACL Anthology Network: http://clair.si.umich.edu/clair/anthology/index.cgi (Last visit: May 2nd, 2012)

[218] Federal Ministry of Education and Research: http://www.bmbf.de/en/ (Last visit: May 2nd, 2012)

[219] While a *lt:KOS* annotation of a *Publication* through the range of the *dcterms:relation* construct were formally possible, it is not recommended to use field-specific annotation with non-field-specific properties.

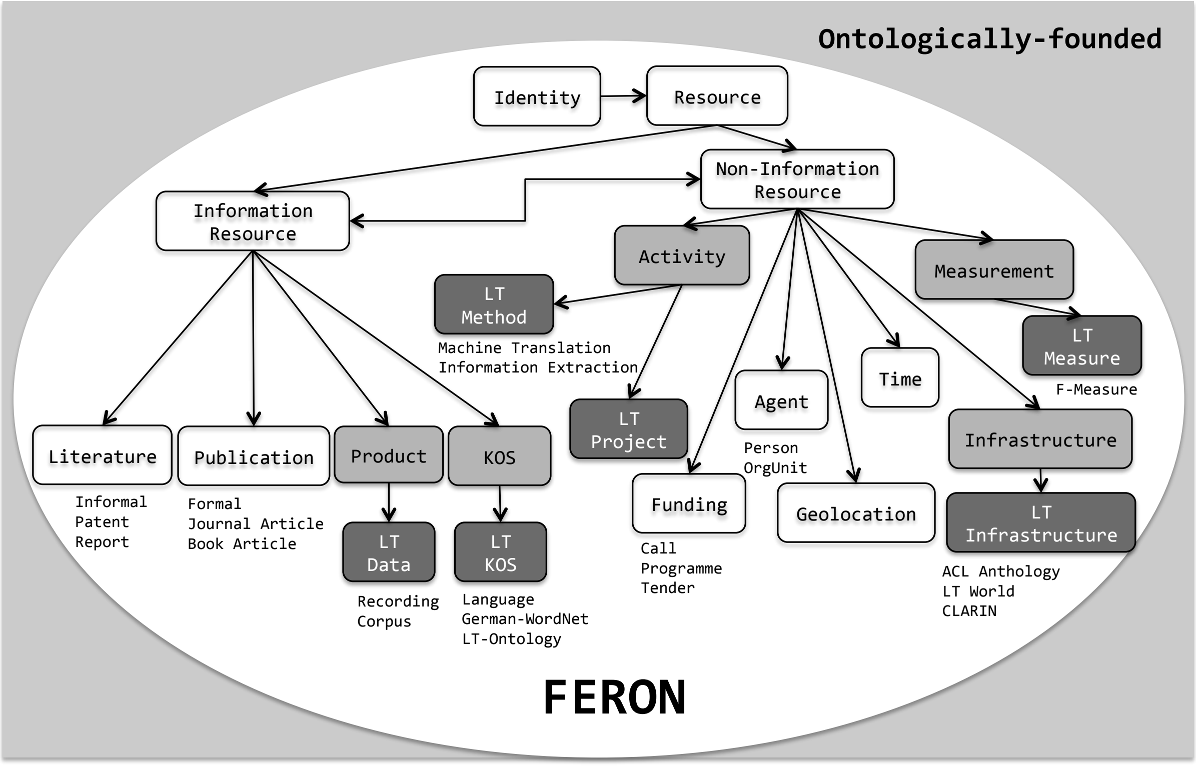


Figure 57: FERON classes featuring field-extensible classes in grey and extensions in dark grey

Information systems such as *LT World*, *META-SHARE* or *CLARIN* are perceived as services *lt:service* subsumed under infrastructure. For the work with FERON, the analysed LT entity descriptions are use-cases to identify and demonstrate field-specific subareas and finally, to evaluate and validate field-extensibility. Figure 57 indicates LT field extensions in dark grey. Formally these are enabled through *lt:Classes* featuring the *lt:relation* property (Figure 57, see additional sections 7.8 and 7.10). E.g. *Language Technology* is perceived as a *lt:Method* labelled *Language Technology*, where the *lt:relation* property emerges. In subsequent sections, LT specific entities are investigated by anticipating FERON.

6.2 LT Methods

The *Survey of the State of the Art* in *Human Language Technology* [HLT Survey 1997][220] gives an overview of the different technologies or methods researched and applied in the field. With LT World, these have been incorporated under the *Technologies* section (Figure 58) with some additions over the years. In total, the *Technologies*' area in LT World counted

[220] HLT Survey: http://www.lt-world.org/hlt-survey/master.pdf (Last visit: June 15th, 2012)

up to more than 100 kinds. For readability, accessibility and usability reasons, these have been further grouped into abstract classes, such as:

- Authoring Tools subsumes e.g. Automatic Hyperlinking, Language Checking, etc.
- Discourse and Dialogue subsumes e.g. Dialogue Modeling, Discourse Modeling, etc.
- Coding and Compression subsumes e.g. Speech Coding, Speech Enhancement, Text Encryption, etc.
- Information Extraction subsumes e.g. Answer Extraction, Relation Extraction, Text Data Mining, Summarisation, etc.
- Information Retrieval subsumes e.g. Categorisation, Clustering, Topic Detection, Relevance Ranking, Speech Retrieval, etc.
- Language Analysis subsumes e.g. Categorial Grammer, Dependency Grammar, Binding Theory, Grammar Formalisms, Lexical Functional Grammar, Head-driven Phrase Structure Grammar, etc.
- Mathematical Methods subsumes e.g. Connectionist Techniques, Conditional Random Fields, Finite State Technology, Hidden Markov Models, etc.

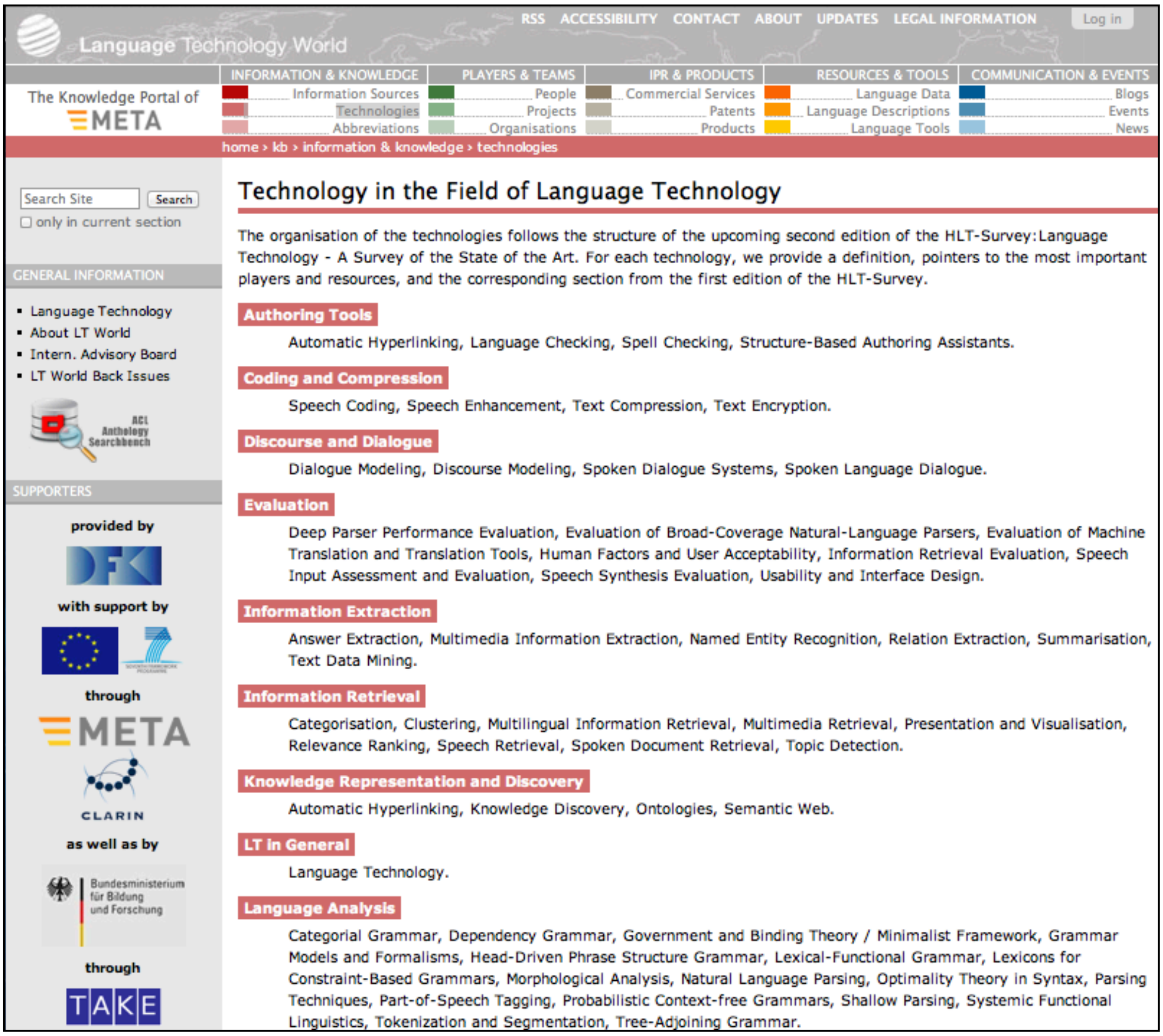
The listed examples show only a few methods[221], but indicate their dependencies, and application relevance in other fields or disciplines – e.g. Language Analysis depends on particular Grammars; e.g. Hidden Markov Models originate from Mathematics, Grammars from Linguistics. FERON, does do not go into the details of each method but is more interested in their ontological or semantic embedding; i.e. to present *lawul* LT-Functions.

In LT World, a *Technology* instance is described by intrinsic properties such as the name, *ltw:technologyName*, abbreviation *ltw:technologyNameAbbreviation*, name variants *ltw:technologyNameVariant*, definition *ltw:technologyDefinition* and by mutual properties such as part *ltw:partOf*, a reference to the HLT Survey book chapter *ltw:hltLabel*, the URL of the HLT Survey book chapter *ltw:hltReference*, but also event references *ltw:relevantEvent*, project references, *ltw:relevantProject*, or organisation references *ltw:relevantOrganisation*, etc.[222]. In LT World, the *Technologies* section under *Information & Knowledge* is special; in fact it is additionally designed as the functional range of the *lt-world:technologicalMethod* property; i.e. one dimension of the LT Ontology [Uszkoreit et al. 2003]. *Technologies* are

[221] The full list of LT World Technologies and their categorization is available from: http://www.lt-world.org/kb/information-and-knowledge/technologies/ (Last visit: June 16th, 2012)

[222] Note: the popular SKOS Concept ontology at the time of initial LT World model design was not yet a W3C recommendation.

thus not only a sub-class of *Information & Knowledge* but additionally applied as a controlled vocabulary KOS through the LT Ontology, although they feature their own mutual properties and in that go beyond typical KOS descriptions. E.g. a *LT World* technology record refers to relevant projects, organisations, people, etc., but also features intrinsic properties such as acronym, name variants and textual description. In *LT World*, *Technologies* are considered a central KOS [Uszkoreit et al. 2003] and as such visible from multiple records, e.g. with project, organisation, or resources and tool instances, etc.[223].



Technology in the Field of Language Technology

The organisation of the technologies follows the structure of the upcoming second edition of the HLT-Survey:Language Technology - A Survey of the State of the Art. For each technology, we provide a definition, pointers to the most important players and resources, and the corresponding section from the first edition of the HLT-Survey.

Authoring Tools

Automatic Hyperlinking, Language Checking, Spell Checking, Structure-Based Authoring Assistants.

Coding and Compression

Speech Coding, Speech Enhancement, Text Compression, Text Encryption.

Discourse and Dialogue

Dialogue Modeling, Discourse Modeling, Spoken Dialogue Systems, Spoken Language Dialogue.

Evaluation

Deep Parser Performance Evaluation, Evaluation of Broad-Coverage Natural-Language Parsers, Evaluation of Machine Translation and Translation Tools, Human Factors and User Acceptability, Information Retrieval Evaluation, Speech Input Assessment and Evaluation, Speech Synthesis Evaluation, Usability and Interface Design.

Information Extraction

Answer Extraction, Multimedia Information Extraction, Named Entity Recognition, Relation Extraction, Summarisation, Text Data Mining.

Information Retrieval

Categorisation, Clustering, Multilingual Information Retrieval, Multimedia Retrieval, Presentation and Visualisation, Relevance Ranking, Speech Retrieval, Spoken Document Retrieval, Topic Detection.

Knowledge Representation and Discovery

Automatic Hyperlinking, Knowledge Discovery, Ontologies, Semantic Web.

LT in General

Language Technology.

Language Analysis

Categorial Grammar, Dependency Grammar, Government and Binding Theory / Minimalist Framework, Grammar Models and Formalisms, Head-Driven Phrase Structure Grammar, Lexical-Functional Grammar, Lexicons for Constraint-Based Grammars, Morphological Analysis, Natural Language Parsing, Optimality Theory in Syntax, Parsing Techniques, Part-of-Speech Tagging, Probabilistic Context-free Grammars, Shallow Parsing, Systemic Functional Linguistics, Tokenization and Segmentation, Tree-Adjoining Grammar.

Figure 58: Screenshot of the LT World portal – Technologies

Figure 58, presents *Technologies* (one dimension of the LT Ontology (see section 6.3.3.1); i.e. the range of *lt-world:technologicalMethod*) within the *LT World* portal structure. These

[223] In a later version of LT World, the *Systems & Resources* class inherent in Figure 56 was transformed into *Resources & Tools* (see Figure 58) to reflect the growing importance of *LT Resources*. This change (and other implied changes) were supported by META-NET, for which LT World became the so-called “Knowledge Portal”, to build a bridge with META-SHARE – a multi-layer infrastructure for *Language Resources*.

overlap conceptually with FERON, in that LT is perceived a *lt:method*. In FERON, *lt:method* features an object-property *lt:relation* that allows for LT dimensions through the *lt:Relation* class with a *bww:function* ranging over LT descriptions, i.e. *lt:KOSs*. The *lt:KOS* class in FERON imports the LT Ontology *lt:Ontology* [Uszkoreit et al. 2003] with sub-class dimensions (such as listed above: *Information Extraction*, *Summarisation*, etc.) under e.g. *lt:technologicalMethod* or *lt:technologicalApplication*, etc., see section 6.3.3.1.

In the next section LT Resources such as *Language Data*, *Language Tools* and *Language Descriptions* are investigated.

6.3 LT Resource

The CLARIN[224] project delivered a survey[225] [CLARIN 2010] as an overview of available Language Resources and Tools. The initial survey did not differentiate between resources and tools, but due to quite heterogenous descriptions the two have hence been treated separately. The distinction is also supported by OLAC where a language resource "is any kind of DATA, TOOL, or ADVICE*"*. The Linguistic Data Consortium (LDC) distinguishes *Language Data* from *Language Tools* and *Language Standards*. The ELRA Universal Catalogue does not distinguish such kinds, but from the types of Resources offered in the catalogue, it seems obvious that ELRA currently maintains more *Data* than *Tools*. With updates in LT World, the area *Resources & Tools* subsumed *Data*, *Tools*, and *Descriptions*. FERON follows the LT World structure for the subsequent investigation of *LT Resources*. *Language Data lt:Data* are subsumed under *vivo:Dataset*, whereas *Language Tools* are considered *services* under *cerif:Infrastructure*; i.e. a *cerif:Service* class subsumes *lt:Tool* and *lt:Service* classes. In FERON, *Language Descriptions* are considered KOSs; e.g. the LT Ontology *lt:Ontology* is an *lt:KOS* subsumed under *skos:KOS* (corresponding to knowledge technologies as seen in Figure 55 [Uszkoreit 2006, p. 1]).

[224] CLARIN project: http://www.clarin.eu/external/ (Last visit: June 4th, 2012)

"The ultimate objective of CLARIN is to create a European federation of existing digital repositories that include language-based data, to provide uniform access tot he data, wherever it is, and to provide existing language and speech technology tools as web services to retrieve, manipulate, enhance, explore and exploit the data. The primary target audience is researchers in the humanities and social sciences and the aim is to cover all languages relevant for the user community." [CLARIN 2010]

[225] CLARIN Survey (A pan-European up-to-date picture can be assumed of the language technology resources available for research and development, given that 188 institutions contributed to the survey.): http://www-sk.let.uu.nl/u/D5C-2.pdf (Last visit: January 8th, 2012)

6.3.1 Language Data

The CLARIN survey identified mostly external metadata features with available *Language Data* description, where internal metadata features such as format, or licensing information (less than 10%), have been clearly underrepresented [CLARIN 2010]. In FERON, internal metadata are perceived as intrinsic properties and external metadata as mutual properties implemented through object-type properties (formally supported by RDF/OWL (see 3.2.6 Formal Ontology) and reflecting functions (3.1.1 Bunge-Wand-Weber Ontology (BWW)), and). OLAC applies the fifteen Dublin Core elements "plus the refinements and encoding schemes of the DCMI Metadata Terms" for *Data and Tools* and adds additional encoding schemes "designed specifically for describing *language resources*, such as subject language and linguistic data type" described in the OLAC Metadata Usage Guidelines[226]. Following the survey results, CLARIN aimed at improvements of metadata coverage by crosswalks over more widely used schemes "The first draft taxonomy was compiled by analyzing the structure of the existing registries (ELRA/ELDA, DFKI, LT-World etc.) and metadata standards for descriptive elements (TEI-Header, IMDI), and the resource descriptions in existing repositories. All together [the resulting survey's LT] 'Resources' comprises nine subcategories: *Aligned Corpus, Multimodal Corpus, Spoken Corpus, Written Corpus, Treebank, Lexicon/Knowledge Source, Grammar, Terminological Resource and other*" [CLARIN 2010] (see also Figure 59).

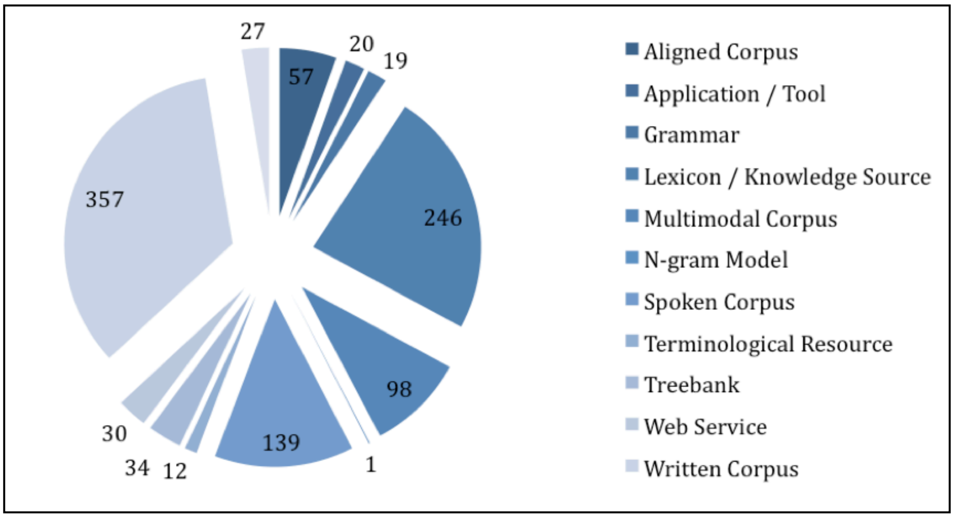


Figure 59: CLARIN Resource Types Structure [CLARIN 2010]

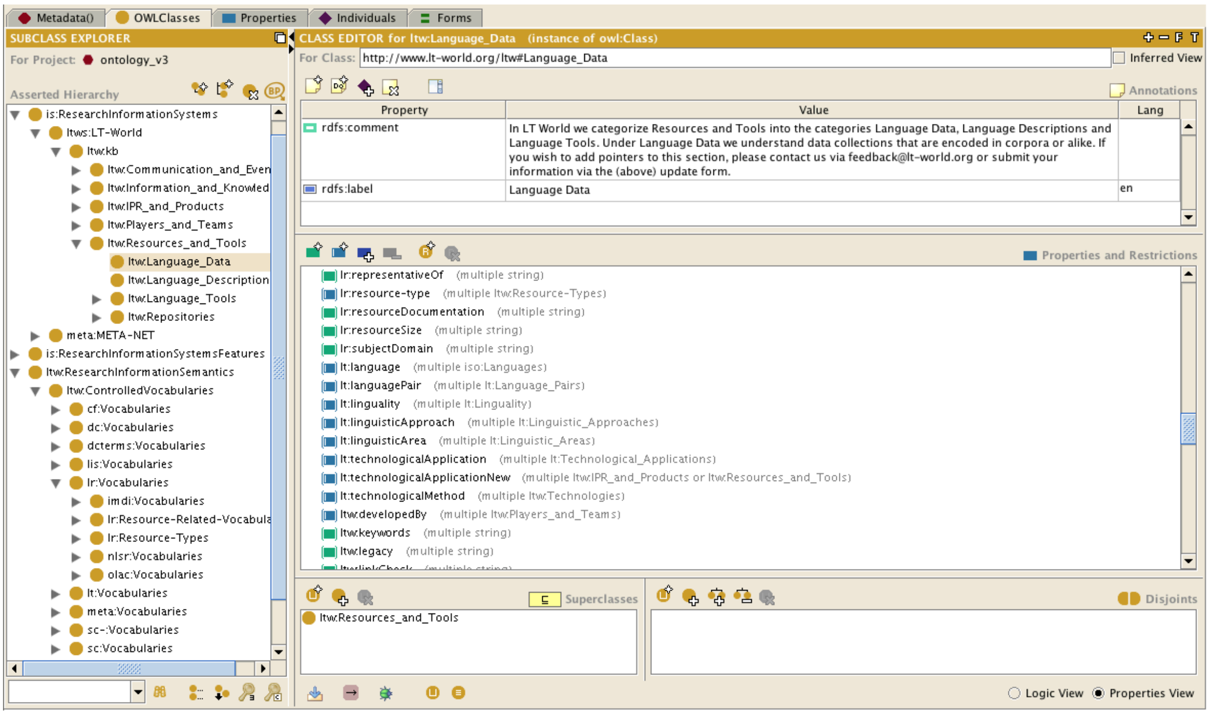
[226] OLAC Metadata Usage Guidelines (2008-07-11): http://www.language-archives.org/NOTE/usage.html (Last visit: June 4th, 2012)

From the given results in Figure 59 it is recognised, that some of the identified types originate from the ACL NLSR hosted at DFKI[227] and indeed not all of the resulting *concepts* are perceived as types of the same kind (the [CLARIN 2010] Resource Types structure is most possibly an unavoidable consequence of the comparison and subsequent integration of investigated heterogeneous underlying LT knowledge organisation systems). According to the definition as given in LT World, *Language Data* are "data collections that are encoded in corpora or alike".

