Open-Access-Statistik

# Alternative impact measures for open access documents? An examination how to generate interoperable usage information from distributed open access services

**IFLA 2010**

**Session 72 - Statistics and Evaluation**

**Göteborg, 11.08.2010**

Initiated by:

DINI – Deutsche Initiative für Netzwerkinformation e.V.

**Ulrich Herb**
Saarland University and State Library, Germany
u.herb@sulb.uni-saarland.de

Funded by:

DFG

# Overview

- Impact measures:
  - relevance
  - a categorisation
- Usage-based impact measures: standardisation?
- Project: Open Access Statistics
  - Aims
  - Technical infrastructure
  - Results
  - Outlook

Open-Access-Statistik

# Impact Measures

„The ‚impact factor' is the most commonly used assessment aid for deciding which journals should receive a scholarly submission or attention from research readership. It is also an often misunderstood tool."
Dong et al. 2005

Open-Access-Statistik

# Impact measures: relevance

- Individual level: *publish or perish*
  - If a scientist does not publish she/he does not have any scientific capital, reputation or impact
  - Without any impact, she/he won't make her/his career
- Organisational level: evaluation
  - Evaluation results determine prospective resources of institutes *and* the future main research
  - Criteria: number of doctoral candidates, amount of third party funds, publications

Open-Access-Statistik

# From publications to impact

- Scientific reputation (or scientific capital) is derived from publication impact

- Impact is calculated mostly by citation measures
  - Journal impact factor (JIF)
  - Hirsch-index (h-index)

  *Especially within the STM domain*

Open-Access-Statistik

# Citation impact: calculation

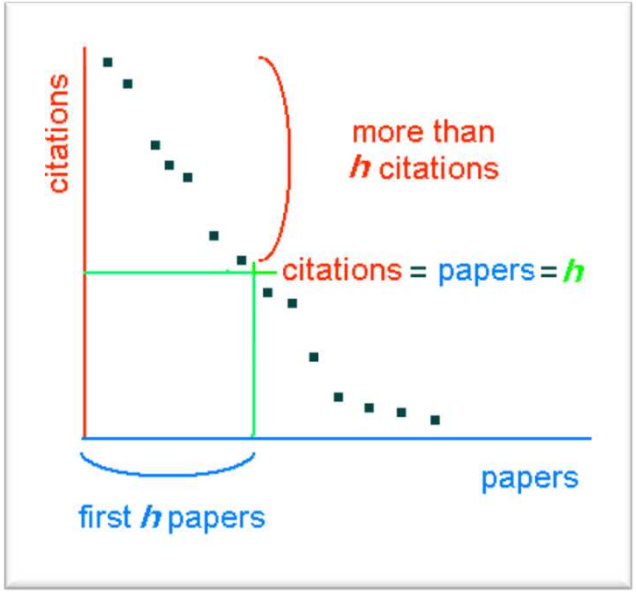
**JIF**

In year X, the impact factor of a journal Y is the average number of citations to articles that were published in Y during the two years preceding X

Garfield: „*We never predicted that people would turn this into an evaluation tool for giving out grants and funding.*" From: Richard Monastersky (2005), The Number That's Devouring Science *The Chronicle of Higher Education*

**H-index**

A scientist has index h if h of N papers have at least h citations each, and the other (N − h) papers have less than h citations each




http://de.wikipedia.org/wiki/H-Index

Open-Access-Statistik

# Citation impact: critical points

- Restricted scope, exclusion of many publication types
- Based exclusively on journal citation reports / web of science (JIF) or other databases
- Language bias: items in English language are overrepresented within the database, so they reach higher citation scores
- JIF focuses on journals: few articles evoke most citations
- JIF discriminates disciplines with lifecycles of scientific information > 2 years

