Open data in global environmental research: Findings from the community

Dr. Birgit Schmidt
Niedersächsische Staats- und Universitätsbibliothek
Göttingen
Established to foster global environmental change research; initiated June 2009 by NSF and NERC, building on the work of the IGFA.
“....the need to address global environmental challenges requires a more coordinated approach to the planning, implementation, and management of data, analytics and E-Infrastructures” through international collaboration.

Belmont Forum, New Delhi, February 2013
1. Data Integration for Multidisciplinary Research
2. Improved Interface Between the Computation and Data Infrastructures
3. Harmonization of Global Data Infrastructure for Sharing Environmental Data
4. Data Sharing
5. Open Data
6. Capacity Building (spans Work Packages 1–5)
The main aim was to learn about

- Key open data initiatives in communities dealing with global environmental change
- Reasons where users’ desire to share can be enhanced
- Barriers to open data sharing from a user perspective
Dissemination of the survey

- c 20 disciplinary mailing lists (geosciences, life sciences, social sciences, etc.)
Open Data Survey: Responses

- 16 September – 12 November 2014
- 1330 responses (1253 used for analysis) from 80 countries
- 70.1% (878) academic, 17.9% (224) government, 5.6% (70) non-profit institutions, 2.6% (32) business, 0.3% (4) media, 3.6% (45) other roles
- 82.3% (1025) data users, 57.6% (718) data providers, 25.3% (315) data managers (multiple answers allowed), 5.3% (66) other roles

Countries with more than 20 answers

Which country are you from?

- Germany
- United States
- Italy
- United Kingdom
- France
- Australia
- Spain
- China
- Netherlands
- Canada
- Norway
- Switzerland
- Belgium
- Japan
- Greece
- India
- Sweden
Responses by discipline (N=1253, multiple answers allowed)

- earth and environmental sciences (68.7%, 846 answers)
- climate and atmospheric sciences (31.3%, 386 answers)
- biological sciences (20.6%, 258 answers)
- physical sciences (12.9%, 162 answers)
- engineering (7.1%, 88 answers)
- computer sciences (6.9%, 85 answers)
- social sciences (5.4%, 66 answers)
- agricultural and veterinary sciences (4.3%, 53 answers)
- chemical sciences (4.1%, 50 answers)
- other discipline (3.2%, 40 answers)
- health sciences (1.8%, 22 answers)
- economics (1.7%, 21 answers)
What properties do they expect for open data?

Which attributes do you think are most important to open data?

- Well defined quality information: 1% Not Important, 17% Intermediate Important, 82% Very Important
- Well defined metadata: 1% Not Important, 20% Intermediate Important, 78% Very Important
- Easily accessible: 1% Not Important, 23% Intermediate Important, 76% Very Important
- Ability to re use and re publish data: 4% Not Important, 26% Intermediate Important, 70% Very Important
- Provision of unrestricted data: 3% Not Important, 30% Intermediate Important, 67% Very Important
- Free of cost: 5% Not Important, 31% Intermediate Important, 64% Very Important
- Well defined license terms: 5% Not Important, 38% Intermediate Important, 57% Very Important
- Low cost: 6% Not Important, 38% Intermediate Important, 57% Very Important
- Potential value for research: 8% Not Important, 40% Intermediate Important, 52% Very Important
- Timely data: 8% Not Important, 41% Intermediate Important, 51% Very Important
- Ability to restrict access: 17% Not Important, 44% Intermediate Important, 38% Very Important
Access and licensing conditions

What access or licensing approaches do you consider most useful for open data?

- Public Domain:
  - Not Useful: 2%
  - Somewhat Useful: 24%
  - Very Useful: 74%

- Attribution:
  - Not Useful: 4%
  - Somewhat Useful: 43%
  - Very Useful: 53%

- Non Commercial:
  - Not Useful: 10%
  - Somewhat Useful: 48%
  - Very Useful: 42%

- Open Government License:
  - Not Useful: 12%
  - Somewhat Useful: 48%
  - Very Useful: 41%

- Attribution Share Alike:
  - Not Useful: 8%
  - Somewhat Useful: 54%
  - Very Useful: 38%

