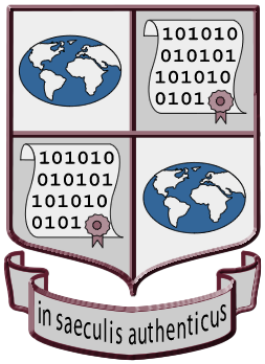
InterPARES 2 Project

International Research on Permanent Authentic Records in Electronic Systems



Accessing Scientific Data in the Future?

Tracey P. Lauriault
Geomatics and Cartographic Research Centre, Carleton University, *tlauriau@magma.ca*

Barbara L. Craig
Faculty of Information Studies, University of Toronto, barbara.craig@utoronto.ca
Research Assistant: Sherry Xie, UBC

Outline

- Portal and Case Studies
- Why care about data?
- Data
- Portals
- Authenticity, Accuracy, Reliability
- Metadata
- Records
- Preservation
- Concluding remarks

Why care about data?

Why Preserve Scientific Data?

"Scientific data represent both the organization and chaos of the natural world. They stimulate us to develop new concepts, theories, and models to make sense of the patterns they represent. The resulting abstractions are the product of scientific endeavour, the goal being to develop the formal and systematic ideas that constitute the understanding of relationships between causes and consequences and perhaps may enable prediction of future sequences of events. Because scientists transform data from the material world into ideas, the observations of objects and processes in the physical world are the stimuli for scientific thought. **Data are thus the seeds of scientific thought**."

National Research Council: Commission on Physical Sciences, Mathematics, and Applications. Preserving Scientific Data on Our Physical Universe: A New Strategy for Archiving the Nation's , Scientific Information Resources

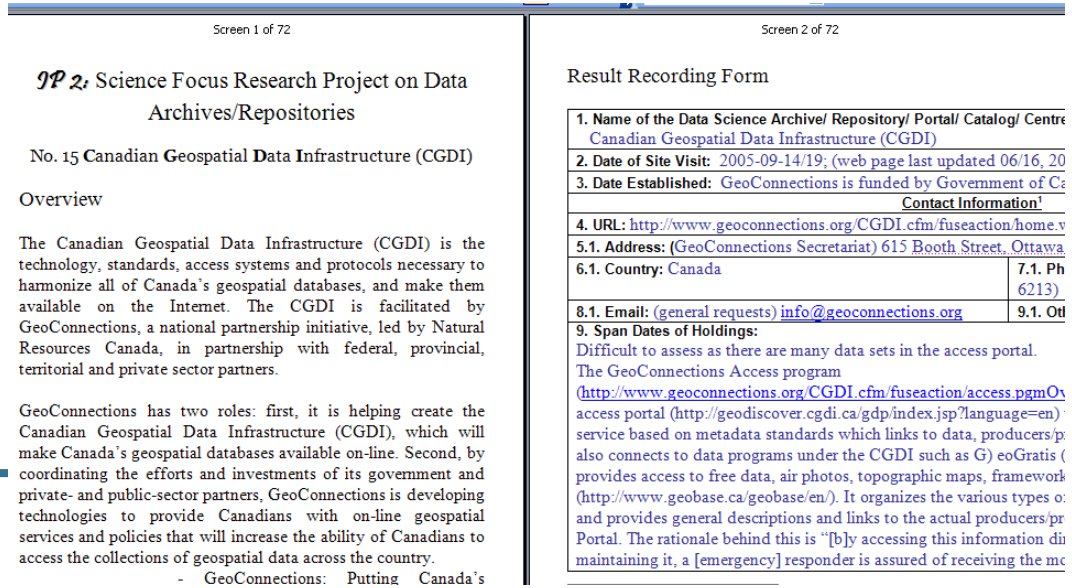
Portal & Case Studies

Portal Selection

- IP2 Focus 2 members proposed 2 – 5 portals each
  - computer science, computer engineering, geomatics, space sciences, astronomy, chemistry, and archives
- 72 portals from a variety of scientific and research disciplines were provided
- 32 short listed

Data Collection

- Structured information was collected in a survey
- RAs @ UBC School of Library, Archival and Information Studies completed the surveys
- Each survey was reviewed
- Some agencies were contacted for additional information or clarification



Screen 1 of 72

IP 2: Science Focus Research Project on Data Archives/Repositories

No. 15 **C**anadian **G**eospatial **D**ata **I**nfrastructure (CGDI)

Overview

The Canadian Geospatial Data Infrastructure (CGDI) is the technology, standards, access systems and protocols necessary to harmonize all of Canada's geospatial databases, and make them available on the Internet. The CGDI is facilitated by GeoConnections, a national partnership initiative, led by Natural Resources Canada, in partnership with federal, provincial, territorial and private sector partners.