→ Mixture of quality and popularity

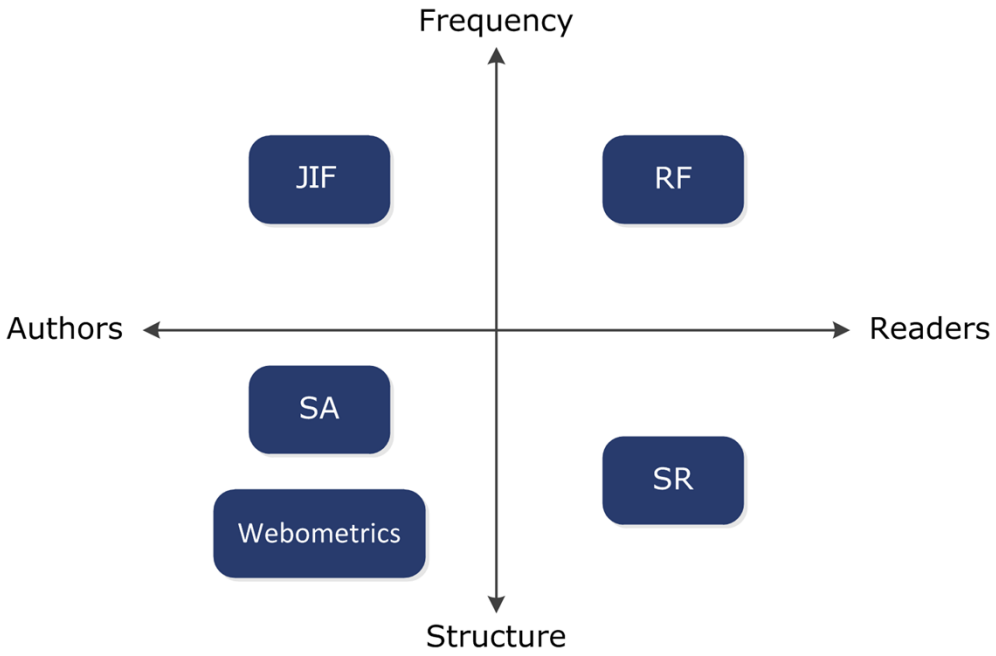
Open-Access-Statistik

# Impact measures: a categorisation

- Citation based measures
  - Author-centred
  - Delayed measurement: at first in the following generation of publications
  - Impact of a separate object is mostly not described
- Usage based measures
  - Reader-centred
  - Measuring: on-the-fly and consecutive
  - Impact of a separate object can be described
  - Automated measurement is possible

Open-Access-Statistik

# Impact measures: a categorisation, pt. II




**JIF = Journal Impact Factor**

**RF = Reading Factor**

**SA = Structure Author**

- **based on networks built by authors and their activities, e.g. Google PageRank, citation graphs, webometrics**

**SR = Structure Reader**

- **based on document usage and its contextual information, e.g. recommenders, download graphs**

Bollen, J. 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.
Preprint Online: http://arxiv.org/abs/cs.DL/0503007

Open-Access-Statistik

# Standards

*„An important issue, however, was the lack of standards on how to produce and report the usage data in a way that could be compared*"
Baker et al. 2008

Open-Access-Statistik

# Usage based impact: standardisation?

- COUNTER
  Counting Online Usage of NeTworked Electronic Resources

  http://www.projectcounter.org

- LogEc

  http://logec.repec.org/

- IFABC

  http://www.ifabc.org/

Open-Access-Statistik

# Usage based impact: standardisation?

- The models mentioned differ in many aspects
  - Detection and elimination of non-human access (robots, automatic harvesting)
  - Definition of double click intervals
  - …
- General problems
  - Ignorance of context information
  - Detection of duplicate users
  - Detection of duplicate information items
  - Ignorance of philosophical questions like: “What degree of similarity makes two files the same document?”

Open-Access-Statistik

# Alternative impact measures: conclusion

- Alternative impact measures are possible
- But: very little standardisation
- Promising, but complex examples/models like MESUR
  http://www.mesur.org
- Requirement: sophisticated infrastructure to generate and exchange interoperable usage information within a network of several different servers

Open-Access-Statistik

# Project:
# Open Access Statistics

Open-Access-Statistik

# Open Access Statistics (OAS)

- 07/2008 – 02/2010
- Project partners:

Universität Stuttgart

HUMBOLDT-UNIVERSITÄT ZU BERLIN

NIEDERSÄCHSISCHE STAATS- UND UNIVERSITÄTSBIBLIOTHEK GÖTTINGEN

SAARLÄNDISCHE UNIVERSITÄTS-UND LANDESBIBLIOTHEK

Initiated by:

DINI – Medienzentren, Bibliotheken, Rechenzentren, Wissenschaft – DEUTSCHE INITIATIVE FÜR NETZWERKINFORMATION E.V.

