Towards a Google for Data

R. Duerr, Soren Scott, Luis Lopez, Siri Jodha Khalsa
NSIDC affiliations and sponsorship

Cooperative Institute for Research in Environmental Sciences

World Data System

Main sponsors:

- NASA
- National Science Foundation
- National Oceanographic and Atmospheric Administration

United States

National Snow & Ice Data Center

1540 30th St
Boulder, CO 80303
(303) 492-6028
nsidc.org

Towards a Google for Data
The National Snow and Ice Data Center...

- Manages and distributes scientific data
- Provides tools for data access
- Researches the cryosphere and data science
- Supports local and traditional knowledge
- Supports data users
- Educates the public about the cryosphere
Repositories

- NSIDC
- PDB
- tDAR
- ICPSR
- NODC

Portals & Tools

Towards a Google for Data
Towards a Google for Data
So what are we finding?

<table>
<thead>
<tr>
<th>Type of Item Found</th>
<th>Number of that Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMS Server</td>
<td>408</td>
</tr>
<tr>
<td>WFS Server</td>
<td>95</td>
</tr>
<tr>
<td>OPeNDAP</td>
<td>1866</td>
</tr>
<tr>
<td>THREDDS</td>
<td>1014</td>
</tr>
<tr>
<td>SOAP</td>
<td>42</td>
</tr>
<tr>
<td>OpenSearch</td>
<td>1935</td>
</tr>
<tr>
<td>OAI feeds</td>
<td>107</td>
</tr>
<tr>
<td>KML</td>
<td>443</td>
</tr>
</tbody>
</table>

NOTE: Some items actually are in multiple categories.
Philosophical questions

• What is the data? Do PDF's really count?
• What are Data Sets? Should data acquired at one of possibly hundreds or thousands of sites count as its own data set or is really just a piece of something much larger (that can be subset and therefore cited by site)?
Are we finding all of it? NO!!!

- Is it mentioned anywhere on the web?
- Are robots allowed at all?
- What's in your Robots.txt file?
  - Do you prevents crawls of your datasets or service description directories?
  - Do you have unreasonable crawl delays (e.g., 3600)
- Do you make humans fill out a form before they can even see the service or data set metadata? Before they can access the data?
- Do you have a sitemap.xml file accessible?
We found it but is it usable?

• Do your services actually produce valid responses?
• Do your services have empty metadata elements - especially required elements?
• Do your services rely on default entries from the base server (things like OWS title defaults to OGC:WMS, etc)
• If your service supports sample queries and templates, do their contents match?
Standards issues

• Limited schemas make service automation problematic
  • Nowhere to put enough service metadata to be able to autonomously call it
  • We have self-describing data, why don't we have self-describing services?
  • Inconsistent parameter descriptions, if any parameter description are available
    • Same parameter used differently in different services or metadata records
    • Different parameters used for the same thing in different services or metadata records
  • No place to put information about the range of appropriate values for a parameter (e.g., DayNightFlag, DatasetName)
So what could RDA do to help?

- Organize groups to agree on best practices for various standards
- Help develop easy to use compliance and validation tools
- Help disseminate data advertising tools out to the community
What relationship do you have to current RDA activities?

- Co-chair two Interest Groups
Would collaboration within an RDA Interest Group make sense?

• Yes!