GeoConnections has two roles: first, it is helping create the Canadian Geospatial Data Infrastructure (CGDI), which will make Canada's geospatial databases available on-line. Second, by coordinating the efforts and investments of its government and private- and public-sector partners, GeoConnections is developing technologies to provide Canadians with on-line geospatial services and policies that will increase the ability of Canadians to access the collections of geospatial data across the country.

- GeoConnections: Putting Canada's

Screen 2 of 72

Result Recording Form

1. Name of the Data Science Archive/ Repository/ Portal/ Catalog/ Centre Canadian Geospatial Data Infrastructure (CGDI)
2. Date of Site Visit: 2005-09-14/19; (web page last updated 06/16, 20
3. Date Established: GeoConnections is funded by Government of Ca
Contact Information[1]
4. URL: http://www.geoconnections.org/CGDI.cfm/fuseaction/home.v
5.1. Address: (GeoConnections Secretariat) 615 Booth Street, Ottawa
6.1. Country: Canada / **7.1. Ph** 6213)
8.1. Email: (general requests) info@geoconnections.org / **9.1. Otl**
9. Span Dates of Holdings: Difficult to assess as there are many data sets in the access portal. The GeoConnections Access program (http://www.geoconnections.org/CGDI.cfm/fuseaction/access.pgmOv access portal (http://geodiscover.cgdi.ca/gdp/index.jsp?language=en) service based on metadata standards which links to data, producers/p also connects to data programs under the CGDI such as G) eoGratis (provides access to free data, air photos, topographic maps, framework (http://www.geobase.ca/geobase/en/). It organizes the various types o and provides general descriptions and links to the actual producers/pr Portal. The rationale behind this is "[b]y accessing this information di maintaining it, a [emergency] responder is assured of receiving the mo

GS10 – Science Data Portals

IP2SF1 - **British Atmospheric Data Centre** (BADC)	IP2SF9 - **Animal Cognition Laboratory, department of Physics, University of Georgia Data Archive**	IP2SF19 - **National Virtual Observatory** (NVO)	IP2SF27 - **Antarctic Digital Database** (ADD)
IP2SF2 - **NASA Life Sciences Archive**	IP2SF10 - **World Data Center for Solar Terrestrial Physics**	IP2SF20 - **TeraGrid Data Collections**	IP2SF28 - **National Snow and Ice Data Center** (NSIDC), NASA
IP2SF3 - **University of Washington: Electrical Engineering Circuits Archive** (EECA)	IP2SF13 - OBIS-SEAMAP (**Ocean Biogeographic Information System** - Spatial Ecological Analysis of Megavertebrate Populations)	IP2SF21 - **Joint Center for Structural Genomics** (JCSG)	IP2SF29 - **U.S. Antarctic Resource Center** (USARC)
IP2SF4 - **Cambridge Crystallographic Data Centre**	IP2SF14 - **Canadian Institute for Health Information** (CIHI)	IP2SF22 - **San Diego Supercomputing Centre** (SDSC)	IP2SF30 - **British Antarctic Survey** - **Antarctic Environmental Data Centre**
IP2SF5 - **IU** (Indiana University) **Bio Archive**	IP2SF15 - **Canadian Geospatial Data Infrastructure** (CGDI) Access Portal	IP2SF23 - **Long Term Ecological Research** (LTER)	IP2SF32 - **Global Change Master Directory** – Global Change Data Center
IP2SF6 - **Computational Chemistry Archives**/ Computational Chemistry List	IP2SF16 - **The National Cancer Registry** (NCR) Ireland	IP2SF24 - **Southern California Earthquake Center** (SCEC)	IP2SF35 - **Community Data Portal at NCAR**
IP2SF7 - The **FMRI Data Center** (fMRIDC) [Functional MRI]	IP2SF17 - **National Institutes of Health** (NIH)	IP2SF25 - **International Comprehensive Ocean Atmosphere Data Set** (ICOADS)	IP2SF36 - **Earth Systems Grid** (ESG) portal
IP2SF8 - NIST (National Institute of Standards and Technology) **StRD Statistical Reference Data Sets** (Dataset Archives)	IP2SF18 - **Statistics Canada**	IP2SF26 - **National Geophysical Data Center** (NGDC - NOAA)	IP2SF37 - **USGS** Data Portals - **GEO-DATA Explorer** (GEODE)