Funded by:

DFG

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

oas Open-Access-Statistik

# Open Access Statistics: motivation

- open access publications are often excluded from citation based impact measures
  - repository documents by definition
  - articles in open access journals due to their short citation history and often also due to their language
- citation based impact measures are revealing several deficiencies
- citation based impact measures should be complemented by usage based impact measures
  - because a multi-faceted approach could remedy some of their deficiencies
  - because the latter ones could create an incentive to use open access services

Open-Access-Statistik

# OAS: aims

- A common standard to exchange usage data between different services
- An infrastructure to collect, process and exchange usage information between different services
- Usage information should be processed according to the standards of COUNTER, LogEc and IFABC
- Additional service for repositories
- Implementation guidelines

Open-Access-Statistik

# OAS: associated projects

- Open Access Statistics


- DOARC
  (Distributed Open Access Reference and Citation Services)

- Open Access Network


Open-Access-Statistik

# OAS: associated Projects

- Open Access Statistics addresses usage description
- DOARC address the issue of tracking citations between electronic publications
- Open Access Network
  - intends to build a network of repositories
  - will bundle the results of DOARC and Open Access Statistics in one user interface
  - offers services for DOARC and Open Access Statistics, e.g. deduplication of documents (based on a asymmetric similarity of fulltext documents)

Open-Access-Statistik

# Technical Infrastructure

*„Collecting, processing, and interpreting usage data is a challenge for libraries, big and small"*
Manoff et al. 2006

