Data Citation

Mark A. Parsons
and the ESIP Preservation and Stewardship Cluster, esp. Ruth Duerr, Curt Tilmes, and Bruce Barkstrom.

GeoData 2011
Broomfield, CO
3 March 2011

Fair warning: you will move. please interrupt.
Tweet: #geodata2011


Text: 303-351-1650
The National Snow and Ice Data Center...

Manages and distributes scientific data

Performs scientific and informatics research

Supports data users

Educates the public about the cryosphere

http://nsidc.org

Images:

Top left: Near Real-Time SSM/I EASE-Grid Daily Global Ice Concentration and Snow Extent, 3 February 2008

Top right: Pancake ice seen off the bow of the research vessel Aurora Australis, in the Southern Ocean near Antarctica. Photo courtesy Ted Scambos.

Bottom right: Data from the Sea Ice Index, shown on Google Earth and distributed from our Web site in the form of a Quick Time movie. The left image is an animated time series of sea ice extent from 1979 to 2006; the static image on the right compares the extent in 2007.


Center: A section of the online data set documentation from Sea Ice Charts of the Russian Arctic in Gridded Format, 1933-2006 (http://nsidc.org/data/g02176.html).
ESIP Federation

- Consortium of 118 Earth Science-Related Partners
  - Type I (data centers)
  - Type II (researchers & science tool developers)
  - Type III (application developers – commercial & nonprofit)
  - Type IV (sponsors)
  - Neutral forum for community networking, collaboration & problem solving
  - Limited international participation

- NRC seeded the idea of a Federation
  - NASA created in 1998
  - Support from NASA, NOAA, and EPA
  - With growing engagement from: USGS, NSF, DOE and USDA

- Contact Carol Meyer (carolbmeyer@esipfed.org) for more info.

ESIP Federation is Very broadly based

While historically centered in NASA, it has evolved to have very strong NOAA and EPA participation

Type Is 18
Type IIs 48
Type IIIIs 50
Type IVs 2
PHILOSOPHICAL
TRANSACTIONS:
GIVING SOME
ACCOUNT
OF THE PRESENT
Undertakings, Studies, and Labours
OF THE
INGENIOUS
IN MANY
CONSIDERABLE PARTS
OF THE
WORLD.

Vol I.
For Anno 1665, and 1666.

In the SAVOY,
Printed by T. N. for John Martyn at the Bell, a little with-
out Temple-Bar, and James Alleby in Duck-Lane:
Printers to the Royal Society.
• What is a data citation?
• I know what a data author is
• I know what a DOI is
• I know how to apply a DOI (or other identifier) to a particular data set or data collection
• I know how to precisely describe data set versions and associate the versions with an identifier(s)
• I understand the difficulties of referring to the precise files/granules/values used

Data Citation Knowledge Continuum

Novice ← Data Citation Knowledge Continuum → Expert

continuum of knowledge break out
Test with
Purpose of Data Citation

- Credit for data creators and stewards
- Track impact of data set
- Accountability for creators and stewards
- Aid reproducibility through direct, unambiguous connection to the precise data used

- A location/reference mechanism not a discovery mechanism per se.
How data citation is currently done

- Citation of traditional publication that actually contains the data, e.g. a parameterization value.
- Not mentioned, just used, e.g., in tables or figures
- Reference to name or source of data in text
- URL in text (with variable degrees of specificity)
- Citation of related paper (e.g. CRU Temp. records recommend citing two old journal articles which do not contain the actual data or full description of methods)
- Citation of actual data set typically using recommended citation given by data center
- Citation of data set including a persistent identifier/locator, typically a DOI
The National Snow and Ice Data Center distributes a variety of different snow cover products derived from the Moderate Resolution Imaging Spectroradiometer (MODIS). The results of a quick analysis of how many scientific papers mention use of "MODIS Snow Cover Data" (according to Google Scholar) and how often the data sets themselves are formally cited show a huge disparity, illustrating the infrequency of proper data citation in practice. Moreover, the lack of data citation standards introduces the possibility that informal references to data do not point to the exact data set actually used.
• International Polar Year — http://ipydis.org/data/citations.html
• DataCite—a consortium of libraries and related organizations working to define a citation approach around DOIs. New schema just out.
• New CODATA Task Group in collaboration with ICSTI.
• DataVerse Network Project—a standard from the social science community using a Handle locator and “Universal Numerical Fingerprint” as a unique identifier.
• Most NASA DAACs and other data centers but with great variation in approach.
• USGS generally requests acknowledgement, but maps are cited, and a more formal approach was proposed in 2000.
• NOAA National Data Centers simply request acknowledgement.
• Overall approaches range from specific data citation, to general acknowledgement, to recommending citing a journal article or even a presentation.
“We found that few policies recommend robust data citation practices: in our preliminary evaluation, only one-third of repositories (n=26), 6% of journals (n=307), and 1 of 53 funders suggested a best practice for data citation. We manually reviewed 500 papers published between 2000 and 2010 across six journals; of the 198 papers that reused datasets, only 14% reported a unique dataset identifier in their dataset attribution, and a partially-overlapping 12% mentioned the author name and repository name. Few citations to datasets themselves were made in the article references section.”