Different types of *Language Data* were discussed, but so far there was no concern about their intrinsic or mutual properties. The *Language Data* class of DFKI's LT World ontology, imports properties from multiple *skos:KOS* and *lt:KOS* and applies its own internal namespaces such as *leg* for legal, *tech* for technical, *sc* for the sciences, or *ltw* for LT World, *lt* for the LT ontology, and *lr* for language resource, or *plone* for system-specific information, etc. The class also imports properties from Dublin Core *dcterms*, Open Language Archives Community *olac*, and the Natural Language Software Registry *nlsr*. The LT World ontology class for *ltw:Language_Data* Figure 60 is a sub-class of *ltw:Resources_and_Tools* as a sub-class of LT World knowledge base *ltw:kb* and *lr:Resources*, and where *ltw:kb* is a sub-class of *ltw:LT-World-LT-Ontology*, *plone:ContentManagementSystem* and *dc:DublinCore*. The *lr:Resources* class is under *lt:Entities*, *sc:Resources*, *lr:LT-World-LR-Ontology*, *nlsr:ACL-Natrural_Language_ Software_Registry*. The *ltw:Language_Data* class thus inherits multiple intrinsic and mutual properties (the property type – datatype or object-type is visible in Figure 65 through ranges – string for the first and class names for the latter), such as e.g.: *dc:title*, *dcterms:abstract*, *dc:language*, *dc:rights*, *dcterms:accessRights*, *dcterms:alternative*, *dcterms:audience*, *dcterms:available*, *dcterms:dateCreated*, *dcterms:dateAccepted*, *dcterms:dateCopyright*, *dcterms:dateSubmitted*, *dcterms:valid*, *dc:publisher*, *dc:source*, *nlsr:title*, *nlsr:url*, *nlsr:academicPricing*, *nlsr:commercialPricing*, *nlsr:multiplePricing*, *nlsr:currency*, *nlsr:ftp*, *nlsr:input-mimetype*, *nlsr:institute*, *nlsr:license*, *nlsr:mail*, *nlsr:mainSection*, *nlsr:supportedLanguage*, *nlsr:supportedPlatform*, *olac:discourse-type*, *olac:format-cpu*, *olac:format-encoding*, *olac:format-markup*, *olac:format-os*, *olac:format-sourcecode*, *olac:funtionality-type*, *olac:linguistic-type*, *olac:role*, *olac:subject-language*, *lr:subjectDomain, lr:resourceSize*, *lr:resourceDocumentation*, *lr:representativeOf*, *ltw:annotationStyle*, *ltw:annotationFormat*, *ltw:annotationContent*, *ltw:annotationAnnotator*, *ltw:annotationValidatedBy*, *plone:title*, *plone:workflowState*, *plone:hasPreferredSuperclass*, *plone:creationTime*, *plone:isHiddenClass*, *plone:lastModified*, *sc:suggestedPurpose*,

[227] ACL Natural Language Software Registry (NLSR): http://registry.dfki.de/ (Last visit: June 4th, 2012)

sc:resource-type, etc. Amongst the mentioned, the class inherits mutual properties from the LT Ontology - namely *lt:language*, *lt:languagePair*, *lt:linguality*, *lt:linguisticApproach*, *lt:linguisticArea*, *lt:technologicalMethod*, and *lt:technologicalApplication* (see also extract in Figure 65).



For Class: http://www.lt-world.org/ltw#Language_Data □ Inferred View

Property	Value	Lang
rdfs:comment	In LT World we categorize Resources and Tools into the categories Language Data, Language Descriptions and Language Tools. Under Language Data we understand data collections that are encoded in corpora or alike. If you wish to add pointers to this section, please contact us via feedback@lt-world.org or submit your information via the (above) update form.	
rdfs:label	Language Data	en

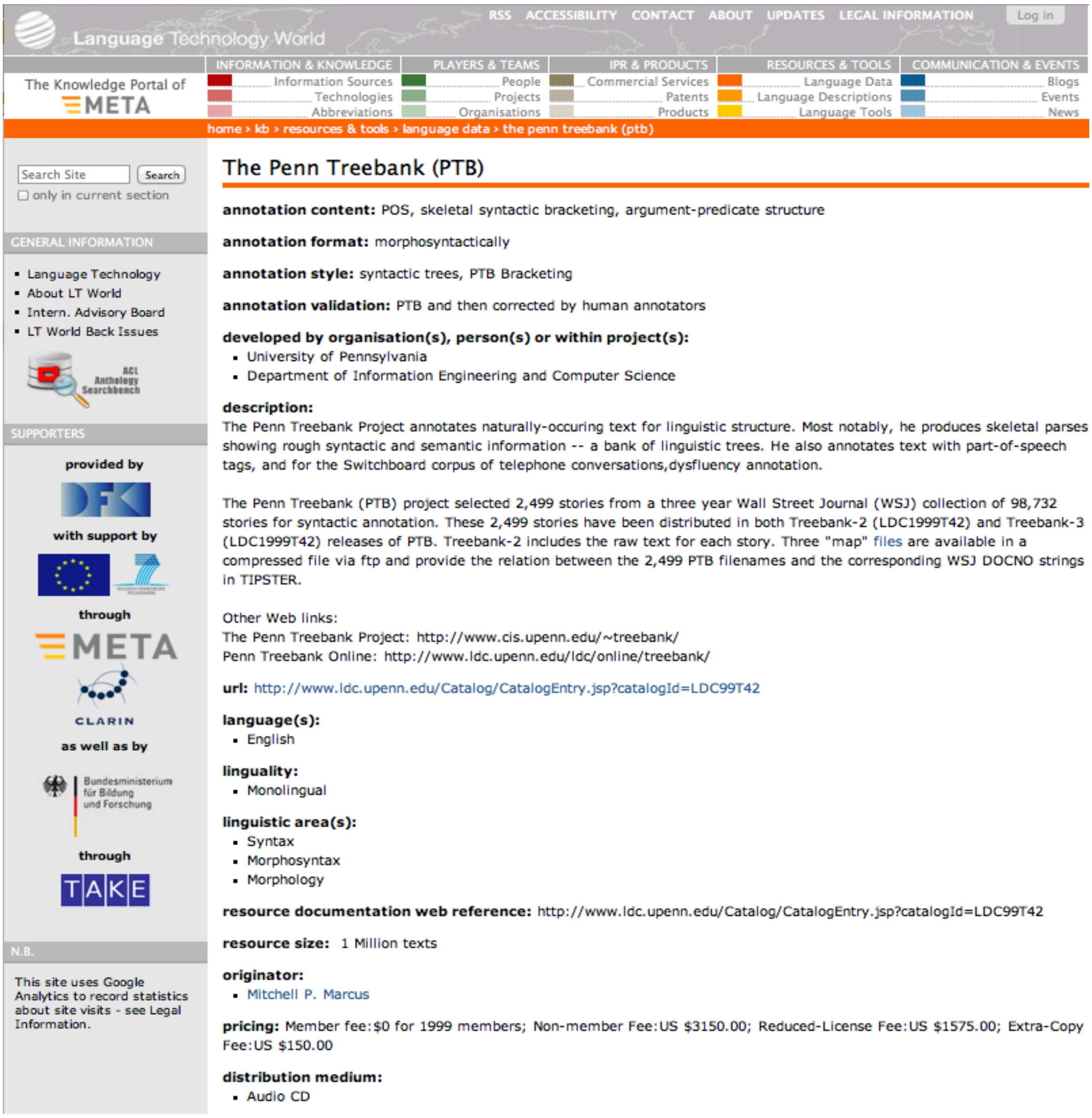
lr:representativeOf (multiple string)
lr:resource-type (multiple ltw:Resource-Types)
lr:resourceDocumentation (multiple string)
lr:resourceSize (multiple string)
lr:subjectDomain (multiple string)
lt:language (multiple iso:Languages)
lt:languagePair (multiple lt:Language_Pairs)
lt:linguality (multiple lt:Linguality)
lt:linguisticApproach (multiple lt:Linguistic_Approaches)
lt:linguisticArea (multiple lt:Linguistic_Areas)
lt:technologicalApplication (multiple lt:Technological_Applications)
lt:technologicalApplicationNew (multiple ltw:IPR_and_Products or ltw:Resources_and_Tools)
lt:technologicalMethod (multiple ltw:Technologies)
ltw:developedBy (multiple ltw:Players_and_Teams)
ltw:keywords (multiple string)
ltw:legacy (multiple string)

Superclasses: ltw:Resources_and_Tools

○ Logic View ● Properties View

Figure 60: LT World Ontology – Language Data Properties View

In FERON, the *lt:Data* class inherits properties from *vivo:Dataset*, a sub-class of *cerif:Product*. It is described by intrinsic properties *dcterms:identifier*, *federated-identifier*, *cerif:keywords*, *dcterms:language*, *cerif:description*, *cerif:title*, *cerif:publication-date*, *datacite:format*, *datacite:size*, *datacite:version*, and by multiple mutual properties such as *dcterms:relation* and *lt:relation*. FERON does not further subclassify field-specific *lt:*classes such as *lt:Data*, *lt:Tools*, or *lt:Service*, but allow for incorporation of multiple *lt:KOS* or other field-specific descriptions under the *skos:KOS* class. This allows for type-independence with instances and for multiple and scalable LT functions through *lt:Relation* (see section 7.6).



This site uses Google Analytics to record statistics about site visits - see Legal Information.

The Penn Treebank (PTB)

annotation content: POS, skeletal syntactic bracketing, argument-predicate structure

annotation format: morphosyntactically

annotation style: syntactic trees, PTB Bracketing

annotation validation: PTB and then corrected by human annotators

developed by organisation(s), person(s) or within project(s):
- University of Pennsylvania
- Department of Information Engineering and Computer Science

description:
The Penn Treebank Project annotates naturally-occuring text for linguistic structure. Most notably, he produces skeletal parses showing rough syntactic and semantic information -- a bank of linguistic trees. He also annotates text with part-of-speech tags, and for the Switchboard corpus of telephone conversations,dysfluency annotation.

The Penn Treebank (PTB) project selected 2,499 stories from a three year Wall Street Journal (WSJ) collection of 98,732 stories for syntactic annotation. These 2,499 stories have been distributed in both Treebank-2 (LDC1999T42) and Treebank-3 (LDC1999T42) releases of PTB. Treebank-2 includes the raw text for each story. Three "map" files are available in a compressed file via ftp and provide the relation between the 2,499 PTB filenames and the corresponding WSJ DOCNO strings in TIPSTER.

Other Web links:
The Penn Treebank Project: http://www.cis.upenn.edu/~treebank/
Penn Treebank Online: http://www.ldc.upenn.edu/ldc/online/treebank/

url: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC99T42

language(s):
- English

linguality:
- Monolingual

linguistic area(s):
- Syntax
- Morphosyntax
- Morphology

resource documentation web reference: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC99T42

resource size: 1 Million texts

originator:
- Mitchell P. Marcus

pricing: Member fee:$0 for 1999 members; Non-member Fee:US $3150.00; Reduced-License Fee:US $1575.00; Extra-Copy Fee:US $150.00

distribution medium:
- Audio CD

Figure 61: LT World portal instance view of Language Data "The Penn Treebank"

The properties of *Language Data* will not be further investigated, but for additional information and examples the LT World portal is a rich information source, from where multiple real records can be retrieved featuring intrinsic and mutual LT properties as revealed in Figure 61[228].

[228] LT World Language Data area: http://www.lt-world.org/kb/resources-and-tools/language-data/
LT World Language Data Instance as in Figure 61: http://www.lt-world.org/kb/resources-and-tools/language-data/ltw_x3alanguage_x5fdata_.2010-09-22.7925619210 (Last visit: June 4th, 2012)

6.3.2 Language Tool

In previous sections it was revealed that although *Language Data* and *Language Tools* have many properties in common, they are best treated separately. FERON does not aim at an exhaustive description of the LT domain to the very details, but aims at demonstrating its field-extensibility. *Language Tool* is thus to a large extent similar with *Language Data* from an extension perspective. However, the *lt:Tool* class is subsumed under *cerif:Service* and therefore perceived as an infrastructure, i.e. a *non-information resource*. The class *w3c-tag:NonInformationResource* does not feature an intrinsic *language* property but *language* is a function featured through the *lt:relation* property upon the range of applicable *lt:KOSs* (e.g. a language dimension as defined in the LT Ontology through the *lt-world:Language* class). In LT World *Language Tools* are defined as "computational tools that support the processing of language data, potentially using language descriptions".

The NLSR[229] taxonomy to classify NLP software collections is largely based on [HLT Survey 1997]. In addition to the Dublin Core-based OLAC format, the NLSR extended its metadata scheme with the following attributes:

> *creatorInstitute; creatorMail; dateCode; description; descriptionAbstract; License; relationAcademicPricing; relationCommercialPricing; relationFTP; relationMainSection; relationMainSectionID; relationMultiplePricing; relationSubSection; relationSubSectionID; relationSupportedLanguage; relationSupportedLanguageID; relationSupportedPlatform; relationURL; titleEn*

The properties of *Language Tool* will not be further investigated, but for additional information and examples the LT World[230] is a rich information source, from where multiple real records can be retrieved (imported from the NLSR); these employ intrinsic and mutual LT properties.

[229] The ACL Natural Language Software Registry (NLSR) http://registry.dfki.de is a concise summary of the capabilities and sources of a large amount of natural language processing (NLP) software available to the NLP community. It comprises academic, commercial and proprietary software with specifications and terms on which it can be acquired. While the NLSR concentrates on listing NLP software, it does not exclude the listing of Natural Language Resources (NLR), i.e. Data. Since there is interest in Resources strongly related to the listed tools.

[230] LT World Language Tools area: http://www.lt-world.org/kb/resources-and-tools/language-tools/ (Last visit: June 4th, 2012)

6.3.3 Language Descriptions (LT KOS)

The concept of knowledge organisation systems (KOS) was introduced in section 5.2.4. These are systems such as thesauri, classification schemes, taxonomies, topic maps, or similar types of controlled vocabularies in support of organising knowledge. In LT Research and thus especially, with *Language Data* and *Language Tools*, the *Language Descriptions* may be additionally employed during natural language processing. In Figure 55 by [Uszkoreit 2006, p. 1] they have been identified at the intersection of multimedia, multimodality, speech or text technologies (labelled knowledge technologies), and are perceived as critically important resources – especially with integration and interchange. FERON, does not sub-type field-specific lt:Classes such as *lt:Method*, *lt:Data*, *lt:Tools*, *lt:Service*, *lt:Project*, *lt:Facility*, *lt:Equipment* or *lt:Measurement*, but allows for importing of multiple *lt:KOS* or other field-specific KOS under the *skos:KOS* class, applicable as functional references upon *lt:Relation*. Under *lt:KOS*, it distinguishes – inline with the parent class *skos:KOS* and ontological foundations – between *lt:Law* and *lt:FunctionalScheme*, however these will not be further defined within this work. Below *lt:FunctionalScheme* FERON incorporates[231] some *lt:KOSs* even though not all are formally and ontologically specified. E.g. under *clarin:ResourceTypes* [CLARIN 2010]) it subsumes the *clarin:WrittenCorpus*, *clarin:SpokenCorpus*, *clarin:MultimodalCorpus, clarin:AlignedCorpus*. Compared with the ELRA categories in the Universal Catalogue, the CLARIN types are perceived as sub-kinds thereof, and are therefore employed as *elra:TerminologicalLR*, *elra:SpokenLR*, *elra:WrittenLR*, and *elra:MultimodalMultimediaLR*. Some NLSR Resource categories have been incorporated (in Figure 56 under *Systems & Resources*, now in LT World re-structured under *Resources & Tools*), e.g. the *nlsr:Grammars*, *nlsr:GrammarResources*, *nlsr:Lexica*, *nlsr:TerminologySystems*, *nlsr:TerminologyTools*, *nlsr:MultimodalCorpora* to explain FERON's capability of multiple KOS implementation with equal concept names, but conceptually distinguished through namespace-prefixes.

Some known *lt:KOS*s will subsequently be presented, at the same time additional insight into the LT field provided by exploration of the diversity of 'knowledge technology' approaches. However, neither of them will be exhaustively modelled, but only some concepts employed as they were perceived within the structure and semantics of FERON. In FERON, the *lt:KOS* class features inherited intrinsic properties such as *dcterms:identifier, cerif:title*, *cerif:keywords*, *dcterms:language, cerif:term*, *cerif:definition*, *cerif:example*, *cerif:source* as

[231] This thesis does not talk of import because some underlying structures and constructs have been defined and designed even before formal ontologies have become popular and where the structure is not easily transformed in a 1:1 manner. Furthermor this work is not aimed at a mapping – but more concerned with identifying perceived substantial things.

well as inherited mutual properties such as *federated-identifier* and *dcterms:relation*. Additionally, *lt:KOS* features a *lt:relation* property anticipating LT relationships within the range *lt:KOS* concepts. The properties are propagated to subclasses and thus inherent with each incorporated LT *knowledge organisation system*.

6.3.3.1 LT Ontology

The LT Ontology was developed for organising LT Resources within the LT World portal. It was considered the field-specific core of the Virtual Information Center *LT World* through which information was structured alongside multiple dimensions "The novel core of our conceptual structure is an ontology of language technology. Since there was no ontology or systematics for our young discipline that we could have adopted, we designed a new multidimensional core ontology [...] The assumed ontology of science and technology assumes that a core role of any R&D activity is dedicated to the research themes [...] A research theme is now construed as an instance of a relational type with six roles" [Uszkoreit et al. 2003, p. 3].

- Application (e.g. Grammar Checking, Text Translation, Speech Dialogue Systems, ...)
- Linguality (monolingual, bilingual, multilingual, translingual, language-independent)
- Languages / Language Pairs (e.g. Romanian, Thai, ... / <en-fr>, <de-gr>,...)
- Technologies (e.g. Hidden Markov Model, Linear Programming, ...)
- Linguistic Area (e.g. Morphology, Syntax, Pragmatics, ...)
- Linguistic Approach (e.g. Two-Level Morphology, Systemic Functional Grammar, ...)

In [Jörg & Uszkoreit 2005, pp. 3 ff.] the ontology is explained as follows: "The LT classification scheme is propagated to multiple sub concepts. Its roles (attributes) – just as many other roles in the ontology – are set-values and not only point to individual targets. Books, articles, projects etc. often are dedicated to several applications, even a software product can bundle several applications.

The first three dimensions of the LT classification scheme depend on the application. They describe the type of application, the linguality and the covered languages. The attribute *Application* takes as value a set of application types. The attribute *Linguality* describes the dependency of an application on a specific set of languages. Applications can be monolingual such as a grammar checker designed just for Finnish. They can be multilingual such as a text-to-speech product for Italian, French and Spanish. Translingual applications cross language boundaries. This is always the case for machine translation. However, there are also other applications carrying information across languages. An example is cross-lingual information retrieval, where a query is formulated in one language but relevant documents are (also) returned in other languages. Finally there exists a large number of language independent applications such as generic search engines or most speech compression programs.

The attribute *Technologies* takes values from a set of methods or techniques origi- nating in computer science, mathematics, or electrical engineering. Linguistic Area is another attribute that adopts from the discipline of linguistic the levels of linguistic description in order to specify which aspects of language are covered by some project, publication, etc. The last attribute specifies the applied *Linguistic Approach* such as theories, models, or methods. The ontology is not only concerned with terminological coverage and data organisation but used as a formal specification for the whole information system and as such is involved with all related functions and processes." [Jörg & Uszkoreit 2005, p. 4] In FERON, the LT Ontology *lt:Ontology* is a sub-class of *lt:KOS* under *lt:FunctionalScheme*.

6.3.3.2 Open Language Archive Community (OLAC)

The Open Language Archives Community[232] is an international partnership of institutions and individuals creating a worldwide library of language resources. In August 2010, the OLAC Archives contained approximately 35.000 records. The latest OLAC Metadata standard release presents a format to describe language resources and to provide associated services within the framework of the Open Archives Initiative (OAI)[233] based on the Dublin Core metadata element set, from where it used all fifteen elements. Additionally, OLAC follows the DC recommendation for qualifying elements by means of refinements or encoding schemes. The OLAC metadata scheme is thus an application profile to incorporate elements from simple and qualified DC. OLAC specific extensions are defined in [Simons & Bird 2008][234] as follows:

[232] Open Language Archives Community (OLAC): http://www.language-archives.org/ (Last visit: January 8th, 2012) On the website OLAC's mission is described as an „international partnership of institutuions and individuals who are creating a worldwide virtual library of language resources by: (i) developing consensus on best practice for the digital archiving of language resources, and (ii) developing a network of interoperating repositories and services for housing and accessing such resources.“ OLAC provides Standards for Language Resources, that is, the OLAC Metadata Set

[233] Open Archives Initiative (OAI): http://www.openarchives.org/ (Last visit: June 4th, 2012)

[234] Recommended OLAC Metadata extensions: http://www.language-archives.org/REC/olac-extensions.html (Last visit: January 8th, 2012)

- Name: olac:discourse-type; Applies to: dc:type, dc:subject
- Name: olac:language; Applies to: dc:language, dc:subject
- Name: olac:linguistic-field; Applies to: dc:subject
- Name: olac:linguistic-type; Applies to: dc:type
- Name: olac:role; Applies to: dc:contributor

In FERON, the OLAC scheme is subsumed as *olac:MetadataScheme* under *lt:KOS* below *lt:FunctionalScheme*. It employs the above named extension sub-classes, where upon rules or constraints, e.g. the DC references such as *dc:type* (now called *dcterms:Type*) could be employed. In FERON, the *DublinCore* Metadata Element Set as *dcterms:DublinCore* is partly incorporated through concept classes under *bww:FunctionalScheme* to support DC functions from within the *Relation* construct, such as with *dcterms:Subject*, or *dcterms:Type*. However in FERON, some DC elements are perceived as *dcterms:Resource* properties – such as *dcterms:identifier* or *dcterms:relation* – and these propagate to underlying resources, including *lt:KOS*. Furthermore, DC elements such as *dcterms:Creator* or *dcterms:Contributor* are perceived as roles, under *InformationResource-Person*.

