Managing Research Data 101

Workshop PhD-Day, 23.11.09

Dr. Jens Klump - GFZ/CeGIT
Roland Bertelmann - GFZ/Library (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 we are here:

Experience and cooperation in projects dealing with information management of scientific data, e.g.

- **Publication and Citation of Scientific Primary Data**
- **Open Access**
- **Repositories and Metadata**
- **Network of Expertise in long-term storage of digital resources** (nestor)

Tell us, why you are here:

Why should we talk about data?

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?

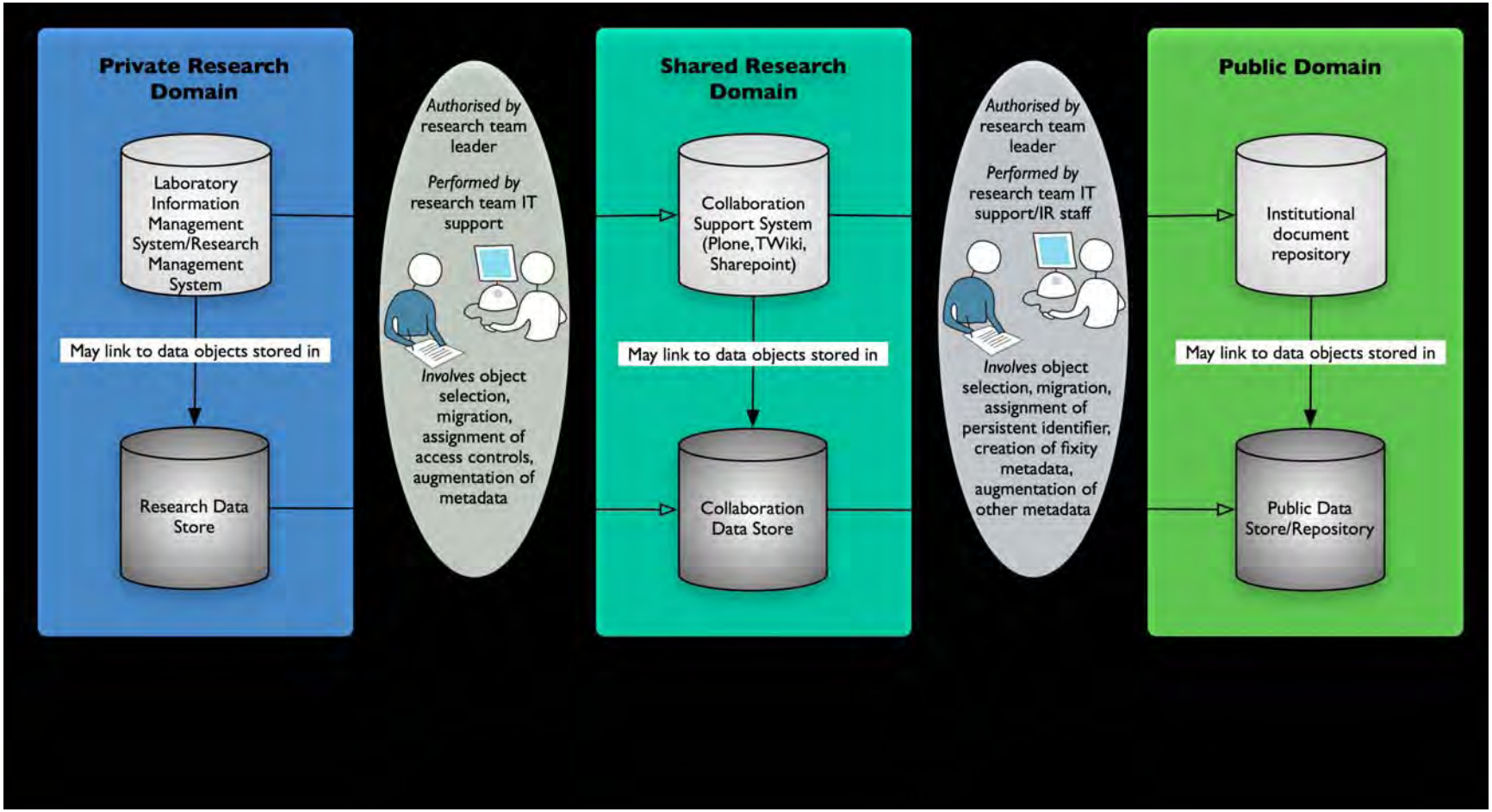
http://scienceblogs.com/bookoftrogool/2009/11/_and_then_what.php

