

## Statistiken für Repositories

#### 4. Leipziger Kongress für Information und Bibliothek

15.03.2010

#### **Ulrich Herb**

Saarländische Universitäts- und Landesbibliothek

<u.herb@sulb.uni-saarland.de>



## Überblick

- Anwendungsszenarien
- State of the Art
- Einige Umfrageergebnisse
- Open Access Statistik



## Unterscheidungen

- Akteure
  - Leser
  - Autor
  - Repositoryadminstrator
  - Verwaltung
- Untersuchungsobjekte
  - Repository
  - Volltexte
- Komplexität der Auswertungen
  - Reine Nutzungshäufigkeiten
  - Clickstreamanalyse
  - Netzwerkanalyse



## Anwendungsszenarien

Statistiken geben Auskunft über ...

- Relevanz eines Dokuments, Rankingelement (Leser)
- Nutzung/ Sichtbarkeit/ Impact eines Repositories (Leser, Autor, Admin, Verwaltung)
- Effiziente Gestaltung eines Repositories (Admin)
- Relevanz eines Forschungsthemas (Autor, Verwaltung)
- Impact/ Qualität eines Dokuments (Leser, Autor, Verwaltung)

### Relevanz eines Dokuments



Advanced Web Statistics 6.5 (build 1.857) - Created by awstats



## Ranking

 Statistik für:
 psydok.sulb.uni-saarland.de

 Zuletzt aktualisiert:
 10.03.2010 - 04:05

 Zeitraum:
 Jahr - ▼ 2010 ▼ 0K

Zurück zur Hauptseite

Filter: pdf Exklusiv-Filter: OK

Seiten-URL		
Filter <b>pdf</b> : 1058 Unterschiedliche Seiten	Zugriffe	durchschnitt. Größe
/volltexte/2006/583/pdf/pdf15.pdf	1376	1.08 MB
/volltexte/2004/321/pdf/BoBe1997_1.pdf	732	107.13 KB
/volltexte/2004/151/pdf/schriftenreihe_42.pdf	562	213.19 KB
/volltexte/2005/447/pdf/Herzog_Brunner_Text04.pdf	491	206.17 KB
/volltexte/2007/981/pdf/e_Buch_MENSCHENBILDER_JFahrenberg_2007	489	1.90 MB
/volltexte/2005/567/html/KapA5.pdf	462	246.91 KB
/volltexte/2004/309/pdf/FAGS-AMSE_Fragebogen.pdf	430	114.69 KB
/volltexte/2003/22/pdf/Funke_2000_Kreativitaet.pdf	425	367.81 KB
/volltexte/2009/2358/pdf/deubel.pdf	425	929.76 KB
/volltexte/2005/448/pdf/Miklautz_Jenull.pdf	412	156.75 KB
/volltexte/2004/383/pdf/Statistik_Skript_2004.pdf	392	456.39 KB
/volltexte/2006/572/pdf/4-Fachtagung-AO-Programmheft-Abstractban	372	2.63 MB
/volltexte/2006/772/pdf/ANDROGYN.pdf	351	327.95 KB
/volltexte/2004/350/pdf/report_psychologie_06-2003_1.pdf	331	284.99 KB
/volltexte/2004/315/pdf/Diplomarbeit.PDF	319	888.55 KB
/volltexte/2004/102/pdf/02_194.pdf	299	1.12 MB
/volltexte/2007/1007/pdf/Fb1033xvvt.pdf	289	1.09 MB
/volltexte/2009/2363/pdf/SPF_Artikel.pdf	273	212.59 KB
/volltexte/2006/775/pdf/Verkehrstherapie_01_2005.pdf	258	6.63 MB
/volltexte/2004/153/pdf/beitraege_is_21.pdf	248	837.87 KB
/volltexte/2007/985/pdf/Studieren_lernen.pdf	247	884.85 KB



## Ranking

#### Ranking

Bitte wählen Sie das gewünschte Rankingkriterium im Menü rechts.