6.3.3.3 Text Encoding Initiative (TEI)

According to their statement on the public website, the Text Encoding Initiative (TEI) is a consortium to develop and maintains a standard for the representation of texts in digital form. Its chief deliverable is a set of guidelines, which specify the encoding methods for machine-readable texts, chiefly in the humanities, social sciences and linguistics. Since 1994, the TEI Guidelines[235] have been widely used by libraries, museums, publishers, and individual scholars to present texts for online research, teaching, and preservation. In addition to the Guidelines themselves, the Consortium provides a variety of supporting resources, including resources for learning TEI, information on projects using the TEI, TEI-related publications, and software developed for or adapted to the TEI.[236]

The Guidelines divide the TEI tag set into two broad categories, those used to capture ‘metadata’ about the text being encoded (authorship and responsibility, bibliographical information, manuscript description, revision history, etc.), and those used to encode the structural features of the document itself, such as sections, headings, paragraphs, quotations,

[235] TEI Guidelines (version P5): http://www.tei-c.org/Guidelines/ (Last visit: January 8th, 2012)

[236] Text Encoding Initiative (TEI): http://www.tei-c.org/ (Last visit: January 15th, 2012)

highlighting, and so on. The Guidelines additionally inform "that there is no single DTD or schema which is the TEI" and one is supposed to choose from available modules, those that one wants, ensuring "the three modules core, header and textstructure (and tei, when using RELAX NG) should always be chosen unless one is certain to know what one is doing".[237]

Table 9: Text Encoding Initiative (TEI) Modules.

analysis	Simple analytic mechanisms
certainty	Certainty and uncertainty
core	Elements common in all TEI documents
corpus	Header extensions for corpus texts
declarefs	Feature system declarations
dictionaries	Printed dictionaries
drama	Performance texts
figures	Tables, formulae, and figures
gaiji	Character and glyph documentation
header	The TEI Header
iso-fs	Feature structures
linking	Linking, segmentation and alignment
msdescription	Manuscript Description
namespaces	Names and dates
nets	Graphs, networks and trees
spoken	Transcribed Speech
tagdocs	Documentaiton of TEI modules
tei	Declarations of datatypes, classes, and macros available to all TEI modules
textcrit	Text criticism
textstructure	Default text structure
transcr	Transcription of primary sources
verse	Verse structure

In FERON, the *tei:Modules* are subsumed below *lt:KOS* under *lt:FunctionalScheme* as classes, and therefore applicable from e.g. *lt:Data* resources through *lt:Relation*.

[237] Although there is no default schema, there are a number of example customizations which may very well meet one's needs, which can be downloaded from the TEI web site or from within the Roma interface.

6.3.3.4 FrameNet

The FrameNet[238] project is building a lexical database of English that is both human- and machine readable by annotating examples of how words are used in actual texts "based on a theory of meaning called **Frame Semantics**, deriving from the work of Charles J. Fillmore and colleagues (Fillmore 1976, 1977, 1982, 1985, Fillmore and Baker 2001, 2010). The basic idea is straightforward: that the meanings of most words can best be understood on the basis of a **semantic frame**: a description of a type of event, relation, or entity and the participants in it [...] Formally, FrameNet annotations are sets of triples that represent the FE [frame element] realizations for each annotated sentence, each consisting of a frame element name (for example Food), a grammatical function (say, Object), and a phrase type (say, noun phrase (NP)) [...] The FrameNet team have defined more than 1,000 semantic frames and have linked them together by a system of frame relations, which relate more general frames to more specific ones and provide a basis for reasoning about events and intentional actions."

Compared to FERON the FrameNet approach is close to the *Roles* concept under *skos:KOS*. Some role schemes have been incorporated such as "Person Employment Types" or "Activity Structure" from the CERIF Vocabulary, where vocabulary terms are introduced through link entities such as e.g. cfPerson_OrganisationUnit. In FERON, any class subsumed under *Roles* below *bww:FunctionalScheme* could be perceived as a semantic frame because roles – contrary to types – are always mutual and thus situational or contextual, i.e semantic. Where CERIF entities (including Semantic Layer entities) are semantically and syntactically declared (e.g. a role is described through a functional cfClass reference) in a top-down manner, FrameNet represents triples evoked from Lexical Units[239] through text extractions, keeping "syntactic realizations and valence patterns".

In FERON, the FrameNet concept intersects highly with functional roles. Due to its field-affiliation, a *frameNet:LexicalUnits* class is subsumed under the *lt:KOS* class. FrameNet is not dedicated to a particular domain, the index of its Lexical Units ranges over a diversity of terms such as *Activity*, *Age*, *Artifact*, *Awareness* to *Emotions*, *Goal*, *Graph_shape*, *Information* or *Infrastructure*, etc. FrameNet annotations of information resources could thus be realized through the *lt:relation* construct, the same holds for *frame element*, *syntactic*

[238] FrameNet: https://framenet.icsi.berkeley.edu/fndrupal/home (Last visit: July 1st, 2012)

[239] Each LU name is followed by the part of speech, the name of the relevant frame, and its status. If a lexical unit has the status "Finished_initial" (meaning it was annotated in FN2) or "FN1_sent" (meaning annotated in FN1), it will be followed by links to the HTML files for the lexical entry and the annotated sentences. Lexical units on which work has not been completed may have only a link for the lexical entry, or no link at all. The lexical entry provdes two tables with information about the LU:Frame Elements and their Syntactic Realizations; and Valence Patterns.

realization or e.g. *valence pattern*. These could be either enabled within the current structure of subsumption classes or by explicit LT properties with LT information resources. The (about 40) FrameNet Semantic Types and their linkage with SUMO, which was aimed to "imporove semantic parsing and ontology lexicalization"[240], have not been invesitgate in this work. However, FERON is aware of the SUMO top-level ontology (see Figure 6) in section 3.1.3 Suggested Upper Merged Ontology (SUMO).

6.3.3.5 ISOcat Data Category Registry (ISOcat DCR)

The ISOcat Data Category Registry[241] provides a framework for defining data categories of widely accepted linguistic concepts. ISOcat is an ISO standard 12620 and compliant with the ISO/IEC 11179 family of standards. According to the ISOcat DCR model "each data category is assigned a unique administrative identifier, together with information on the status or decision-making process associated with the data category. In addition, data specifications in the DCR contain linguistic descriptions such as data category definitions, statements of associated value domains, and examples. Data category specifications can be associated with a variety of data element names and with language-specific versions of definitions, names, value domains and other attributes."

Within ISO, the technical committees TC37/SC2 work concentrates on terminological and lexicographical working methods[242]. ISO 10241-1:2011 is the latest published version revising ISO 10241:1992. ISOcat distinguishes 14 thematic domain groups:

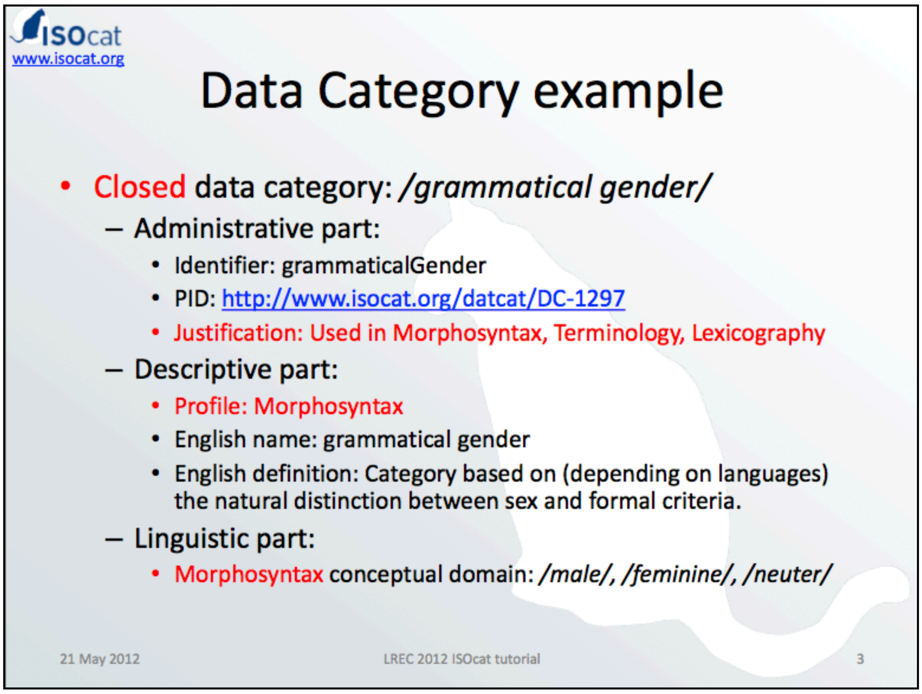
[240] Related FrameNet Projects: https://framenet.icsi.berkeley.edu/fndrupal/related_projects (Last visit: May 3rd, 2012)

[241] ISOcat Data Category Registry: http://www.isocat.org/ (Last visit: January 8th, 2012) The registration authority of the so-called TC37 DCR is the Max Planck Institute for Psycholinguistics.

[242] The ISO Technical Subcommittees/Working Groups are divided into Language Coding, Terminography, Lexicography, Source identification for language resources, requirements and certification schemes for cultural diversity management, Translation and interpretation processes: http://www.iso.org/iso/iso_technical_committee.html?commid=48124 (Last visit: January 8th, 2012) "ISO 10241:2011 specifies requirements for the drafting and structuring of terminological entries in standards, exemplified by terminological entries in ISO and IEC documents. Terms and other designations occurring in terminological entries can include letters, numerals, mathematical symbols, typographical signs and syntactic signs (e.g. punctuation marks, hyphens, parentheses, square brackets and other connectors or delimiters), sometimes in character styles (i.e. fonts and bold, italic, bold italic or other style conventions) governed by language-, domain- or subject-specific conventions. Terms can also include standardized symbols (which can be language independent or internationally harmonized, such as symbols for quantities and units as well as graphical symbols) which are under the responsibility of different committees in ISO and IEC. ISO 10241-1:2011 is based on the principles and methods given in ISO 704 and provides rules for both monolingual and multilingual terminological entries in standards and their indexes. ISO 10241-1:2011 is applicable to all standards that contain terminological entries. It does not deal with the administrative procedures nor the technical specifications required by standardizing bodies for the preparation of terminology standards. Since presentation and layout rules by nature are very much tied to the script and to the publishing rules of the standardizing body, they are dealt with only on an abstract level in ISO 10241-1:2011." Unfortunately the relevant specification document is only available for 158 CHF.

- TDG 1: Metadata
- TDG 2: Morphosyntax
- TDG 3: Semantic Content Representation
- TDG 4: Syntax
- TDG 5: Machine Readable Dictionary
- TDG 6: Language Resource Ontology
- TDG 7: Lexicography
- TDG 8: Language Codes
- TDG 9: Terminology
- TDG 11: Multilingual Information Management
- TDG 12: Lexical Resources
- TDG 13: Lexical Semantics
- TDG 14: Source Identification

An ISOcat data category example is provided with Figure 62.



ISOcat
www.isocat.org

Data Category example

- Closed data category: */grammatical gender/*
  - Administrative part:
    - Identifier: grammaticalGender
    - PID: http://www.isocat.org/datcat/DC-1297
    - Justification: Used in Morphosyntax, Terminology, Lexicography
  - Descriptive part:
    - Profile: Morphosyntax
    - English name: grammatical gender
    - English definition: Category based on (depending on languages) the natural distinction between sex and formal criteria.
  - Linguistic part:
    - Morphosyntax conceptual domain: */male/, /feminine/, /neuter/*

21 May 2012 LREC 2012 ISOcat tutorial 3

Figure 62: ISOcat data category example (LREC 2012 ISOcat tutorial)[243]

FERON includes *isocat:ThematicDomainGroup* under *lt:KOS* within *lt:FunctionalScheme*, but does not further elaborate on properties.

[243] ISOcat tutorial at LREC 2012 http://www.isocat.org/2012-LREC-ISOcat/material/ISOcat-3-DC-specifications.pdf (Last visit: May 15th, 2012)

6.3.3.6 Machine-Readable Terminology Interchange Format (MARTIF)

MARTIF was developed in cooperation with the Text Encoding Initiative (TEI) and the Localisation Industry Standards Association (LISA) with the goal to be platform independent and publicly available. MARTIF is also known as ISO (FDIS) 1220, where 150 categories are standardised as ISO 12620[244]. The categories resulted from different needs and approaches of different working groups and a merge of synonymous names was achieved[245]. MARTIF with Specified Constraints (MSF) has later been reflected in the CLS Framework[246], dealing with the structure and content of terminological databases (also called termbases) for representation, with new termbase design and for the sharing of terminological data. MARTIF was mostly applicable with XML formats but also dealt with representing terminological data in a relational database. The graphical representation of the CLS framework is shown in the following graphs (Figure 63 and Figure 64)[247].

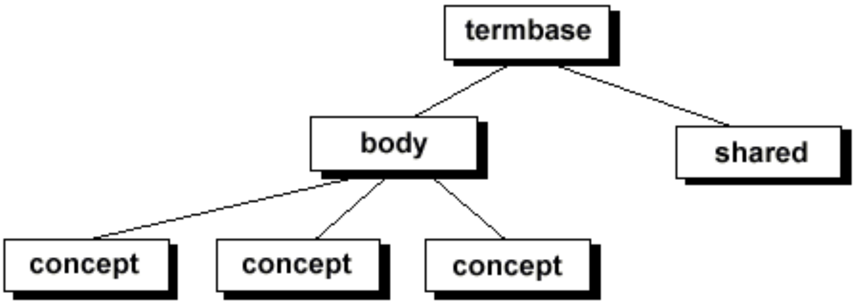


Figure 63: MARTIF termbase description

Each concept entry follows the structure in Figure 64. One thing not explicitly shown is the fact that there can be many language sections, each containing one ore more terms.

[244] ISO 12620 provides an inventory of types of data items, each type being called a "data category". A terminological concept entry (term entry, for short) is composed of data items (an item being a field, a cell, or an element, depending on the representation); each item is an instance of a data category, but 12620 does not specify the structure of a term entry, i.e., it does not specify the relationships among data items in an entry. The framework provides (1) an approach to structuring the items in a term entry in a manner consistent with current theory and practice in concept-oriented terminology and (2) provides a set of data categories taken from ISO 12620.

[245] Interestingly, as MARTIF was concept-centered rather than word-centered, it was not considered appropriate for NLP needs. [MARTIF] does not match the needs of non-concept-oriented approaches to terminology, i.e. lexicographic and NLP approaches, because MARTIF presupposes a concept orientation rather than a word orientation" http://www.creativyst.com/cgi-bin/M/Glos/st/Glossary.pl?TermList=M (Last visit: May 15th, 2012)

[246] The CLS (Concept-oriented with Links and Shared references) Framework is the result of a joint effort of the Brigham Young University Translation Research Group (BYU TRG) and the Kent State University Institute for Applied Linguistics (KSU IAL). : http://www.ttt.org/clsframe/overview.html (Last visit: May 15th, 2012)

[247] The images (Figure 63 and Figure 64) have been extracted from the ttt.org: http://www.ttt.org/clsframe/graphic.html (Last visit: May 15th, 2012)

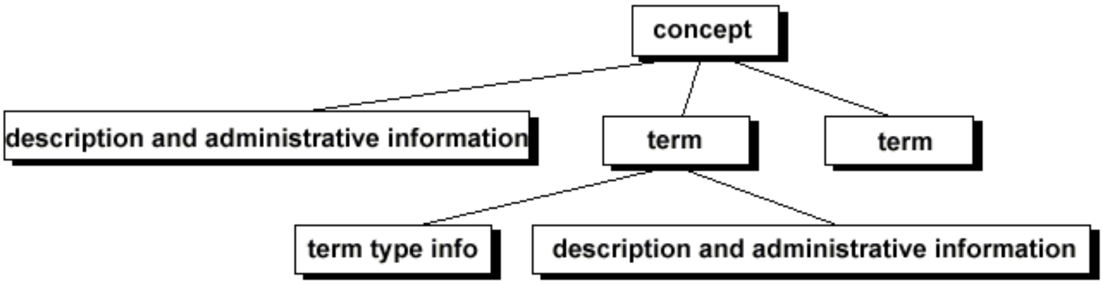

concept
description and administrative information
term
term
term type info
description and administrative information

Figure 64: MARTIF concept description

Each piece of contextual information has a language code associated with it. A link can be attached to any item; a link can be to another concept entry or to a reference. The MARTIF draft XML specification is still available online[248]. However, the development of MARTIF stopped in September 2000.

6.3.3.7 ISLE Meta Data Initiative (IMDI)

The ISLE Meta Data Initiative (IMDI) is a proposed metadata standard to describe multimedia and multi-modal language resources, and written language corpora. The initiative has been motivated by the desire to enable not only resource discovery of major resources but also resources within resources, and descriptions of resources. IMDI provides interoperability for browsable and searchable corpus structures and resource descriptions with the help of specific tools [IMDI I3.0.3. 2003][249]. IMDI foresees metadata transcriptions to only contain references to real language resources, accompanied by a structure to specify the access restrictions for these resources. It considers metadata transcriptions always free, where access to resources themselves may need restrictions. IMDI metadata descriptions are characterised by formal identification implying IMDI-particular structure, and IMDI elements involved. The elements required for error-free tool usage should be mandatory, a session name to distinguish between sessions within a corpus or sub-corpus seems sufficient. The IMDI website[250] hosts Metadata Elements for Session Descriptions, IMDI Metadata Elements for Catalogue Descriptions, Metadata Elements for Lexicon Descriptions, Vocabulary Taxonomy Structure, and Mapping IMDI Session Descriptions with OLAC. The documents were last

[248] Default XLT Format (DXLT): http://www.ttt.org/oscar/xlt/DXLTspecs.html (Last visit: May 15th, 2012)

[249] IMDI: http://www.mpi.nl/IMDI/ (Last visit: July 2nd, 2012)

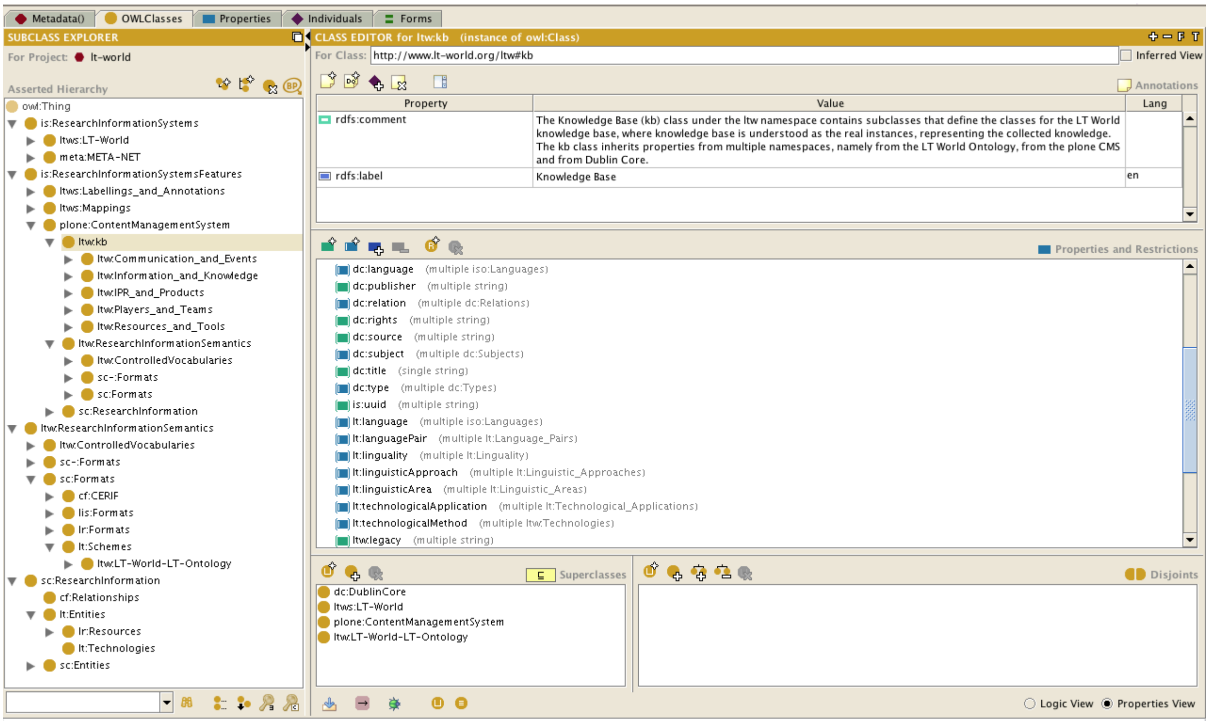
[250] ISLE Meta Data Initiative (IMDI): http://www.mpi.nl/imdi/ (Last visit: May 15th, 2012)

updated in 2003. In FERON, each of the IMDI Metadata Elements could be subsumed under *lt:KOS* as a *lt:functionalScheme*.

6.4 LT Infrastructure

A few examples of an LT infrastructures as perceived inline with FERON will now be presented. E.g., the *LT World* portal www.lt-world.org is perceived as a service and incorporated as a *lt:Service* under *cerif:Service* and thus under *cerif:Infrastructure*. The LT World Ontology always provided a holistic view over the LT domain. Although developed pragmatically it was aware of the Research context (Figure 56): "We also sketched a more generic ontology of research and technology in order to make sure that our ontology of LT and the resulting decisions for LT World can be compatible with the ontologies of the fields. On the other side, we integrated the ontology of LT with specialized ontologies for systems, projects, publications, etc." [Uszkoreit et al. 2003, p. 3] (see also [Jörg et al. 2010]). The LT World Ontology represents explicit properties with classes; it inherits LT properties from the LT Ontology *ltw:LT-World-LT-Ontology* (Figure 60 and Figure 65). FERON is generic at class level and reflects anticipated contextual and field functions through *skos:KOS* and *lt:KOS* via time-aware lawful *Relation* instances from generic *dcterms:relation* and LT *lt:relation* properties. The snapshots in Figure 65 and Figure 60 of the LT World portal and ontology reveal the LT World intension, granularity and types – it is truely a representation of "an information system as a thing" (see *3.1.1 bullet (2)*), and implemented as such [Burt and Jörg 2008].

FERON's complementing *skos:KOS* structure is similar in the LT World Ontology; which separates *ltw:ResearchInformationSemantics* from *is:ResearchInformationSystems* and thus distinguishes the information system features *is:ResearchInformationSystemFeatures* from its knowledge organisation *ltw:ResearchInformationSemantics* (Figure 65) and from the domain *sc:ResearchInformation*. Figure 65 describes the LT World portal (in FERON it is a *lt:service*) and gives insight into its contextual semantics, e.g. by *ltw:ControlledVocabularies*. Figure 66 shows the infrastructure class *cerif:Infrastructure* in FERON with designed sub-classes, where with lt:Classes an additional property *lt:relation* emerges.