So it’s pretty clear from all this that the author is not being fairly credited.
Some measurement issues  (Chen and Downs 2010)

- Data are not used in isolation – often different data are combined, used with models and other analytic techniques
- Impacts may be indirect, i.e., resulting from development of information, papers, tools, etc. that relied on derived data or products
- Impacts may be delayed, i.e., months or years for a peer-reviewed publication to be released, or a decision to be made and implemented
- Impacts may be unexpected, e.g., a new scientific discovery or a novel application of data collected for a different purpose
- Impacts may be hard to compare, e.g., in scientific, economic, or ethical terms

But it is still important to try:
- Need to justify investment in data acquisition, maintenance, distribution and long-term stewardship!
- Need to help community become more effective and efficient in data management and use!

regarding, tracking the impact...

Possible citation metrics

- Qualitative
  - Examples of data use and impacts in key papers, discoveries, decisions
  - Assessment of broader impacts such as influence of data on attitudes and thinking (e.g., Apollo 8 image!)
- Quantitative
  - Counts of papers that cite data in peer-reviewed journals
  - Weighted indicators of data citations (e.g., type/quality of citation, impact of journal)
- Quantitative and Qualitative:
  - Number of data citations in top peer-reviewed scientific journals and “key” reports by decision-makers
  - Data usage in other peer-reviewed journals, textbooks, reports, magazines, documentary films, online tools, maps, blogs, twitters, etc.
"Tracking Dataset Citations Using Common Citation Tracking Tools Doesn’t Work"

—Heather Pinowar, DataONE

- Traditional fields such as author and date too imprecise
- Web of Science, Scopus, and other tools don’t handle identifiers
Accountability

- A new standard of accountability in a post-climategate world
- Data “publication” needs to be tied to promotion, tenure, etc.
- Implies peer review — See AGU Position Statement on Data
- What is peer-review?
  - An assertion of accuracy or validity?
  - An audit of complete documentation and sound practice?
  - Related to but different than QA and automated tests.
  - How does it overlap with curation and stewardship?
- *Earth System Science Data* one approach, but not universally applicable.
- Open or informal review or usage comments within the metadata
- Versioning and transparency are essential!

Peer-review is normally the process of having independent, anonymous reviewers read papers and present "book-review"-like reports on papers under review. Reviewers are unpaid and unsupported, so quality is variable, although journals usually provide some guidance. Reviewers are expected to be acquainted with the relevant literature.

This may differ from approaches in other scholarly disciplines. In mathematics, there may be formal panels assigned by the relevant professional society to verify the correctness of a proof outline (two years and several dozens of people - e.g. four-color proof).

In Earth science, and particularly for carefully produced data, QA is embedded into the software and reviewed by the science team as part of producing data. If produced data fails QA, a team will normally review the failure and attempt to come up with appropriate remedial action.

Also, there are intercomparisons of data, although these would not usually be identified as "peer-reviews". The ones I'm familiar with in cloud property retrievals involve huge amounts of data.

Reviews are anonymous - to avoid conflicts of interest. These comments are NOT placed in metadata. Expect much criticism from the scientific communities that produce and use data.
Need two strategies

- Publishers and educators (incl. style guides) need to provide guidelines and requirements on how to cite data, esp. historical data.
- Data Centers need to provide consistent, precise citation recommendations for their data.
Basic data citation form and content

Per DataCite:
Creator. PublicationYear. Title. [Version]. Publisher. [ResourceType]. Identifier.