IP2 Case Studies

- Each Foci conducted case studies
  - activities creating the records,
  - their purpose,
  - their phases and the component actions,
  - their by-products
  - and their structure,
  - and their context,
  - their technological environment
  - their use.
    - Semi structured interviews
    - 23 Question survey

Case Studies

CS06 - **Cybercartographic Atlas of Antarctica** (Lauriault and Hackett 2005)	CS14 - **Coalescent Communities in Arizona** (O'Meara, Pearce-Moses & Preston 2004)	CS24 - City of Vancouver Geographic Information System **(VanMap)** (McLellan 2005)
CS08 - **Mars Global Surveyor Data Records in the Planetary Data System** (Underwood 2005)	CS18 - **Computerization of Alsace-Moselle's Land Registry** (Blanchette 2004)	CS26 - **Most Satellite Mission: Preservation of Space Telescope** Data (Ballaux 2005)
	CS19 - **Authenticating Engineering Objects for Digital Preservation** (Hawkins 2005)	

Scientific Data

What are scientific data?

- "numerical quantities or other factual attributes generated by scientists and derived during the research process (through observations, experiments, calculations and analysis)" *(CODATA 2007).*
- "facts, ideas, or discrete pieces of information, especially when in the form originally collected and unanalyzed" *(Pearce-Moses 2005)*
- "numbers, images, video or audio streams, software and software versioning information, algorithms, equations, animations, or models/simulations" *(NSF 2005).*

Types of data

- Raw or Level 0 data
- Processed Data
- Refined data / Synthesized data
- Intermediate data

Data - characteristics

- Data can be distinguished by how they were collected
  - Observational data
  - Computational data
  - Experimental data

NOAA Example

1) Original Data;

2) Synthesized Products;

3) Interpreted Products;

4) Hydrometeorological, Hazardous Chemical Spill, and Space Weather Warnings, Forecasts, and Advisories;

5) Natural Resource Plans;

6) Experimental Products; and

7) Corporate and General Information.

Observation

- Science is a heterogeneous discipline
- Scientific data are complex
- Data are more than organized discrete facts and observations
- Some data are inseparable from their proprietary software
- It is not possible for one person to embody the knowledge required to manage data from all of the sciences

Portals

Portals *(gen. def.)*

- **Content aggregators:**
  - services that collect and provide access to content created and produced by others. Content aggregators are often sector specific and collect either a single type of content or content in a given discipline (McDonald & Shearer 2005).
- **A web portal:**
  - a website that provides a gateway to other resources on the Internet or an intranet (in the case of enterprise information portals). Unlike aggregators, portals do not house content, but links to content held elsewhere (McDonald & Shearer 2005).