For Class: http://www.lt-world.org/ltw#kb

```
owl:Thing
▼ is:ResearchInformationSystems
    ▶ ltws:LT-World
    ▶ meta:META-NET
▼ is:ResearchInformationSystemsFeatures
    ▶ ltws:Labellings_and_Annotations
    ▶ ltws:Mappings
    ▼ plone:ContentManagementSystem
        ▼ ltw:kb
            ▶ ltw:Communication_and_Events
            ▶ ltw:Information_and_Knowledge
            ▶ ltw:IPR_and_Products
            ▶ ltw:Players_and_Teams
            ▶ ltw:Resources_and_Tools
        ▼ ltw:ResearchInformationSemantics
            ▶ ltw:ControlledVocabularies
            ▶ sc-:Formats
            ▶ sc:Formats
        ▶ sc:ResearchInformation
▼ ltw:ResearchInformationSemantics
    ▶ ltw:ControlledVocabularies
    ▶ sc-:Formats
    ▼ sc:Formats
        ▶ cf:CERIF
        ▶ lis:Formats
        ▶ lr:Formats
        ▼ lt:Schemes
            ▶ ltw:LT-World-LT-Ontology
▼ sc:ResearchInformation
    cf:Relationships
    ▼ lt:Entities
        ▶ lr:Resources
        lt:Technologies
    ▶ sc:Entities
```

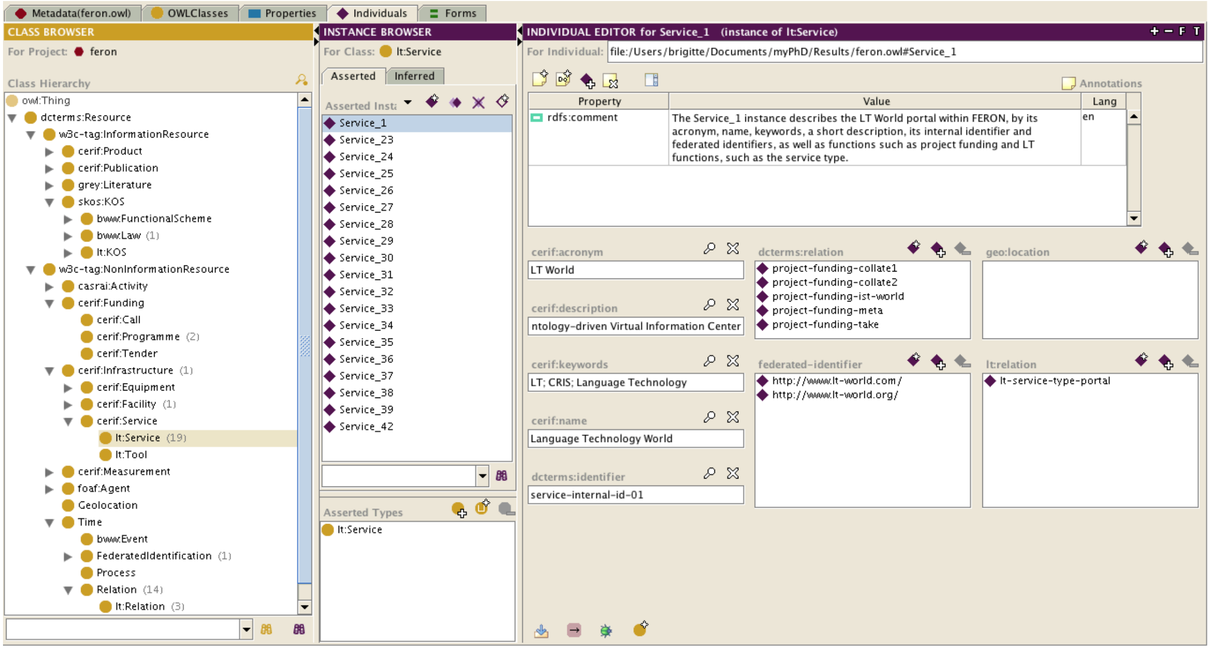
Property	Value	Lang
rdfs:comment	The Knowledge Base (kb) class under the ltw namespace contains subclasses that define the classes for the LT World knowledge base, where knowledge base is understood as the real instances, representing the collected knowledge. The kb class inherits properties from multiple namespaces, namely from the LT World Ontology, from the plone CMS and from Dublin Core.	
rdfs:label	Knowledge Base	en

```
dc:language   (multiple iso:Languages)
dc:publisher   (multiple string)
dc:relation   (multiple dc:Relations)
dc:rights   (multiple string)
dc:source   (multiple string)
dc:subject   (multiple dc:Subjects)
dc:title   (single string)
dc:type   (multiple dc:Types)
is:uuid   (multiple string)
lt:language   (multiple iso:Languages)
lt:languagePair   (multiple lt:Language_Pairs)
lt:linguality   (multiple lt:Linguality)
lt:linguisticApproach   (multiple lt:Linguistic_Approaches)
lt:linguisticArea   (multiple lt:Linguistic_Areas)
lt:technologicalApplication   (multiple lt:Technological_Applications)
lt:technologicalMethod   (multiple ltw:Technologies)
ltw:legacy   (multiple string)
```

Superclasses:

dc:DublinCore
ltws:LT-World
plone:ContentManagementSystem
ltw:LT-World-LT-Ontology

Figure 65: LT World ontology – Knowledge Base View



```
owl:Thing
▼ dcterms:Resource
    ▼ w3c-tag:InformationResource
        ▶ cerif:Product
        ▶ cerif:Publication
        ▶ grey:Literature
        ▼ skos:KOS
            ▶ bww:FunctionalScheme
            ▶ bww:Law (1)
            ▶ lt:KOS
    ▼ w3c-tag:NonInformationResource
        ▶ casrai:Activity
        ▼ cerif:Funding
            cerif:Call
            cerif:Programme (2)
            cerif:Tender
        ▼ cerif:Infrastructure (1)
            ▶ cerif:Equipment
            ▶ cerif:Facility (1)
            ▼ cerif:Service
                lt:Service (19)
                lt:Tool
        ▶ cerif:Measurement
        ▶ foaf:Agent
        Geolocation
        ▼ Time
            bww:Event
            ▶ FederatedIdentification (1)
            Process
            ▼ Relation (14)
                lt:Relation (3)
```

Service_1
Service_23
Service_24
Service_25
Service_26
Service_27
Service_28
Service_29
Service_30
Service_31
Service_32
Service_33
Service_34
Service_35
Service_36
Service_37
Service_38
Service_39
Service_42

Asserted Types: lt:Service

For Individual: file:/Users/brigitte/Documents/myPhD/Results/feron.owl#Service_1

Property	Value	Lang
rdfs:comment	The Service_1 instance describes the LT World portal within FERON, by its acronym, name, keywords, a short description, its internal identifier and federated identifiers, as well as functions such as project funding and LT functions, such as the service type.	en

cerif:acronym: LT World

cerif:description: ntology-driven Virtual Information Center

cerif:keywords: LT; CRIS; Language Technology

cerif:name: Language Technology World

dcterms:identifier: service-internal-id-01

dcterms:relation: project-funding-collate1, project-funding-collate2, project-funding-ist-world, project-funding-meta, project-funding-take

geo:location:

federated-identifier: http://www.lt-world.com/, http://www.lt-world.org/

lt:relation: lt-service-type-portal

Figure 66: FERON – cerif:Infrastructure View

FERON perceives the DFKI-hosted ACL Natural Language Software Registry (NLSR)[251], META-SHARE[252], the ELRA Universal Catalogue[253], the Linguistic Data Consortium

[251] Natural Language Software Registry (NLSR): http://registry.dfki.de (Last visit: June 4th, 2012)

[252] META-SHARE: http://www.meta-share.eu/ (Last visit: June 4th, 2012)

(LDC)[254], or the Language Grid[255] [e.g. Hayashi et al. 2007] portals as LT infrastructures, i.e. services *lt:Service*, which will not be further investigated in this work. In FERON, a *lt:Service* instance features the *lt:relation* to enable time-aware LT functions and inherits properties from the *cerif:Service* class such as *cerif:acronym*, *cerif:name*, *cerif:description*, *federated-identifier*, *geo:location* and *dcterms:identifier* and *dcterms:relation*. Here services are described as an LT infrastructure, section 6.3.3 Language Descriptions (LT KOS) introduces applicable knowledge organisation systems, e.g. for services or other LT entities or infrastructures.

6.5 LT Measurement

LT measurement is a section in this work and a class in FERON, because it is truely a field-specific area. However, where generic models of measurements have just started as to being formally developed within the Research domain (see section 5.2.2.4 Measurement) – a field specific measurement extension is assumed to be of a high granularity and complexity and requires deep domain knowledge; i.e. it is far beyond the scope of this work.

6.6 Summary

With chapter 6 Analysis of Language Technology Entities the field was briefly introduced and its entities and relationships investigated. For FERON, LT was a use-case to evaluate and validate its field extensibility. In FERON, field classes *lt:Classes* are subsumed under more generic Research classes without further sub-classes and feature a *lt:relation* property upon the range of a *lt:Relation* class inline with FERON's structure. Field-specific KOSs *lt:KOS* functions are enabled through time-aware lawful relationships *lt:Relation*. The structure of FERON thus allows for extensions in any field through incorporation of specific KOSs and their time-aware functional applications in relationships. That is, FERON scales with the collection and incorporation of multiple available (including field) descriptions *skos:KOS*.

[253] European Language Resources Association (ELRA) Universal Catalogue: http://universal.elra.info/ (Last visit: June 4th, 2012)

[254] Linguistic Data Consortium (LDC) supports language-related education, research and technology development by creating and sharing linguistic resources: data tools and standards: http://www.ldc.upenn.edu/ (Last visit: June 4th, 2012)

[255] Language Grid: http://langrid.org/en/index.html (Last visit: June 4th, 2012)

Because FERON is at first domain-agnostic, it only indicates (LT) sub-areas for field-extensions in *lt:Classes* with an emerging property *lt:relation*.

An awareness of subsequent system implementation however, reminds of the requirements with explicit standards or formats and the support from constraints or rules to manage applications and interfaces, which will be elaborated in section 7.8 Field Extensibility.

7 FERON – *F*ield-extensible *R*esearch *ON*tology

"On the Shoulders of Giants"

[Bernard of Chartres, 12th Century]

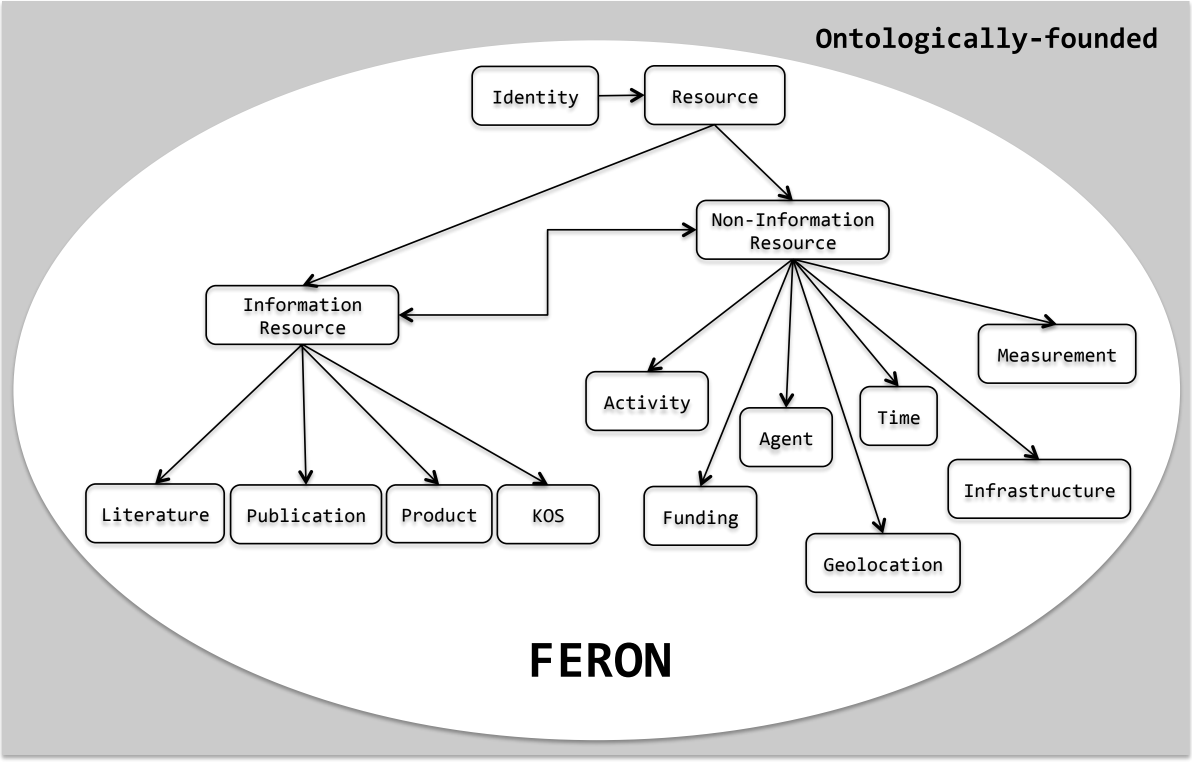


Figure 67: Field-extensible Research Ontology – A Top Level View

According to Bunge the world is made up of *substantial things*. With chapters 5 and 6 those *things* perceived substantial in Research and the Language Technology field were thoroughly analysed – both investigations anticipated FERON (Figure 29). With this chapter, the final FERON model (Figure 67)[256] is presented by summarising the analysis and design process and by elaborating on its ontological commitments, language, namespaces, field extensibility, time-awarenes in relationships, geographic location and furthermore constraints and extended applicability.

[256] FERON was formally modelled with Protégé and is available in OWL as *feron.owl*. However, it is not exhaustively designed but only prototypically populated and labelled. Its classes and properties are enriched by annotation properties *rdfs:label*, *rdfs:isDefinedBy*, *rdfs:comment* as supplied with RDF.

7.1 Analysis and Design

FERON was at first aimed at human readability, and may be extended for machine reading and processing with concrete applications. The modeling of FERON followed a hybrid approach and was guided by ontological foundations – the Bunge-Wand-Weber Ontology. At first, FERON was therefore developed intensionally in a *top-down* manner (Figure 29) before it was subsequently refined *bottom-up* through an analysis of openly available descriptions from recognised authorities and standards bodies. The result is as presented with Figure 67. The *canonical* domain model FERON – as initially developed (Figure 29) – guided the entire entity analysis process and supported with comparison of the investigated descriptions. The modelling approach is explicitly explained with Figure 68.

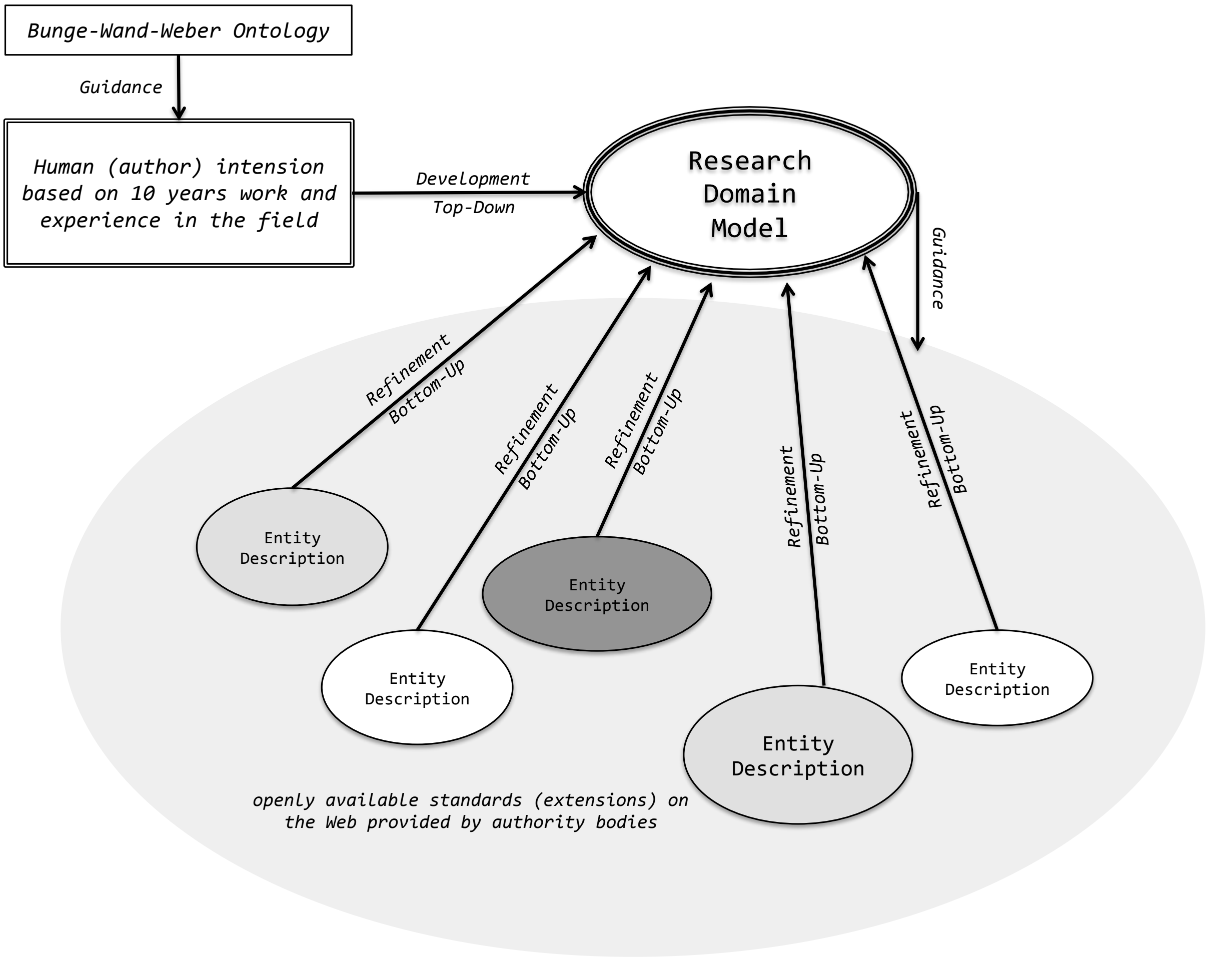
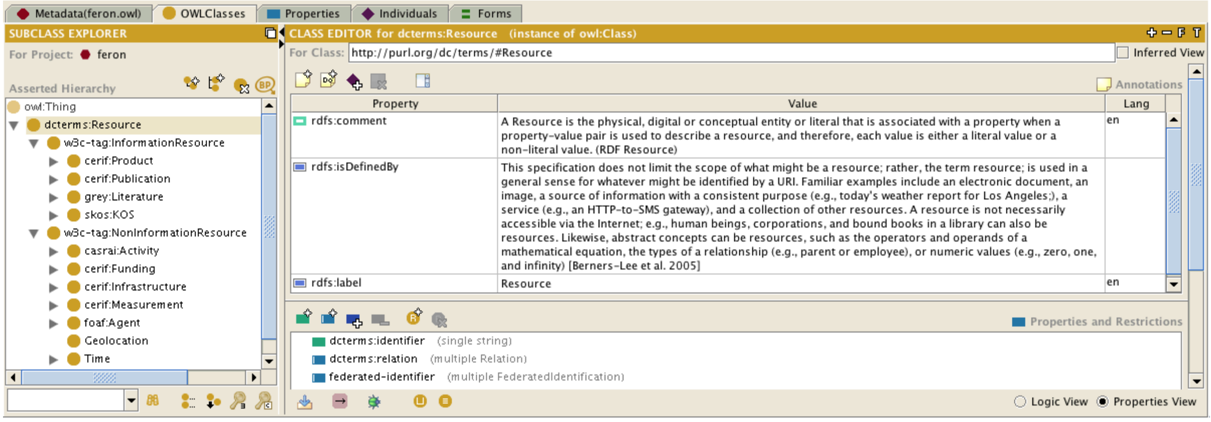


Figure 68: Hybrid Modelling Approach: First Top-Down – Second Bottom-Up

The integration of Language Technology entities demonstrated the field-extensibility. The top-down anticipation of FERON (see also Figure 29, Research Entities embedded in Society) was based on a more than ten years experience of the author (see Motivation), and took into account ontological foundations from the start (partly these are inherent in technologies and tools). During investigations of the two domains, FERON was *faced* with available formats and models (*extensions*), and changes were then adopted where necessary. Figure 67, shows FERON – *F*ield-*e*xtensible *R*esearch *ON*tology at top level; *Research* is perceived as to being constituted of identifiable *Resources* – either *Information* or *Non-Information* resources. Where the former distinguishes *Literature*, *Publication*, *Product*, and *Knowledge Organisation System (KOS)*, the latter recognises *Activity*, *Funding*, *Agent*, *Time*, *Geolocation*, *Infrastructure*, and *Measurement*. This top level view of FERON is formally represented as shown in Figure 69, within Protégé, where the *dcterms:Resource* class features the intrinsic property *dcterms:identifier*, and mutual properties such as *dcterms:relation*, and *federated-identifier*. The *owl:Thing* class is supported by Protégé and provided through OWL. The namespace prefixes indicate the origin of entities at class or property levels.



For Project: feron

- owl:Thing
  - dcterms:Resource
    - w3c-tag:InformationResource
      - cerif:Product
      - cerif:Publication
      - grey:Literature
      - skos:KOS
    - w3c-tag:NonInformationResource
      - casrai:Activity
      - cerif:Funding
      - cerif:Infrastructure
      - cerif:Measurement
      - foaf:Agent
      - Geolocation
      - Time

For Class: http://purl.org/dc/terms/#Resource

□ Inferred View

Property	Value	Lang
rdfs:comment	A Resource is the physical, digital or conceptual entity or literal that is associated with a property when a property-value pair is used to describe a resource, and therefore, each value is either a literal value or a non-literal value. (RDF Resource)	en
rdfs:isDefinedBy	This specification does not limit the scope of what might be a resource; rather, the term resource; is used in a general sense for whatever might be identified by a URI. Familiar examples include an electronic document, an image, a source of information with a consistent purpose (e.g., today's weather report for Los Angeles;), a service (e.g., an HTTP-to-SMS gateway), and a collection of other resources. A resource is not necessarily accessible via the Internet; e.g., human beings, corporations, and bound books in a library can also be resources. Likewise, abstract concepts can be resources, such as the operators and operands of a mathematical equation, the types of a relationship (e.g., parent or employee), or numeric values (e.g., zero, one, and infinity) [Berners-Lee et al. 2005]	
rdfs:label	Resource	en

dcterms:identifier (single string)
dcterms:relation (multiple Relation)
federated-identifier (multiple FederatedIdentification)

○ Logic View ● Properties View

Figure 69: Field-extensible Research Ontology – dcterms:Resource view in Protégé

A visualisation of FERON within Protégé is possible through OntoGraf (Figure 70 presents the graph without namespaces). A *Resource* is thus a *Thing* composed of two kinds – namely *InformationResource* and *Non-InformationResource*, and where *InformationResource* is associated with *Non-InformationResource* through *Relation* – a sub-class of *Time* to feature time properties in addition to a *function* property upon the range of *KOS* and two connecting properties *links-to* and *is-linked-by* upon the range of *Resource*. The distinguished types below *InformationResource* and *Non-InformationResource* are as shown in Figure 67, where *Information* features an intrinsic object-property *language* in the range of the *Language* class.

