Paradigm shifts in Information Access - beyond classical scholarly publication

Jan Brase,
DataCite - TIB

GL14 Conference
November 29th
Rome

DataCite
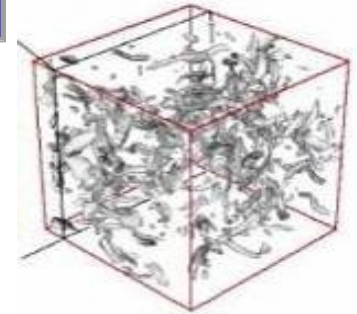
Science Paradigms

Thousand years ago:
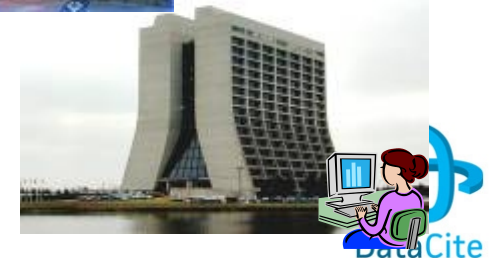 science was **empirical**
      describing natural phenomena

Last few hundred years:
  theoretical branch
      using models, generalizations

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - \mathrm{K}\frac{c^2}{a^2}$$

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

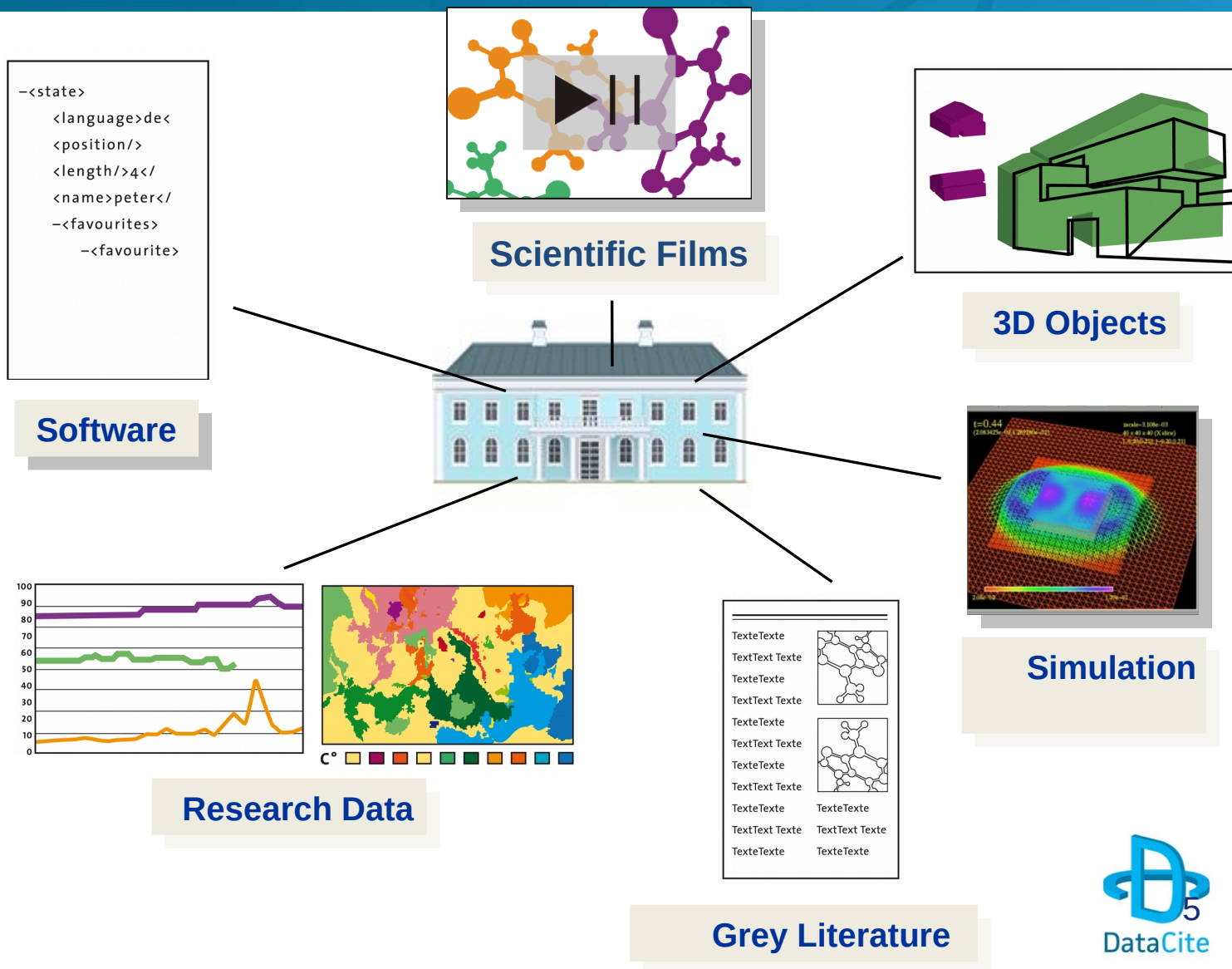Scientific Information is more than a journal article or a book