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

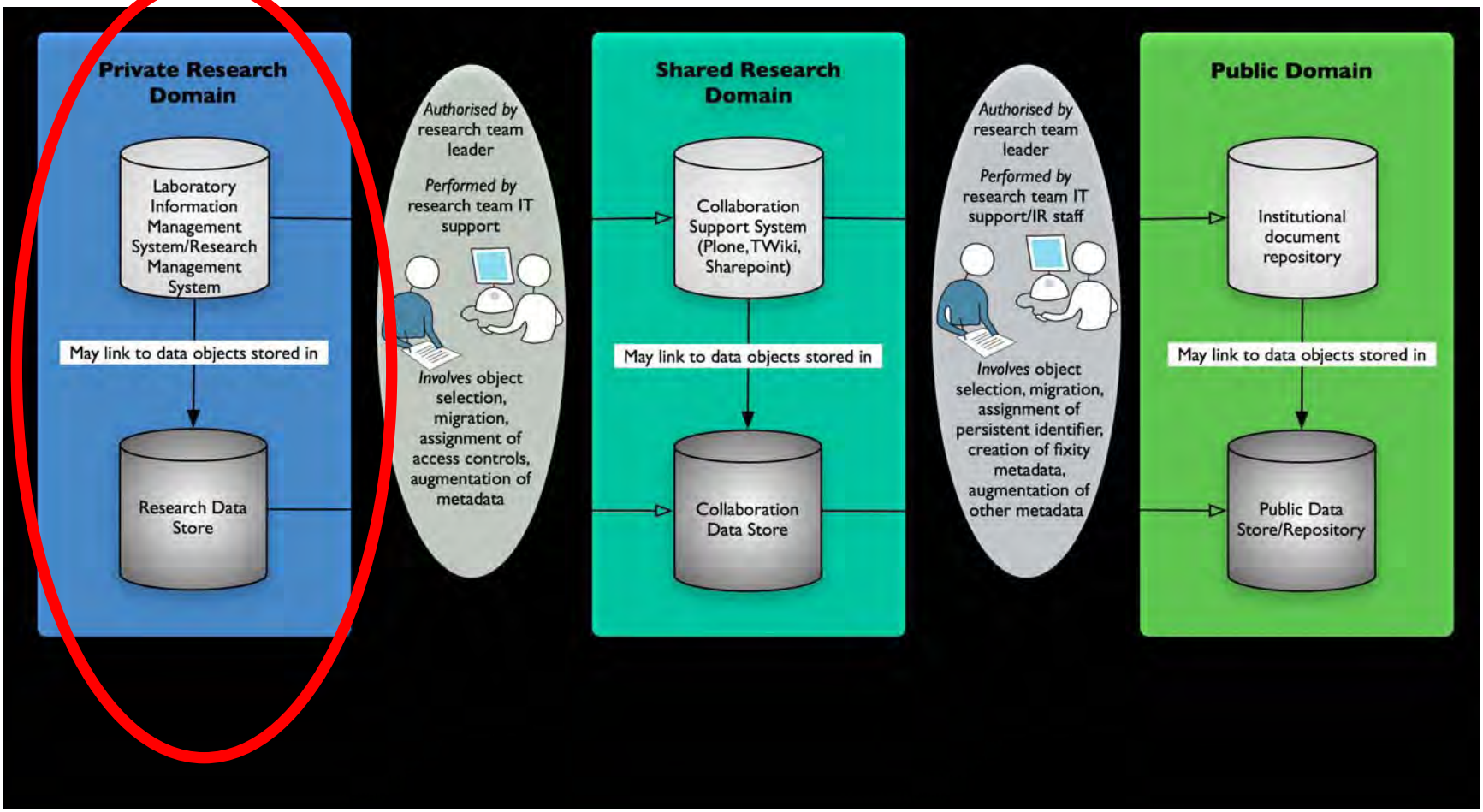
- **What else?**

Data Curation Continuum



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

Documentation and Metadata

Your 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'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

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.