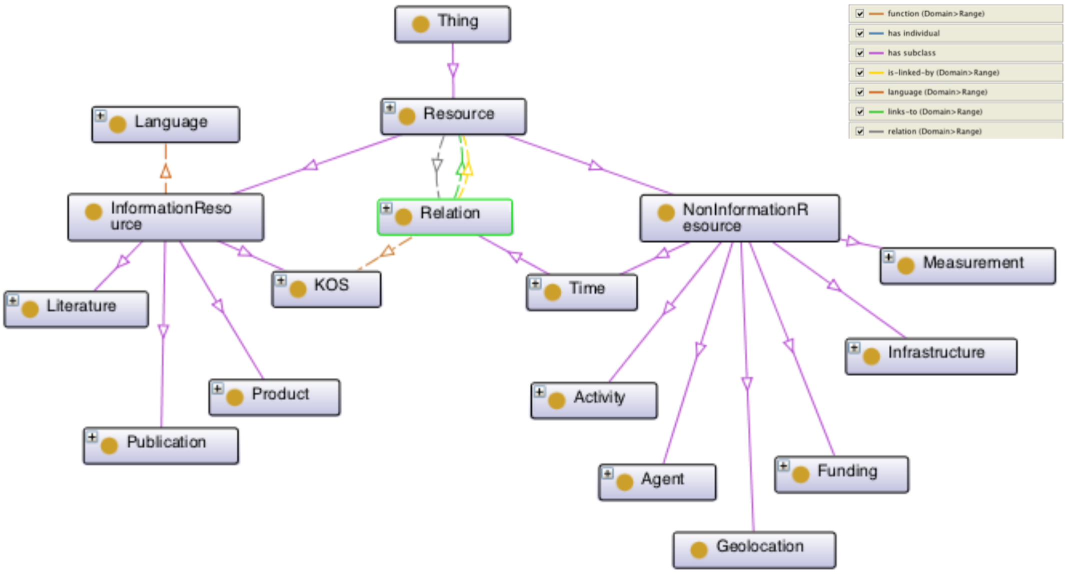


Figure 70: Field-extensible Research Ontology – indicating properties (visualised with OntoGraf)

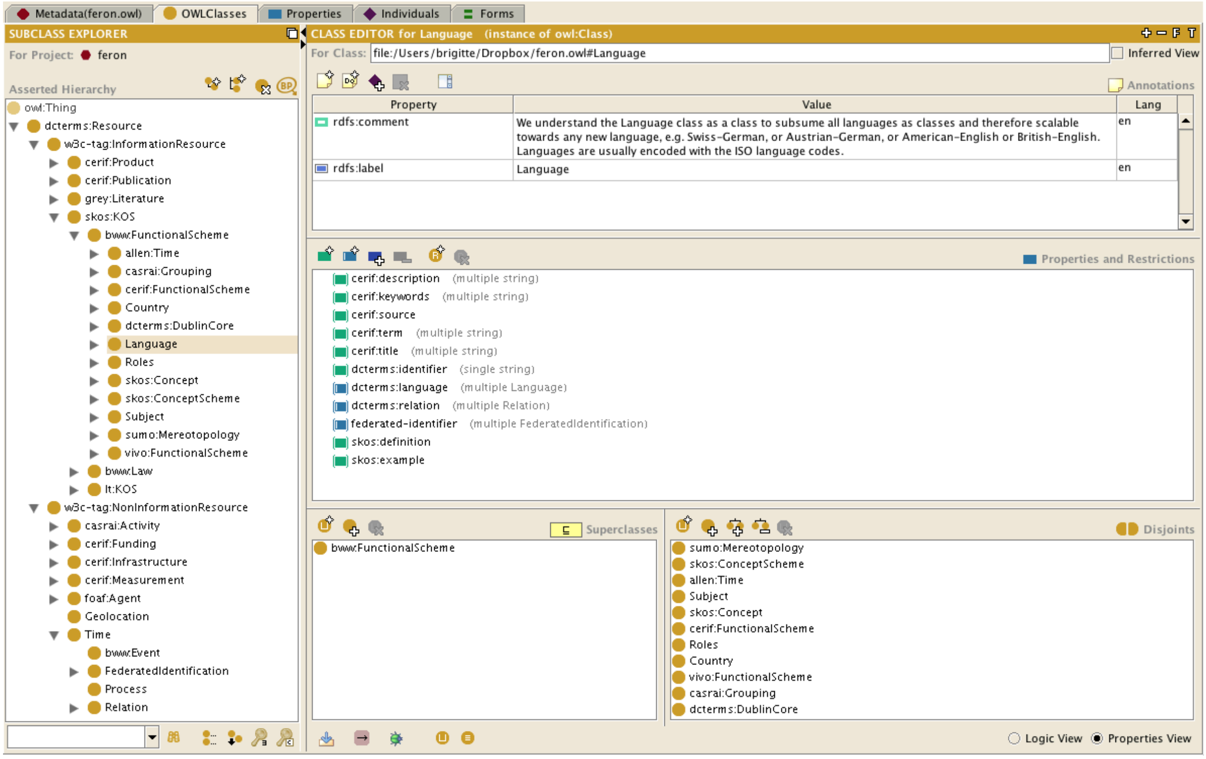
It is a paradigm in the Linked Open Data world to re-use existing elements from well-known and globally accepted ontologies and vocabularies. For FERON, those that were perceived most meaningful were selected and then incorporated. FERON was not aimed at a mapping of available formats but at a human-readable formal Research domain description – open for field extensions – by employing existing descriptions to a maximum extent. The intensional top-down approach with incorporated amendments from investigated extensions (bottom-up) implies FERON to being hybridly designed, inline with the proposed architectures and styles in sections 4.2 and 4.3 and under the paradigm that form follows function.

7.2 Ontological Commitments

According to [Gruber 1993, p. 909], ontological commitments should be minimal. This is considered particularly true with designing open worlds – and this is what FERON is supposed to be. Ontological commitments in FERON are mostly reflected through the *Relation* construct. FERON's domain entities are designed rather flat and not in very deep hierarchies, i.e. a minimal number of abstract classes (abstract classes do not feature own properties). An exception is the KOS class, where it is anticipated to allow concepts being imported from e.g. hierarchies or inline with their originating structural embeddings. FERON features only a small number of *functional*[257] domain properties but allows for multiple unary, binary and even n-ary *Relation functions* and is highly scalable under *skos:KOS*.

7.3 Language

Language is perceived to be inherent in information resources and FERON therefore features an intrinsic language property *dcterms:language* upon the range of the *Language* class with *w3c-tag:InformationResource*. As a feature, *language* is thus propagated to all sub-classes. Where language is clearly an intrinsic property of information sources, it is obviously also a functional feature for many fields and areas, not least Language Technology. Language as a function is recognised e.g. with interfaces and with any formal objects description – and not only in Language Technology – but also in fields such as Medicine, Physics, Biology etc.



For Class: file:/Users/brigitte/Dropbox/feron.owl#Language

Property	Value	Lang
rdfs:comment	We understand the Language class as a class to subsume all languages as classes and therefore scalable towards any new language, e.g. Swiss-German, or Austrian-German, or American-English or British-English. Languages are usually encoded with the ISO language codes.	en
rdfs:label	Language	en

```
owl:Thing
▼ dcterms:Resource
  ▼ w3c-tag:InformationResource
    ▶ cerif:Product
    ▶ cerif:Publication
    ▶ grey:Literature
    ▼ skos:KOS
      ▼ bww:FunctionalScheme
        ▶ allen:Time
        ▶ casrai:Grouping
        ▶ cerif:FunctionalScheme
        ▶ Country
        ▶ dcterms:DublinCore
        ▶ Language
        ▶ Roles
        ▶ skos:Concept
        ▶ skos:ConceptScheme
        ▶ Subject
        ▶ sumo:Mereotopology
        ▶ vivo:FunctionalScheme
      ▶ bww:Law
      ▶ lt:KOS
  ▼ w3c-tag:NonInformationResource
    ▶ casrai:Activity
    ▶ cerif:Funding
    ▶ cerif:Infrastructure
    ▶ cerif:Measurement
    ▶ foaf:Agent
      Geolocation
    ▼ Time
        bww:Event
      ▶ FederatedIdentification
        Process
      ▶ Relation
```

```
cerif:description   (multiple string)
cerif:keywords   (multiple string)
cerif:source
cerif:term   (multiple string)
cerif:title   (multiple string)
dcterms:identifier   (single string)
dcterms:language   (multiple Language)
dcterms:relation   (multiple Relation)
federated-identifier   (multiple FederatedIdentification)
skos:definition
skos:example
```

Superclasses:

```
bww:FunctionalScheme
```

Disjoints:

```
sumo:Mereotopology
skos:ConceptScheme
allen:Time
Subject
skos:Concept
cerif:FunctionalScheme
Roles
Country
vivo:FunctionalScheme
casrai:Grouping
dcterms:DublinCore
```

Figure 71: FERON's Language class viewed from within Protégé

To reflect the various functional applications possible through language, the *Language* class is subsumed under *bww:FunctionalScheme* (Figure 71) as a sub-class of *skos:KOS* and in that

[257] A functional property is a property that can have only one (unique) value y for each instance x [...] Both object properties and datatype properties can be declared as "functional" http://www.w3.org/TR/owl-ref/ (Last visit: May 1st, 2012)

accounts for its applicability from within any FERON relationship. Through *Relation* and *lt:Relation* classes with *bww:function* defined upon the range of *skos:KOS* and *lt:KOS*, FERON allows for any time-aware functional language information with information sources and with non-information resources.

Language as a class or concept thus inherits properties such as *cerif:term* applicable with the language name in natural language, and further properties applicable in *skos:KOS* such as *cerif:title*, *cerif:source*, *cerif:keywords*, *skos:example*, and *skos:definition*.

7.4 Namespaces

The notion of namespaces was introduced in section 3.3.3 Naming Conventions or Standards. Namespaces are grounded in XML and supported by Protégé. For FERON, the entity descriptions from multiple sources were investigated and prototypically incorporated. The following list gives an overview of the selected formats or models incorporated in FERON. Where there was no formal namespace available, FERON designed its own internal representation:

- `allen: http://www.James-F-Allen.Time.net#`
- `bibo: http://purl.org/ontology/bibo/#`
- `bww: http://www.bunge-wand-weber-ontology.virt/bww#`
- `casrai: http://www.casrai.org/1.1.0#`
- `cerif: http://www.eurocris.org/cerif1.4#`
- `clarin: http://www.clarin-project.eu/#`
- `cyc: http://sw.opencyc.org/#`
- `datacite: http://schema.datacite.org/2.2/#`
- `dcterms: http://purl.org/dc/terms/#`
- `elra: http://catalog.elra.info/retd/#`
- `event: http://purl.org/NET/c4dm/event.owl#`
- `foaf: http://xmlns.com/foaf/0.1/#`
- `frameNet: https://framenet.icsi.berkeley.edu#`
- `geo: http://aims.fao.org/aos/geopolitical.owl#`
- `grey: http://www.greynet.org/#`
- `ieee-lom: http://www.ieee-lom.virt/#`
- `isocat: http://www.isocat.org/#`
- `lt: http://www.lt-world.org/#`

- lt-world: http://www.lt-world.org/ontology#
- nlsr: http://registry.dfki.de/#
- olac: http://www.language-archives.org#
- orcid: http://www.orcid.org/#
- owl: http://www.w3.org/2002/07/owl#
- protege: http://protege.stanford.edu/plugins/owl/protege#
- rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#
- rdfs: http://www.w3.org/2000/01/rdf-schema#
- skos: http://www.w3.org/2008/05/skos#
- sumo: http://www.ontologyportal.org/#
- tei: http://www.tei-c.org#
- unisist: http://www.unesco.org/#
- vivo: http://vivoweb.org/ongology/core#
- voa3r: http://voa3r.eu#
- w3c-tag: http://w3.org/tag#

Namespaces enable the disambiguation of concepts by indicating their originating source, and are applicable in class names, property names or instance names.

7.5 Identities and Entities

FERON does not collect duplicates but only incorporates entities perceived most meaningful from available descriptions; it avoids equal or same-as statements. This design decision is owned to the fact that FERON is not aimed at a mapping of available descriptions but at first dedicated to human domain understanding. To support this goal, the complexity of the domain was reduced by identifying substantial things, these were then incorporated into the anticipated conceptual structure of FERON – following the LOD paradigm in that matching available descriptions were re-used – and changes adopted where necessary.

FERON does not employ sub-classes with field-extensible classes, but envisions scalable class sub-typing under *bww:FunctionalScheme* below *skos:KOS* only.

Research Information is often managed from within heterogeneous systems (see section 4.1 Information System Kinds), but for information integration and exchange the preservation of entities and their identities is crucially important (see 3.3.4 Entities and Identification). In addition to URIs – the default identifying mechanism in the Semantic and Linked Open Data Web, and supported by Protégé through RDF/OWL as a means for unique resource identification (i.e. the name of the class, property or instance) – FERON features a *federated-identifier* property emergent with the *dcterms:Resource* class to range over a *FederatedIdentification* class, and to allow for preservation of multiple identifiers in addition to URIs (it features the *dcterms:identifier* property for the preservation of system-internal (closed-world) identifiers). The *federated identifier* supports open-world assumptions in that it anticipates interlinkage with any kind of resource. It is considered critically important with respect to the setting up of new kinds of Research Information Systems [Jörg et al. 2012c].

7.6 Classes

FERON applies classes and properties for domain description, and separates them from instances, which are perceived as data. It conceptually distinguishes three kinds of classes: (1) classes identifying *substantial Research* domain *things;* (2) classes describing *functions*, i.e. organised in functional schemes; (3) field-extensible classes.

FERON describes Research through identified entities in a class hierarchy (Figure 72) with *Research* domain sub-classes, e.g. under *OrganisationUnit*, *Event*, etc. The *Research* classes feature intrinsic properties and inherit the semantically-neutral *dcterms:relation* property (chapter 5 Analysis of Research Entities) upon the range of *Relation*.

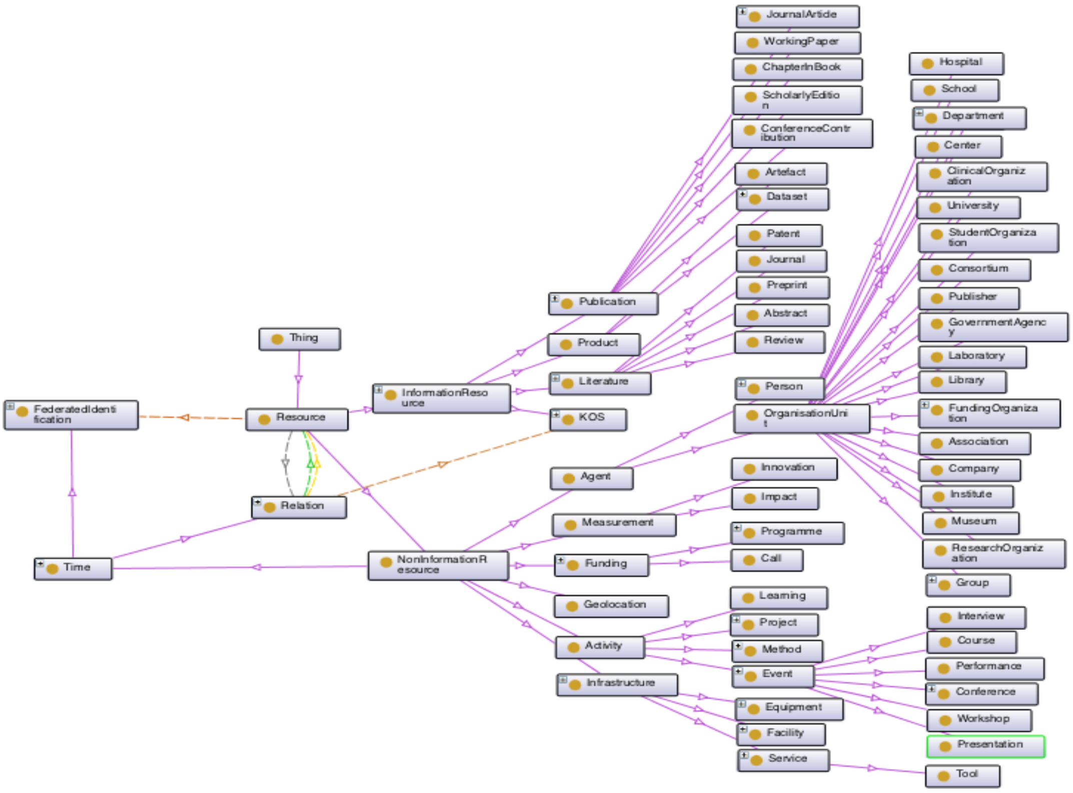


Figure 72: FERON – Selected Domain Classes (visualised with OntoGraf)

In FERON, Research domain classes are conceptually distinguished from functional scheme classes and from field-extensible classes – but such a distinction is not possible in the visualisation with OntoGraf.

7.7 Functional Schemes

In FERON, functional schemes are subsumed under KOS, i.e. they are information resources. Besides the *FunctionalScheme* class there is a *Law* class to account for the lawfulness of functions in the spirit of Bunge. The KOS class is meant to be open and thus scales with respect to knowledge organisation system imports of any kind. [Evermann & Wand 2001, p. 356-357] define a functional scheme according to Bunge: “A set of attributes used to describe a set of things with common properties is called a *functional schema*. Depending on which aspects one is interested in, there can be different schemas describing the same thing. The *state* of a thing is a *complete* assignment of values to all state functions in the functional schema.” In FERON, the *FunctionalScheme* class currently incorporates values from e.g. Allen Time (before, during, equal, finishes, meets, overlaps, starts), CASRAI Grouping (AbridgedCV, ResearchActivity, PersonnelProfile, StudentCV), SKOS Concept (broader, inScheme, Member, narrower, etc), SKOS ConceptScheme (Collection, OrderedCollection).

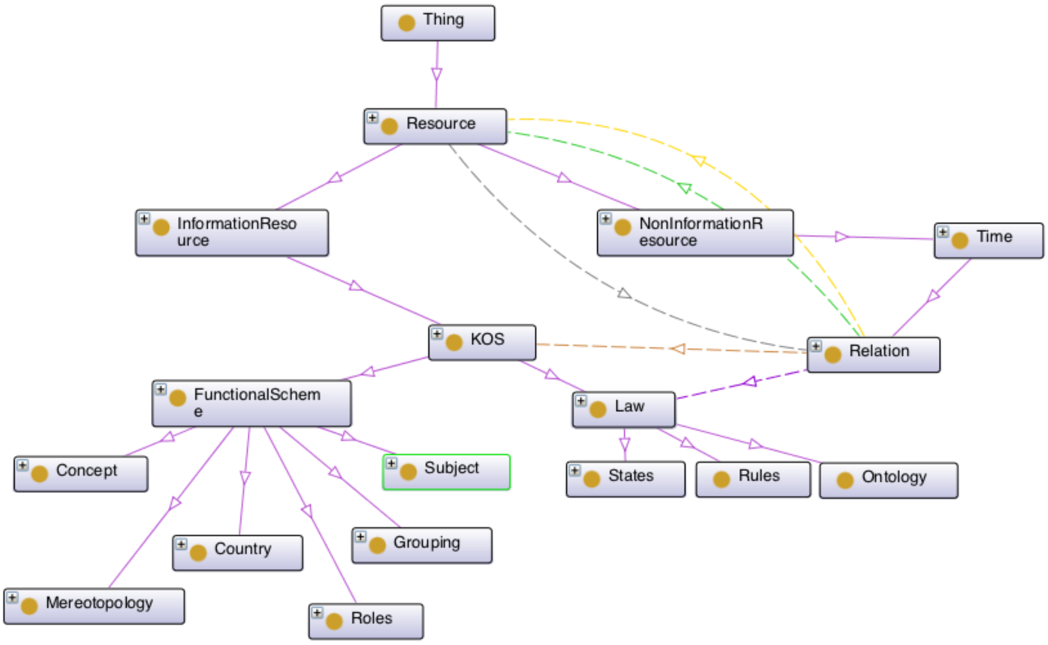


Figure 73: FERON - Functional Scheme Classes (visualised with OntoGraf)

Under *KOS* within *bww:FunctionalScheme*, FERON allows for fully-connected graph representations. The *Law* class is aimed at incorporating applicable descriptions of e.g. states, rules or ontological constraints (these may include fixpoint definitions), to ensure lawful-ness with applications.

7.8 Field Extensibility

FERON features field-extensions mostly at class level and at property level only employs one single object property *lt:relation*. With Figure 34, field-extensible classes were identified in light grey. With Figure 57 concrete LT examples were included to demonstrate how explicit field-extensions are perceived to happen – with *non-information* and *information resources*:

- **Activity**: e.g. project, method, event, learning
- **Product**: e.g. dataset, database, recording, corpus
- **Method:** e.g. machine translation, information exraction
- **Infrastructure**: e.g. equipment, facility, service, tool
- **KOS**: e.g. language specification, ontology
- **Measurement**: e.g. lt measurement

FERON subsumes field classes below field-extensible classes. E.g. *lt:Project* under *cerif:Project*, *lt:Data* under *vivo:Dataset*, *lt:Method* under *cyc:Method*, *lt:Equipment* under *cerif:Equipment*, *lt:Facility* under *cerif:Facility*; *lt:Service* under *cerif:Service*, *lt:KOS* under *skos:KOS*, *lt:Relation* under *Relation*. The generic relationship construct is adopted for field relationships; i.e. the *lt:relation* property emerges at LT classes upon the range of *lt:Relation* to account for multiple time-aware field functions – functional values are supplied through controlled (functional scheme) vocabularies – upon the range of *lt:KOS* as a subclass of *skos:KOS*. The *lt:relation* object property is the only field property in FERON, and time-aware functions are entirely managed through the *lt:Relation* class construct by employing the pre-defined functional scheme values.