Libraries should open their cataolgues to any kind of 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 know where you can find

We do not have it

And here is the link to it!

BUT

Including non-classical publications

Scientific Films

3D Objects

```
−<state>
    <language>de<
    <position/>
    <length/>4</
    <name>peter</
    −<favourites>
        −<favourite>
```

Software

Simulation

Research Data

Grey Literature

TIB

German National Library of Science and Technology

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

Global Supplier for scientific and technical information

Financed by Federal Government and all Federal States

- € 18 mio. annual acquisition budget
- 18,500 journal subscriptions
- 7,0 mio. items

Examples

GetInfo

Grey Literature at TIB

Fulfilling ist national mission, TIB holds large collections of:

- Project reports
- Conference proceedings
- Scientific research reports
- Preprints
- Patents & Standards

DataCite

DataCite

What if any kind of scientific content would be citable?

High visability of the content

Easy re-use and verification.

Scientific reputation for the collection and documentation of content (Citation Index)

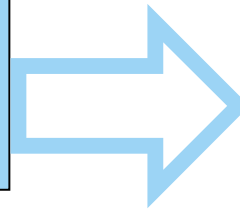Encouraging the *Brussels declaration on STM publishing*

Avoiding duplications

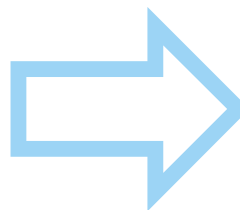Motivation for new research

DataCite

DOI names for citations

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).

The page cannot be found

The page you are looking for might have been removed, had its name changed, or is temporarily unavailable.

Please try the following:

- If you typed the page address in the Address bar, make sure that it is spelled correctly.
- Open the httpd.apache.org home page, and then look for links to the information you want.
- Click the ⇦ Back button to try another link.
- Click ⊗ Search to look for information on the Internet.

HTTP 404 - File not found
Internet Explorer

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

Dataset
Yancheva et al (2007). Analyses on sediment of Lake Maar. PANGAEA.
doi:10.1594/PANGAEA.587840

DataCite

How to achive this?

Science is global

- it needs global standards
- Global workflows
- Cooperation of global players

Science is carried out locally

- By local scientist
- Beeing part of local infrastrucures
- Having local funders