Per Parsons:
Author(s). Year. Title, [version]. [editor(s)]. Publisher. Location. [date accessed]. [subset used].

resource type is a short controlled list of “The general type of a resource.”
many other optional fields possible

Authors are those who put the intellectual effort into creating the data set. Study designers, algorithm developers, instrument designers, field team leaders, etc.
Gary King; Langche Zeng, 2006, "Replication Data Set for 'When Can History be Our Guide? The Pitfalls of Counterfactual Inference'" hdl:1902.1/DXRXCFAWPK UNF:3:DaYIT6QSX9r0D50ye+tXpA== Murray Research Archive [distributor]
Identifier vs. Locator

- **Human ID**: Mark Alan Parsons (son of Robert A. and Ann M., etc.)
  - every term defined independently (only unique in context/provenance)

- **Human Locator**: 1540 30th St., Room 201, Boulder CO 80303.
  - every term has a naming authority

- **Data Set IDs**: data set title, filename, database key, object id code (e.g. UUID), etc.

- **Data set Locators**: URL, directory structure, catalog number, registered locator (e.g. DOI), etc.

They’re different, but sometimes locator can be used as an ID (The person working in this position at this address). Hence the general use of the term “identifier “such as in DOI, which is better described a locator.
One of the main purposes of assigning DOI names (or any persistent identifier) is to separate the location information from any other metadata about a resource. Changeable location information is not considered part of the resource description. Once a resource has been registered with a persistent identifier, the only location information relevant for this resource from now on is that identifier, e.g., http://dx.doi.org/10.xx.

— DataCite Metadata Scheme for the Publication and Citation of Research Data, Version 2.0, January 2011 (my emphasis).
An assessment of identification schemes for digital Earth science data

<table>
<thead>
<tr>
<th>ID Scheme</th>
<th>Unique Identifier</th>
<th>Unique Locator</th>
<th>Citable Locator</th>
<th>Scientifically Unique ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>URL/N/I</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PURL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XRI</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Handle</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DOI</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARK</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LSID</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OID</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UUID</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


Debate in LSID community weakens it. - - an LSID is a locator; but also the ObjectID part of it is an Identifier and most people use a UUID for the ObjectID part of it

OID problem

ARK is a bit better than the rest of the locators because it has additional trust value ... maybe the color should have been more orange than yellow but I didn’t want to add more colors.
An assessment of identification schemes for digital Earth science data

<table>
<thead>
<tr>
<th>ID Scheme</th>
<th>Unique Identifier</th>
<th>Unique Locator</th>
<th>Citable Locator</th>
<th>Scientifically Unique ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>URL/N/I</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PURL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XRI</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Handle</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DOI</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARK</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LSID</td>
<td></td>
<td><strong>Green</strong></td>
<td><strong>Green</strong></td>
<td></td>
</tr>
<tr>
<td>OID</td>
<td></td>
<td><strong>Green</strong></td>
<td><strong>Green</strong></td>
<td></td>
</tr>
<tr>
<td>UUID</td>
<td></td>
<td><strong>Green</strong></td>
<td><strong>Green</strong></td>
<td></td>
</tr>
</tbody>
</table>

Debate in LSID community weakens it. - - an LSID is a locator; but also the ObjectID part of it is an Identifier and most people use a UUID for the ObjectID part of it

OID problem

ARK is a bit better than the rest of the locators because it has additional trust value ... maybe the color should have been more orange than yellow but I didn’t want to add more colors.
Why the DOI?

- Not perfect but well understood by publishers
- DataCite working with Thomson Reuters to get data citations in their index.

But...
- What is the citable unit?
- How do we handle different versions?
- What about “retired” data?

What is the citable unit with a DOI? A file? A collection of files? How many? Further, it is important to note that data products can be purged from an archive; such deleted information still needs to be able to be referenced. Even if the products themselves are not preserved, the raw data must be preserved along with detailed documentation describing how the product was created, and that documentation must be citable.