Die Daten zu Nutzungs- und Zitationsrelevanz stammen jeweils aus sämtlichen verfügbaren Quellen: lokal, national\*, international und Verlagsdaten\*\*

\* Beteiligte Betreiber: Berlin, DNB, Göttingen, Saarbrücken, Stuttgart

\*\* Beteiligte Verlage: BioMed

\*\*\* In einem zweiten Schritt werden Sie gebeten die zugrundezulegende Metrik auszuwählen.

#### # Titel

- Schallehn, Volker (2007) 4 Jahre E-Books an der LMU München Erfahrungen und Perspektiven
- Schäffler, Hildegard (2007) Checkliste für Standards aus bibliothekarischer Sicht: Vorstellen der Ergebnisse der AG E-Books im Bibliotheksverbund Bayern
- Giebenhain, Sabine (2007) E-Book-Angebot in der Universität Stuttgart Zwischenbilanz nach 1 1/2 Jahren

Nutzungsrelevanz\*\*\*
Nutzungsrelevanz\*\*\*
Zitationsrelevanz
Downloads\*\*\*

9.053



## Sichtbarkeit/ Impact eines Repositories

- Indizierung durch relevante Suchmaschinen und Datenbanken?
- Ausmaß der Vernetzung/ Bestimmung der Position innerhalb der Informationsangebote: Woher kamen und gingen Nutzer?
- Inlinks: Grad der Vernetzung mit anderen Fachinformationsangeboten und Repositories?
- Inlinks (von Fachseiten): Akzeptanz in den Communities
- Besucher: Internationaliät? (sofern relevant)



## Effizienz eines Repositories

- Anzahl der wiederkehrenden Nutzer
- Einstiegs- und Ausstiegsseiten: Was lockt Nutzer? Was verprellt sie? Sind Ein- und Ausstiegsseiten identisch?
- Wie navigieren Nutzer durch ein Repository? Bruchstellen im System?
- Suchbegriffe: Zu welchen Themen finden Nutzer Informationen?

Nötig: Clickstream-Analyse, Anonymisierung/Pseudonymisierung, Beachtung des Datenschutzes



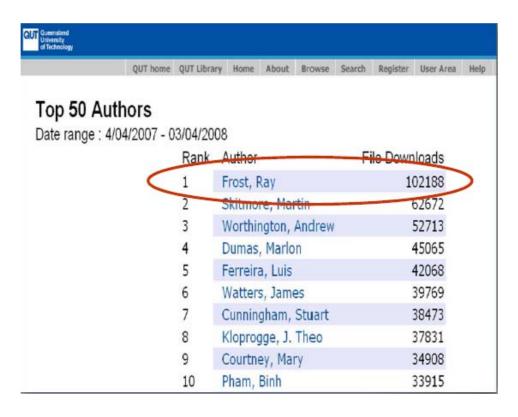
## Relevanz eines Forschungsthemas

- Autor/ Lehrstuhl/ Forschungseinrichtung: Dokumentnutzung gibt Feedback über Publikations- und Forschungsthemen
- Verwertung durch Verwaltung (Benchmarking) erfordert Sensibilität und Beachtung der Rahmenbedingungen
  - Technikaffinität der Disziplin
  - Publikations- und Rezeptionspräferenzen der Disziplin
  - Alter und Verwertungszyklus der Dokumente
  - Dokumenttyp

- ..



## Sind alle Auswertungen sinnvoll?



Tom Cochrane, University of Queensland. *Mandates: an Australian example of the Queensland University of Technology, CERN Workshop in Innovations in Scholarly Communication (OAI6)* 



## Impact eines Dokuments

#### Zitationsbasierte Maße

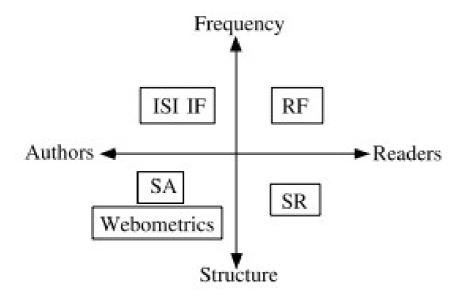
- autorenzentriert
- Messung erst in folgenden Publikationsgenerationen möglich
- Messung erfolgt i.d.R. auf Journal- oder Autorenebene, Impact eines Textes/Objekts wird nicht abgebildet

