Managing Research Data 101

Workshop

GFZ PhD-Day, 22.11.10

Dr. Jens Klump - GFZ/CeGIT

Roland Bertelmann - GFZ/Library and Information Services (LIS)

Managing Research Data 101

Agenda:

With a little help from: Managing Research Data 101, MIT Libraries, MacKenzie Smith (2009)
http://libraries.mit.edu/guides/subjects/data-management/Managing%20Research%20Data%20101.pdf

Why should we talk about data?

You have digital data. You think they are important.

Some questions:

- Your grant runs out… and then what?
- You have been doing all the data-management and then you leave with Ph.D. in hand… and then what?
- Your favorite grant agency institutes a data-sustainability requirement for all grants… and then what?
- Your lab's PI retires… and then what?
- Your instrument manufacturer or favorite software's developer goes out of business… and then what?

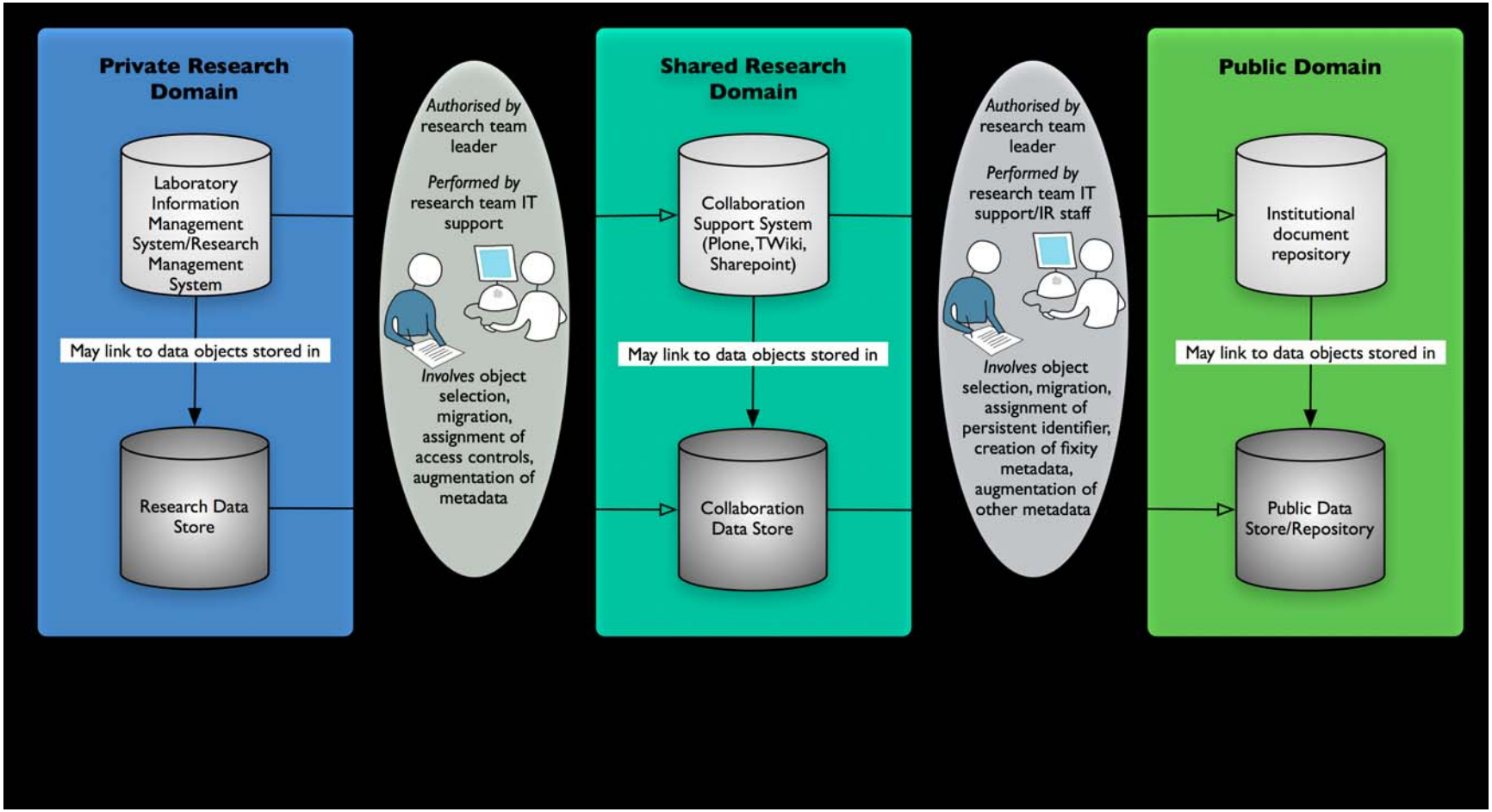
http://scienceblogs.com/bookoftrogool/2009/11/_and_then_what.php

What do you expect today?

- You’re managing research data
- You’re not sure how to do that
- You’re not sure if you should worry about it
- You want some clues and pointers