Note the serious concerns within the library community regarding the economic stranglehold publishers wish to exert over publications - although much of the data have been provided at public expense and are usually "published" through publically funded archives and repositories.
Versioning and locators: some initial suggestions from NSIDC

- major version.minor version.[archive version]
- Individual stewards need to determine which are major vs. minor versions and describe the nature and file/record range of every version.
- Assign DOIs to major versions.
- Old DOIs should be maintained and point to some appropriate page that explains what happened to the old data if they were not archived.
- A new major version leads to the creation of a new collection-level metadata record that is distributed to appropriate registries. The older metadata record should remain with a pointer to the new version and with explanation of the status of the older version data.
- Major version (after the first version) should be captured in the data set title.
- Minor versions should be explained in documentation, ideally in file-level metadata.
- Minor versions should be exposed in the data set title and recommended citation.
- Applying UUIDs to individual files upon ingest aids in tracking minor versions and historical citations.
The hardest piece of all and the one most critical to sci. repeatability. One might think of this as citing a passage in a publication-- a page #
Page Numbers for Kindle Books an Imperfect Solution

Neither solution is perfect—“locations” or page numbers—because the problem is unsolvable. The best we can hope for is a choice...

Amazon’s Kindle will have page numbers that correspond to real books.

Chapter and Verse

• Bible

• Koran

• Bhagavad-Gita and Ramayana

• other sacred texts

• A “structural index”

A “structural index (chapter-verse)” vs. “internal location id (page number)”

structural index useful for when many different versions but with fairly consistent structure. Contrast Origin of Species–version must be noted.

“When, finally, in the nineteenth century, Dr. F. Scrivener, a scholar working to modern standards, attempted to collate all the editions of the King James Bible then in circulation, he found more than 24,000 variations between them. The curious fact is that no one such thing as ‘The King James Bible’–agreed, consistent and whole–has ever existed.”


The difficulty with the Unique Numerical Fingerprint approach is that it assumes that there is an immutable order to the data elements. This is not always the case therefore c-v needs to refer to “equivalence classes” not canonical versions.

But you can’t deny the human readability of the approach.

I guess my point with all this is that we probably need both approaches. We need the chapter and verse that makes sense to people and is easily conceived and communicated between people, but then we still need the precise location and identity of that mutable verse represented in a way that computers can readily understand and be precise about, i.e. the identifier.

And then we can’t forget the fact that we have billions if not trillions of “verses” or “granules” that we’re dealing with. Our human approach needs to make sense at a high level of aggregation, while the computer approach needs to handle the volumes and precision.

another option is to capture query and date—this relies on really good provenance tracking.

The person credited with dividing the Bible into chapters is Stephen Langton, the Archbishop of Canterbury from 1207–1228. While Langton’s isn’t the only organizational scheme that was devised, it is his chapter breakdown that has survived.

On p. 221, Nicolson comments that “There is, on the whole, no telling that this text [the KJV] has been assembled like a mosaic floor, every tessera gauged and weighed, held up, examined, placed, replaced, rejected, reabsorbed, a winnowing of exactness from a century of scholarship.”

On p. 224, he notes that the text “they delivered to Roger Barker was not entirely good. The Hebrew and particularly the Greek texts they were working from were not the most accurate, even by the standards of their own time.”

Chaos in printing means that no copy of the 1611 Bible is like any other.

The books of the Bible are considered canonical number 24 for Jews, 66 for Protestants, 73 for Catholics, and 78 for most Orthodox Christians not to mention the apocrypha. Some chapter divisions also occur in different places,

e.g. 1 Chronicles 5:27-41 in Hebrew Bibles is numbered as 1 Chron 6:1-15 in Christian
The “Archive Information Unit”

“An Archival Information Package whose Content Information is not further broken down into other Content Information components, each of which has its own complete Preservation Description Information. It can be viewed as an ‘atomic’ AIP”

“From an Access viewpoint, new subsetting and manipulation capabilities are beginning to blur the distinction between AICs and AIUs. Content objects which used to be viewed as atomic can now be viewed as containing a large variation of contents based on the subsetting parameters chosen. In a more extreme example, the Content Information of an AIU may not exist as a physical entity. The Content Information could consist of several input files (or pointers to the AIPs containing these data files) and an algorithm which uses these files to create the data object of interest.”