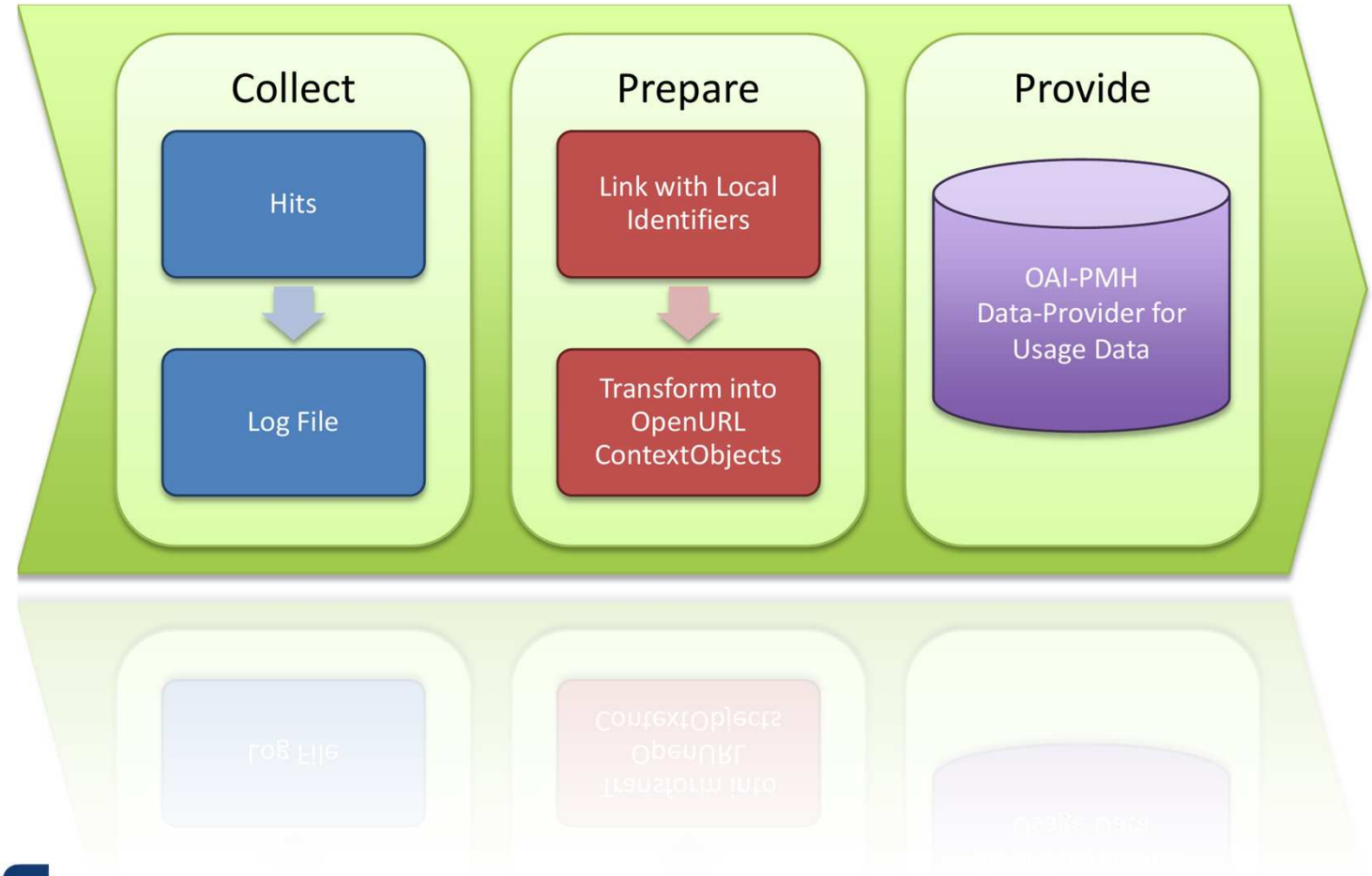
Open-Access-Statistik

# OAS: background

- Data pools at partner institutions
- Aggregation of usage events in a central service provider
- Services provided by the central service provider
- Usage data will be retransferred to the local data pools and to the Open Access Network Service