DataCite

Global consortium carried by local institutions

focused on improving the scholarly infrastructure around datasets and other non-textual information
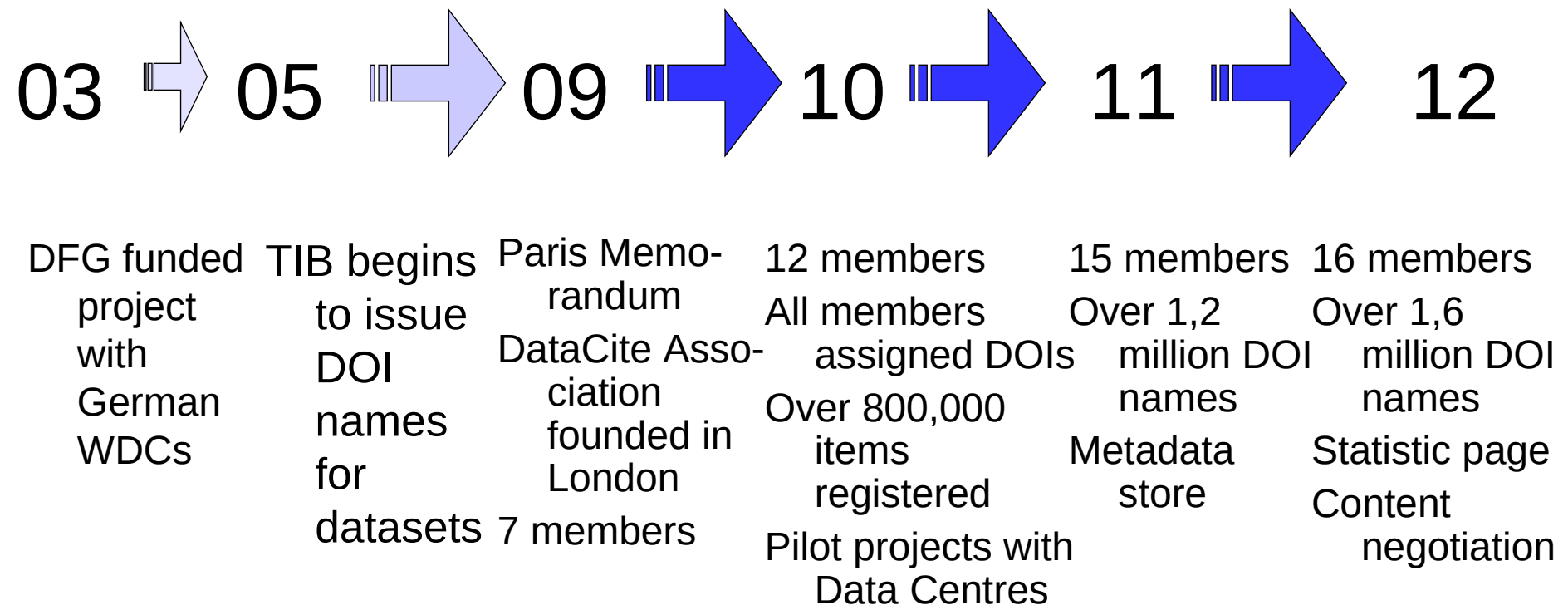
focused on working with data centres and organisations that hold content