- **What else?**

Data Curation Continuum

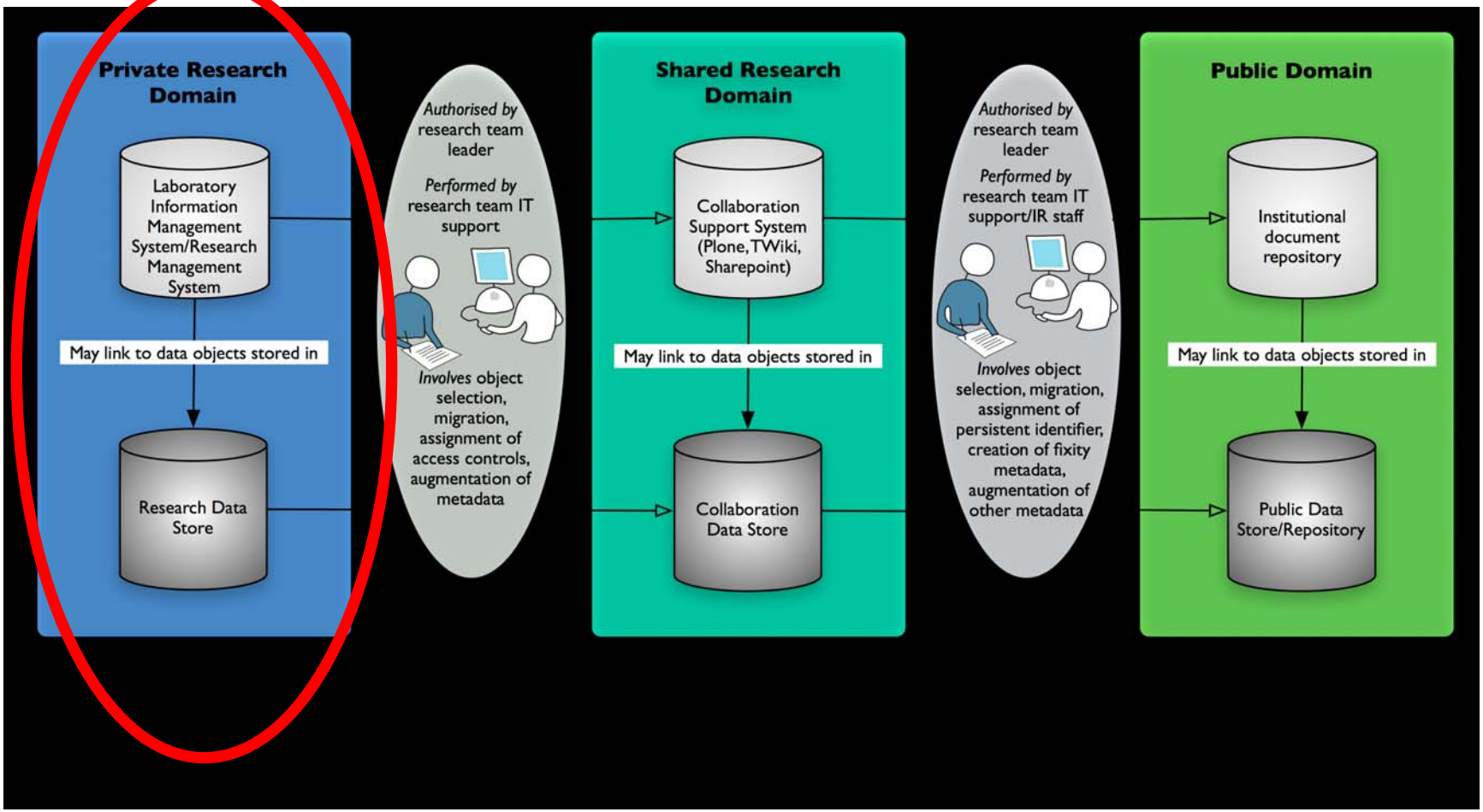


Andrew.treloar.net

Basics

Basics for the Private Domain

Aim: make your data reusable



What are Data?

What Are Data?

Observational data captured in real-time

-- Usually irreplaceable

Experimental data from lab equipment

-- Often reproducable, but can be expensive

What Are Data?

Simulation data

-- Model and metadata inputs are more important than outputs

Derived and compiled data

-- Reproducable (expensive)

What Are Data?

- **Text** e.g. flat text files, Word, PDF
- **Numerical** e.g. SPSS, STATA, Excel, Access, MySQL
- **Multimedia** e.g. jpeg, tiff, mpeg, quicktime
- **Models** e.g. 3D, statistical
- **Software** e.g. Java, C
- **Domain-specific** e.g. OGC, SEED
- **Instrument-specific** e.g. a certain Microscope Data Format

A planning checklist

Start: a Data Planning Checklist

- What type of data will be produced?
- How much of it, and at what growth rate?
- Will it change frequently?
- Who is it for?
- Who controls it (you, your group, your PI)?
- How long should it be retained?

How long should it be retained?

"Digital information lasts forever –
or five years, whichever comes first."

(Jeff Rothenberg, RAND Corp., 1997)

Choose:
3-5 years, 10-20 years, permanently

Data Planning Checklist / 2

- Are there tools or software needed to create/process/visualize the data?
- Any privacy requirements from the funders or lab?
- Any sharing requirements from the funders or lab?
- Any other funder requirements?

Documentation and Metadata

Project Documentation

- **Title**

  name of the dataset or research project that produced it
- **Creator**

  names and addresses of the organization or people who created the data, including all significant contributors
- **Identifier**

  The identification number used to identify the data, even if it it's just an internal project reference number
- **Subject**

  keywords or phrases describing the subject or content of the data

Project Documentation

- **Dates**

  key dates associated with the data,

  including: project start and end date; release date;

  other dates associated with the data lifespan, e.g. maintenance cycle, update schedule
- **Funders**

  organizations or agencies who funded the research
- **Language**

  language(s) of the intellectual content of the resource, when relevant

Project Documentation

- **Location**

  where the data relates to a physical location, record information about its spatial coverage
- **Rights**

  description of any known intellectual property rights held for the data
- **List of file names and relationships**

  list of all digital files in the archive, with their names and file extensions

More Metadata

- **Formats**

  format(s) of the data, e.g. SPSS, HTML, JPEG
- **Methodology**

  how the data was generated, including equipment or software used, experimental protocol, other things you would include in your lab notebook. Reference a published article, if it covers everything
- **Sources**

  references to source material for data derived from other sources, including details of where the source data is held, how identified and accessed

More Metadata

- **Versions**
  date/time stamped, and use a separate ID for each version!
- **Checksums**
  to test if your file has changed over time
- **Explanation of codes used in file names and files**
  list of codes used in file names
  list of any special values used in the data

Metadata

At least:

Store (appropriate) metadata in a readme.txt file together with the data

And:

Ask for data management tools!

Storage

Security and Backups

What do you do?

Storage Options

- Personal PC
  not recommended
- External Drives
- GFZ network
  Backup!
- Subject Archive
  e.g. GFZ Scientific Drilling Data Base SDDB, other: WDC-RSAT, Pangaea
- Personal: Cloud storage (e.g. Amazon S3)

What else?

- Lots of copies keep stuff safe!