### Nutzungsbasierte Maße

- leserzentriert
- Messung live und fortlaufend möglich
- Impact eines Textes/Objekts wird abgebildet
- Stark automatisierte Messung möglich



## Impact eines Dokuments



ISI IF = Journal Impact Factor

RF = Reading Factor

SA = Structure Author

- basiert auf Netzwerk, das durch Autorenhandlung gebildet wird
- •Bsp: Google PageRank, Zitationsgraphen, Webometrics

SR = Structure Reader

- basiert auf Kontextinformationen der Dokumentnutzung
- Recommendersysteme
- •Bsp: Downloadgraphen

Bollen, Johan et al. (2005): Toward alternative metrics of journal impact: A comparison of download and citation data. In: Information Processing and Management 41(6): S. 1419-1440.



## Impact eines Dokuments

"Our results indicate that the notion of scientific impact is a multidimensional construct that can not be adequately measured by any single indicator, although some measures are more suitable than others. The commonly used citation Impact Factor is not positioned at the core of this construct, but at its periphery, and should thus be used with caution. "

"Usage-based measures such as Usage Closeness centrality may in fact be better ,consensus' measures."

Bollen, J.; Van De Sompel, H.; Hagberg, A.; Chute, R.: A principal component analysis of 39 scientific impact measures. In: PloS one 4 (2009), Issue 6, e6022. DOI: 10.1371/journal.pone.0006022.



# Einige Umfrageergebnisse



## Einige Umfrageergebnisse

Drei Erhebungen im Rahmen des Projekts Open Access Statistik

- Experteninterviews 05-06/2009
- Onlinebefragung unter Repositorynutzern 10-12/2009
- Mailbefragung unter Repository-Betreibern 09-10/2009



# Experteninterviews



## Experteninterviews 05-06/2009

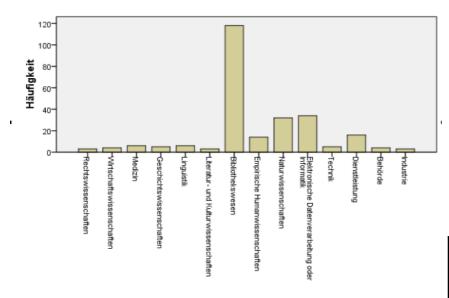
### Gewünschte nutzungsdatenbasierte Funktionalitäten

- Relevanzranking und Recommender
- Recommender nach Nutzungsmuster (implizit) und Bewertung anderer (explizit)
- Recommender nach Nutzungmuster, Zitationen und eigenen Bewertungen
- Messung der Nutzung und der in die Dokumentnutzung investierte Zeit
- Recommender anhand Verhalten anderer Nutzer und derer Profile
- Zentralität von Dokumenten in Nutzungsnetzwerken





#### In welchem thematischen Bereich sind Sie zur Zeit tätig?



Teilnehmer: 253 Personen

Bibliothekswesen	118	46,6 %
EDV/Informatik	34	13,4 %
Naturwissenschaften	32	12,6 %
Dienstleistungssektor	16	6,3 %
Sozialwissenschaften	14	5,5 %



 Bewerten Sie den Nutzen einer Anzeige der Nutzungshäufigkeit bei jedem einzelnen Dokument

		Häufigkeit	Prozent
Gültig	gar nicht nützlich	15	5,9
	selten nützlich	46	18,2
	manchmal nützlich	87	34,4
	häufig nützlich	59	23,3
	immer nützlich	35	13,8
	Gesamt	242	95,7
Fehlend	99	11	4,3
Gesamt		253	100,0



Bewerten Sie den Nutzen einer Darstellung der Nutzung aller Dokumente eines Autors

			_
		Häufigkeit	Prozent
Gültig	gar nicht nützlich	18	7,1
	selten nützlich	44	17,4
	manchmal nützlich	91	36,0
	häufig nützlich	58	22,9
	immer nützlich	31	12,3
	Gesamt	242	95,7
Fehlend	99	11	4,3
Gesamt		253	100,0



Bewerten Sie den Nutzen einer Anzeige der Gesamtnutzung eines Repositoriums

		Häufigkeit	Prozent
Gültig	gar nicht nützlich	17	6,7
	selten nützlich	60	23,7
	manchmal nützlich	83	32,8
	häufig nützlich	49	19,4
	immer nützlich	33	13,0
	Gesamt	242	95,7
Fehlend	99	11	4,3
Gesamt	<u> </u>	253	100,0



Für wie nützlich halten Sie es, wenn das System aus den Abrufzahlen die Qualität der Dokumente schätzt?