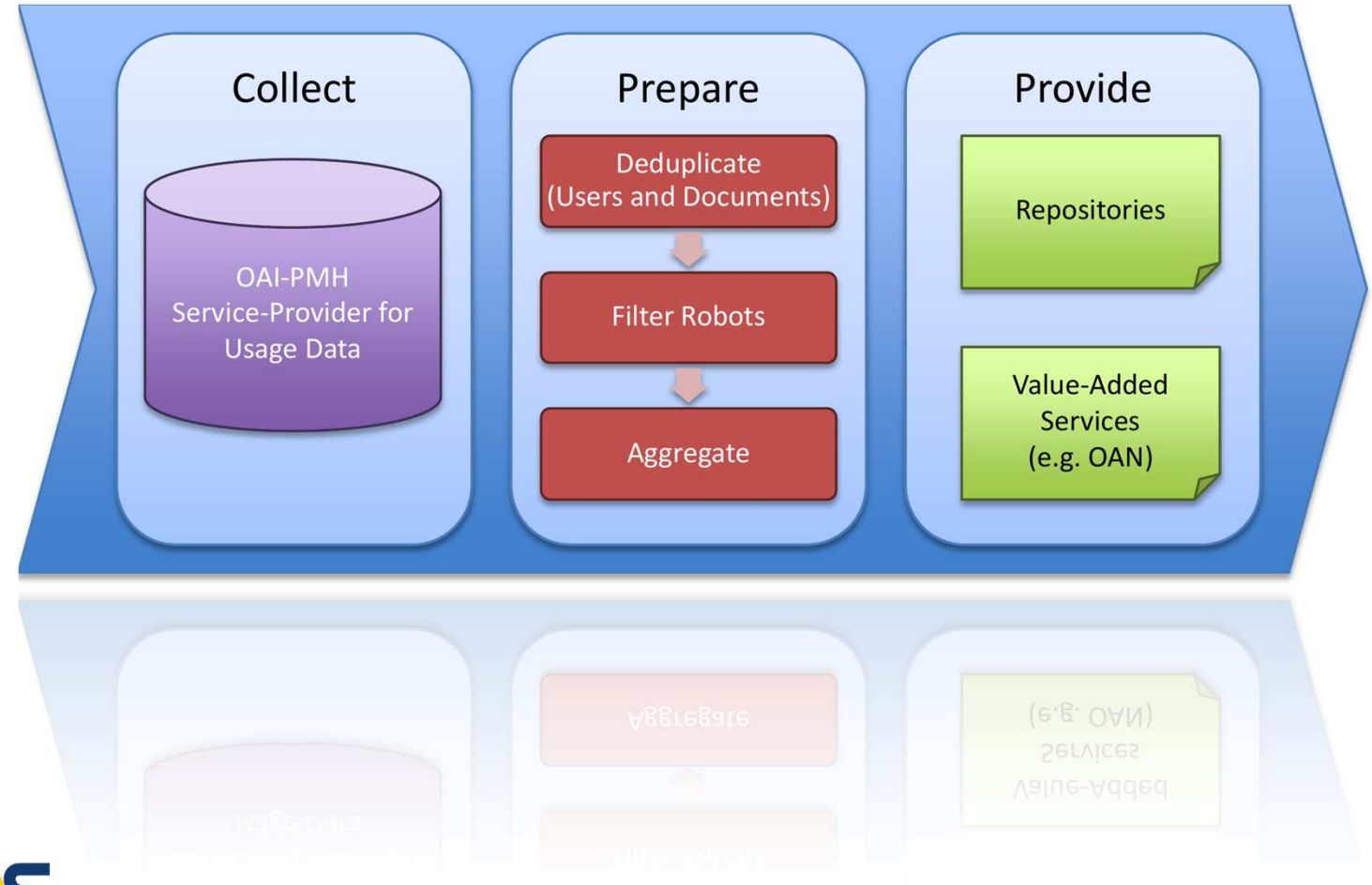
Open-Access-Statistik

# OAS: data provider



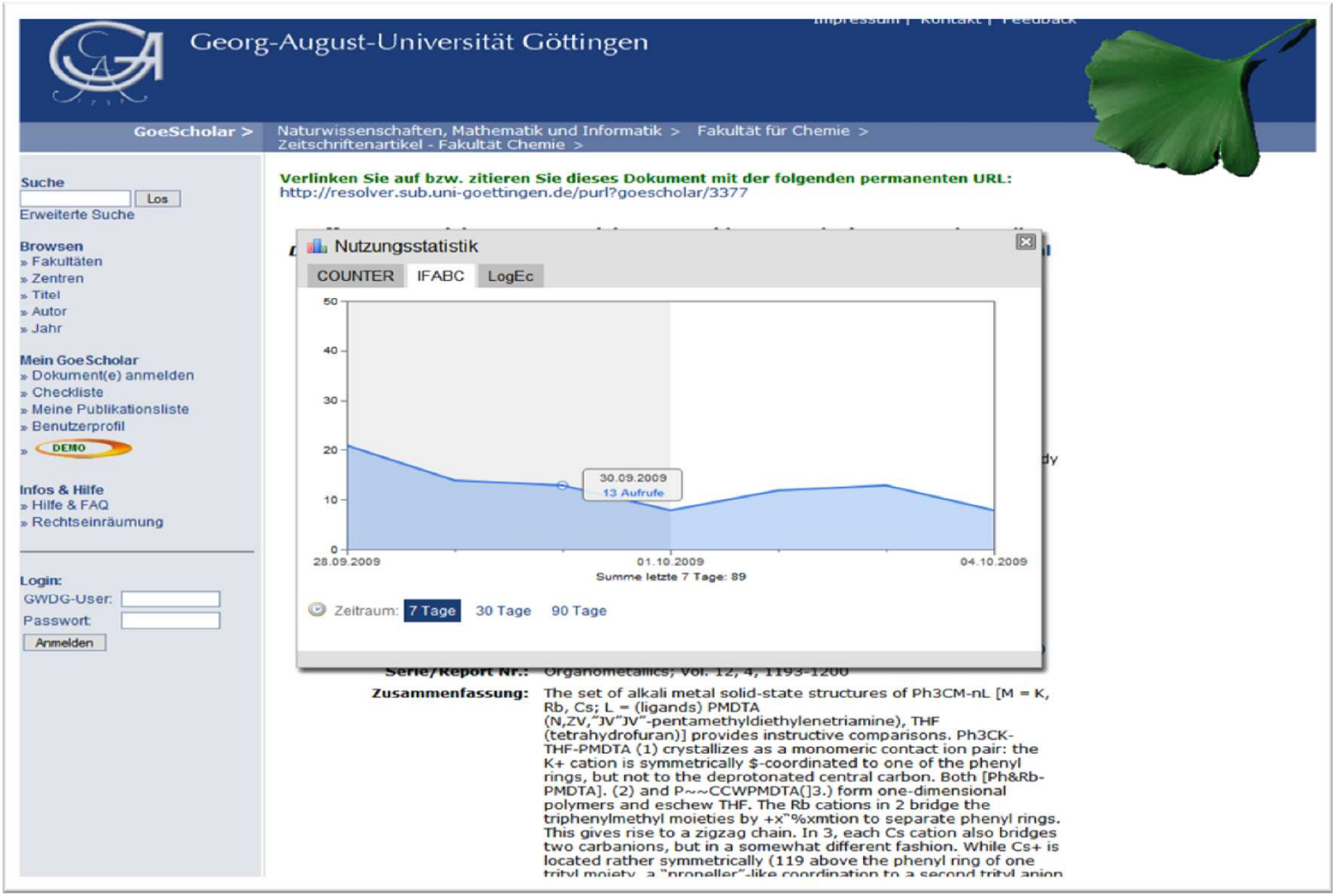

Open-Access-Statistik

# OAS: service provider




Open-Access-Statistik

# OAS: repository integration




Open-Access-Statistik

# OAS: usage scenarios

data may be used

- from an user perspective as a criterion to estimate the relevance of a document (e.g. rankings)
- from an author perspective as an indicator for the dissemination of a concept
- from a service provider perspective:
  - as additional metadata for search engines, databases …
  - as a recommender service
- from a repository perspective:
  - as a recommender service
  - as additional metadata for users