- Test File Recovery!
  At setup time, and on a regular schedule

- To secure data
  Protect your hardware
  Use file encryption (e.g. PGP)
  keep passwords and keys on paper (2 copies) and in a PGP encrypted digital file

Directory Structures and Naming Conventions

Good Directory Structure

- Directory top-level folder should include the **project title**, **unique identifier**, and **date** (e.g. year)
- Substructure should have clear, documented naming convention
  - e.g. each run of an experiment, each version of a dataset, each person in the group

File Naming Conventions

- Reserve the 3-letter file extension for application-specific codes, e.g. formats like WRL, MOV, TIF.
- Identify the activity or project in the file name, e.g. use the unique project name or identifier.
- Example:
  - Project_instrument_location_YYYYMMDD[hh][mm][ss][_extra].ext

File Naming Conventions

- Many academic disciplines have specific recommendations, e.g.
- DOE's Atmospheric Radiation Measurement (ARM) Program
  - http:// www.arm.gov/data/plan.stm
- GIS datasets from Massachusetts StateGIS State
  - http://www.mass.gov/mgis/dwn-name.htm

File Renaming

- Use free tools to help you!
  - http://www.bulkrenameutility.co.uk/
  - http://renamer4mac.com/
  - http://www.powersurgepub.com/products/psrenamer.html

File Version Control

- Strategies include:
  - file-naming conventions
  - standard file headers (inside the file) listing creation date, version number, status
  - log files
  - version control software (e.g. SVN)
  - Always record every change to a file no matter how small.
  - Discard obsolete versions after making back-ups.

Data Identifiers

- Must be globally unique, persistent
- Many different schemes:
  - PURL http://purl.org/
  - DOI http://www.doi.org/
  - Handle http://www.handle.net/
  - ACCESION http://www.ncbi.nlm.nih.gov/
  - InChI http://www.iupac.org/inchi/
  - URI http://www.ietf.org/rfc/rfc2396.txt
  - URN http://nbn-resolving.de/

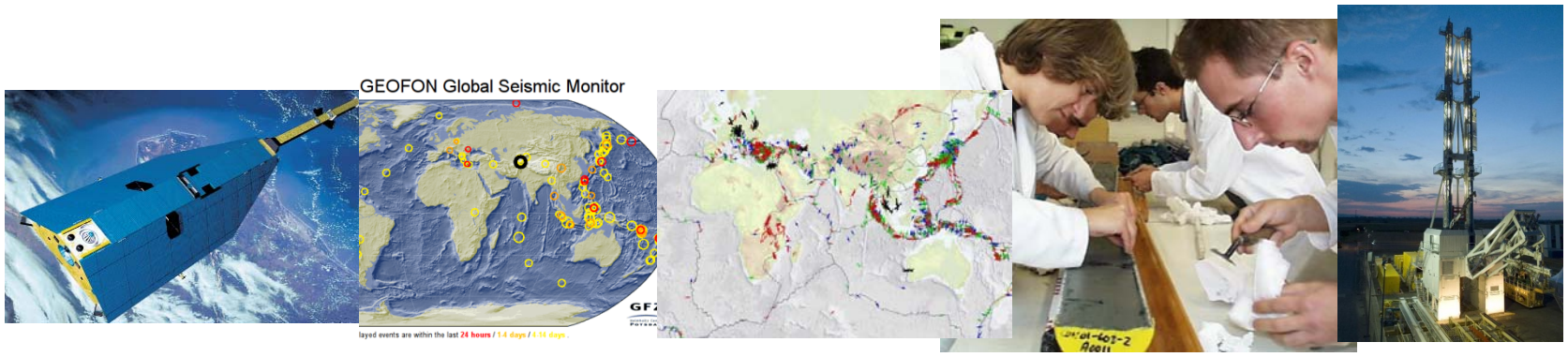
- GFZ offers DOI, Handle, and URN

Search and Find

Data Portals

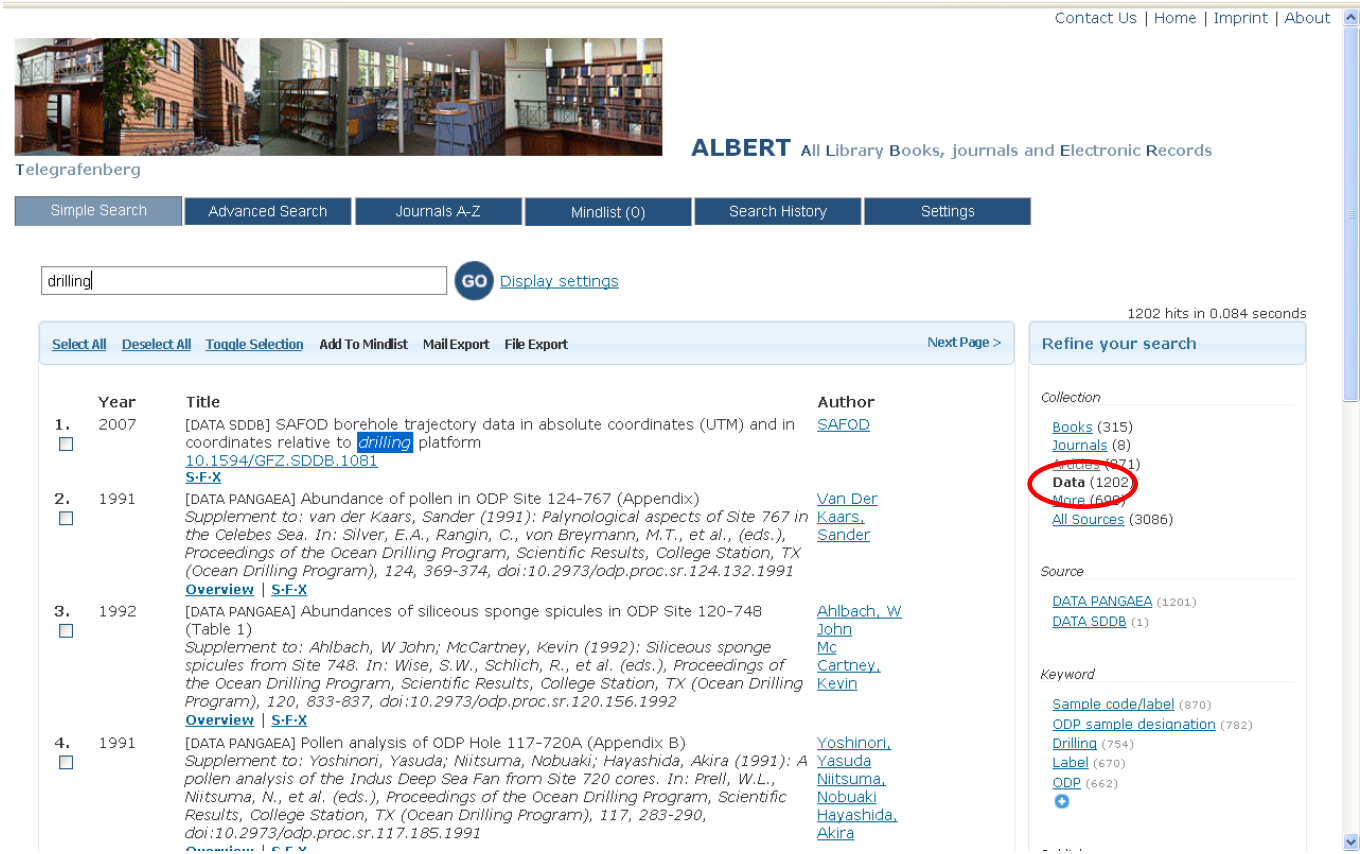
Examples at GFZ and GFZ cooperations:

- Geodetic Satellites (ISDC) – CHAMP, GRACE
- Magnetic Observatories - NDC Boulder, Colorado
- Gravity Field Models – ICGEM
- Seismic Network (GEOFON) – IRIS, EMSC
- Scientific Drilling (SDDB) – ICDP [includes DOI]
- World Stress Map – ICSU World Data System [includes DOI]



World Data System: Example

Library


Contact Us | Home | Imprint | About
Telegrafenberg
ALBERT All Library Books, journals and Electronic Records
Simple Search
Advanced Search
Journals A-Z
Mindlist (0)
Search History
Settings
drilling
GO
Display settings
1202 hits in 0.084 seconds
Select All
Deselect All
Toggle Selection
Add To Mindlist
Mail Export
File Export
Next Page >
Year
Title
Author
1. 2007 [DATA SDDB] SAFOD borehole trajectory data in absolute coordinates (UTM) and in coordinates relative to drilling platform 10.1594/GFZ.SDDB.1081 S·F·X SAFOD
2. 1991 [DATA PANGAEA] Abundance of pollen in ODP Site 124-767 (Appendix) Supplement to: van der Kaars, Sander (1991): Palynological aspects of Site 767 in the Celebes Sea. In: Silver, E.A., Rangin, C., von Breymann, M.T., et al., (eds.), Proceedings of the Ocean Drilling Program, Scientific Results, College Station, TX (Ocean Drilling Program), 124, 369-374, doi:10.2973/odp.proc.sr.124.132.1991 Overview | S·F·X Van Der Kaars, Sander
3. 1992 [DATA PANGAEA] Abundances of siliceous sponge spicules in ODP Site 120-748 (Table 1) Supplement to: Ahlbach, W John; McCartney, Kevin (1992): Siliceous sponge spicules from Site 748. In: Wise, S.W., Schlich, R., et al. (eds.), Proceedings of the Ocean Drilling Program, Scientific Results, College Station, TX (Ocean Drilling Program), 120, 833-837, doi:10.2973/odp.proc.sr.120.156.1992 Overview | S·F·X Ahlbach, W John Mc Cartney, Kevin
4. 1991 [DATA PANGAEA] Pollen analysis of ODP Hole 117-720A (Appendix B) Supplement to: Yoshinori, Yasuda; Niitsuma, Nobuaki; Hayashida, Akira (1991): A pollen analysis of the Indus Deep Sea Fan from Site 720 cores. In: Prell, W.L., Niitsuma, N., et al. (eds.), Proceedings of the Ocean Drilling Program, Scientific Results, College Station, TX (Ocean Drilling Program), 117, 283-290, doi:10.2973/odp.proc.sr.117.185.1991 Yoshinori, Yasuda Niitsuma, Nobuaki Hayashida, Akira
Refine your search
Collection
Books (315)
Journals (8)
Articles (871)
Data (1202)
More (6082)
All Sources (3086)
Source
DATA PANGAEA (1201)
DATA SDDB (1)
Keyword
Sample code/label (870)
ODP sample designation (782)
Drilling (754)
Label (670)
ODP (662)

Data Sharing and Citation

Data Sharing

- As a member of the Helmholtz Association GFZ is committed to further the aims of the „Berlin Declaration“.
- Open Access (to data) is part of the GFZ publication guidelines, as are the DFG „Rules for Good Scientific Practice“.
  - PS: This is part of your employment contract with GFZ.
  - http://www.gfz-potsdam.de/portal/cms/Bibliothek/Publizieren/H-Publizieren+am+...

- German Science Organisations: Grundsätze zum Umgang mit Forschungsdaten
- DFG also asks, that research data should be made accessible.

IPR and data licenses

- Most data NOT copyrightable
  - facts cannot be copyrighted
  - limited protection for databases in EU
- But: Licenses (e.g. CC licenses) provide a work-around.
  - http://www.gfz-potsdam.de/portal/cms/Bibliothek/Publizieren/J-Urheberrecht

