Research data in library context

Dr Jan Brase, Head of R&D

10th Anniversary IGeLU Conference in Budapest, Hungary. September 2nd
Science Paradigms

• Thousand years ago: science was empirical describing natural phenomena

• Last few hundred years: theoretical 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 journal article or a book
- Libraries should open their catalogues 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 do not have it

BUT

We know where you can find

And here is the link to it!
Simulation
Scientific Films
3D Objects
Software
Research Data
Grey Literature

Including non-classical publications
Why is this a role for libraries?

• Libraries have a history in bringing information to the public
• Libraries have a tendency to be persistent
  – A project will be forgotten in 40 years, the library will very likely still exist then
• Library are very trustworthy organisations
What type of data are we talking about?

- Earth quake events => doi:10.1594/GFZ.GEOFON.gfz2009kciu
- Climate models => doi:10.1594/WDCC/dphase_mpeps
- Sea bed photos => doi:10.1594/PANGAEA.757741
- Distributes samples => doi:10.1594/PANGAEA.51749
- Medical case studies => doi:10.1594/eaacinet2007/CR/5-270407
- Computational model => doi:10.4225/02/4E9F69C011BC8
- Audio record => doi:10.1594/PANGAEA.339110
- Grey Literature => doi:10.2314/GBV:489185967
- Videos => doi:10.3207/2959859860

 Anything that is the foundation of further research is research data

Data is evidence
Examples
SAFOD borehole trajectory data in absolute coordinates (UTM) and in coordinates relative to drilling platform
SAFOD
von Deutsches GeoForschungsZentrum GFZ; 2007

Abstract
SAFOD is motivated by the need to answer fundamental questions about the physical and chemical processes controlling faulting and earthquake generation within a major plate-bounding fault, SAFOD will drill and instrument an inclined borehole across the San Andreas Fault Zone to a depth of 3.2 km, targeting a repeating microearthquake source. The drill site is located west of the vertical San Andreas Fault on a segment of the fault that moves through a combination of seismic creep and repeating microearthquakes. It lies at the extreme northern end of the rupture zone of the 1966, Magnitude 6 Parkfield earthquake, the most recent in a series of events that have ruptured the fault five times since 1857. The Parkfield region is the most comprehensively instrumented section of a fault anywhere in the world, and has been the focus of intensive study for the past two decades. This data set contains SAFOD borehole trajectory data.

Jan Brase – Keynote IgeLu, Budapest
SAFOD borehole trajectory data in absolute coordinates (UTM) and in coordinates relative to drilling platform

Cite as:
SAFOD (2007): SAFOD borehole trajectory data in absolute coordinates (UTM) and in coordinates relative to drilling platform. Deutsches Geoforschungszentrum GFZ. http://dx.doi.org/10.1594/GFZ.SDOB.0191

Data Files
- data.csv 117724 bytes
  License: cc-by

Abstract
SAFOD is motivated by the need to answer fundamental questions about the physical and chemical processes controlling faulting and earthquake generation within a major plate-bounding fault. SAFOD will drill and instrument an inclined borehole across the San Andreas Fault Zone to a depth of 3.2 km, targeting a repeating microearthquake source. The drill site is located west of the vertical San Andreas Fault on a segment of the fault that moves through a combination of aseismic creep and repeating microearthquakes. It lies at the extreme northern end of the rupture zone of the 1966, Magnitude 6 Parkfield earthquake, the most recent in a series of events that have ruptured the fault five times since 1857. The Parkfield region is the most comprehensively instrumented section of a fault anywhere in the world, and has been the focus of intensive study for the past two decades. This data set contains SAFOD borehole trajectory data.

Keywords
- Solid Earth, Deep Drilling, Geology, Azimuth, Dog Leg Severity, Existing (Departure), Existing UTM NAD27, Inclination, Land based, Northing (Latitude), Northing UTM NAD27, San Andreas Fault Zone Observatory at Depth, True Vertical Depth, Vertical Section

GCMD Science Keywords
- EARTH SCIENCE > Solid Earth > Deep Drilling > Well-Logging

More Metadata
- iso19115: view inline / download xml
datacite: view inline / download xml
diff: view inline / download xml
escidoc: view inline / download xml

Location
- Northern Latitude: 35.9713  Southern Latitude: 35.9712
- Eastern Longitude: -120.5512  Western Longitude: -120.5513
### Publications

<table>
<thead>
<tr>
<th>Mark</th>
<th>Type</th>
<th>Title</th>
<th>Authors</th>
<th>Year</th>
<th>DOI</th>
<th>PubMed ID</th>
</tr>
</thead>
</table>

### Search Publications

Search

Go!