Open-Access-Statistik

# Results and Outlook

Open-Access-Statistik

# OAS: lessons learned

linkresolvers are rarely offering suitable information

- external services (ovid) don’t offer usage information
- SFX-logs are very heterogenous
  - target may be a splash page or a fulltext
- hardly any information about open access documents

document deduplication seems difficult

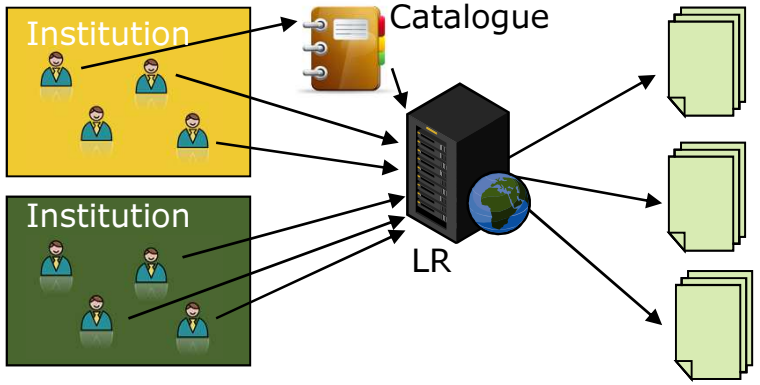
- a given document may have more than one IDs
  cause: multiple fulltext deposit on several repositories
- a given document may have several splash pages on different servers pointing at one fulltext on one single server
  cause: metadata harvesting
- …

Open-Access-Statistik

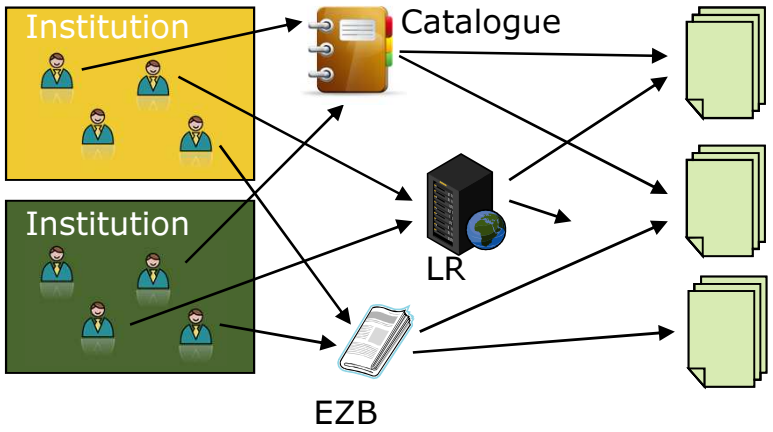
# OAS: lessons learned

- The requirement for a central clearing house
- A lot of unnecessary data (OpenURL CO)
  → increase of the data size by factor ~10
- Different situation with Linkresolver

**USA**




**Germany**




Open-Access-Statistik

# OAS: results

- Infrastructure for exchange usage statistics
- Modules for OPUS- and DSpace-based repositories, other products can be configured easily (http://www.dini.de/projekte/oa-statistik/english/software/)
- Specification of the data format and exchange
- Online demo (http://oa-statistik.sub.uni-goettingen.de/statsdemo)
- Website with further information (http://www.dini.de/projekte/oa-statistik/english/)

Open-Access-Statistik

# OAS: further plans → OAS 2

Aims for a possible second funding:

- Opening the OAS infrastructure to offer standardised usage statistics
- Evaluation of metrics more sophisticated than the calculation of pure usage frequencies
- Cooperation for international comparable usage statistics
- Offer a suitable service infrastructure

Open-Access-Statistik

# OAS: international cooperation

- SURFSure
- COUNTER
- PIRUS
- Knowledge Exchange – Usage Statistics Group
- NEEO
- PEER
- OAPEN

Open-Access-Statistik

Open-Access-Statistik

# Thanks for your attention!

**IFLA 2010**

**Session 72 - Statistics and Evaluation**

**Göteborg, 11.08.2010**

Initiated by:

DINI – Deutsche Initiative für Netzwerkinformation e.V.

**Ulrich Herb**
Saarland University and State Library, Germany
u.herb@sulb.uni-saarland.de

Funded by:

DFG