- DFG also asks, that research data should be made accessible.

IPR and data licenses

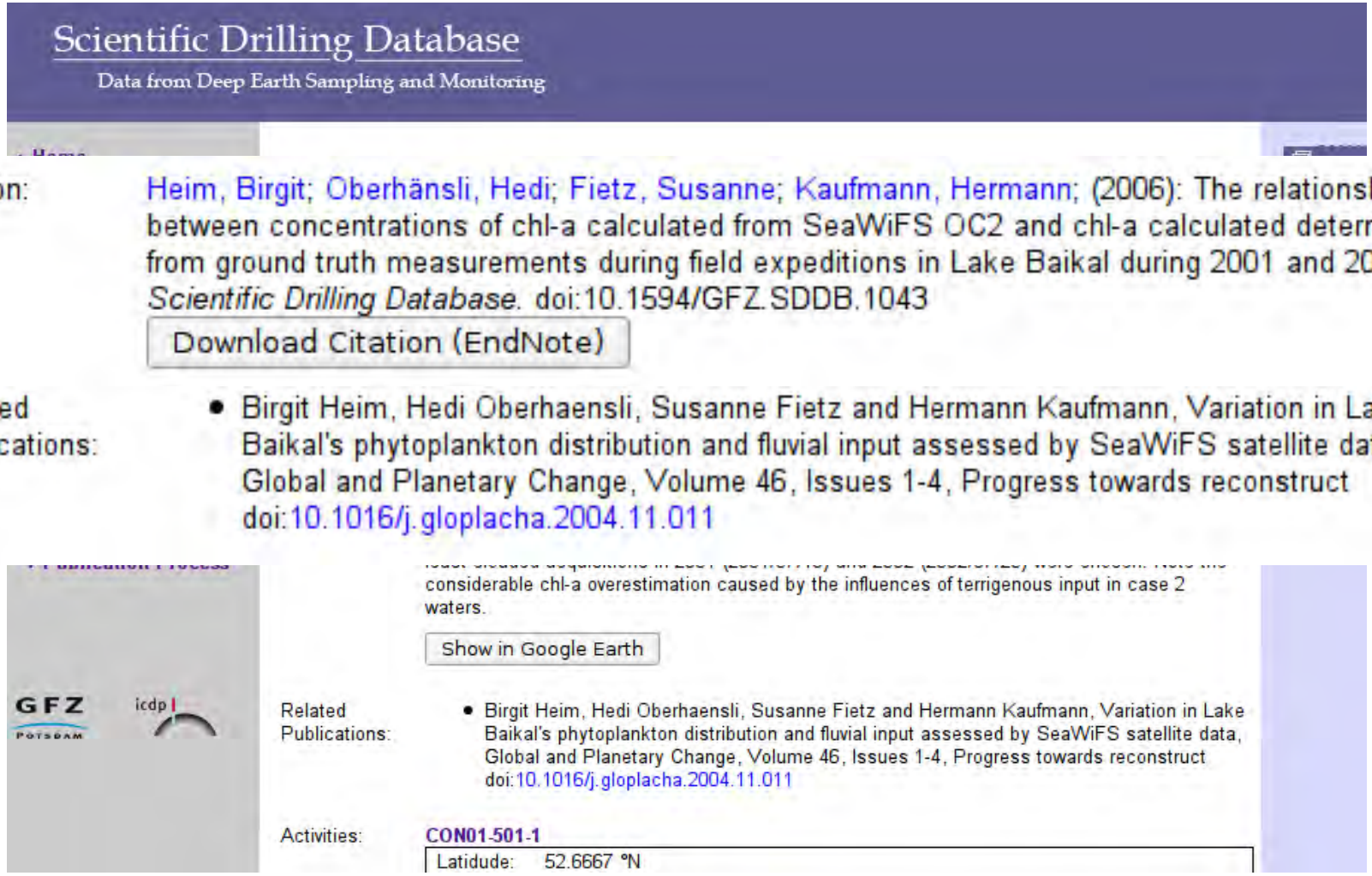
- Most data NOT copyrightable
  - facts cannot be copyrighted
  - limited protection for databases in EU
- Licenses (e.g. CC licenses) provide a work-around.