Providing standards, workflows and best-practice

Initially, but not exclusivly based on the DOI system

Founded December 1st 2009 in London

History

03 → **05** → **09** → **10** → **11** → **12**

DFG funded project with German WDCs

TIB begins to issue DOI names for datasets

Paris Memo-randum

DataCite Association founded in London

7 members

12 members

All members assigned DOIs

Over 800,000 items registered

Pilot projects with Data Centres

15 members

Over 1,2 million DOI names

Metadata store

16 members

Over 1,6 million DOI names

Statistic page
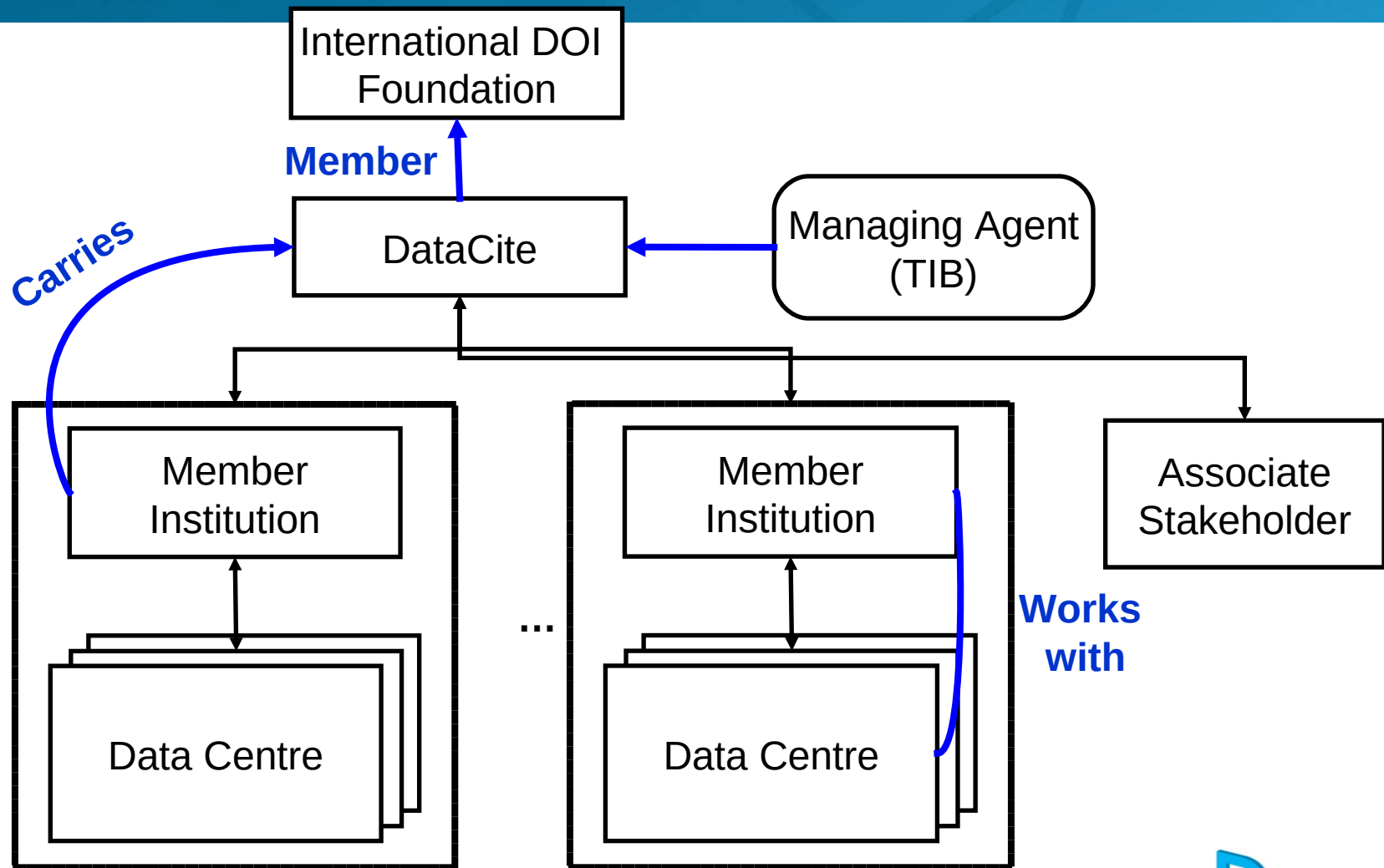
Content negotiation

DataCite

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)
Conferenza dei Rettori delle Università Italiane (CRUI)
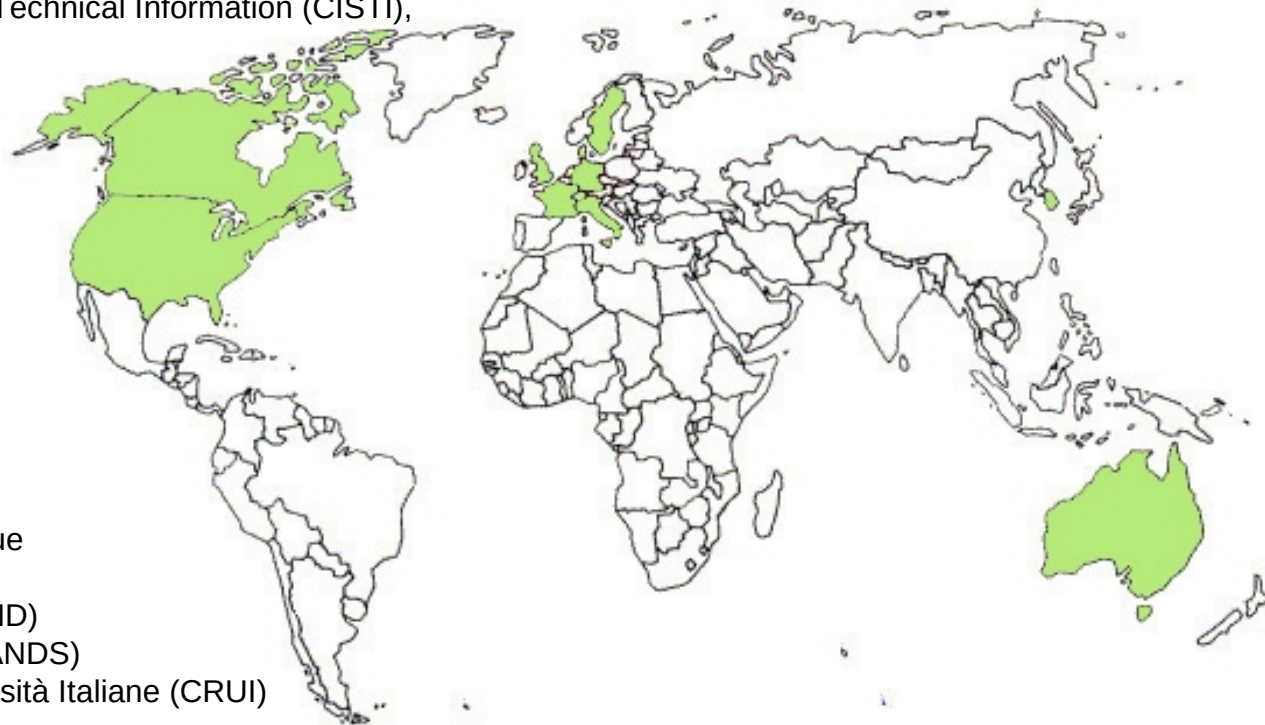
Affiliated members:
Digital Curation Center (UK)
Microsoft Research
Interuniversity Consortium for Political and Social Research (ICPSR)
Korea Institute of Science and Technology Information (KISTI)