Citation scenarios and production patterns

- What kind of “atomic” item is being cited—the “Archive Information Unit (AIU)” (e.g., a data file, a data element within a file, a relational (or other) database, a job “residue”)?

- How many AIUs items are in a typical citation for the scenario being considered?

- What other digital or physical objects need to be available to make the unit usable—the “Preservation Description Information (PDI)”?

**Key Question:**

- What structure or structures can we use to organize data collections that might be common across Earth sciences?

It would probably be wise to distinguish between cases where the file collections are open, but adding relatively stable items, from ones where the files themselves are mutable.
An example production pattern for Cline et al. (2003).
A production pattern for Cline et al., 2003
A production pattern for Cline et al., 2003

<table>
<thead>
<tr>
<th>TRANSECT</th>
<th>DATE</th>
<th>TIME</th>
<th>UTME</th>
<th>UTMN</th>
<th>DEPTH</th>
<th>SWET</th>
<th>SRUF</th>
<th>CNPY</th>
<th>TEMP</th>
<th>SURVEYOR</th>
<th>QC</th>
<th>COMMENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>FAA01.1</td>
<td>iop4</td>
<td>2003-03-25</td>
<td>1017</td>
<td>425941</td>
<td>4410860</td>
<td>104</td>
<td>d</td>
<td>y</td>
<td>n</td>
<td>-999</td>
<td>&quot;Fitzgerald, Matous, Dundas&quot;</td>
<td>QC(000)</td>
</tr>
<tr>
<td>FAA01.2</td>
<td>iop4</td>
<td>2003-03-25</td>
<td>1017</td>
<td>425956</td>
<td>4410868</td>
<td>13</td>
<td>d</td>
<td>y</td>
<td>n</td>
<td>-999</td>
<td>&quot;Fitzgerald, Matous, Dundas&quot;</td>
<td>QC(000)</td>
</tr>
<tr>
<td>FAA04.1</td>
<td>iop4</td>
<td>2003-03-25</td>
<td>1221</td>
<td>425938</td>
<td>4411193</td>
<td>325</td>
<td>d</td>
<td>y</td>
<td>n</td>
<td>-999</td>
<td>&quot;Fitzgerald, Matous, Dundas&quot;</td>
<td>QC(000)</td>
</tr>
</tbody>
</table>

- Field notebook
- Excel v1 printout
- Excel v2

- Distributed data set
- Camera
- Interim jpgs
- Collated and named jpgs

- Born digital
- Analog to digital with QC
- Ascii files
- Shapefiles
- Camera
- Interim jpgs
- Collated and named jpgs

- Camera
- Interim jpgs
- Collated and named jpgs
<table>
<thead>
<tr>
<th>TRANSECT</th>
<th>DATE</th>
<th>TIME</th>
<th>UTME</th>
<th>UTMN</th>
<th>DEPTH</th>
<th>SWET</th>
<th>SRUF</th>
<th>CNPY</th>
<th>TEMP</th>
<th>SURVEYOR</th>
<th>QC</th>
<th>COMMENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>FAA01.1</td>
<td>iop4, 2003-03-25, 1017</td>
<td>425941, 4410860</td>
<td>104</td>
<td>d, y, n</td>
<td>-999</td>
<td>&quot;Fitzgerald, Matous, Dundas&quot;</td>
<td>QC(000)</td>
<td>Couldn't find post, used GPS 5940 1197; FAA4.4 and FAA4.5 unsafe, avalanche area!</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FAA01.2</td>
<td>iop4, 2003-03-25, 1017</td>
<td>568, 44</td>
<td>d, y, n</td>
<td>-999</td>
<td>&quot;Fitzgerald, Matous, Dundas&quot;</td>
<td>QC(000)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FAA02.1</td>
<td>iop4, 2003-03-25, 1017</td>
<td>364, 44</td>
<td>d, y, n</td>
<td>-999</td>
<td>&quot;Fitzgerald, Matous, Dundas&quot;</td>
<td>QC(000)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FAA02.2</td>
<td>iop4, 2003-03-25, 1017</td>
<td>364, 44</td>
<td>d, y, n</td>
<td>-999</td>
<td>&quot;Fitzgerald, Matous, Dundas&quot;</td>
<td>QC(000)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FAA03.1</td>
<td>iop4, 2003-03-25, 1017</td>
<td>364, 44</td>
<td>d, y, n</td>
<td>-999</td>
<td>&quot;Fitzgerald, Matous, Dundas&quot;</td>
<td>QC(000)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FAA03.2</td>
<td>iop4, 2003-03-25, 1017</td>
<td>364, 44</td>
<td>d, y, n</td>
<td>-999</td>
<td>&quot;Fitzgerald, Matous, Dundas&quot;</td>
<td>QC(000)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**A production pattern for Cline et al., 2003**