### Filter Publications

- Publication Type
- Uni-BI Co-authors
- Publishing Year

### Display & Export Publications

- Citation Style: default
- Sorting

RSS Feed
University Library of Bielefeld
Panoramic high dynamic range images in diverse environments

Bielefeld University. doi:10.4119/unibi/2689637

Abstract

This database contains 421 panoramic high dynamic range images recorded in diverse environments. The images are panoramic in full 360° in azimuth and between -90° below and 47° above the horizon in elevation. We used a spectral filter to limit the camera’s spectral sensitivity to wavelengths in the range of 480-580 nm (green). This filtering mimics the spectral sensitivity of photoreceptors R1-R6 of the fly that provide the input of the motion vision system. As a consequence, the mapping of colors to gray values in these images is similar to the green color channel in RGB images. The raw images have a resolution of approximately 1 Megapixel (920x520) and 12-bit. The images have high dynamic range covering the entire brightness range encountered in natural environments (excluding the solar disc). After linearization the resulting image values had a dynamic range of 1:23,900 covering 3,555 intensity steps. Note, however, that the pixel brightness values cannot be recalculated to a SI unit like candela, though the values are proportional to luminance in the green spectral range. For more technical details about the recording of the image sequence see Meyer et al. (2014) in addition to the raw camera images.
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
Indexing and search

visual search

content based indexing
content based indexing

- segmentation with form-primitives
- extraction of room connectivity graphs

visual search

- attributed graph
- 3D sketch

classification of floor types
→ machine learning

result visualization
Chemical search
Content based search
Visual Search in Time series

- Query-by-Example, Query-by-Sketch
- Visual Catalog as result list
- Colormaps for the indication of similarity
Video cataloging with automatic indexing
What if any kind of scientific content would be citable?

- High visibility 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
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

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
DataCite

- Global consortium carried by local institutions mostly libraries
- 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 exclusively based on the DOI system
- Founded December 1st 2009 in London
DataCite structure

International DOI Foundation

Member

DataCite

Member Institution

Data Centre

Associate Stakeholder

Member Institution

Data Centre

Works with

Carries
DataCite members

1. Technische Informationsbibliothek (TIB), Germany
2. Göttingen State and University Library (SUB), Germany
3. ZB MED, Germany
4. ZBW, Germany
5. Gesis, Germany
7. Technical Information Center of Denmark
8. The British Library
9. Library of ETH Zürich
10. L’Institut de l’Information Scientifique et Technique (INIST), France
11. Swedish National Data Service (SND)
12. Australian National Data Service (ANDS)
13. Conferenza dei Rettori delle Università Italiane (CRUI)
14. National Research Council of Thailand (NRCT)
15. The Hungarian Academy of Sciences
16. University of Tartu, Estonia
17. Bibsys, Norway
18. Canada Institute for Scientific and Technical Information (CISTI), Canada
19. California Digital Library, USA
20. Purdue University, USA
21. Office of Scientific and Technical Information (OSTI), USA
22. Japan Link Center (JaLC)
23. South African Environmental Observation Network (SAEON)
24. European Organisation for Nuclear Research (CERN)

Affiliated members:
1. Digital Curation Center (UK)
2. Microsoft Research
3. Interuniversity Consortium for Political and Social Research (ICPSR)
4. Korea Institute of Science and Technology Information (KISTI)
5. Beijing Genomic Institute (BGI)
6. IEEE
7. Harvard University Library
8. GWDG
Example

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
More Data example

- **Higgs particle**
  ATLAS Collaboration (2013) HepData,
  [http://doi.org/10.7484/INSPIREHEP.DATA.A78C.HK44](http://doi.org/10.7484/INSPIREHEP.DATA.A78C.HK44)

- **ECOLI outbreak**
  Li, D et al (2011):
  *Genomic data from Escherichia coli O104:H4 isolate TY-2482.*
  BGI Shenzhen.
  [http://dx.doi.org/10.5524/100001](http://dx.doi.org/10.5524/100001)
Now what?
The wave

Growth of Information –

User requirements – e.g.: Science 2.0, collaborative networks, social media

Diversity of media types and formats
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.

• To develop and maintain these tools is a classical tasks for libraries!

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

… it is an opportunity

Libraries should ride the wave …