		Häufigkeit	Prozent
Gültig	gar nicht nützlich	63	24,9
	selten nützlich	69	27,3
	manchmal nützlich	73	28,9
	häufig nützlich	30	11,9
	immer nützlich	9	3,6
	Gesamt	244	96,4
Fehlend	99	9	3,6
Gesamt		253	100,0



■ Für wie nützlich halten Sie eine Sortierung der Trefferliste danach wie oft die Dokumente von anderen zitiert wurden?

			_
		Häufigkeit	Prozent
Gültig	gar nicht nützlich	6	2,4
	selten nützlich	26	10,3
	manchmal nützlich	86	34,0
	häufig nützlich	86	34,0
	immer nützlich	43	17,0
	Gesamt	247	97,6
Fehlend	99	6	2,4
Gesamt		253	100,0



Bewerten Sie den Nutzen von Empfehlungen, die darauf basieren welche Autoren beteiligt sind

		Häufigkeit	Prozent
Gültig	gar nicht nützlich	6	2,4
	selten nützlich	27	10,7
	manchmal nützlich	113	44,7
	häufig nützlich	63	24,9
	immer nützlich	37	14,6
	Gesamt	246	97,2
Fehlend	99	7	2,8
Gesamt	·	253	100,0



Wie nützlich ist es, wenn Ihnen Dokumente ausgehend von Ihrem bisherigen Leseverhalten empfohlen werden?

		Häufigkeit	Prozent
Gültig	gar nicht nützlich	8	3,2
	selten nützlich	23	9,1
	manchmal nützlich	104	41,1
	häufig nützlich	94	37,2
	immer nützlich	24	9,5
	Gesamt	253	100,0





Betreiber von 138 Repositories angeschrieben

Rücklauf: 16 Antworten

- 14 Repositories erheben Daten über Benutzerzugriffe
- 14 Repositories erheben Daten über Dokumentzugriffe, 9 zeigen die Daten intern an, 7 extern



- Benutzte Standards: COUNTER (2), AWSTATS, IRS, direkt aus Apache-Logfiles erstellt
- Wie erfolgt die Bereinigung von maschinellen Zugriffen, falls kein Standard Anwendung findet?
  - Auswertung des User-Agents in Logfiles
  - Bereinigung durch AWStats
  - Bereinigung findet statt, Antworten geben keine Auskünfte über das *wie*.



- Falls kein Standard Anwendung findet: Wie werden Doppelklicks bestimmt?
  - 30 Minuten-Intervall (1)
  - Software erlaubt keine exakte Identifikation von Doppelklicks (5)
  - Frei konfigurierbar (2)
  - Eliminierung unmittelbar aufeinander folgender Zugriffe (2)



- Software zur Generierung der Daten
  - AWStats (6, davon einmal eigene Code-Erweiterungen)
  - Apache Log Files (5)
  - Logfiles + eigene Scripte (2)
  - DSpace (sessionbasierte Statistik) (1)
  - Eigenentwicklung, nicht auf Logfiles basiert (1)
  - OPUS-Statistik (1)
  - EPrints-Statistik (1)



- Mindestanforderungen an Statistiken auf Dokumentebene aus Sicht der Repositorybetreiber
  - Verlässliche Aussagen über Dokumentnutzung (13)
  - Eliminierung von maschinellen Zugriffen (3) und Doppelklicks (1)
  - Unterscheidung der Dokumentnutzung nach Campus-Nutzern und Nicht-Campus-Nutzern (1)
  - Beachtung datenschutzrechtlicher Bestimmungen (1)
  - Geringere Manipulierbarkeit, hohe Granularität (1)