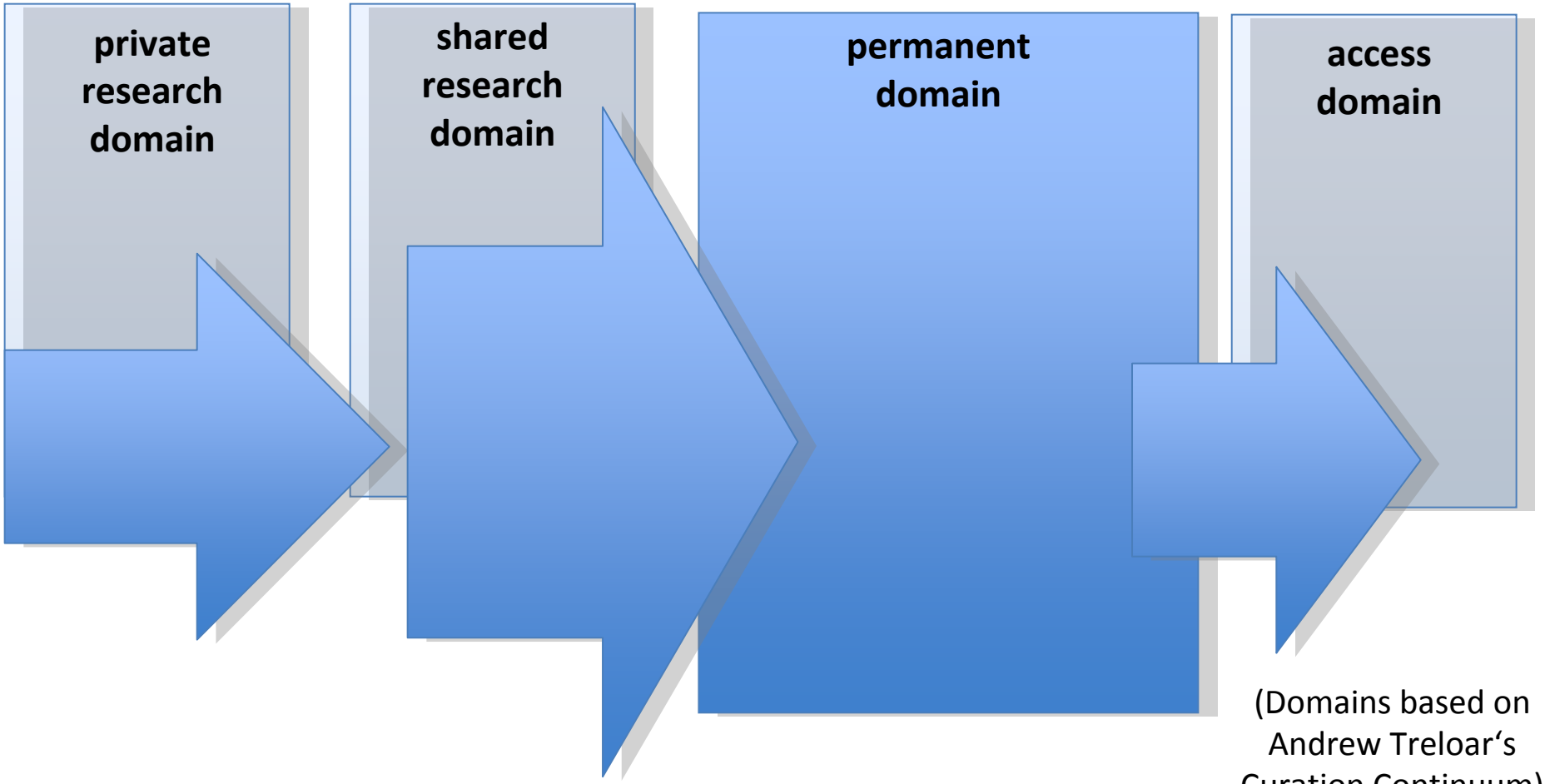
- Also: Data from external sources might be covered by licence agreements.

Citing Data

- ISO 690-2
- Can include
  - Author
  - Title
  - Size
  - Edition
  - Language
  - Publisher
  - publication date
  - publication place
- Assumes a unique identifier for the dataset
- Like citing a publication.

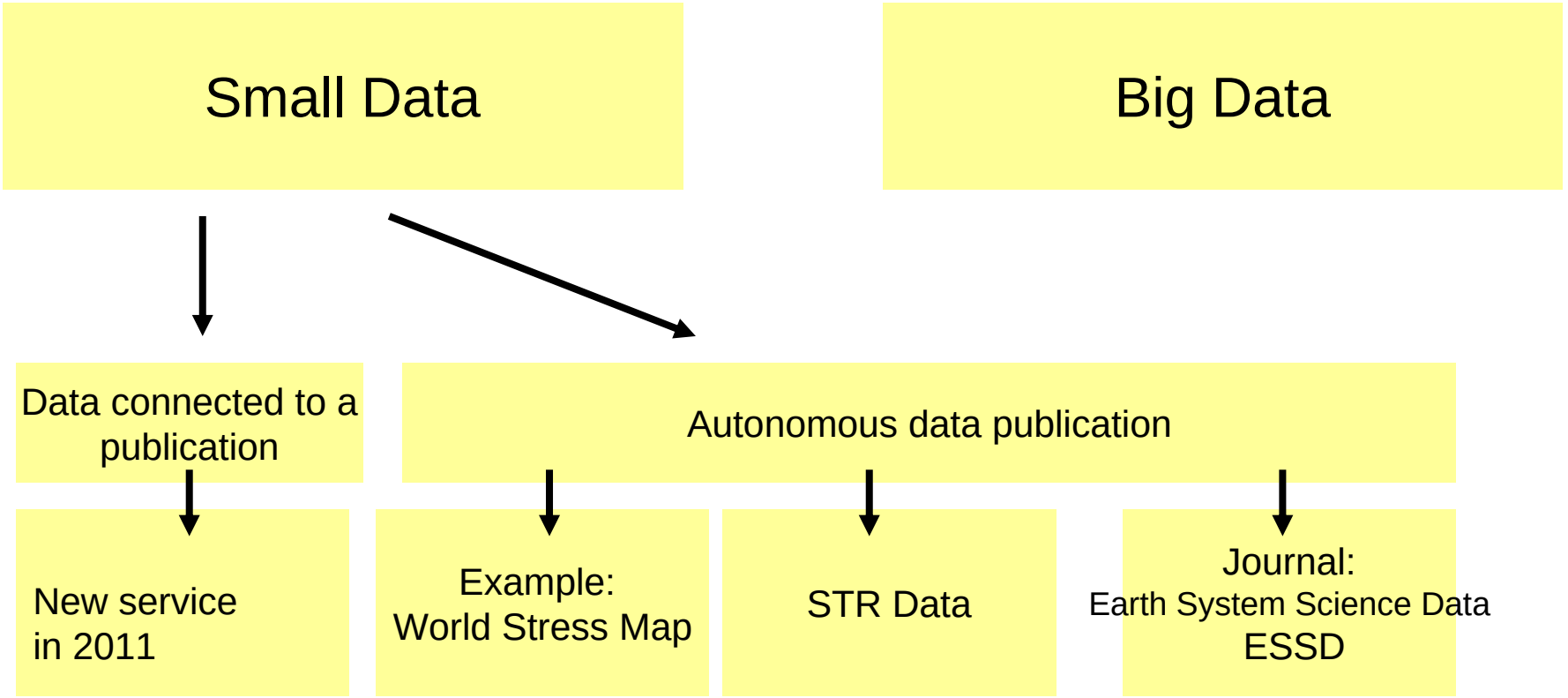
Publish Data

Publish Data <–> Access Domain



(Domains based on Andrew Treloar's Curation Continuum)