- Field notebook
- Collated and named jpgs
- Interim jpgs
- Shapefiles
- Ascii files
- Excel v1
- Excel v2
- 100s
- 1000s
- Digital to analog w/ QC
## A production pattern for Cline et al., 2003

### Field notebook

- **Excel v1**
  - **Printout**
  - **Excel v2**

### HTML Doc.

- Analog to digital w/ QC
- Born digital

### Camera

- Interim jpgs
- Collated and named jpgs

### Data Conversion

<table>
<thead>
<tr>
<th>Transect</th>
<th>Sub</th>
<th>Date</th>
<th>Time</th>
<th>UTME</th>
<th>UTMN</th>
<th>Depth</th>
<th>SWET</th>
<th>SRUF</th>
<th>CNPY</th>
<th>TEMP</th>
<th>SURVEYOR</th>
<th>QC</th>
<th>COMMENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>FAA01.1</td>
<td>iop4</td>
<td>2003-03-25</td>
<td>1017</td>
<td>425941</td>
<td>4410860</td>
<td>104</td>
<td>d</td>
<td>y</td>
<td>n</td>
<td>-999</td>
<td>Fitzgerald, Matous, Dundas</td>
<td>QC</td>
<td>Couldn't find post, used GPS 5940 1197; FAA4.4 and FAA4.5 unsafe, avalanche area!</td>
</tr>
<tr>
<td>FAA01.2</td>
<td>iop4</td>
<td>2003-03-25</td>
<td>1017</td>
<td>425956</td>
<td>4410860</td>
<td>13</td>
<td>d</td>
<td>n</td>
<td>n</td>
<td>-999</td>
<td>Fitzgerald, Matous, Dundas</td>
<td>QC</td>
<td></td>
</tr>
</tbody>
</table>

### Files

- HTML Doc.
- Distributed data set
- ASCII files
- Shapefiles
- Camera
- Excel v1
- Excel v2
- Interim jpgs
<table>
<thead>
<tr>
<th>TRANSEC</th>
<th>CODE</th>
<th>DATE</th>
<th>TIME</th>
<th>UTME</th>
<th>UTMN</th>
<th>DEPTH</th>
<th>SWET</th>
<th>SRUF</th>
<th>CNPY</th>
<th>TEMP</th>
<th>SURVEYOR</th>
<th>COMMENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>FAA01.1</td>
<td>iop4</td>
<td>2003-03-25</td>
<td>1017</td>
<td>425941</td>
<td>4410860</td>
<td>104</td>
<td>d</td>
<td>y</td>
<td>n</td>
<td>-999</td>
<td>Fitzgerald, Matous, Dundas</td>
<td>QC(000)</td>
</tr>
<tr>
<td>FAA01.2</td>
<td>iop4</td>
<td>2003-03-25</td>
<td>1017</td>
<td>425956</td>
<td>4410860</td>
<td>13</td>
<td>d</td>
<td>n</td>
<td>n</td>
<td>-999</td>
<td>Fitzgerald, Matous, Dundas</td>
<td>QC(000)</td>
</tr>
<tr>
<td>FAA04.1</td>
<td>iop4</td>
<td>2003-03-25</td>
<td>1221</td>
<td>425938</td>
<td>4411193</td>
<td>325</td>
<td>d</td>
<td>n</td>
<td>n</td>
<td>-999</td>
<td>Fitzgerald, Matous, Dundas</td>
<td>Couldn’t find post, used GPS 5960 1197; FAA4.4 and FAA4.5 unsafe, avalanche area!</td>
</tr>
</tbody>
</table>