- Von der Einführung dokumentbezogener Statistiken erhoffte Effekte
  - Nutzung als Rankinginformation (6)
  - Teaser für Autoren (6)
  - Analyse der Repositorynutzung (2)
  - Steigerung der Attraktivität des Repositories (2)
  - Entwicklung von Mehrwertdiensten (1)
  - Nutzung als Qualitätskriterium (1)
  - Transparente Information über Nutzung (1)
  - Organisationsinterne Argumentationshilfe (1)



#### Erwünschte Zusatzfunktionen

- Analyse von Nutzungswegen (5)
- Einblendung der meistgenutzten Dokumente (3)
- Nutzungsanalyse: Browser, Herkunft, Tageszeit (3)
- Schaffen eines Standards (2)
- Statistiken auf Autoren- und Artikelebene (2)
- Unterstützung des VG Wort Mechanismus (2)
- Recommenderdienste (2)
- Nutzungsbasierter Impact-Factor (1)



#### **Fazit**

Die Verwendung von Nutzungsstatistiken zur Bestimmung der

- der Relevanz eines Dokuments
- der Beliebtheit/Sichtbarkeit/ Impact eines Repositories
- der effizienten Gestaltung eines Repositories
- der Relevanz eines Forschungsthemas
- des Impacts/ der Qualität eines Dokuments
- ... benötigt standardisierte und verlässliche Daten



#### Standards?

COUNTER

**PIRUS** 

LogEc

**IFABC** 

**AWStats** 

Mesur





# LogEc





Counting Online Usage of NeTworked Electronic Resources



## Open Access Statistik

- gefördert durch die DFG (05/2008 - 02/2010)
- DINI



- initiiert von DINI
- Projektpartner

Niedersächsische Staats- und Universitätsbibliothek Göttingen Saarländische Universitäts- und Landesbibliothek Universitätsbibliothek Stuttgart

Humboldt Universität zu Berlin (Computer- und Medienservice)

http://www.dini.de/oa-statistik



# Partnerprojekte

DOARC (Distributed Open-Access Reference Citation services): Zitationsanalyse

#### Open Access Netzwerk

- Bildung eines Repository-Netzwerks
- bündelt die Ergebnisse von OA-S und DOARC
- bietet Services für OA-S und DOARC (z.B. Dokumentdeduplizierung)



#### Skizze

- Aufbau einer dauerhaften Infrastruktur zur Erfassung und Verarbeitung von Nutzungsdaten
- Entwicklung und Etablierung eines einheitlichen Standards zur Ermittlung von Zugriffszahlen und Statistiken
- Anbieten von Serviceleistungen für Repositorien-Betreiber und Nutzer
- Entwicklungen von Richtlinien zur Implementierung der Services
- Empfehlungen im DINI-Zertifikat



#### Skizze

■Aggregation von Nutzungsdaten verschiedener Open-Access-Angebote/ Server in einem Serviceprovider (v.a. Webserverlogs und Linkresolverlogs), Psyeudonymisierung

■Bereinigung von Verzerrungen (maschinelle Zugriffe durch Spider, Dublettenerkennung etc.)

■Aufbereitung der Nutzungsdaten anhand verschiedener Standards (COUNTER, LogEc, IFABC)

□Rückspielen der ermittelten Kennwerte in die verteilten Angebote und externen Services (OA Netzwerk), Verwendung als Metadatum



# Repository-Ansicht

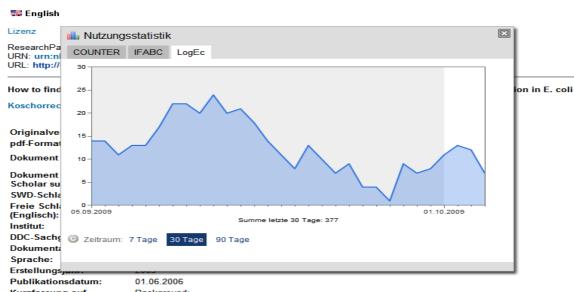




Home | Recherche | Veröffentlichen | Elektronische Dissertationen | Hilfe | Dokumentation | English