Small Data at GFZ



Data publication through SDDB

Scientific Drilling Database

Data from Deep Earth Sampling and Monitoring

Citation: Heim, Birgit; Oberhänsli, Hedi; Fietz, Susanne; Kaufmann, Hermann; (2006): The relationship between concentrations of chl-a calculated from SeaWiFS OC2 and chl-a calculated determined from ground truth measurements during field expeditions in Lake Baikal during 2001 and 2002. *Scientific Drilling Database*. doi:10.1594/GFZ.SDDB.1043

Download Citation (EndNote)

Related Publications:

- Birgit Heim, Hedi Oberhaensli, Susanne Fietz and Hermann Kaufmann, Variation in Lake Baikal's phytoplankton distribution and fluvial input assessed by SeaWiFS satellite data, Global and Planetary Change, Volume 46, Issues 1-4, Progress towards reconstruct doi:10.1016/j.gloplacha.2004.11.011

considerable chl-a overestimation caused by the influences of terrigenous input in case 2 waters.

Show in Google Earth

GFZ POTSDAM icdp

Related Publications:

- Birgit Heim, Hedi Oberhaensli, Susanne Fietz and Hermann Kaufmann, Variation in Lake Baikal's phytoplankton distribution and fluvial input assessed by SeaWiFS satellite data, Global and Planetary Change, Volume 46, Issues 1-4, Progress towards reconstruct doi:10.1016/j.gloplacha.2004.11.011

Activities: CON01-501-1

Latidude: 52.6667 °N

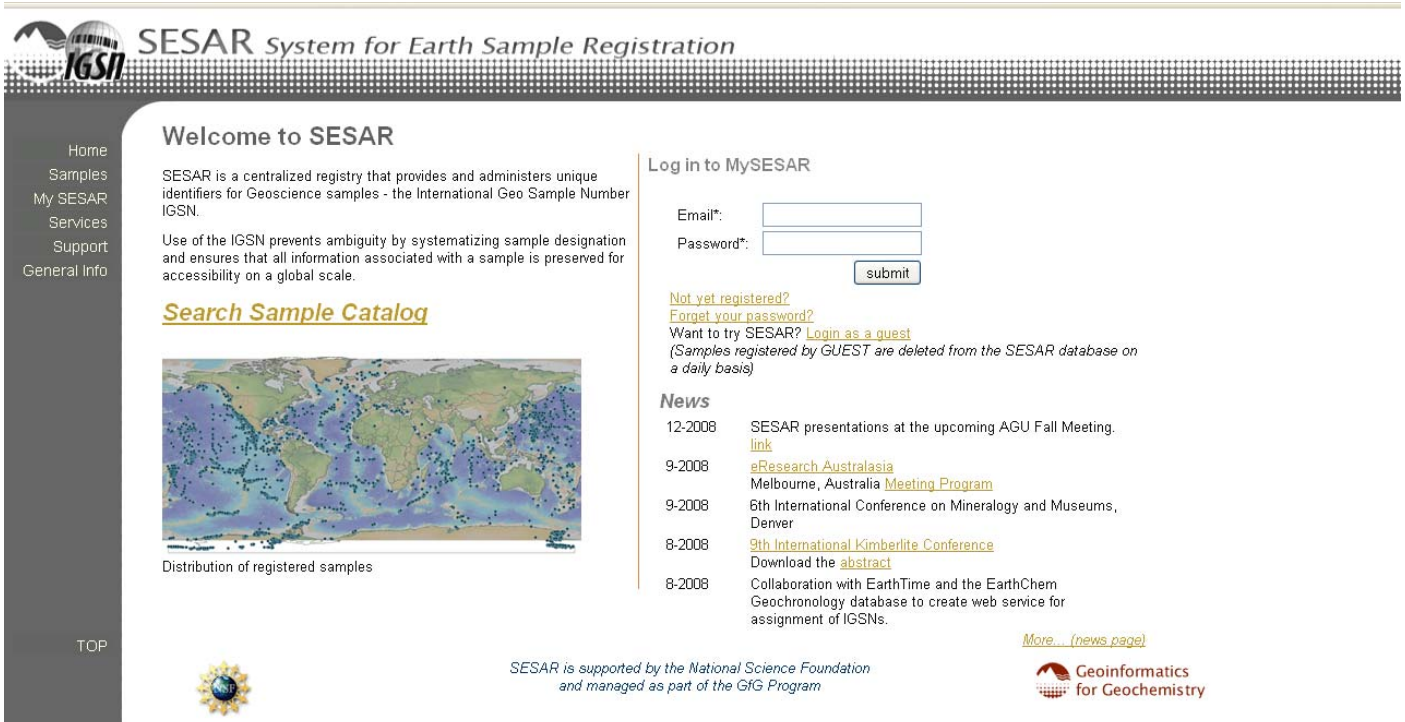
Data Integration

Data Integration

- Semantic Web or Linked Open Data Web
- Requires URI for each Resource, e.g. distinct data entry.
- Requires RDF encoding of data
- Ideally has an "ontology" for the data model
- Alternatives include,
  - Manually map different database or XML schemas
  - Develop “über-ontology”and map data to that
  - Many gotchas (e.g. different metrics, synonyms)