- 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

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

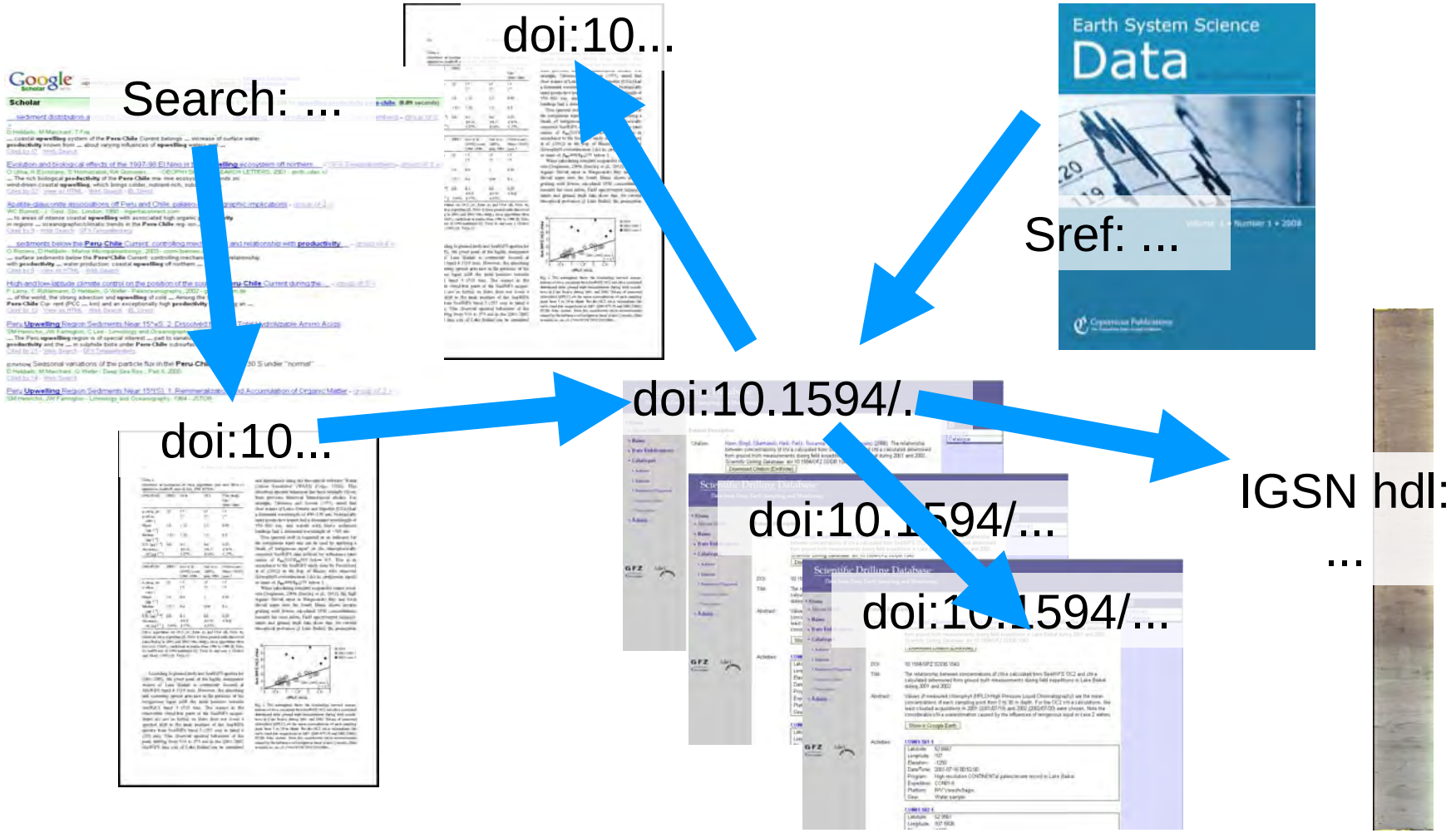
Latidude: 52.6667 °N

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)

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