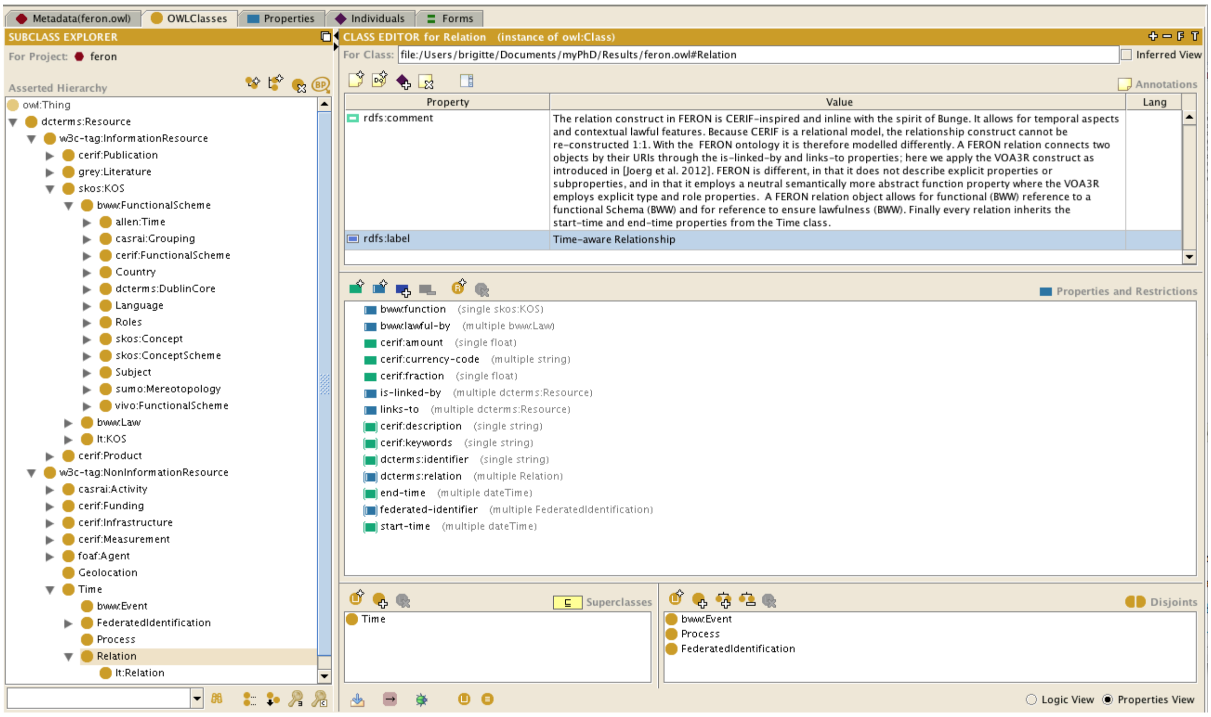
7.9 Properties

As indicated, FERON does not explicitly model functional properties but mostly manages relationship semantics through functional scheme class values. Because FERON is at first aimed at the human understanding of *Research* – generically and without a particular view – the functional schemes are more perceived as to being situational or contextual applicable and thus relevant only within particular use-cases or implementations[258]. The reflection of semantics through subsumption classes is processable by common reasoning engines and accounts for the required scalability and extensibility of the ontology and its technological applicability. Another reason for the preference of functional scheme classes over properties in FERON is owed to the need of time-awareness with relationships.

7.10 Time-aware Relationships

In FERON time is not modeled as a class under which all others are subsumed anticipating all things inherit, i.e. relate to time, but it is perceived, that temporally-changing relevant aspects are inherent mostly in relationships.

[258] In e.g. [Jörg et al. 2012c] the aim was to publish data for usage by clients with limited reasoning capabilities following the *Materialize Inferences* pattern [Heath & Bizer 2011] in exposing both, the original and the inferred triples as linked data, e.g. hosting properties and sub-properties of different vocabularies.


Metadata(feron.owl)
OWLClasses
Properties
Individuals
Forms
SUBCLASS EXPLORER
For Project: feron
Asserted Hierarchy
owl:Thing
dcterms:Resource
w3c-tag:InformationResource
cerif:Publication
grey:Literature
skos:KOS
bww:FunctionalScheme
allen:Time
casrai:Grouping
cerif:FunctionalScheme
Country
dcterms:DublinCore
Language
Roles
skos:Concept
skos:ConceptScheme
Subject
sumo:Mereotopology
vivo:FunctionalScheme
bww:Law
lt:KOS
cerif:Product
w3c-tag:NonInformationResource
casrai:Activity
cerif:Funding
cerif:Infrastructure
cerif:Measurement
foaf:Agent
Geolocation
Time
bww:Event
FederatedIdentification
Process
Relation
lt:Relation
CLASS EDITOR for Relation (instance of owl:Class)
For Class: file:/Users/brigitte/Documents/myPhD/Results/feron.owl#Relation
Inferred View
Annotations
Properties and Restrictions
bww:function (single skos:KOS)
bww:lawful-by (multiple bww:Law)
cerif:amount (single float)
cerif:currency-code (multiple string)
cerif:fraction (single float)
is-linked-by (multiple dcterms:Resource)
links-to (multiple dcterms:Resource)
cerif:description (single string)
cerif:keywords (single string)
dcterms:identifier (single string)
dcterms:relation (multiple Relation)
end-time (multiple dateTime)
federated-identifier (multiple FederatedIdentification)
start-time (multiple dateTime)
⊑ Superclasses
Time
Disjoints
bww:Event
Process
FederatedIdentification
○ Logic View ● Properties View

Property	Value	Lang
rdfs:comment	The relation construct in FERON is CERIF-inspired and inline with the spirit of Bunge. It allows for temporal aspects and contextual lawful features. Because CERIF is a relational model, the relationship construct cannot be re-constructed 1:1. With the FERON ontology it is therefore modelled differently. A FERON relation connects two objects by their URIs through the is-linked-by and links-to properties; here we apply the VOA3R construct as introduced in [Joerg et al. 2012]. FERON is different, in that it does not describe explicit properties or subproperties, and in that it employs a neutral semantically more abstract function property where the VOA3R employs explicit type and role properties. A FERON relation object allows for functional (BWW) reference to a functional Schema (BWW) and for reference to ensure lawfulness (BWW). Finally every relation inherits the start-time and end-time properties from the Time class.	
rdfs:label	Time-aware Relationship	

Figure 74: FERON Relationship Class modeled in Protégé

In FERON, the time-aware *Relation* class (Figure 74) reflects the entire dynamics[259]. As a sub-class of *Time* it inherits two functional intrinsic datatype properties – *starttime* and *endtime* – and upon these, instantiated relationship records converge to *state functions* (conceptually resembling mutual properties). The *Relation* class inherits *dcterms:identifier, dcterms:relation* and *federated-identifier* properties and is as such a *Resource* on its own. This view and construct is inline with the definition of *Resource* by [Berners-Lee et al. 2005] where types of relationships can be perceived as resources. *Relation* hence features recursion through the *dcterms:relation* property inherited from *dcterms:Resource* for inter-*Relation* linkage. Formally, a *Relation* instance can thus be linked to any FERON *dcterms:Resource* instance.

The relationship construct in FERON was inspired by CERIF [Jörg et al. 2012] and VOA3R – e.g. [Jörg et al. 2012c] – but goes beyond the two with respect to openness. It currently allows for n-ary relationships (see Figure 74) through the properties *links-to* and *is-linked-by*; they are not defined as functional, i.e. can have multiple values[260]. It is recommended to use *Relation* with binary relationships (in FERON unary relationships are perceived as types, i.e.

[259] Date properties such as publication date or registration date are not considered as dynamic.

[260] If applied n-ary, then the corresponding FERON *bww:function* needs to be n-ary too. Within FERON, some examples were instantiated, where n-arity is currently indicated through the instance's URI name and therefore imposed to subsumption classes (functions) under a semantic frame-like functional scheme *bww:functionalScheme*, where *n* contexts can be subsumed under the *Roles* class.

sub-classes, and currently there are no means in FERON to guide n-ary constraints except through naming). In addition to the above properties, *Relation* features a semantically-neutral *bww:function* property upon the range of the *skos:KOS* class. The preference for the *Relation* class over explicitly modeled mutual properties is owed to the need for time-awareness and to the dynamics in relationship qualities[261]. The *Relation* construct is justified, because it is semantically neutral and consistently applied throughout FERON, and therefore managable (scalable). Each *Relation* instance maintains its own intrinsic *dcterms:identifier*.

The *Relation* class allows for inferencing over timepoint or time interval values from instantiated functions (called facts in YAGO2), and updates according to rules, as envisioned in [Hoffart et al. 2010, p. 15] "The principle for handling these situations is to use rules that propagate the begin or end of an entity's existence time to the occurrence time of a fact, where the entity occurs as a subject or object." With FERON, the subject and object are available through the *links-to* or *is-linked-by* identifier values, and e.g. during population with instances, the time-types as identified by [Allen 1983] and condensed by [Correndo 2010] (see again 3.3.6 Temporal Aspects) may be automatically deduced or inferred from inherent time values, according to functional (e.g. controlled vocbabulary) and lawful time or rule schemes. In FERON, a time scheme is a functional scheme and hence classified as *skos:KOS* with sub-classes such as: *before*, *equal*, *meets*, *overlaps*, *during*, *starts*, *finishes* [Correndo et al. 2010, p. 3, fig 1].

It is anticipated, that most queries are aimed at relationship retrieval and the time-aware functional values are not only supportive with recall and precision but equally important for data and thus system quality [Hoffart et al. 2010, pp. 4 ff.].

[261] N-ary relationships (http://www.w3.org/TR/swbp-n-aryRelations/) versus Qualified Relation Pattern: The N-ary Relation pattern is similar to the Qualified Relation pattern as both involve the same basic solution: modelling a relationship as a resource rather than a property. They differ in their context. In the Qualified Relation pattern the desire is to annotate a relationship between two resources, whereas in the N-ary Relation pattern the goal is to represent a complex relation between several resources. http://patterns.dataincubator.org/book/nary-relation.html (Last visit: June 4th, 2012)

The main reason for not modeling relationships as classes is, "that it causes explosion in the number of terms in a vocabulary, e.g. each predicate is replaced with two predicates and a class. A vocabulary can quickly become unwieldy, so the value of the extra modelling structure needs to be justified with clear requirements for needing the extra complexity." http://patterns.dataincubator.org/book/qualified-relation.html (Last visit: June 3rd, 2012)

7.11 Geographic Location

FERON follows the VIVO approach and employs a *Geolocation* class, which is not further specified with properties. AIMS defined a geopolitical ontology[262], a public source for geographical names is Geonames[263]. Furthermore, the Food and Agriculture Organization of the United Nations (FAO) presents country profiles according to the AIMS geopolitical ontology, which itself employs known related formats and standards (see public website[264]). In FERON, *cerif:Organisation*, *Artefact* and *event:Event* feature the *geo:location* property upon the range of *Geolocation*. This overlaps with YAGO2 location assignments to *Groups*, *Artifacts* and *Events* [Hoffart et al. 2010].

[262] The Agricultural Information Management Standards (AIMS) developed a geopolitical ontology http://aims.fao.org/aos/geopolitical.owl (Last visit: June 4th, 2012)

[263] The GeoNames geographical database covers all countries and contains over eight million placenames that are available for download free of charge: http://www.geonames.org/ (Last visit: June 4th, 2012)

[264] http://www.fao.org/countryprofiles/geoinfo/en/ The use of the information presented on the FAO Country Profiles portal is governed by FAO's copyright reservation. Any queries regarding the content, sources or use, please contact FAO-country-profiles@fao.org or visit http://www.fao.org/countryprofiles. (Last visit: June 21st, 2012)

8 Conclusion

FERON was developed following basic notions, with a knowledge of conceptual modeling and the awareness of architectural styles and information systems structures in support of functional needs. FERON is modelled as a frame, formally described in RDF/OWL and allows for a fully-connected graph representation of the underlying information. The frame version was chosen to improve its human readability. FERON was not aimed at a mapping of investigated formats and models, but at a field-extensible formal description of the *Research* world, as perceived through an analysis of its constituting *substantial* domain entities. The method to design FERON is hybrid in that it was at first modeled *top-down* and subsequently fine-tuned through an analysis of openly available formats and descriptions *bottom-up*. To indicate the origin or source of concepts FERON employs namespaces. FERON accounts for closed-world systems as well as for the open-world assumption even beyond the Semantic Web through federated identifiers. An intrinsic language property emerges in information resources, whereas functional language information is featured in *Relations* by reference to the Language class. Non-information resources are linked with information resources through the same mentioned *Relation* construct emerging at *dcterms:Resource*. FERON conceptually distinguishes three kinds of classes:

- Research classes to constitute the domain
- Functional Scheme classes to constitute domain functions
- Field-extensible Research classes to subsume field classes

FERON only employs one single field-specific (LT) property – namely *lt:relation* to allow for LT functions. It commits to a *Relation* class construct for all functional features and in that prefers a functional classes biased model over explicitly modeled functional properties – for reasons of openness, scalability, time-awareness and data quality. Through its consistency with this appraoch it copes with the imposed complexity.

FERON is designed at the highest perceived level of generality to account and allow for the required extensibility with fields. It is at first addressed at the human reader although also available machine-readable, and in that it also supports the GDI.

With applications of FERON e.g. at subsequent architectural (system) levels towards implementations, constraints or rules need to be defined. This is necessary for a provision of fix-points (e.g. *skos:hasTopConcept*) – and requires particular investigations of and decisions over use cases, standards support, user needs or system application. Because FERON as it is, only defines and re-uses a minimum set of properties, there is room for application-specific property extensions, especially, if time-awareness and additional qualities are not that critical. Because user-needs, application semantics or functional decisions are very tightly linked to particular requirements and stakeholders, FERON does not commit to any but accounts for the openness through functional scheme classes from re-used namespaces.

Additional field-specific extensions such as functional linguistic information are enabled in that way and would allow for the overcoming of concept centered-approaches as identified to be problematic with natural language processing. FERON is not the model of an information system as a thing, but a perceived world that information systems ought to be able to model. Information system or application setups imply inferencing machines' or logic constraints' definitions, which are considered contextually bound and therefore far beyond the scope of this work. For a validation of FERON, available domain models were investigated and compared, as well as relevant domain and field entities (see list of namespaces). The analysis of system architectures and styles, but especially the history of conceptual modeling revealed an urgent need for ontological foundations, and FERON was therefore BWW guided from its start.

FERON as a result is generic enough to scale beyond the validated LT domain into other domains such as Computer Science, Math, Medicine, Physics. The clarity that was stressed, and the consistency in structure was intended to support this. It was shown how with minimal ontological commitments, multiple identifiers, with an elaborated relationship construct, and also with the employed SKOS structure for functions, the intended scalability is enabled, and Language Technology as a highly cross-disciplinary field can surely be considered a suitable extension example. FERON as a *canonical* formal Research domain description *feron.owl*[265] will contribute and guide future community activities – especially the author's forthcoming activities in the CRIS community, but also within specific sub-communities. In the end it should be clear, that technology on its own is not enough to enable integration and to grant information access.

FERON as a formal domain model is available with this work as a valid FERON.owl file, and may as such be considered as a template for further population, where formal restrictions and rules are directly applicable and where field-specific as well as terminological extensions are immediately possible. A few instances (data) are delivered with the formal FERON.pprj Protégé file.

[265] FERON feron.owl as a frame was realised with Protégé 3.5 alpha. The Ontograf pictures as a visualisation thereof have been produced within Protégé 4.1, in a non-frame manner. Both versions of Protégé are freely available and documented in a Wiki: http://protegewiki.stanford.edu/wiki/Main_Page (Last visit: September 16th, 2012)

Bibliographic References

[Allen1983] J.F. Allen: *Maintaining Knowledge about temporal intervals*. Communications of the ACM. Vol. 26 (11), pp. 832-843, November 1983.

[Asserson et al. 2002] A. Asserson, K. Jeffery, A. Lopatenko: *CERIF: Past Present and Future: An Overview.* pp. 33-40. In Proceedings: Sixth International Conference on Current Research Information Systems (CRIS 2002), University of Kassel, Kassel 2002.

[Asserson & Jeffery 2004] A. Asserson, K.Jeffery: *Research Output Publications and CRIS*. In Proceedings: Seventh International Conference on Current Research Information Systems (CRIS 2004), May 13-15, 2004, pp. 29-40, Leuven University Press.

[Aßmann et al. 2006] U. Aßmann, S. Zschaler, G. Wagner: *Ontologies, Metamodels, and the Model-Driven Paradigm*. Computer and Information Science: Ontologies for Software Engineering and Software Technology. pp. 249-273, 2006.

[Atkinson & Kühne 2003] C. Atkinson, T. Kühne: *Model-Driven Development: A Metamodeling Foundation*. IEEE Software. pp. 37-41, September/October 2003.

[Atkinson et al.1989] M. Atkinson, F. Bancilhon, D. DeWitt, K. Dittrich, D. Maier; and S. Zdonik; *The Object-Oriented Database System Manifesto*. Proceedings: First International Conference on Deductive and Object-Oriented Databases. pp. 223-240, Kyoto, Japan, December 1989.

[Bachman 1973] C.W. Bachman: *The Programmer as Navigator*. ACM Turing Award Lecture, 1973. Communications of the ACM. Vol. 16 (11), pp. 653-658, November 1973.

[Baker 2012] D. Baker: *CASRAI and Research Impacts.* In: S. Bittner, S. Hornbostel (Eds.): Forschungsinformation in Deutschland: Anforderungen, Stand und Nutzen existierender Forschungsinformationssysteme. Workshop Forschungsinformationssysteme 2011, IFQ Working Paper No. 10, pp. 127–128, May 2012.

[Berners-Lee et al. 2005] T. Berners-Lee; R. Fielding; L. Masinter: *IETF RFC 3986 Uniform Resource Identifier (URI): Generic Syntax*. Online: http://www.ietf.org/rfc/rfc3986.txt (Last visit: October 24th, 2012)

[Bianchini & Guerrini 2009] C. Bianchini, M. Guerrini: *From Bibliographic Models to Cataloguing Rules: Remarks on FRBR, ICP, ISBD, and RDA and the Relationships Between them*. Cataloging and Classification. Quaterly, Vol. 47, pp. 105-124, 2009.

[Bird et al. 2008] S. Bird, R. Dale, B.J. Dorr; G.B. Joseph, M. T Kan, M-Y. Lee, D. Powley, D.R. Radev, Y.F. Tan: *The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics.* In Proceedings: International Conference on Language Resources and Evaluation (LREC 08), Marrakesh, Morocco, May 2008.

[Bizer et al. 2009] C. Bizer, T. Heath, T. Berners-Lee: *Linked Data - The Story So Far. Preprint* International Journal on Semantic Web and Information Systems (IJSWIS). T. Heath, M. Hepp, C. Bizer (Eds.), Vol. 5 (3), 2009.

[Björk 2007] B.-C. Björk: *A model of scientific communication as a global distributed information system*. Information Research. Vol. 12 (2), pp. 307-355, January 2007.

[Le Bœuf 2003] P. Le Bœuf: *Brave new FRBR world* (Version unknown). Prepared for the 4th IFLA Meeting of Experts on an International Cataloguing Code (IME ICC 4), Seoul, South Korea, August 16-18, 2003.

[Booch et al. 1998] G. Booch, J. Rumbaugh; I. Jacobson: *Unified Modeling Language User Guide, The.* Addison Wesley 512 pages, First Edition October 20, 1998.

[Borgman 2011] C.L. Borgman: *The Conundrum of Sharing Research Data*. Journal of the American Society for Information Science and Technology. Working Paper Series, pp. 1-40, 2011.

[Bouquet et al. 2007] P. Bouquet, H. Stoermer, D. Giacomuzzi: *OKKAM: Enabling a Web of Entities.* In Proceedings: World Wide Web Conference (WWW2007). Banff, Canada, May 8-12, 2007,

[Bouquet et. al. 2006] P. Bouquet, H. Stoermer, M. Mancioppi, D. Giacomuzzi: *OKKAM: Towards a solution to the "Identity Crisis" on the Semantic Web.* In Proceedings: Third Italian Semantic Web Workshop (SWAP 2006), G. Tummarello, P. Bouquet, O. Signore (Eds.), Vol 201, Pisa, Italy, December 18-20, 2006.

[Bray et al. 2009] T. Bray, D. Hollander, A. Layman, R. Tobin, H.-S. Thomson: *Namespaces in XML 1.0 (Third Edition).* W3C Recommendation 8 December 2009. Online: http://www.w3.org/TR/REC-xml-names/ (Last visited: October 24th, 2012)

[Bray et al. 2008] T. Bray, J. Paoli, C.M. Sperberg-McQueen; E. Maler, F. Yergeau: *Extensible Markup Language (XML) 1.0 (Fifth Edition)* W3C Recommendation 26 November 2008. Online: http://www.w3.org/TR/REC-xml/ (Last visit: October 24th, 2012)

[Brickley & Guha 2000] D. Brickley and R.V. Guha: *Resource Description Framework (RDF) Schema Specification 1.0*. W3C Candidate Recommendation 27 March 2000.
Online: http://www.w3.org/TR/2000/CR-rdf-schema-20000327/ (Last visited: October 24th, 2012)

[Brown 2004] A.W. Brown: *Model driven architecture: Principles and practice.* Software and System Modelling. Vol. 3, pp. 314-327, 2004.

[Buckland 1991] M.K. Buckland: *Information as Thing.* Journal of the American Society for Information Science (ASIS). Vol. 42 (5), pp. 351-360, 1991.

[Buitelaar et al. 2009] P. Buitelaar, P. Cimiano, P. Haase, M. Sintek: *Towards Linguistically Grounded Ontologies*. In Proceedings: Sixth European Semantic Web Conference: Research and Applications (ESWC2009). pp. 111-125, Springer-Verlag Berlin, Heidelberg, 2009.

[Bunge 1977] M. Bunge: *Ontology* - III: *The Furniture of the World.* Treatise on Basic Philosophy. Dordrecht: Reidel, 1977.

[Bush 1945] V. Bush: *As We may think*. The Atlantic. 1945.

[Brachman 1976] R.J. Brachman: *What's in a Concept: Structural Foundations for Semantic Networks*. International Journal of Man-Machine Studies. Vol. 9 (2), pp. 127-152, 1976.

[Calero et al. 2006] C. Calero, F. Ruiz, M. Piattini (Eds.): *Ontologies for Software Engineering and Software.* Springer-Verlag, Berlin, Heidelberg, 2006.

[Candela et al. 2011] L. Candela, G. Athanasopoulos, D. Castelli, K. El Raheb, P. Innocenti, Y. Ioannidis, A. Katifori, A. Nika, G. Vullo, S. Ross: *The Digital Library Reference Model.* Version 0.98. Deliverable.

[Capstick et al. 2001] J. Capstick, T. Declerck, G. Erbach, A. Jameson, B. Jörg, R. Karger, H. Uszkoreit, W. Wahlster, T. Wegst. *COLLATE: Competence Center in Speech and Language Technology.* InProceedings: 3rd International Conference on Language Resources and Evaluation (LREC 2002), May 28-31, Las Palmas, Canary Islands, Spain, 2002.

[Capurro & Hjørland 2003] R. Capurro: *The Concept of Information.* Annual Review of the Information Science and Technology. R. Capurro (Ed.), Vol. 37 (8), pp. 343-411, 2003. Online: http://arizona.openrepository.com/arizona/html/10150/105705/infoconcept.html (Last visited: October 24th, 2012) "The state-of-the-art report "The Concept of Information" was published in the *Annual Review of Information Science and Technology Ed. B. Cronin, Vol. 37 (2003) Chapter 8, pp. 343-411.* The draft version below is not identical with the published version."

[Carnap1947] R. Carnap: *Meaning and Necessity – A Study in Semantics and Modal Logic.* The University of Chicago Press, Chicago, Illinois, 1947.