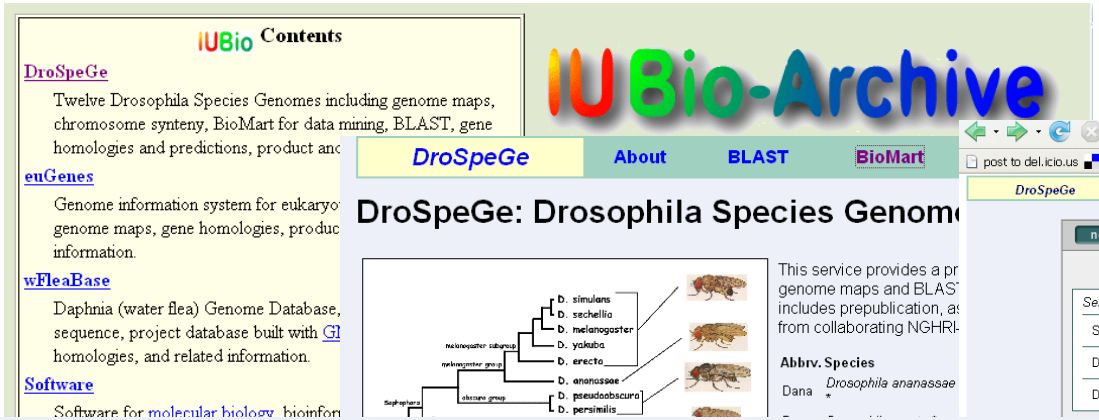
Portal Services

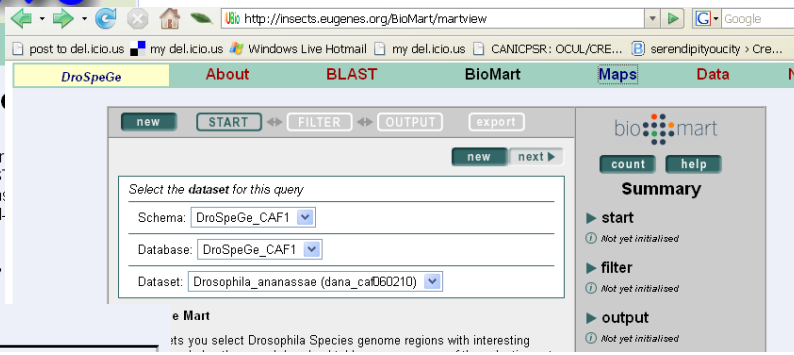
- Discovery and access to data,
- Item descriptions – metadata,
- Data registration,
- Display services,
- Data processing,
- Data dissemination,
- Data integration,
- The platform to share models and simulations,
- The collection and maintenance of data.

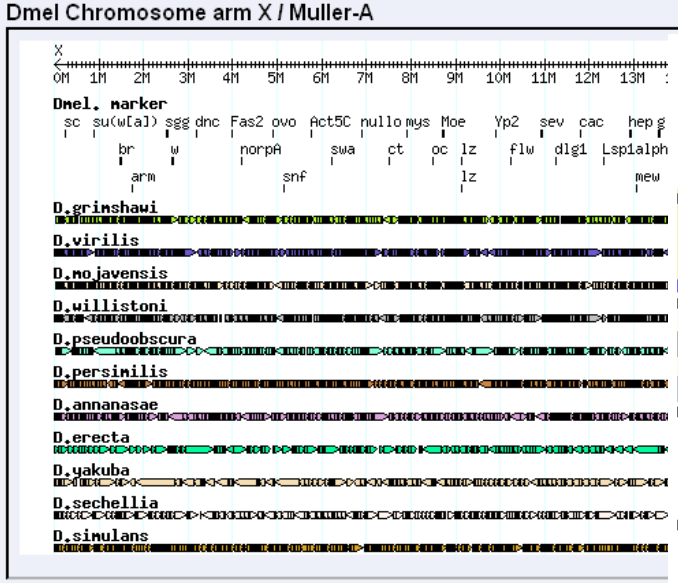
Types of Portals

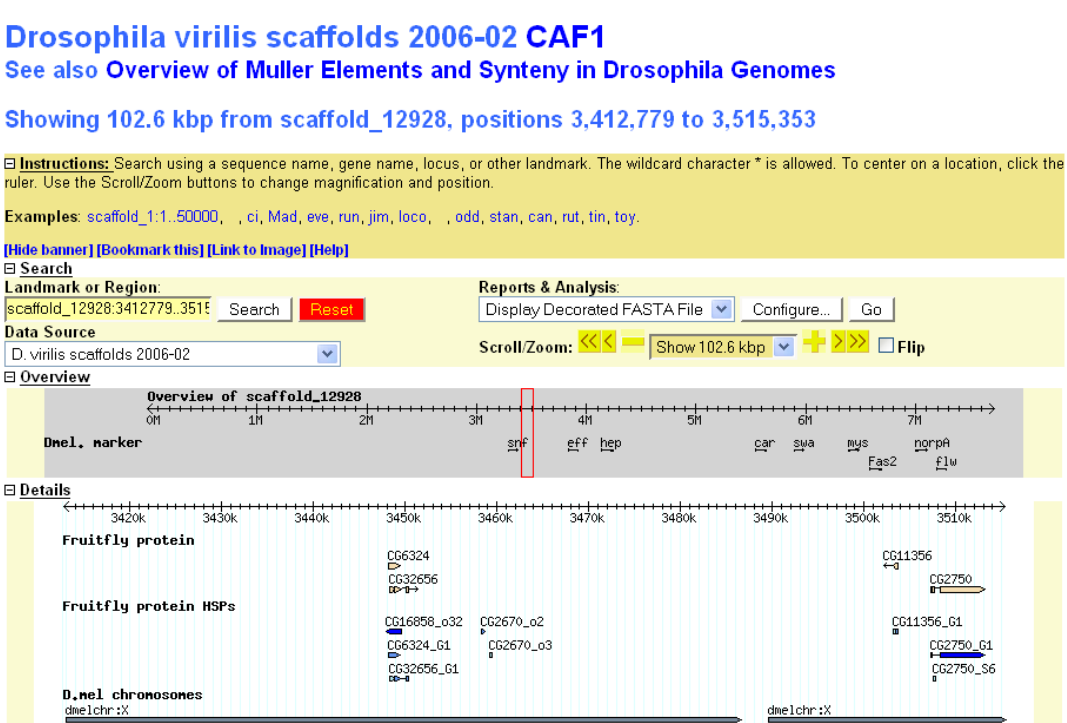
- Distributed portal
- Collection level catalog/portal
- Unified catalogue