DataCite structure

International DOI
Foundation

Member

Carries

DataCite

Managing Agent
(TIB)

Member
Institution

Member
Institution

Associate
Stakeholder

...

**Works
with**

Data Centre

Data Centre

DataCite

DataCite's main goals

Act as DOI registration agency

Actively involved in developing standards and workflows
> CODATA-TG, STM, ICSTI, Data citation index

Central portal allowing access to the metadata from all registered objects. (OAI)
> ISI, Scopus, Microsoft Academic search

Community for exchange of all relevant stakeholders in the area access to and linking of data (data centers, publishers, libraries, research organisation, science unions, funders)

DataCite

DataCite in 2012

Over 1,600,000 DOI names registered so far

DataCite Metadata schema published (in cooperation with all members) http://schema.datacite.org

DataCite MetadataStore
http://search.datacite.org

OAI Harvester
http://oai.datacite.org

DataCite statistics (resolution and registration)
http://stats.datacite.org

What type of data are we talking about?

Earth quake

Climate mo

Sea bed ph

Distributes

Medical cas

Computatio

Audio recor

Grey Litera

Videos =>

Anything that is the foundation of further reserach is research data

Data is evidence

So isn't grey literature also data?
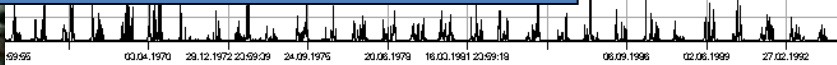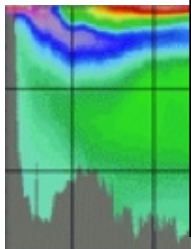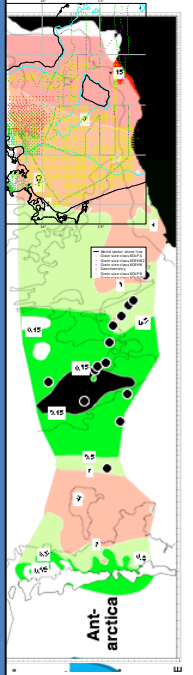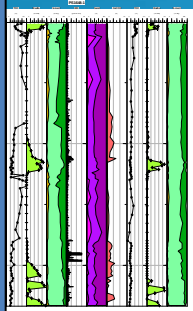
DataCite search

Searchterm: *

Searchterm: uploaded:[NOW-7DAY TO NOW]

Searchterm: relatedIdentifier:*

Searchterm:
   relatedIdentifier:issupplementto\:10.1029*

Searchterm:relatedIdentifier:*\:10.1055*

Citation

The dataset:

Storz, D et al. (2009):

Planktic foraminiferal flux and faunal composition of sediment trap L1_K276 in the northeastern Atlantic.

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

DataCite Content Service

Service for displaying DataCite metadata

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

Content Negotation (through MIME-Typ)

- Access through DOI proxy (http://dx.doi.org)

- First implemented by CNRI and CrossRef:

Documentation:

http://www.crosscite.org/cn/

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
    Emperor penguin. GigaScience. http://dx.doi.org/10.5524/100005*

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
=> *?*

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

DataCite

Future

Try to measure various kinds of use:

- Resolution
- Downloads
- Mentions
- Citations
- Other types of linking

DataCite

A threat?

Information overload is only a problem for manual curation.

Google is not complaining 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…

DataCite