Metadata in the Wild

An Empirical Survey of OPeNDAP-accessible Metadata and its Implications for Discovery

Darren Hardy
University of California, Santa Barbara

and the DAPADL project: Greg Janée & James Frew at UCSB, James Gallagher & Peter Cornillon at OPeNDAP, Inc. & U Rhode Island. Funding from NSF.

2006 AGU Fall Meeting, San Francisco, CA, 15 December 2006
Example
OPeNDAP

• What is OPeNDAP?
  • Access remote data over the Internet
  • Standard, protocol, and software

Example
Problem
Problem

- Where are the data available via OPeNDAP?
- Discovery via hand-crafted directories (e.g., NASA’s GCMD)
- But, no global search service for data
Approach

Unified search service for OPeNDAP
Approach

Unified search service for OPeNDAP

1. Text
   “sea surface temperatures”
Approach

**Unified search service** for OPeNDAP

1. Text
   “sea surface temperatures”

2. Spatiotemporal coverage
   “in North Atlantic during last year”
Approach

Unified search service for OPeNDAP

1. Text
   “sea surface temperatures”

2. Spatiotemporal coverage
   “in North Atlantic during last year”

3. Spatiotemporal resolution
   “daily averages at 6km”
Design

Crawl

Google  GCMD

Discover

OPeNDAP servers
Survey

Crawl

Google

GCMD

Discover

Harvest

Search

OPeNDAP servers

Metadata

Index
Survey

- Goal: Characterize existing metadata
- Can we automate...
  - Text search?
  - Spatiotemporal coverage?
  - Spatiotemporal resolution?
- Are metadata conventions helping?
## Sample size

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Study Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Servers</td>
<td>213</td>
<td>162</td>
</tr>
<tr>
<td>Datasources</td>
<td>1,408,996</td>
<td>396,638</td>
</tr>
<tr>
<td>Variables</td>
<td>49,711,772</td>
<td>18,359,268</td>
</tr>
<tr>
<td>Attributes</td>
<td>349,319,571</td>
<td>57,833,469</td>
</tr>
</tbody>
</table>
Distribution of servers by domain (n = 213)

gov 31%
org 15%
others 21%
edu 34%
Suitability Analysis

Text search  Good  80% have some general description (e.g., title, institution, etc.)

Coverage  Not Bad  Lat / Lon / Time identifiable 80% of the time, but reliable bounds <10%

Resolution  Poor  Spatial 40%, Time <10%, and requires data access
Datasources in study set (n=396,638)

- None: 60%
- Stated: 35%
- Unknown: 5%

Stated Conventions

- COARDS: 74%
- CF: 16%
- GDT: 10%
- Unknown: 5%
- None: 60%

Stated: 35%
Results

• Survey: 200+ servers, 1.4M datasources
• Suitability of existing metadata
  • For text search? OK.
  • With coverage/resolution? Not really.
• Metadata conventions (e.g., COARDS)
  • 35% stated, plus 30% possible
  • But, not simplifying fundamental problems of heterogeneity
Future work

- Propose “summarization” convention
- Digital library (ADL) search service
- Compliance to conventions as feedback
Summary

- Global search service for OPeNDAP
- Survey of existing metadata
- Found support for text search, but only limited support for spatiotemporal coverage / resolution
- Suggest better “search-aware” metadata
Thank you.