Research Data Collection e.g.









InterPARES 2 Project, Science Focus
Tracey P. Lauriault, Barbara L. Craig
Session 306, SAA, Chicago 2007

Resource/community data collections

welcome to the southern california earthquake data center

faults of Southern California

Recent Earthquakes in

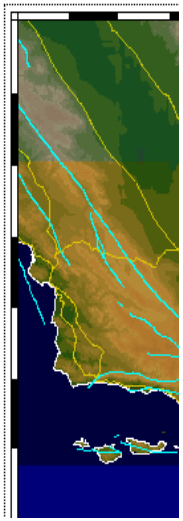
Searchable Earthquake

Clickable Fault Map of

Historic Earthquakes i

Below is a map of sout
1. **Southern Coast Ra**
2. **Sierra Nevada an**
3. **Mojave region** is y
4. Extreme **southern**
5. **Los Angeles** area i

This map is clickable. C
area, with local faults h
information about these
which is visible will be
and highway labels will



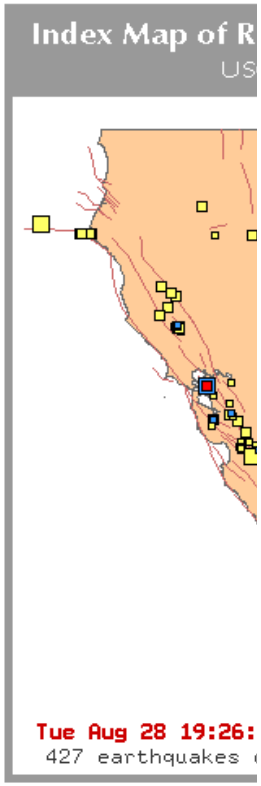
Supported by:

SCEC

an NSF+USGS center

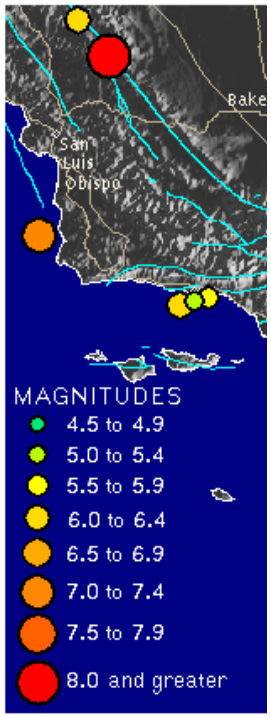
scie

recent earthquakes in California and Nevada



historic earthquakes in Southern California

Below is a clickable map of
back as 1812) of particular
major highways (in tan) an



southern california catalogs

1932–Present* Earthquake Catalog

Use the form below to extract data for selected events from this catalog.
Please limit your search to smaller time periods.