[Chalmers 1999] A.F. Chalmers: *What is this thing called Science.* Open University Press. February 1999. Online: http://www.scribd.com/doc/12834586/Chalmers-AF-1999-What-is-This-Thing-Called-Science-3e-0872204537 (Last visit: October 24th, 2012)

[Chen 1976] P. P.-S. Chen: *The Entity-Relationship Model - Toward a Unified View of Data.* ACM Transactions on Database Systems. Vol. 1 (1), pp. 9-36, March, 1976.

[CCSDS Blue Book 2002] Consultative Committee for Space Data Systems: *Reference Model for an Open Archival Information System (OAIS)*. CCSDS 650.0-B-1 Blue Book, Issue 1, CCSDS Secretariat, Program Integration Division (Code M-3), National Aeronautics and Space Administration, Washington, DC 20546, USA, January 2002.

[Chen et al. 1999] P.P. Chen, B. Thalheim, L.Y. Wong: *Future Directions of Conceptual Modeling*. Conceptual Modeling, LNCS 1565, pp. 287-301, Springer-Verlag Berlin, Heidelberg 1999.

[CLARIN 2010] E. Hinrichs (Responsible): *Language Resources and Tools Survey and Taxonomy and Criteria for the Quality Assessment*. Deliverable D5C-2. R. Carlson, T. Caselli, K. Elenius, B. Gaiffe, D. House, E. Hinrichts, V. Quochi, K. Simov, I. Vogel (Eds.), version 1, 2010.

[Clements & Lockhart 2010] A. Clements, N. Lockhart: *Using CERIF-XML to integrate heterogeneous research information from several institutions into a single portal.* JISC-funded CRISPool project. Final Report. University of St. Andrews, September 2010.

[Codd 1970] E.F. Codd: *A relational model of data for large shared data banks.* ACM Transactions on Database Systems. Vol. 13 (6), pp. 377-387, June 1970.

[Codd 1980] E.F. Codd: *Data Models in Database Management*. In Proceedings: Workshop on Data Abstraction, Databases and Conceptual Modelling, ACM, pp. 112-114, New York, United States, 1980.

[Correndo et al. 2010] G. Correndo, M. Slavadores, I. Millard, N. Shadbolt: *Linked Timelines: Temporal Representation and Management in Linked Data.* First International Workshop on Consuming Linked Data (COLD 2010), Shanghai, China, 2010.

[Corson-Rikert et al. 2012] J. Corson-Rikert, D.B. Krafft, B.J. Lowe: *VIVO: Semantic Network of Researchers and Research Information as Linked Open Data.* In: S. Bittner, S. Hornbostel (Eds.): Forschungsinformation in Deutschland: Anforderungen, Stand und Nutzen existierender Forschungsinformationssysteme. Workshop Forschungsinformationssysteme 2011, IFQ Working Paper No. 10, pp. 139–154, May 2012.

[Cox et al. 2011] M. Cox, R. Gartner, K. Jeffery; B. Jörg *Measuring Impact under CERIF (MICE) Project*. Project report from JISC-funded MICE project. Online: http://mice.cerch.kcl.ac.uk/wp-uploads/2011/06/ImpactUnderCERIF.doc (Last visit: October 24th, 2012)

[Cross et al. 2000] P. Cross, D. Brickley, T. Koch: *Conceptual Relationships for Encoding Thesauri, Classification Systems and Organised Metadata Collections and a Proposal for Encoding a Core Set of Thesaurus Relationships using an RDF Schema*. 2000. DESIRE II report: Online: http://ilrt.org/discovery/2001/01/rdf-thes/ (Last visit: October 24th, 2012).

[Dameron et al. 2005] O. Dameron, D.L. Rubin, M.A. Museen: *Challenges in Converting Frame-Based Ontology into OWL: the Foundational Model of Anatomy Case-Study*. AMIA Annual Symposium Proceedings 2005, pp. 181-185, 2005

[DataCite 2011] *DataCite Metadata Schema for the Publication and Citation of Research Data.* DataCite - International Data Citation, Version 2.2, July 2011. Online: http://schema.datacite.org/meta/kernel-2.2/doc/DataCite-MetadataKernel_v2.2.pdf (Last visit: October 24th, 2012)

[van Dijk et al. 2010] E. van Dijk, A. Hogenaar, M. van Meel: *Users in the spotlight: study on the use of the Dutch scientific portal NARCIS.* In Proceedings: Tenth International Conference on Current Research Information Systems (CRIS 2010), Aalborg, Denmark, June 2010.

[Dublin Core 2010] *Dublin Core Metadata Element Set, Version 1.1.* Dublin Core Metadata Initiative: Online: http://dublincore.org/documents/2012/06/14/dces/ (Last visit: October 24th, 2012)

[Ducloy et al. 2010] J. Ducloy, T. Daunois, A. Hermann, J.-C. Lamirel, C. Vanoirbeek, M. Fouloneau, S. Sire: *Metadata for Wicri, a Network of Semantic Wikis for Communities in Research and Innovation*. Proceedings: International Conference on Dublin Core and Metadata Applications (DC-2010), Pittsburgh, October 20-22, 2010.

[Dunsire et al. 2011] G. Dunsire, D. Hillmann, J. Phipps, K. Coyle: *A Reconsideration of Mapping in a Semantic World.* In Proceedings: Int. Conference on Dublin Core and Metadata Applications. pp. 26-35, 2011.

[EC Report 2010] Expert Group on Assessment of University-Based Research: *Assessing Europe's University-Based Research.* Publications Office of the European Union, Brussels, 2010.

[EC-SDI Report 2010] High level Expert Group on Scientific Data. *Riding The Wave*. European Commission 2010. Final Report, October 2010.

[EDM Primer 2010] *Definition of the Europeana Data Model Elements (Version 4.11)*. 2010.

[Enserink 2009] M. Enserink: *Are You Ready to Become a Number?* Science, pp. 1662-1664, 2009.

[Esswein et al. 2010] W. Esswein, J. Stark, H. Schlieter: *The Selection of Modeling Grammars*. In Proceedings: Modellierung betrieblicher Informationssysteme (MobIS 2010). Lecture Notes in Informatics. Vol. P-171, pp. 13-28, September 2010.

[EUROHORCS-ESF 2008] *EUROHORCS and ESF Vision on a Globally Competitive ERA and their Road Map for Actions.* Online: http://www.eurohorcs.org/SiteCollectionDocuments/ESF_Road%20Map_long_0907.pdf (Last visit: October 24th, 2012)

[Evermann 2009] J. Evermann: *A UML and OWL description of Bunge's Upper Level Ontology Model.* Software and System Modelling. Vol. 8 (2), pp. 235-249, April 2009.

[Evermann & Wand 2001] J. Evermann, Y. Wand: *Towards Ontologically Based Semantics for UML Constructs.* In Proceedings: 20th International Conference on Conceptual Modeling: Conceptual Modeling (ER'01), pp. 354-367, H.S.Kunii, S. Jajodia, A. Sølvberg (Eds.), LNCS 2224, Springer-Verlag London, 2001.

[Fettke & Loos 2003] P. Fettke and P. Loos: *Ontological Evaluation of the Specification Framework Proposed by the "Standardized Specification of Business Components" memorandum - some preliminary results.* In Proceedings: First International Workshop on Component Engineering Methodology. S. Overhage and K. Turowski (Eds.), pp. 1-12, 2003.

[Fielding 2000] R.T. Fielding: *Architectural Styles and the Design of Network-based Software Architectures*. Doctoral Dissertation. University of California, Irvine, 2000.

[Fielding et al. 1999] R.T. Fielding, J. Gettys, J. Mogul, H. Frystyk, T. Berners-Lee: *IETF RFC 2616 Hypertext Transfer Protocol - HTTP 1.1.* Network Working Group. Internet Draft. 1999.

[Floridi 2005] L. Floridi: *Is Information Meaningful Data? preprint.* Philosophy and Phenomenological Research. Vol. 70 (2), pp. 351-370, 2005. Online: http://philsci-archive.pitt.edu/2536/1/iimd.pdf (Last visited: October 24th, 2012)

[Franklin et al. 2005] M. Franklin, A. Halevy, D. Maier: *From Databases to Dataspaces: A New Abstraction for Information Management*. ACM SIGMOD Record. Vol 34 (4), pp. 27-33, December 2005.

[Frege 1892] G. Frege: *Über Begriff und Gegenstand.* Vierteljahresschrift für wissenschaftliche Philosophie. Vol. 16, pp. 192-205, 1892.

[FRBR 1997/2009] IFLA Study Group: *Functional Requirements for Bibliographic Records.* Final Report. 1997 (as amended and corrected through February 2009). Online: http://www.ifla.org/files/assets/cataloguing/frbr/frbr_2008.pdf (Last visit: October 24th, 2012)

[Gangemi et al. 2002] A. Gangemi, N. Guarino, C. Masolo, A. Oltramari, L. Schneider: *Sweetening Ontologies with DOLCE*. In Proceedings: Thirteenth Int. Conference on Knowledge Engineering and Knowledge Management (EKAW'02). Ontologies and the Semantic Web. pp.166-181, Springer-Verlag, London 2002.

[Gartner 2008] R. Gartner: *Metadata for the Digital Libraries: State of the Art and Future Directions (1.0).* JISC Technology & Standards Watch, Peer-reviewed Report, 2008. Online: http://www.jisc.ac.uk/media/documents/techwatch/tsw_0801pdf.pdf (Last visited: October 24th, 2012)

[Gehlert & Esswein 2005] A. Gehlert, W. Esswein: *Towards a formal Research Framework for ontological analyes.* Advanced Engineering Informatics. Vol. 21 (2), pp. 119-131, 2005.

[Gemino & Wand 2005] A.Gemino and Y. Wand: *Complexity and Clarity in Conceptual Modeling: Comparison of Mandatory and Optional Properties*. Data and Knowledge Engineering. Vol. 55 (3), pp. 301-326, December 2005.

[Gibbons et al. 1994] M. Gibbons, C. Limoges, H. Nowotny, S. Schwartzman, P. Scott, M. Trow: *The New Production of Knowledge: Dynamics of Science and Research in Contemporary Societies.* SAGE Publications Inc., London, California, New Delhi, 1994.

[Ginty et al. 2012] K. Ginty, S. Kerridge, P. Fairley, R. Henderson, P. Cranner, A. Bokma, S. Garfield: *CERIF for Datasets (C4D) - An Overview.* Proceedings: 11th International Conference on Current Research Information Systems (CRIS 2012). K. Jeffery, J. Dvorak (Eds.): pp. 53-60, June 2012.

[van Godtsenhoven et al. 2008] K. van Godtsenhoven, M.K. Elbæk, G. Schmeltz Pedersen, B. Sierman, M. Bijsterbosch, P. Hochstenbach, R. Russell, M. Vanderfeesten: *The European Repository Landscape 2008 - Survey on Technology.* M. Vernooy-Gerritsen (Ed.). Amsterdam University Press, Amsterdam 2008.

[van der Graaf & van Eijndhoven 2008] M. van der Graaf and K. van Eijndhoven: *The European Repository Landscape 2008 – Inventory of Digital Repositories for Research Output in the EU.* Amsterdam University Press, Amsterdam 2008.

[Gruber 1993] T.R. Gruber: *Toward Principles for the Design of Ontologies Used for Knowledge Sharing*. International Journal Human-Computer Studies. Vol. 43, pp. 907-928, Revision August 23, 1993.

[Gruber 1993a] T.R. Gruber: *A Translation Approach to Portable Ontology Specifications*. Knowledge Acquisition. Volume 5 (2), pp. 199-220, Academic Press Ltd. London, United Kingdom, June 1993.

[Guarino 1998] N. Guarino (Ed): *Formal Ontology and Information Systems.* In Proceedings: Formal Ontology and Information Systems (FOIS98), Trento, Italy, 6-8 June, 1998, pp. 3-15, IOS Press Amsterdam, 1998.

[Guarino & Giaretta 1995] N.Guarino, P. Giaretta: *Ontologies and Knowledge Bases - Towards a Terminological Clarification*. Towards Very Large Knowledge Base: Knowledge Building and Knowledge Sharing. pp. 25-32, IOS Press Amsterdam, 1995.

[Guarino & Guizzardi 2006] N. Guarino and G. Guizzari: *In the Defense of Ontological Foundations for Conceptual Modeling.* Scandinavian Journal of Information Systems. Vol 18 (1), Debate Forum in reply to the article entitled *On Ontological Foundations of Conceptual Modeling* by B. Wyssusek. Online: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.86.6281 (Last visited: October 24th, 2012)

[Guizzardi & Halpin 2008] G. Guizzard, T. Halpin: *Ontological Foundations for Conceptual Modelling*. Applied Ontology. Vol 3 (1-2), pp. 1-12, 2008.

[Guizzardi & Wagner 2008] G. Guizzardi, G. Wagner: *What's in a Relationship: An Ontological Analysis*. In Proceedings: 27th International Conference on Conceptual Modeling. pp. 83-97, Springer-Verlag Berlin, Heidelberg 2008.

[Halpin et al. 2009] H. Halpin, V. Presutti, A. Gangemi: *An Ontology of Resources: Solving the Identity Crisis.* In Proceedings: European Semantic Web Conference (ESWC 2009), pp. 121 – 140. Heraklion, Crete, Greece, May 31 – June 4, 2009.

[Halpin 2006] H. Halpin: *Identity, Reference, and Meaning on the Web*. In Proceedings: World Wide Web Conference (WWW2006), May 22-26, 2006, Edinburgh, UK.

[Heath & Bizer 2011] T. Heath and C. Bizer: *Linked Data: Evolving the Web into a Global Data Space.* Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan & Claypool 2011

[Herrera et al. 2005] S. Herrera, D. Pallioto, G. Tkachuk, P.A. Luna: *Ontological Modelling of Information Systems from Bunge's Contributions.* In Proceedings: CAiSE workshop, Porto, Portugal, 2005. pp. 571-582. Online: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.101.7744 (Last visited: October 24th, 2012)

[Hirwade 2011] M.A. Hirwade: *A Study of Metadata Standards.* Library High Tech News. Vol. 28 (7), pp. 18-25, 2011.

[HLT Survey 1997] R. Cole, J. Mariani, H. Uszkoreit, G.B. Varile, A. Zaenen, A. Zampolli, V. W. Zue (Eds.): *Survey of the State of the Art in Human Language Technology*. Cambridge University Press. 1997.

[Hoffart et al. 2010] J. Hoffart, F.M. Suchanek, K. Le Bœuf erberich, G. Weikum: *YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia.* Max Planck Institute for Informatics. Campus E14, Saarbrücken, 2010.

[Houssos et al. 2012] N. Houssos, B. Jörg, B. Matthews. *A Multi-Level Metadata Approach for a Public Sector Information Data Infrastructure.* In Proceedings: Eleventh Int. Conference on Current Research Information Systems (CRIS 2012), Prague, Czech Republic, June 6-9, 2012.

[IMD I3.0.3. 2003] *Metadata Elements for Session Descriptions*. IMDI Team (Draft Proposal), Max-Planck Institute, Part 1 July 2003.

[ISBD 2007] *International Standard Bobliographic Description (ISBD).* IFLA 2007. Online: http://www.ifla.org/publications/international-standard-bibliographic-description (Last visit: October 24th, 2012)

[Jacobs & Walsh 2004] I. Jacobs and N. Walsh: *Architecture of the World Wide Web, Volume One*. W3C Recommendation, 15 December 2004. Online: http://www.w3.org/TR/webarch/ (Last visit: October 24th, 2012)

[Jeffery & Asserson 2010] K. Jeffery and A. Asserson: *CERIF-CRIS for the European E-Infrastructure*. Data Science Journal. Vol. 9, pp. 1–6, July 24, 2010.

[Jeffery & Asserson 2006] K. Jeffery and A. Asserson: *CRIS Central Relating Information System*. In Proceedings: 8th Int. Conference on Current Research Information Systems (CRIS2006): Enabling Interaction and Quality: Beyond the Hanseatic League, pp.109-120, Leuven University Press, 2006.

[Jeffery 2010] K. Jeffery: *The CERIF Model As the Core of a Research Organisation*. Data Science Journal. Vol. 9, pp. 7-13, July 24, 2010.

[Jeffery & Asserson 2004] K. Jeffery and A. Asserson: *Relating Intellectual Property Products to the Corporate Context*. Research Publication Quaterly. Vol. 21 (1), 2004.

[Jeffery 1999] K. Jeffery: *An Architecture for Grey Literature in a R&D Context*. In Proceedings: Grey Literature Conference. Washington D.C., USA, October 4-5, 1999.

[Jeffery et al. 1989] K. Jeffery, J. O. Lay, J.F. Miquel; S. Zardan; F. Naldi; and I. Vannini Parenti: *IDEAS: A System for International Data Exchange and Access for Science.* Information Processing and Management. Vol. 25 (6), pp. 703-711, 1989.

[Jörg 2012] B. Jörg: *Übersicht Systeme Europa.* In: S. Bittner, S. Hornbostel (Eds.): Forschungsinformation in Deutschland: Anforderungen, Stand und Nutzen existierender Forschungsinformationssysteme. Workshop Forschungsinformationssysteme 2011, IFQ Working Paper No. 10, pp. 103–114, May 2012.

[Jörg et al. 2012] B. Jörg, K. Jeffery, J. Dvorak, N. Houssos, A. Asserson, G. van Grootel, R. Gartner, M. Cox, H. Rasmussen, T. Vestdam, L. Strijbosch, A. Clements, V. Brasse, D. Zendulkova, T. Höllrigl, L. Valkovic, A. Engfer, M. Jägerhorn, M. Mahey, N. Brennan, M-A. Sicilia, I. Ruiz-Rube, D. Baker, K. Evans, A. Price, M. Zielinski: *CERIF2008 - 1.3 Full Data Model (FDM) - Model Introduction and Specification*. euroCRIS, January 2012.

[Jörg et al. 2012a] B. Jörg, J. Dvorak, T. Vestdam: *Streamlining the CERIF XML Data Exchange Format: Towards CERIF 2.0.* In Proceedings: 11th International Conference on Current Research Information Systems (CRIS 2012), Prague, Czech Republic, June 6-9, 2012. (Best Paper Award)

[Jörg et al. 2012b] B. Jörg, T. Höllrigl, M.-A. Sicilia. *Entities and Identities in Research Information Systems.* In Proceedings: 11th International Conference on Current Research Information Systems (CRIS 2012), Prague, Czech Republic, June 6-9, 2012.

[Jörg et al. 2012c] B. Jörg, I. Ruiz-Rube, M.A. Sicilia, J. Dvorak, K. Jeffery, T. Höllrigl, H.S. Rasmussen, A. Engfer, T. Vestdam, E. Garcia Barriocanal: *Connecting Closed World Research Information Systems through the Linked Open Data Web*. International Journal of Software Engineering and Knowledge Engineering (IJSEKE), Vol. 22, Consuming and Producing Linked Data on Real World Applications. June, 2012.

[Jörg et al. 2011] B. Jörg, K. Jeffery, G. van Grootel: *Towards a Sharable Research Vocabulary SRV – A Model-driven Approach.* In Proceedings: Metadata & Semantics Research Conference (MTSR11), Yasar University, Izmir, Turkey, October 2011, Springer, Berlin, Heidelberg.

[Jörg 2010] B. Jörg: *CERIF: The Common European Research Information Model*. Data Science Journal. Vol. 9, pp. 24-31, July 24, 2010.

[Jörg et al. 2010] B. Jörg, H. Uszkoreit, A. Burt: *LT World: Ontology and Reference Information Portal.* In Proceedings: 7th Conference on International Language Resources and Evaluation (LREC10), Valetta, Malta, May 19-21, 2010.

[Jörg et al. 2007a] B. Jörg, K. Jeffery, A. Asserson, G. van Grootel, E. Grabczewski: *CERIF2006-1.1 Full Data Model (FDM) - Model Introduction and Specification*. euroCRIS, October 2007.

[Jörg et al. 2007b] B. Jörg, O. Krast, K. Jeffery, G. van Grootel: *CERIF2006XML-1.1 Data Exchange Format Specification*. euroCRIS, April 2007.

[Jörg & Uzkoreit 2005] B. Jörg, H. Uszkoreit: *The Ontology-based Architecture of LT World, a Comprehensive Web Information System for a Science and Technology Discipline.* In: Leitbild Informationskompetenz: Positionen - Praxis - Perspektiven im europäischen Wissensmarkt. 27. Online Tagung (zugleich 57. Jahrestagung) der DGI. Frankfurt am Main, 23-25 Mai, 2005.

[Krafft et al. 2010] D.B. Krafft, N.A. Cappadona, B. Caruso, J. Corson-Rikert, D. Medha, B.L. Lohe: *Vivo: Enabling National Networking of Scientists.* In Proceedings: Web Science Conference (WebSci10). Extending the Frontiers of Society (online), April 26-27, 2010, Raleigh, NC, United States.

[Krause 2002] J. Krause: *Current Research Information as Part of Digital Libraries and the Heterogeneity Problem - Integrated Searches in the Context of Databases with Different Content Analyses.* In Proceedings: 6th International Conference on Current Research Information Systems (CRIS 2002), W. Adamczak, A. Nase (Eds.), pp. 21-31, August 29th – 31st in Kassel, Germany.

[Kripke 1980] S. Kripke: *Naming and Necessity*. Harvard University Press, Cambridge, Massachusetts, United States, 1980.

[Kuhn1962] T. S. Kuhn: *The Structure of Scientific Revolutions.* University of Chicago Press, 1962.

[Lagoze & Van de Sompel 2008] C. Lagoze and H. Van de Sompel: *Open Archives Initiative - Object Reuse and Exchange*. ORE User Guide Primer. Open Archives Initiative, 2008.
Online: http://www.openarchives.org/ore/1.0/primer.html (Last visit: October 24th, 2012)

[Li et al. 2010] Y.-F. Li, G. Kennedy, F. Davies, J. Hunter: *PODD: Towards an Extensible, Domain-agnostic Scientific Data Management System*. In Proceedings: IEEE Sixth International Conference on e-Science (ESCIENCE10), pp. 137-144, IEEE Computer Society, Washington DC, USA.

[Lindland et al. 1994] O.I. Lindland, G. Sindre, A. Solvberg: *Understanding Quality in Conceptual Modeling.* Software IEEE, pp. 42-49, Vol. 11 (2), March 1994

[Lowe et al. 2007] B. Lowe, B. Caruso, J. Corson-Rikert: *VIVO Development Roadmap: Enhancing an Ontology-Based University Research Portal with OWL and Rules*. In Proceedings: Workshop on OWL: Experiences and Directions OWLED 2007. Innsbruck, Austria, June 6-7, 2007.
Online: http://ceur-ws.org/Vol-258/paper05.pdf (Last visited: October 24th, 2012)

[Manola & Miller 2004] F. Manola and E. Miller: *RDF Primer - W3C Recommendation* 2004.

[Miles & Bechhofer 2009] A. Miles, S. Bechhofer *SKOS Simple Knowledge Organisation System Reference*. W3C Recommendation 18 August 2009. Online: http://www.w3.org/TR/skos-reference/ (Last visited: October 24th, 2009).