A production pattern for Cline et al., 2003
<table>
<thead>
<tr>
<th>TRANSECT, IOP, DATE, TIME, UTME, UTMN, DEPTH, SWET, SRUF, CNPY, TEMP, SURVEYOR, QC, COMMENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>FAA01.1, iop4, 2003-03-25, 1017, 425941, 4410860, 104, d, y, n, -999, &quot;Fitzgerald, Matous, Dundas&quot;, QC(000)</td>
</tr>
<tr>
<td>FAA01.2, iop4, 2003-03-25, 1017, 425956, 4410860, 13, d, n, n, -999, &quot;Fitzgerald, Matous, Dundas&quot;, QC(000)</td>
</tr>
<tr>
<td>FAA04.1, iop4, 2003-03-25, 1221, 425938, 4411193, 325, d, y, n, -999, &quot;Fitzgerald, Matous, Dundas&quot;, QC(000)</td>
</tr>
<tr>
<td>FAA04.2, iop4, 2003-03-25, 1221, 425956, 4411193, 325, d, y, n, -999, &quot;Fitzgerald, Matous, Dundas&quot;, QC(000)</td>
</tr>
</tbody>
</table>

Couldn't find post, used GPS 5940 1197; FAA4.4 and FAA4.5 unsafe, avalanche area!
Crude production pattern for MODIS/Aqua Snow Cover Daily L3 Global 500m Grid V005 (Hall et al., 2007)

L# files records interim files not L2.
Crude production pattern for MODIS/Aqua Snow Cover Daily L3 Global 500m Grid V005 (Hall et al., 2007)

L# files records interim files not L2.
Crude production pattern for MODIS/Aqua Snow Cover Daily L3 Global 500m Grid V005 (Hall et al., 2007)

L# files records interim files not L2.
More examples

• Glacier Photos: chemical creation of image, digitization, multi-source, no versioning
• Hurricane Ike: direct digital creation, embedded software, single-source, no versioning
• Rock or Ice Cores: physical specimens with later experimental operations to create data collections – digital results may be stored in databases with complex provenance
• SeaDataNet: Many individual Conductivity, Temperature, Depth measurements from many individual cruises compiled into one database.
• Global Historical Climate Network: In situ measurements of Temp and Humidity, recorded as time series and appended to files (with some quality control) – multi-source, some versioning
• Radiosonde Network: Balloon-borne Temp and Humidity from about 20,000 sites every 12 or 6 hours assimilated into weather forecasts – archives maintained at NCAR, GSFC, and elsewhere – reanalysis can create new versions
• Large-scale Satellite Data Production: complex, large-scale data flows from multiple sources, systematic approaches to versioning where the structure of one version of a data product resembles the next
Doing it as best we can...?


Hypothesis: ~80% of citation scenarios for geospatial data can be addressed with basic citations (Author(s). Year. Title, [version]. [editor(s)]. Publisher. Location. [date accessed]. [geospatial and temporal subset used].) and reasonable due diligence.

Proposal: We need a defined research agenda for the last 20%
Future directions on scientific equivalence
Content equivalence and provenance equivalence serve as a proxy for scientific equivalence.

**Content Equivalence:** Is there an algorithm that can consider the content of a file and come up with a unique identifier that will be the same for objects in the same “content equivalence class”?

- Exact content equivalence can use digital signature or cryptographic techniques MD5, SHA-1, etc.
- Loose or scientific content equivalence. Universal Numeric Footprint (UNF) takes a digital signature of a “canonical” representation of the information, but canonical is a problem.
Content equivalence and provenance equivalence serve as a proxy for scientific equivalence.

**Provenance Equivalence:** A process is reproducible when there is sufficient creation provenance details for someone else to make an equivalent file. If someone follows those provenance details to re-create the object, the resulting object will be equivalent to the original.

- If we can enumerate/list sufficient or “essential” creation provenance details to make an equivalent file, then we can describe an algorithm to produce an identifier that will be the same for files that match those provenance details.

- Algorithm can be similar to the content equivalence algorithm before: take a digital signature of a canonical representation of the information in this case a canonical serialization of processes.
Thank You
parsonsm@nsidc.org

photo courtesy NOAA
Much of this talk comes from:


- A lot of discussion at: http://wiki.esipfed.org/index.php/Preservation_and_Stewardship

photo courtesy NOAA