Search by:

LOCATION, MAG, AND TIME | EVENT ID | POLYGON | RADIUS | MULTI-MAGS | MOMENT TENSORS

CATALOG TO SEARCH: SCSN

Output Format: SCEDC

SEARCH PARAMETERS:

Start date:	Year: 2007	Month: 08	Day: 28	Hour: 00	Min: 00	Sec: 00
End date:	Year: 2007	Month: 08	Day: 29	Hour: 00	Min: 00	Sec: 00

Minimum magnitude:	0.0	Maximum magnitude:	9.0
Minimum depth (km):	0.0	Maximum depth (km):	700.0
Southern latitude:	32.0	Northern latitude:	37.0
West longitude:	-122.0	East longitude:	-114.0

Event type: local, Regional, Teleseism, Quarry blast, nuclear test, Sonic blast, Subnet Trigger

Reference data collections

GS10 - Type of Institution

- Government
- University Collaboration
- Collaborative Research
- NGO

GS10 - Target Users

- Researchers
- Academia
- Students
- General Public
- Government
- Private Sector
- Specialists
- Scientists
- Investigators
- Industry
- NGOs
- Policy Makers
- Planners

GS10 - Funding

- Government
- National Science Foundation
- National Institute for General Medical Sciences
- Bilateral funding with Ministers of Health
- Sales of Services &/or data
- Private Sector
- Project Funding

GS10 - Use rights & fees

- Use Restrictions
  - Strict privacy and confidentiality on health information
  - Copyright permission required
  - Acknowledgements required
  - User restrictions/screened users
  - Use disclaimers
- Access Fees
  - User and Data set dependent
  - Mostly no fees except for special requests
  - No Fees
  - Cost recovery
  - Fees for commercial users

Observation

- Portal is a framework around which archivists can expand policies, standards and metadata.
- Many of the difficult policy and legal issues have already been addressed.
- Scientists, research groups, funding agencies, data organizations and government have already appraised these data.

Data Quality

Data Quality and Science

- Wars, analysis and prediction of calamities, vacationing, real estate transactions, medical research, exploration etc. rely on accurate quality data. For centuries people have been willing to pay high prices for high quality data. Think of spies, planners, construction engineers, and especially those involved in the medical and military sciences who want more exacting data quality.
- Statistics Canada (IP2SF18) for example states that "the confidence of clients in the quality of that information is critical to the Agency's reputation as an independent, objective source of trustworthy information"

Data quality elements

- Positional accuracy,
- Attribute and thematic accuracy,
- Completeness,
- Semantic accuracy,
- Temporal information,
- Reliability,
- Lineage,
- Logical consistency,
- Objectivity.

Appraisal & data quality

"The relevant framework of appraising scientific data sets, thus, is not defined by the business activities or the need for corporate memory of the sponsoring agency, but by the research community. Seeking the input of scientists in the appraisal of the data recognizes that the roles and the actions of academic researchers are at least as important as the functions of the agency that funded the research or launched the satellite."

Kenneth Thibodeau,
"Preserving Scientific Data on Our Physical Universe",
IASSIST Quarterly, Winter (1995)

OMB Guidelines

- OMB Guidelines for Ensuring and Maximizing the Quality,
  - Objectivity,
  - Utility, and
  - Integrity of Information.
- Disseminated by Federal Agencies

Lineage

- Lineage is the history of the dataset, the dataset's pedigree as it changes form, its life cycle from collection to acquisition by a repository, through all the dataset's stages of conversion, correction and transformations, and its parentage.

IP2 Benchmarks of Authenticity

1) Identity and Integrity of a record
2) Access privileges
3) Protective procedures against loss and corruption
4) Protective procedures of media & technology
5) Establishment of documentary forms
6) Authentication Records
7) Identification of Authoritative record
8) Removal and transfer of relevant documentation

Presumption of Authenticity

- "An inference as to the fact of a record's authenticity that is drawn from known facts about the manner in which that record has been created and maintained" *InterPARES 2, Terminology and Glossary*
- For example, if "the authenticity of records and documents is usually presumed, rather than requiring affirmation. Federal rules of evidence stipulate that to be presumed authentic, records and documents must be created in the 'regular practice' of business and that there be no overt reason to suspect the trustworthiness of the record (Uniform Rules of Evidence, as approved July 1999)" *SAA Glossary, 2005*

Portals & Authenticity

- How data are ingested or made accessible in the system
  - peer review
  - control of contributors
  - data only from approved government programs
  - If data fits the mandate
- Once data are in the system
  - user authentication
  - registration
  - only trained personal have access
  - validation processes
- Transfer of data is a weak point

Accuracy

- "If accuracy can be considered to represent distance from the truth, then the truth should be known. But the truth cannot be known; it is instead accepted that the true position that could be obtained using the best available surveying techniques, personnel, uptodateness, etc." *Jane Drummond, 1995.*
- *"to a purist, no number has meaning unless it is accompanied by an estimate of uncertainty" and "at a minimum, the metadata should include general comments on the maximum expected errors, even if a quantitative measure such as standard deviation cannot be given" National Research Council: Commission on Physical Sciences, Mathematics, and Applications*

Accuracy and data life cycle

1) Collection

2) Compilation and

3) Derivation

Many sources of Error!