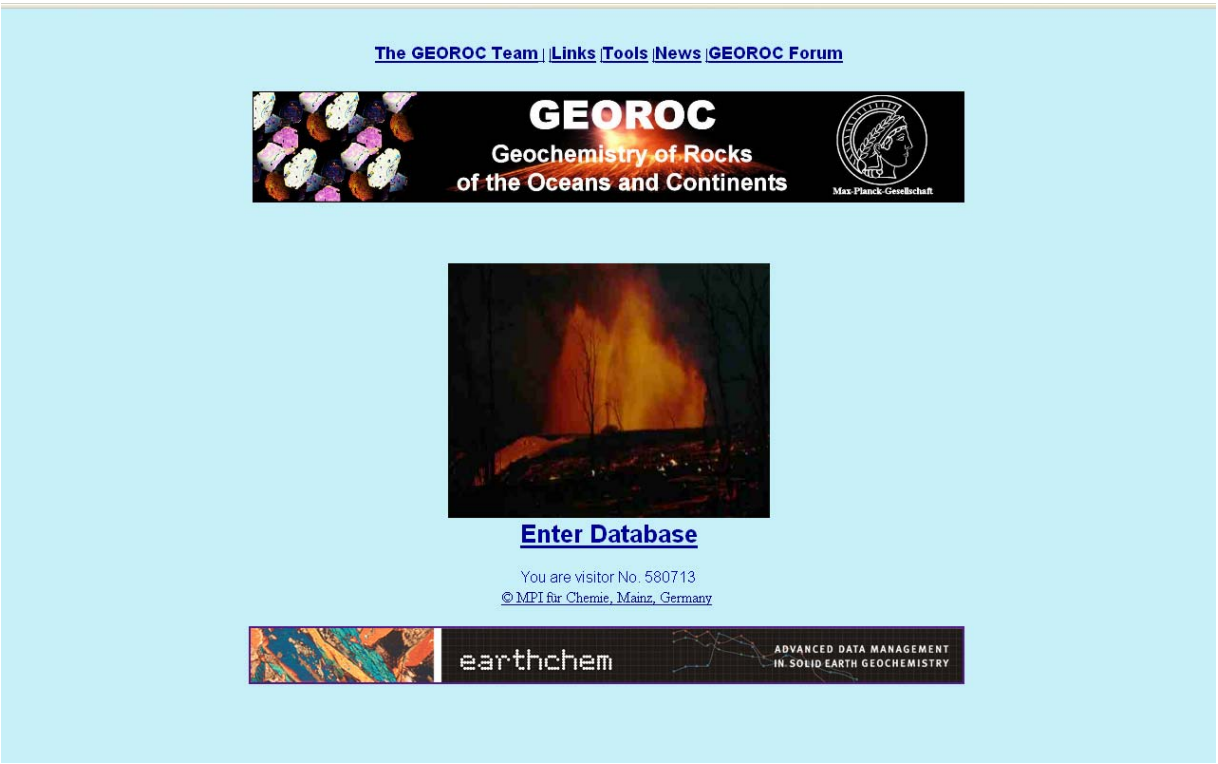
Identifier for Samples

International Geo Sample Number IGSN for solid earth samples)

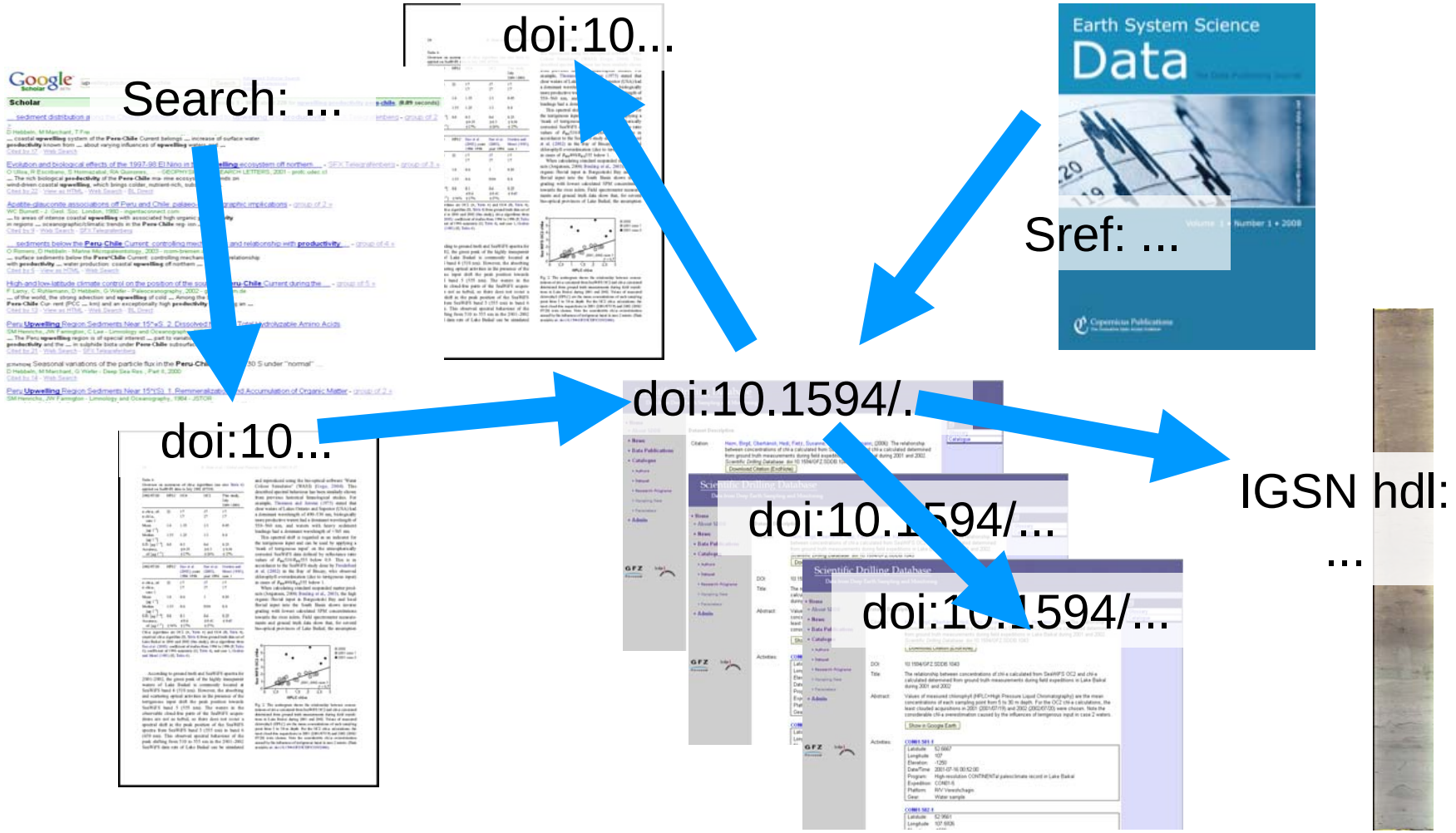


Registries

e.g. for methods and standards



Literature, Data, Objects



File Formats for Long-Term Access

File Formats for Long-Term Access

- Principles:
  - Unencrypted
  - Uncompressed
  - Non-proprietary
  - Open, documented standard
  - Common usage by research community
  - Standard representation (ASCII, Unicode)

File Formats for Long-Term Access

- Examples
  - PDF/A, not Word
  - MPEG-4, not Quicktime
  - TIFF or JPEG2000, not GIF or JPG
  - XML or RDF, not RDBMS
- Discipline Standards, e.g. Environmental data
  - http:// daac.ornl.gov/PI/bestprac.html

Data Retention and Archiving

Data Retention and Archiving

- From the checklist:
  - How permanent are the data?
  - Long-term (e.g. 10 years)? Or Short-term (e.g. 3-5 years)?
  - Should discarded data be destroyed?
- Keep all versions? Just final version? First and last?
  - Depends on re-processing costs. If you can re-process the data, probably better to do so, but keep all the software and protocol info to support that.