**OPUS** 

#### **Eingang zum Volltext in OPUS**



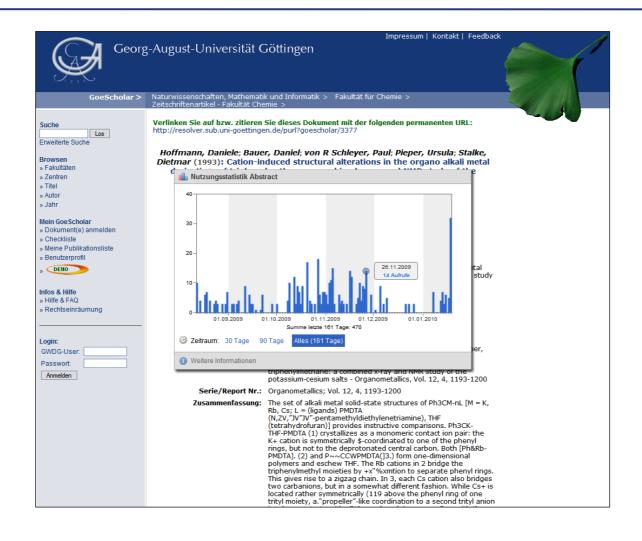
Kurzfassung auf Background: Englisch:

In screening of libraries derived by expression cloning, expression of active proteins in E. coli can be limited by formation of inclusion bodies. In these cases it would be desirable to enrich gene libraries for coding sequences with soluble gene products in E. coli and thus to improve the efficiency of screening. Previously Wilkinson and Harrison showed that solubility can be predicted from amino acid composition (Biotechnology 1991, 9(5):443-448). We have applied this analysis to members of the alpha/beta hydrolase fold family to predict their solubility in E. coli. alpha/beta hydrolases are a highly diverse family with more than 1800 proteins which have been grouped into homologous families and superfamilies.

The predicted solubility in F coli depends on hydrolase size phylogenetic origin of the host organism, the homologous



# Repository-Ansicht





# Teilnahmebedingungen

- Implementierung eines Data-Providers, der die Nutzungsdaten zum Abruf bereithält
  - OPUS- und DSpace-Module werden von OA-Statistik bereitgestellt
- Hilfedokumente werden zur Verfügung gestellt
- Integration der Nutzungsstatistiken in die eigene Umgebung



## Teilnahmebedingungen

- Die Spezifikation liegt in einer ersten Draft-Version vor
  - Specification: Data Format and Exchange for OA Statistics, Version 0.5
- Auf der Projektwebseite:

http://www.dini.de/projekte/oa-statistik/projektergebnisse/



## Internationale Abstimmung

- SURFSure Statistics on the Usage of Repositories
- COUNTER Counting Online Usage of Networked Electronic Resources
- PIRUS Publisher and Institutional Repository Usage Statistics
- NEEO Network of European Economists Online
- PEER Publishing and the Ecology of European Research
- OAPEN
   Open Access Publishing in European Networks



## Open Access Statistik II

Ziele:

- □Internationalisierung
- ■Standardisierung von Kennwerten, die auf reiner Nutzungshäufigkeit basieren
- Integration neuer Services (Journals, Repositories)
- ■Erprobung von Kennwerten, die auf komplexeren Analysen (vornehmlich Netzwerkanalyse der Dokumentnutzung) beruhen
- Implementierung von Zusatzservices (Empfehlungsdienste)

## ... die Anwendungsszenarien

- Relevanz eines Dokuments, Rankingelement (OAS I)
- Nutzung/ Sichtbarkeit/ Impact eines Repositories (auf lokaler Ebene mit OAS I möglich, Pilot in OAS II)
- Effiziente Gestaltung eines Repositories (auf lokaler Ebene mit OAS I möglich, Pilot in OAS II)
- Relevanz eines Forschungsthemas (OAS I)
- Impact/ Qualität eines Dokuments (OAS I, OAS II)
- N.B.: Die Aussagekraft steigt mit breiterer Datenbasis.





#### Vielen Dank für Ihre Aufmerksamkeit!

#### 4. Leipziger Kongress für Information und Bibliothek

15.03.2010

#### **Ulrich Herb**

Saarländische Universitäts- und Landesbibliothek

<u.herb@sulb.uni-saarland.de>

