Practical Issues for Earth Science Data Citations

B. R. Barkstrom

Computational Issues in Data Citations

April, 2014
1. Citing Needles in Haystacks

- Earth science projects may create hundreds of thousands of data objects – usually files
  - 30 years $\approx 10,000$ days, $10,000$ jobs per day $\Rightarrow 10^8$ files

- Precise specification of data used in an experiment may require referencing specific files and even specific subsets of data values
  - Example: Combine 20,000 radiosonde profiles with satellite data within 30 minutes of satellite overpasses when no clouds

- Typical publications have room for only 6 to 12 citations
- Accurate reproduction of results may require specifying many more than 12 objects
1. Citing Needles in Haystacks

- Earth science projects may create hundreds of thousands of data objects – usually files
  - 30 years ≈ 10,000 days, 10,000 jobs per day ⇒ $10^8$ files

- Precise specification of data used in an experiment may require referencing specific files and even specific subsets of data values
  - Example: Combine 20,000 radiosonde profiles with satellite data within 30 minutes of satellite overpasses when no clouds

- Typical publications have room for only 6 to 12 citations
- Accurate reproduction of results may require specifying many more than 12 objects
1. Citing Needles in Haystacks

- Earth science projects may create hundreds of thousands of data objects – usually files
  - 30 years \( \approx 10,000 \) days, 10,000 jobs per day \( \Rightarrow 10^8 \) files

- Precise specification of data used in an experiment may require referencing specific files and even specific subsets of data values
  - Example: Combine 20,000 radiosonde profiles with satellite data within 30 minutes of satellite overpasses when no clouds

- Typical publications have room for only 6 to 12 citations
  - Accurate reproduction of results may require specifying many more than 12 objects
1. Citing Needles in Haystacks

- Earth science projects may create hundreds of thousands of data objects – usually files
  - 30 years $\approx 10,000$ days, $10,000$ jobs per day $\Rightarrow 10^8$ files

- Precise specification of data used in an experiment may require referencing specific files and even specific subsets of data values
  - Example: Combine 20,000 radiosonde profiles with satellite data within 30 minutes of satellite overpasses when no clouds

- Typical publications have room for only 6 to 12 citations
- Accurate reproduction of results may require specifying many more than 12 objects
2. Not Every Object in an Earth Science Archive is Digital

- OAIS RM [p. 1-10]: “Data Object: Either a Physical Object or a Digital Object.”

- Identifiers may not necessarily be opaque [Note label with well permit # on box end in right figure].
2. Not Every Object in an Earth Science Archive is Digital

- OAIS RM [p. 1-10]: “Data Object: Either a Physical Object or a Digital Object.”

- Identifiers may not necessarily be opaque [Note label with well permit # on box end in right figure]
3. Replication and Non Uniqueness

- Archives Create Backups That are Replicas of Originals

**Case 1: Data Producer, Archives, and User**

<table>
<thead>
<tr>
<th></th>
<th>Producer</th>
<th>Archive 1</th>
<th>Archive 2</th>
<th>User</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active Storage</td>
<td><img src="image" alt="Green" /></td>
<td><img src="image" alt="Green" /></td>
<td><img src="image" alt="Green" /></td>
<td><img src="image" alt="Green" /></td>
</tr>
<tr>
<td>Online Backup</td>
<td><img src="image" alt="Green" /></td>
<td><img src="image" alt="Green" /></td>
<td><img src="image" alt="Green" /></td>
<td><img src="image" alt="Green" /></td>
</tr>
<tr>
<td>Offsite Backup</td>
<td><img src="image" alt="Green" /></td>
<td><img src="image" alt="Green" /></td>
<td><img src="image" alt="Green" /></td>
<td><img src="image" alt="Green" /></td>
</tr>
</tbody>
</table>

- Original File and All Copies (files are indistinguishable when tested by comparing cryptographic digests)

- Replicas are only distinguishable by physical location.
- Location may need to be part of the object ID

B. R. Barkstrom

Practical Issues
3. Replication and Non Uniqueness

- Archives Create Backups That are Replicas of Originals

Case 1: Data Producer, Archives, and User

<table>
<thead>
<tr>
<th></th>
<th>Producer</th>
<th>Archive 1</th>
<th>Archive 2</th>
<th>User</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active Storage</td>
<td>Green</td>
<td>Green</td>
<td>Green</td>
<td>Green</td>
</tr>
<tr>
<td>Online Backup</td>
<td>Green</td>
<td>Green</td>
<td>Green</td>
<td>Green</td>
</tr>
<tr>
<td>Offsite Backup</td>
<td>Green</td>
<td>Green</td>
<td>Green</td>
<td>Green</td>
</tr>
</tbody>
</table>

- Original File and All Copies (files are indistinguishable when tested by comparing cryptographic digests)

- Replicas are only distinguishable by physical location.
  - Location may need to be part of the object ID
4. Scientifically Equivalent Arrays - An Example

Consider an array of surface types:

- 18 surface types
- each type represented as a one-byte number
- numbers in a global, equal-area array (each pixel has an area of about 972.9 km$^2$)

**Scientific task:**
- Compute the area of Deciduous Broadleaf Forest

**Algorithm:**
1. Count number of array elements whose token value = 3, the pixel value for Deciduous Broadleaf Forest pixels
2. Multiply total number of pixels by the area of a single pixel
Note on Bitmaps

Classic Windows Bitmaps (.bmp files)

- Have two parts to the file
  - A header containing the array size and a color palette
  - An array of one-byte numbers

- Scientific data values are in the array

- If the array represents geolocated data,
  - Geolocation is implicit in the order of the token string with the data values OR
  - Geolocation is stored somewhere, perhaps outside the file
Note on Bitmaps

Classic Windows Bitmaps (.bmp files)

- Have two parts to the file
  - A header containing the array size and a color palette
  - An array of one-byte numbers
- Scientific data values are in the array
- If the array represents geolocated data,
  - geolocation is implicit in the order of the token string with the data values **OR**
  - geolocation is stored somewhere, perhaps outside the file
Note on Bitmaps

Classic Windows Bitmaps (.bmp files)
- Have two parts to the file
  - A header containing the array size and a color palette
  - An array of one-byte numbers
- Scientific data values are in the array
- If the array represents geolocated data,
  - geolocation is implicit in the order of the token string with the data values OR
  - geolocation is stored somewhere, perhaps outside the file
Example Visualization - Case 1
Example Visualization - Case 2
Scientific Equivalence of the Data Files

- The array sizes are identical: 1024 in x, 512 in y.
- The one-byte values that categorize the IGBP vegetation types use the same indexing scheme.
- The palette in the first visualization is different from the palette in the second visualization.
  - The cryptographic digest of the file with first visualization is not equal to the cryptographic digest of the file with the second visualization.
- Most scientists would treat these files as having scientifically equivalent data:
  
  They produce the same answer to the algorithm.