[Mylopooulos 1992] J. Mylopoulos: *Conceptual Modeling and Telos.* P. Loucopoulos and R. Zicari (Eds.) Conceptual Modeling, Databases and Case. pp. 49-68, Wiley, 1992.

[Nogueras-Iso et al. 2004] J. Nogueras-Iso, F.J. Zarazaga-Soria, J. Lacasta, R.Bejar, P.R. Muro-Medrano: *Metadata Standard Interoperability: Application in the Geographic Information Domain. Preprint.* Computers, Environment and Urban Systems. Vol. 28, pp. 611-634, 2004.

[Oei et al. 1992] J.L.H. Oei, L.J.G.T. ven Hemmen, E.D. Falkenberg, S. Brinkkemper: *The Meta Model Hierarchy: A Framework for Information Systems Concepts and Techniques.* Department of Information Systems, University of Nijmegen, 1992.

[Perry & Wolf 1992] D. E. Perry, A. L. Wolf: *Foundations for the Study of Software Architecture.* Software Engineering Notes. Vol. 17 (4), pp. 40-52, 1992.

[Poole 2001] J.D. Poole: *Model-Driven Architecture: Vision, Standards And Emerging Technologies*. Position Paper from Adaptive Object-Models and Meta Modeling Techniques Workshop, European Conference on Object-Oriented Programming (ECOOP), Budapest, June 2001.

[Mons et al. 2011] B. Mons, H. van Haagen, C. Chichester, P.-B. 't Hoen, J. T. den Dunnen, G. van Ommen, E. van Mulligen, B. Singh, R. Hooft, M. Roos, J. Hammond, B. Kiesel, B. Giardine, J. Velterop, P. Groth, E. Schultes: *The Value of Data.* Nature Genetics. Vol. 43 (4), pp. 281-283, April 2011.

[Patel-Schneider & Horrocks 2006] P.F. Patel-Schneider and I. Horrocks: *Position Paper: A Comparison of Two Modelling Paradigms in the Semantic Web*. Later published in: Journal Web Semantics. Vol. 5 (4), pp. 240-250, 2007. Online: http://www.cs.ox.ac.uk/ian.horrocks/Publications/download/2006/PaHo06a.pdf (Last visit: October 24th, 2012)

[Alexander et al. 2009] K. Alexander, R. Cyganiak, M. Hausenblas, J. Zhao. *Describing Linked Datasets - On the Design and Usage of voiD, the "Vocabulary of Interlinked Datasets".* In Proceedings: Linked Data on the Web Workshop (LDOW09), 18th International World Wide Web Conference (WWW09), 2009.

[Hayashi 2007] Y. Hayashi, T. Declerck, P. Buitelaar, M. Monachini: *Ontologies for a Global Language Infrastructure.* In Proceedings: First International Conference on Global Interoperability for Language Resources (ICGL-2008), January 9-11, Hong Kong, China, pp. 105-112, Online-Proceedings, 1/2008.

[Haase 2004] K. Haase: *Context for Semantic Metadata.* In Proceedings: 12th Annual ACM Int. Conference on Multimedia (MM04), pp. 204-211. 2004.

[Harth et al. 2010] A. Harth, M. Janik, S. Staab: *Semantic Web Architecture*. Karlsruhe Institute of Technology, August, 2010.

[Hendler et al. 2002] J. Hendler, T. Berners-Lee, E. Miller: *Integrating Applications on the Semantic Web.* Journal of the Institute of Electrical Engineers Japan. Vol. 122 (10), pp. 676-680, October 2002.

[Hevner et al. 2004] A.R. Hevner, S.T. March, J. Park, S. Ram: *Design Science in Information Systems Research.* MIS Quarterly. Vol. 28 (1), pp. 75–105, March 2004.

[Hornbostel 2006] S. Hornbostel: *From CRIS to CRIS: Integration and Interoperability.* In Proceedings: 8th Int. Conference on Current Research Information Systems (CRIS 2006), pp. 29-38, Bergen Norway, May 10-13, 2006.

[Horrocks et al. 2003] I. Horrocks, P. Patel-Schneider, F. van Harmelen: *From SHIQ and RDF to OWL: The Making of a Web Ontology Language*. Journal of Web Semantics. Vol. 1, 2003.

[ICP 2009] *Statement of International Cataloguing Principles*. The International Federation of Library Associations and Institutions (IFLA) 2009. Online: http://www.ifla.org/files/assets/cataloguing/icp/icp_2009-en.pdf (Last visit: October 24th, 2012)

[Isaak & Summers 2008] A. Isaac, E. Summers: *SKOS Simple Knowledge Organisation System Primer*. W3C Working Draft 21 February 2008. Online: http://www.w3.org/TR/2008/WD-skos-primer-20080221/ (Last visit: October 24th, 2012)

[Jarrar & Meersmann 2002] M. Jarrar and R. Meersman: *Formal Ontology Engineering in the DOGMA Approach*. Proceedings: International Conference on Ontologies, Databases and Applications of Semantics (ODBase02), LNCS 2519, pp. 1238-1254, Springer-Verlag, London 2002.

[Luzi et al. 2004] D. Luzi, M. Castriotta, R. di Cesare, L. Libutti, M. Manco: *The Communication Flow of Research Project Results*. *Preprint.* In Proceedings: Grey Literature Conference 2004.

[Marshal & Shipman 2003] C.C. Marshall, F. L. Shipman: *Which Semantic Web?* In Proceedings: 14th ACM Conference on Hypertext and Hypermedia, pp. 57-66, ACM Press, August 2003.

[Masolo et al. 2003] C. Masolo, S. Borgo, A. Gangemi, N. Guarino, A- Oltramari: *WonderWeb Deliverable D18*. Ontology Library (final). Version 1.0. Laboratory for Applied Ontology, ISTC-CNR. Trento, Italy, 2003.

[McNeely 2002] I.F. McNeely: *The Unity of Teaching and Research: Humboldt's Educational Revolution*. University of Oregon, 2002. Online: https://scholarsbank.uoregon.edu/xmlui/handle/1794/1456 (Last visit: October 24th, 2012)

[Minsky1974] M. Minsky: *A Framework for Representing Knowledge*. The Psychology of Computer Vision. 1974. *Reprinted.*

[Motik et al. 2009] B. Motik, P.F. Patel-Schneider, B. Parsia: *OWL2 Web Ontology Language - Structural Specification and Functional-Style Syntax*. W3C Recommendation 27 October 2009. Online: http://www.w3.org/TR/owl2-syntax/ (Last visit: October 24th, 2012)

[Di Nitto & Rosenblum 1999] E. Di Nitto, D. Rosenblum: *Exploiting ADLs to Specify Architectural Styles Induced by Middleware Infrastructures*. In Proceedings: 21st International Conference on Software Engineering. pp. 13-22, ACM New York, NY, United States, 1999.

[Niles & Pease 2001] I. Niles, A. Pease: *Towards a Standard Upper Ontology.* In Proceedings: 2nd International Conference on Formal Ontology and Information Systems (FOIS 2001), pp. 2-9. C. Welty, B. Smith (Eds.): Maine, October 17-19, 2001.

[Noy 2004] N.F. Noy: *Semantic Integration: A Survey of Ontology-based Approaches.* ACM SIGMOD Record. Vol. 33 (4), December 2004.

[Nonaka 1991] I. Nonaka: *The Knowledge-Creating Company.* Harvard Business Review, November-December 1991.

[OECD 2002] *Frascati Manual: Proposed Standard Practice for Surveys on Research and Experimental Development*. Organisation for Economic Co-operation and Development (OECD), 2002.

[OMG 2008] *Semantics of Business Vocabulary and Business Rules (SBVR), v1.0*. Open Management Group. 2008. Online: http://www.omg.org/spec/SBVR/1.0/ (Last visit: October 24th, 2012)

[Parsons 1996] S. Parsons: *Current Approaches to Handling Imperfect Information in Data and Knowledge Bases.* Transactions on Knowledge and Data Engineering. Vol. 8 (3), pp. 353-372, 1996.

[Pepe et al. 2010] A. Pepe, M. Mayeink, C.L. Borgman. H. van de Sompel: *From Artifacts to Aggregations: Modeling Scientific Life Cycles on the Semantic Web.* Journal of the American Society for Information Science and Technology. Vol. 61 (3), pp. 567-582, March 2010. John Wiley & Sons, Inc. New York, United States.

[PMSEIC 2006] Working Group on Data for Science: *From Data to Wisdom: Pathways to successful Data Management for Australian Science*. A paper prepared by an independent Working Group for the *Prime Minister's Science, Engineering and Innovation Council (PMSEIC)*, 2006. *Its views are those of the Group, not necessarily those of the Australian Government.*

[Quillian 1969] R. Quillian: *The Teachable Language Comprehender: A Simulation Program and Theory of Language*. Communications of the ACM, Vol. 12 (8), pp. 459-476, August 1969.

[Quine1969] W.V. Quine: *Ontological Relativity & Other Essays*. Columbia University Press, 1969.

[Raymond & Abdallah 2007] Y. Raimond, S. Abdallah. *The Event Ontology.* 2007. Online: http://motools.sourceforge.net/event/event.html (Last visit: October 25th, 2012)

[RIM Report 2010] *Research Information Management. Developing Tools to Inform the Management of Research and Translating Existing Good Practice.* Report published by JISC, Joint Information Systems Committee (JISC), United Kingdom, August 2010.
Online: http://www.jisc.ac.uk/media/documents/programmes/RIM/RIMTNT_FinalReport.pdf (Last visit: October 24th, 2012)

[Rodriguez et al. 2009] E. Rodríguez, J. Conesa, M.-Á. Sicilia: *Clarifying the Semantics of Relationships between Learning Objects*. In Proceedings: Metadata and Semantics Research (MTSR 2009), Communications in Computer and Information Science (CCIS). Vol. 46, 35 pages, Springer Berlin, Heidelberg, 2009.

[Rosemann et al. 2004] M. Rosemann, P. Green, M. Indulska: *A Reference Methodology for Conducting Ontological Analyses*. Lecture Notes in Computer Science (LNCS 3288). pp. 110-121, 2004.

[Sauermann & Gyganiak 2008] W3C Interest Group Note 03 December 2008. *Cool URIs for the Semantic Web*. L. Sauermann, R. Cyganiak (Eds.): Online: http://www.w3.org/TR/cooluris/ (Last visit: October 24th, 2012)

[Schäfer 2012] U. Schäfer: *Satzsemantische Suche - präziser Finden mit der TAKE Searchbench* DOK.magazin, Vol. 2012 (2), pp. 28-31, May 2012.

[Scheer 1991] A.W. Scheer: *Architektur integrierter Informationssysteme*. Springer-Verlag, May 1991.

[Schimank & Winnes 2000] U. Schimank, M. Winnes: *European University Systems.* Science and Public Policy. Vol. 27 (6), pp. 397-408, December 2000.

[Scholze & Maier 2012] F. Scholze and J. Maier: *Establishing a Research Information System as Part of an Integrated Approach to Information Management: Best Practice at the Karlsruhe Institute of Technology (KIT).* Liber Quarterly. Vol. 21 (2), pp. 201-212, 2012.

[Schüette & Rotthowe 1998] R. Schütte, T. Rotthowe: *The Guidelines of Modeling - an Approach to Enhance the Quality in Informaiton Models.* In Proceedings: 17th International Conference on Conceptual Modeling. T.W. Ling, S.Ram, M.L. Lee (Eds.), pp. 240-254, Lecture Notes in Computer Science (LNCS), Springer 1998.

[Shanks et al. 2003] G. Shanks, E. Tansley, R. Weber: *Using Ontology to validate Conceptual Models*. Communications of the ACM. Vol. 46 (10), pp. 85–89, 2003.

[Shotton 2010] D. Shotton: *CiTO, the Citation Typing Ontology*. Journal of Biomedical Semantics, 2010. Vol. 1 (56), (Suppl 1), 2010

[Siau 2002] K. Siau: *The Psychology of Information Modeling.* Advanced Topics in Databases Research. Vol. 1, pp. 106 – 118. Idea Group Publishing, Hershey, PA, USA 2002.

[Sicilia 2010] M-Á Sicilia: *On Modeling Research Work for Describing and Filtering Scientific Information.* In Proceedings: Metadata and Semantic Research (MTSR2010). Communications in Computer and Information Science (CCIS). Vol. 108, pp. 247-254, December 2010.

[Simons & Bird 2008] G. Simons, S. Bird: *Recommended metadata extenxions*. OLAC Open Languange Archives Community. Online: http://www.language-archives.org/REC/olac-extensions.html (Last visit: October 24th, 2012)

[Smith 2003] B. Smith: *Ontology*. Blackwell Guide to the Philosophy of Computing and Information. L. Floridi (Ed.), pp. 155-166 *Preprint*, Oxford, Blackwell 2003.

[de Solla Price 1963] *Little Science, Big Science*. Columbia University Press. 1963.

[Søndergaard et al. 2003] T.F. Søndergaard, J. Andersen, B. Hjørland: *Documents and the Communication of Scientific and Scholarly Information: Revising and Updating the UNISIST Model.* Journal of Documentation. Vol. 59 (3), pp. 278-320, 2003.

[Sowa 2007] J. Sowa: *Fads and Fallacies about Logic. Preprint* Intelligent Systems. Vol. 22 (2), pp. 84-89, March 2007.

[Sowa 1999] John F. Sowa. *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Brooks Cole Publishing Co., Pacific Grove, Canada August 1999.

[Spree 2002] U. Spree: *Information für alle? Seit wann und warum ist der Begriff Information im öffentlichen Diskurs so reizvoll?* Info. Vol. 7 (17), pp. 5-18, 2002.

[Spyns et al. 2002] P. Spyns, R. Meersman, M. Jarrar: *Data Modelling versus Ontology Engineering.* Newsletter. ACM SIGMOD Record, Vol. 31 (4), pp. 12-17, ACM New York, NY, United States, December 2002.

[Stonebraker & Hellerstein 2005] M. Stonebraker, J. M. Hellerstein: *What Goes Around Comes Around.* In Readings In Database Systems (Fourth Edition). M. Stonebraker, J.M. Hellerstein (Eds.), MIT Press, MA, pp. 2-41, 2005.

[Storey 1993] V.C. Storey: *Understanding Semantic Relationships*. VLDB Journal, Vol. 2 (4), pp. 455-488, October 1993.

[Stracke 2010] C. M. Stracke: *The Benefits and Future of Standards: Metadata and Beyond.* In Proceedings: Metadata and Semantics Research (MTSR 2010). Communications in Computer and Information Science. Vol. 108, pp. 354-361, December 2010.

[Swan 2012] A. Swan: *Policy Guidelines for the Development and Promotion of Open Access*. Published by the United Nations Educational, Scientific and Cultural Organization (UNESCO). 7, Place de Fontenooy, Paris, France, 2012

[Technology Watch Report 2008] M. Bijsterbosch, M. K. Elbaek, P. Hochstenbach, J. Ludwig, G. Schmeltz Pedersen, R. Russell, B. Schmidt, B. Sierman, M. Vanderfeesten, K. van Godtsenhoven: *Technology Watch Report. Deliverable D4.3*, Digital Repository Infrastructure Vision for European Research II (DRIVER II). EC-funded project. 2008.

[Thomas 2006] O. Thomas: *Management von Referenzmodellen.* Entwurf und Realisierung eines Informationssystems zur Entwicklung und Anwendung von Referenzmodellen. P. Loos (Eds.): Dissertation zur Erlangung eines Doktors der Wirtschaftswissenschaft der Rechts- und Wirtschaftswissenschaftlichen Fakultät der Universität des Saarlandes.

[Uhlir & Schroeder 2007] P.F. Uhlir and P. Schröder: *Open Data for Global Science*. Data Science Journal. Vol. 6, pp. 36-53, June 17, 2007.

[Uschold & Gruninger 1996] M. Uschold and M. Gruninger: *Ontologies: Principles, Methods and Applications.* Knowledge Engineering Review. Vol. 11 (2), June 1996.

[UNISIST 1974, 1981, 1986] H. Dierickx and A. Hopkinson:. *UNISIST Reference Manual for Machine-Readable Bibliographic Descriptions. 3rd. ed.* Paris: UNESCO. 2nd ed. (1981). 1st ed. (1974). Online: http://unesdoc.unesco.org/images/0000/000062/006279EB.pdf (Last visit: October 24th, 2012)

[UNISIST 1971] United Nations Educational, Scientific and Cultural Organization and the International Council of Scientific Unions: *Study Report on the feasibility of a World Science Information System*. UNESCO 1971, Place de Fontenoy, 75 Paris 7e. Printed by Imprimerie Copedith, 1971 Printed in France.

[Uszkoreit 2007] H. Uszkoreit: *Methods and Applications for Relation Detection.* In Proceedings: Third IEEE International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE 2007). pp. 6-10, Beijing 2007.

[Uszkoreit 2006] H. Uszkoreit: Language Technology. A First Overview. Online: https://helda.helsinki.fi/bitstream/handle/10138/29375/sprakvisreport.pdf?sequence=2 (Last visit: October 25th, 2012).

[Uszkoreit et al. 2003] H. Uszkoreit, B. Jörg, G. Erbach: *An Ontology-based Knowledge Portal for Language Technology*. In Proceedings: ENABLER/ELSNET Workshop International Roadmap for Language Resources, Paris. 2003.

[Uszkoreit 1999] H. Uszkoreit: *Das Wesen der Information.* Sprachtechnologie für die Wissensgesellschaft: Herausforderungen und Chancen für die Computerlinguistik und die theoretische Sprachwissenschaft. F. Meyer-Krahmer and S. Lange (Eds.). Series Geisteswissenschaften und Innovationen, Physica Verlag, 1999.

[Vessey & Glass 1998] I. Vessey, R. Glass: *Strong vs. Weak - Approaches to Systems Development*. Communications of the ACM. Vol. 41 (4), pp. 99-102, April 1998.

[W3C 2009] W3C OWL Working Group (Eds.): *OWL 2 Web Ontology Language Document Overview.* Online: http://www.w3.org/TR/owl2-overview/ (Last visit: April 2nd, 2012)

[W3C 2004] D.L. McGuinness, F. van Harmelen (Eds.): *OWL Web Ontology Language Overview.* W3C Recommendation 10 February 2004. Online: http://www.w3.org/TR/owl-features/ (Last visit: October 24th, 2012)

[Wand et al. 1999] Y. Wand, V.C. Storey, R. Weber: *An Ontological Analysis of the Relationship Construct.* Conceptual Modeling. ACM Transactions on Database Systems. Vol. 24 (4), pp. 494-528, 1999.

[Wand et al. 1995] Y. Wand, D.E. Monarchi, J. Parsons, C. C. Woo. *Theoretical Foundations for Conceptual Modelling in Information Systems Development.* Decision Support Systems. Vol. 15 (4), pp. 285-304, December 1995.

[Wand & Weber 2006] Y. Wand, R. Weber: *On Ontological Foundations of Conceptual Modelling: A Response to Wyssusek.* Scandinavian Journal of Information Systems. Vol. 18 (1), pp. 127-138, 2006.

[Wand & Weber 1993] Y. Wand, R. Weber: *On the Ontological Expressiveness of Information Systems Analysis and Design Grammars.* Information Systems. Vol. 3 (4), pp. 217-237, 1993.

[Wand & Weber 1990] Y. Wand and R. Weber: *An Ontological Model of an Information System.* IEEE Transactions on Software Engineering, Vol. 16 (11), pp. 1282-1292, November 1990.

[Wand & Weber 1990a] Y. Wand, R. Weber: *Mario Bunge's Ontology as a Formal Foundation for Information Systems Concepts.* Studies on Mario Bunge's Treatise. P. Weingartner and G.J.Dorn (Eds.), pp. 123-149, 1990.

[Wand & Weber 1988] Y. Wand, R.Weber: *An Ontological Analysis of Some Fundamental Information System Concepts*. Proceedings: International Conference on Information Systems. DeGross and M.H. Olson (Eds.), pp. 213-225, 1988.

[Welty 2003] C. Welty: *Ontology Research*. AI Magazine; AAAI, Vol. 24 (3), November 2003.

[Welty & Fikes 2006] C. Welty, R. Fikes: *A Reusable Ontology for Fluents in OWL*. In Proceedings: Conference on Formal Ontology in Information Systems (FOIS 2006), pp. 226-235. B. Bennet, C. Fellbaum (Eds.), IOS Press, Amsterdam, 2006.

[Wilks 2008] Y. Wilks: *The Semantic Web: Apotheosis of Annotation, but what are its Semantics?*
IEEE Intelligent Systems. Vol. 23 (3), pp. 41-49, 2008.

[Wolski et al. 2011] M. Wolski, J. Richardson, R. Rebollo; *Shared Benefits from exposing Research Data.* D-Lib Magazine. Volume 17 (5/6), May/June 2011.

[Woods 1975] W.A. Woods: *What's in a Link: Foundations for Semantic Networks*. Bolt Beranek and Newman, Inc. Prepared for the Office of Naval Research, 1975. Distributed by NTIS National Technical Information Service U.S. Department of Commerce.

[Wyssusek 2006] B. Wyssusek: *On Ontological Foundations of Conceptual Modelling*. Scandinavian Journal of Information Systems. Vol. 18 (1), pp. 63-80, 2006.

[Xu 2007] F. Xu: *Bootstrapping Relation Extraction from Semantic Seeds.* PhD-Thesis at Saarland University, 2007.

[Zachman 1987] J.A. Zachman: *A Framework for Information Systems Architecture.* IBM Systems Journal. Vol. 26 (3), pp. 276-292, 1987.

[Zachman 2003] J. A. Zachman: *The Zachman Framework For Enterprise Architecture: Primer for Enterprise Engineering and Manufacturing*. 2003. *Book Excerpt*.

[Zimmermann 2002] E. Zimmermann: *CRIS-Cross: Current Research Information Systems at a Crossroads*.
In Proceedings: Sixth International Conference on Current Research Information Systems (CRIS 2002), August 29-31, 2002, Kassel Germany.

[Zimmermann 2003] H.H. Zimmermann: *Zur Gestaltung eines Internetportals als offenes Autor-zentriertes Kommunikationssystem*, 2003. Online: http://is.uni-sb.de/zimmermann/pdf/2003c.pdf (Last visit: July 18th 2011)

[Zimmermann 1993] H.H. Zimmermann: *Aspektierung von Thesaurus Relationen, Öffnung in universale Anwendbarkeit*. In Proceedings: Deutscher Dokumentartag - Qualität und Information. Friedrich-Schiller-Universität Jena, 28-30 September, 1993.

[Zimmermann et al. 1992] H.H. Zimmermann, H.D. Luckhardt, Angelika Schulz (Hg). *Mensch und Maschine - Informationelle Schnittstellen der Kommunikation.* Hochschulverband für Informationswissenschaft (HI) e.V., Konstanz; Universitätsverlag Konstanz, 1992.