Long-term, in the context of research data, means **well beyond the end of the project**.

Remember

- Documentation **is the most important thing**
- Don’t lose the bits
- Use good hygiene (formats, file names)
- Think about what you want to accomplish

Over Time

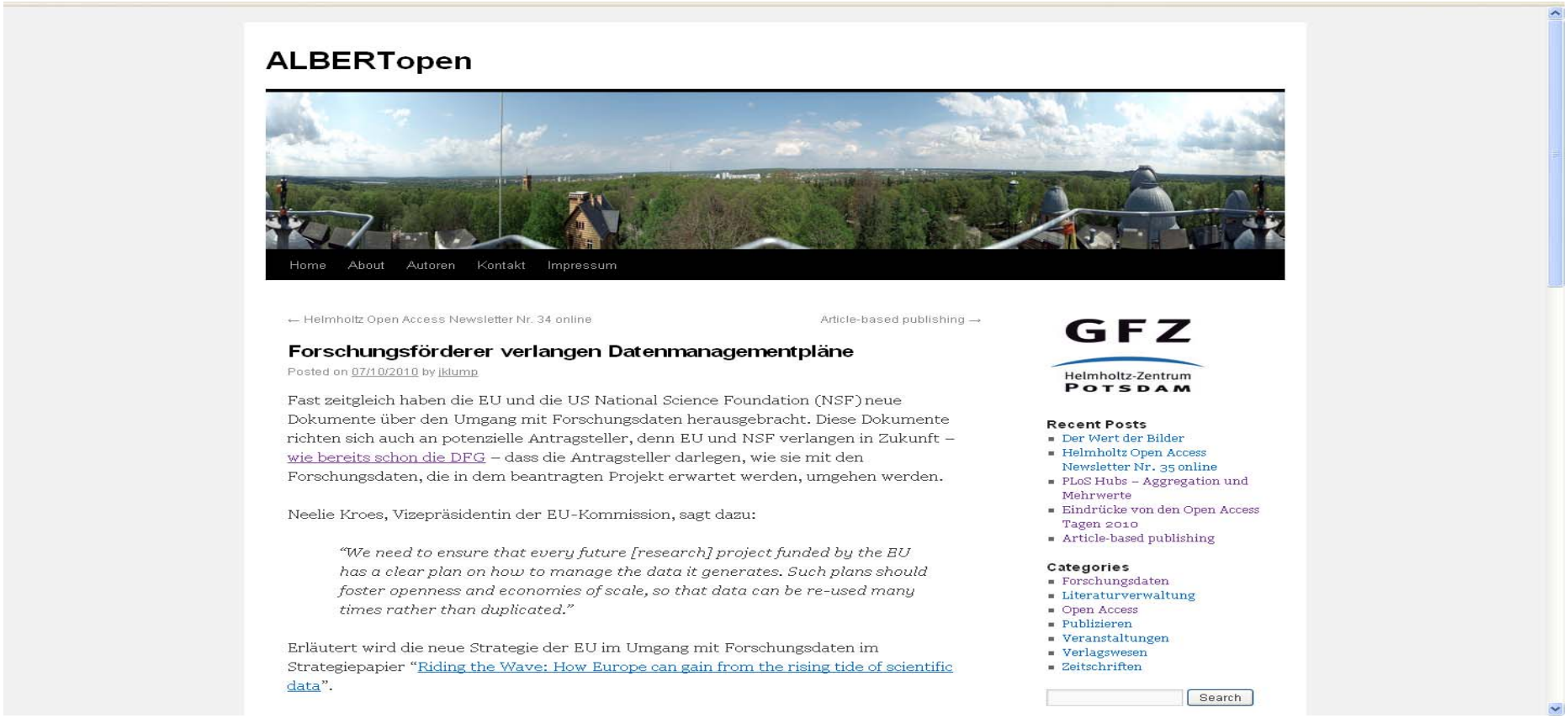
- Test data restore from backup
- Check documentation and metadata
- Are files still readable?
- Still accessible at the published URL?
- Migrate files to newer formats
- Update software to read/write data
- Weed out obsolete data (and destroy where appropriate)

Where data management is concerned …

“Perfection is the Enemy of the Good”

just do the best you can and don't be shy to ask

Stay informed!



ALBERTopen

Home About Autoren Kontakt Impressum

← Helmholtz Open Access Newsletter Nr. 34 online

Article-based publishing →

Forschungsförderer verlangen Datenmanagementpläne

Posted on 07/10/2010 by jklump

Fast zeitgleich haben die EU und die US National Science Foundation (NSF) neue Dokumente über den Umgang mit Forschungsdaten herausgebracht. Diese Dokumente richten sich auch an potenzielle Antragsteller, denn EU und NSF verlangen in Zukunft – wie bereits schon die DFG – dass die Antragsteller darlegen, wie sie mit den Forschungsdaten, die in dem beantragten Projekt erwartet werden, umgehen werden.

Neelie Kroes, Vizepräsidentin der EU-Kommission, sagt dazu:

> *"We need to ensure that every future [research] project funded by the EU has a clear plan on how to manage the data it generates. Such plans should foster openness and economies of scale, so that data can be re-used many times rather than duplicated."*

Erläutert wird die neue Strategie der EU im Umgang mit Forschungsdaten im Strategiepapier "Riding the Wave: How Europe can gain from the rising tide of scientific data".

GFZ Helmholtz-Zentrum Potsdam

Recent Posts

- Der Wert der Bilder
- Helmholtz Open Access Newsletter Nr. 35 online
- PLoS Hubs – Aggregation und Mehrwerte
- Eindrücke von den Open Access Tagen 2010
- Article-based publishing

Categories

- Forschungsdaten
- Literaturverwaltung
- Open Access
- Publizieren
- Veranstaltungen
- Verlagswesen
- Zeitschriften

Search