Reliability

- Associated with the concepts of reproducibility and accuracy in the sciences.
- The degree to which a forecast's or model's probabilities or results match the observed frequencies of an occurrence in the environment or consistently produce the same result.
- "Reliability is considered to be one of the foundations of trustworthiness" *Heather MacNeil, Trusting records: Legal, Historical, and Diplomatic Perspectives, 2000*

Data Quality Disclaimers

- Ironically, while most organizations aim to ensure their data are accurate, reliable and authentic, many of these same organizations will add disclaimers to absolve themselves of any responsibility for damages that may result from the use of their data.

Portals – Data Quality

- Expert reviewers
- Procedures
- Data checks
- Quality rests with data providers
- Quality Guidelines
- Methodology document
- Use will determine the quality
- Appear in the Metadata

- Applicants are screened
- Data validation process
- Technology calibrations
- Work with scientists
- Agency standards
- Data quality manuals
- Data quality assurances

Case Studies – Data Quality

- Trust the professional practices of the data providers
- Only trained professionals
- Data processing plans, manuals, specifications
- Peer review and data validation
- Ground truth
- Data are modelled

Observations

- Accuracy is associated with the risk of having inaccurate data: the more legal requirements, the higher the cost of mistakes and risk to human health there are, the more rigorous are the quality checks.
- Data quality is scientific discipline specific and mandate of the organization specific
- Metadata are critical
- Data quality is complex
- Archivists will need to work with data creators

Metadata

Metadata

- "A data set without metadata, or with metadata that do not support effective access and assessment of data lineage and quality, has little long-term use." *National Science Foundation, Report of the National Science Board, 2005*
- "To make data useable it is necessary to preserve adequate documentation relating to the content, structure, context, and source (e.g., experimental parameters and environmental conditions) of the data collection – collectively called metadata. Ideally, the metadata are a record of everything that might be of interest to another researcher." *National Research Council: Commission on Physical Sciences, Mathematics, and Applications, 1995,*

Portals & Case Studies - Metadata

- Processes but no metadata
- FGDC Content Standard for Digital Metadata
- ISO 19115
- Methodology documents with each dataset
- Resource Metadata VSO
- Registry Service
- Ecological Metadata Language
- Federated of Digital Seismic Network System + other specialized systems
- Metadata Repositories
- Own Standard Described in a Manual
- Peer review article and associated documentation

Records

Record Archival Def.

- "a document made or received in the course of a practical activity as an instrument or a by-product of such activity, and set aside for action or reference" *InterPARES 2, Glossary and Terminology, 2006*
- *Five characteristics are required for a digital entity to be a record:*
  - *stable content and fixed form,*
  - *embedded action,*
  - *archival bond,*
  - *three persons (i.e., author, addressee, writer) and*
  - *an identifiable administrative and documentary context*

Record - Scientists

- Means data,
- databases, and
- related information (e.g., metadata).

Preservation

Does the portal have a statement related to archiving?

- Few have preservation strategies
  - TeraGRID provides archival storage
  - Southern California Earthquake Centre has an online data archive?
  - Archiving only published documents not data
  - International Comprehensive Ocean Atmospheric Data archive some historical and real time data
  - National Geophysical Data Centre archives some data?

Concluding remarks

General Observations

- The longer is the timeline the data set covers, the more robust the record of an event, experiment or simulation is.
- Data are not trusted without metadata
- Metadata include data quality parameters
- Errors and data limitations are implicit in science
- Data come in many formats, distributed
- Data are often intertwined with the proprietary systems that created them
- Interoperability across time and space

Observations

- Many types of portals each requiring specific data preservation strategies
- Portals are data discovery and dissemination systems
- Portals have authenticity measures in place
- Data in portals have already been appraised
- Science is a heterogeneous discipline and archivists will have to work with data creators.
- Definition of a record needs to be revised
- Funded science is not enveloped in preservation policies yet the public has paid for this research
- Digital Science Data archives are needed!