Riding the Wave - Paradigm shifts in Information Access

Jan Brase, DataCite
eResearch Australia Conference
November 9th
Melbourne

Thousand years ago:
science was empirical
describing natural phenomena

Last few hundred years:
thetical branch
using models, generalizations

Last few decades:
a computational branch
simulating complex phenomena

Today:
data exploration (eScience)
unify theory, experiment, and simulation

Jim Gray, eScience Group, Microsoft Research
Consequences for Libraries

Scientific Information is more than a published article or a book
Libraries should open their catalogues to this non-textual information

The catalogue of the future is NOT ONLY a window to the library’s holding, but

A portal in a net of trusted providers of scientific content

We do not have it
And here is the link to it!

BUT
TIB

German National Library of Science and Technology

- Architecture
- Chemistry
- Computer Science
- Mathematics
- Physics
- Engineering technology

Global Supplier for scientific and technical documents

Financed by Federal Government and all Federal States

- € 18 mio. annual acquisition budget
- 18,500 journal subscriptions
- 7.0 mio. items

Including non-textual content

- Scientific Films
- 3D Objects
- Simulation
- Research Data
- Text
- Software
https://getinfo.de/app

Move beyond text - Tools

Make more scientific and technical content searchable

Develop tools to address each type of scientific and technical Information
   ➔ Present systems are designed to handle text formats
Example from architecture

Indexing and search

visual search

content based indexing
How it works

- content based indexing
  - segmentation with form-primitives
  - extraction of room connectivity graphs
- visual search
  - attributed graph
  - 3D sketch

(a) corridor type
(b) distributor type
classification of floor types
→ machine learning

Result visualization

Extracting the information from the text

Chemical Names

Linked entities from the table

Reaction scheme

Table with reaction scheme

2a-i: Derivates from the reaction

Chemical structure
Chemical search

![Chemical search interface](image)

Content based search

![Content based search interface](image)
Visual search approach

Visual Search in Time series
- Query-by-Example, Query-by-Sketch
- Visual Catalog as result list
- Colormaps for the indication of similarity

Data
Problem with data: The research trajectory

Data are synthesised and interpreted. They become Information that is published and becomes Knowledge. However, data are often lost, become inaccessible, or not traceable.

Digital Object Identifiers (DOI names) offer a solution.

- Mostly widely used identifier for scientific articles
- Researchers, authors, publishers know how to use them
- Put datasets on the same playing field as articles

DOI names for data

URLs are not persistent

- (e.g. Wren JD: URL decay in MEDLINE - a 4-year follow-up study. Bioinformatics. 2008, Jun 1;24(11):1381-5).

Dataset

Yancheva et al (2007). Analyses on sediment of Lake Maar. PANGAEA.

doi:10.1594/PANGAEA.587840

Digital Object Identifiers (DOI names) offer a solution

- Mostly widely used identifier for scientific articles
- Researchers, authors, publishers know how to use them
- Put datasets on the same playing field as articles
What if data would be citable?

High visibility of the data
Easy re-use and verification of the data sets.
Scientific reputation for the collection and documentation of data (Citation Index)
Encouraging the Brussels declaration on STM publishing
Avoiding duplications
Motivation for new research

How to achieve this?

Science is global
- It needs global standards
- Global workflows
- Cooperation of global players

Science is carried out locally
- By local scientists
- Being part of local infrastructures
- Having local funders
Global consortium carried by local institutions focused on improving the scholarly infrastructure around datasets and other non-textual information focused on working with data centres and organisations that hold data Providing standards, workflows and best-practice Initially, but not exclusively based on the DOI system Founded December 1st 2009 in London

DataCite

History

- 03 05
  - DFG funded project with German WDCs
- 03. 09
  - TIB begins to issue DOI names for datasets
- 12. 09
  - Paris Memorandum
- 12 members
- 06. 10
  - DataCite Association founded in London
  - 7 members
  - All members assigned DOIs
  - Over 800,000 items registered
  - Pilot projects with Data Centres
  - Over 1.2 million DOI names
  - 15 members
- 06. 11
  - Metadata store
DataCite members

- Technische Informationsbibliothek (TIB)
- Canada Institute for Scientific and Technical Information (CISTI), California Digital Library, USA
- Purdue University, USA
- Office of Scientific and Technical Information (OSTI), USA
- Library of TU Delft, The Netherlands
- Technical Information Center of Denmark
- The British Library
- ZB Med, Germany
- ZBW, Germany
- Gesis, Germany
- Library of ETH Zürich
- L’Institut de l’Information Scientifique et Technique (INIST), France
- Swedish National Data Service (SND)
- Australian National Data Service (ANDS)

Affiliated members:
- Digital Curation Center (UK)
- Microsoft Research
- Interuniversity Consortium for Political and Social Research (ICPSR)
- Korea Institute of Science and Technology Information (KISTI)

DataCite structure
DataCite’s main goals

- Act as DOI registration agency
- Actively involved in developing standards and workflows
- Central portal allowing access to the metadata from all registered objects.
- Community for exchange of all relevant stakeholders in the area access to and linking of data (data centers, publishers, libraries, research organisations, science unions, funders)

DataCite in 2011

- Over 1,000,000 DOI names registered so far
- DataCite Metadata schema published (in cooperation with all members) [http://schema.datacite.org](http://schema.datacite.org)
- Beta Version of DataCite MetadataStore online in July [http://search.datacite.org](http://search.datacite.org)
DataCite Content Service

Service for displaying DataCite metadata

Different formats (BibTeX, RIS, RDF, etc.)

Content Negotiation (through MIME-Typ)

- Access through DOI proxy (http://dx.doi.org)
- First implemented by CNRI and CrossRef:

Alpha available:

http://data.datacite.org

Examples

```
curl -L -H "Accept: application/x-datacite+text"
  "http://dx.doi.org/10.5524/100005"
⇒ Li, J; Zhang, G; Lambert, D; Wang, J (2011): Genomic data from

curl -L -H "Accept: application/rdf+xml"
  http://dx.doi.org/10.5524/100005
⇒ RDF-file

curl -L -H "Accept: application/raw" http://dx.doi.org/10.5524/100005
⇒ ?
```
What type of data are we talking about?

Anything that is the foundation of further research is research data

Data is evidence

Citation

The dataset:
http://dx.doi.org/10.1594/PANGAEA.724325

Is supplement to the article:
Storz, David; Schulz, Hartmut; Waniek, Joanna J; Schulz-Bull, Detlef; Kucera, Michal (2009): Seasonal and interannual variability of the planktic foraminiferal flux in the vicinity of the Azores Current. Deep-Sea Research Part I-Oceanographic Research Papers, 56(1), 107-124,
http://dx.doi.org/10.1016/j.dsr.2008.08.009
Beyond citation

We don’t use the Web. Berners-Lee created the Web as a scholarly communication tool. Today the Web has changed *everything but* scholarly communication. Online journals are essentially paper journals, delivered by faster horses. But journals and citation are technology of the 18th century.

Future

Try to measure various kinds of use:
- Resolution
- Downloads
- Mentions
- Citations
- Other types of linking
A threat?

Information overload is only a problem for manual curation.

Google is not crying about data deluge—they’re constantly trying to get more data.

The more data you throw, the better the filter gets.

Don’t turn off the taps, build boats.
It is not only a challenge …
… it is an opportunity

Let us ride the wave together…