- No Derivatives:
  - Not Useful: 26%
  - Somewhat Useful: 57%
  - Very Useful: 16%
Motivators to publish data as open data

Data managers who contributed to the survey were significantly more committed to open data than all other data professionals (p<0.001).
Barriers to publish data as open data

- Desire to publish results before releasing data: 54% (Major Barrier)
- Legal constraints: 47% (Major Barrier)
- Loss of credit or recognition: 41% (Major Barrier)
- Misinterpretation or misuse: 37% (Major Barrier)
- Loss of control over intellectual property: 34% (Major Barrier)
- Organisational constraints: 32% (Major Barrier)
- Commercial use and exploitation: 32% (Major Barrier)
- Difficulty of clarifying rights multiple inputs or authors: 26% (Major Barrier)
- Concerns about impact of data release: 22% (Major Barrier)
- Concerns about legal liability or release: 21% (Major Barrier)

Percentage

Response: Minor Barrier, Barrier, Major Barrier
Some barriers relate with age / career stage

Desire to publish results before releasing data

31–35 year-olds expressed a significantly higher desire to publish results before releasing data.
Infrastructure expectations

What services or functionalities do you expect from infrastructures?

- Attribution highlighted: 3% Most Important, 23% Intermediate Important, 75% Most Important
- Citable via persistent identifiers: 2% Most Important, 25% Intermediate Important, 73% Most Important
- Links to publications: 2% Most Important, 35% Intermediate Important, 63% Most Important
- Usage conditions available: 4% Most Important, 36% Intermediate Important, 61% Most Important
- Repository certified: 12% Most Important, 47% Intermediate Important, 42% Most Important
- Interoperability: 9% Most Important, 51% Intermediate Important, 40% Most Important
- Visual interfaces: 16% Most Important, 48% Intermediate Important, 35% Most Important
- Information about tools: 11% Most Important, 54% Intermediate Important, 35% Most Important
- Code published or accessible: 8% Most Important, 59% Intermediate Important, 33% Most Important
- Publisher endorsed: 23% Most Important, 50% Intermediate Important, 27% Most Important

Percentage

Response: Not Important, Intermediate Important, Most Important
Discovery of data

How do you normally discover data?

<table>
<thead>
<tr>
<th>Discovery route</th>
<th>Frequency</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>References in journal articles</td>
<td>622</td>
<td>79.8</td>
</tr>
<tr>
<td>Web search engines</td>
<td>549</td>
<td>70.5</td>
</tr>
<tr>
<td>Searching in specific data repositories</td>
<td>492</td>
<td>63.2</td>
</tr>
<tr>
<td>Direct requests to data providers</td>
<td>314</td>
<td>40.3</td>
</tr>
<tr>
<td>Newsletters or other publications</td>
<td>201</td>
<td>25.8</td>
</tr>
<tr>
<td>Government or institutional announcements</td>
<td>164</td>
<td>21.1</td>
</tr>
<tr>
<td>Directories or catalogs</td>
<td>136</td>
<td>17.5</td>
</tr>
<tr>
<td>Social media</td>
<td>70</td>
<td>9.0</td>
</tr>
<tr>
<td>Blogs</td>
<td>63</td>
<td>8.1</td>
</tr>
<tr>
<td>Other discovery</td>
<td>36</td>
<td>4.6</td>
</tr>
</tbody>
</table>

Table shows frequencies and valid percentages for each discovery route, multiple answers were allowed (n = 779).

doi:10.1371/journal.pone.0146695.t009
Data repositories: Lead examples

- For data discovery and use
  e.g. Pangaea, Google / general search engines, Dryad, a range of repositories managed by NASA and the National Oceanic and Atmospheric Administration (NOAA), the Global Biodiversity Information Facility (GBIF, a network of repositories), Genbank, government data, Figshare, journals, etc. – but also: many smaller repositories (e.g. projects), personal websites

>> About 100 repositories are currently added to re3data (out of c 300)
What do they value in data repositories?

- "Good for heterogenous datasets, also long-tail"
- "easy to obtain quality data"
- "great to have automatic DOIs"
- "well-defined data quality and metadata"
- "large data files, managing copies close to computing"
- "visualization interface"
- "independent, trustful, recommended by several publishers"
- "The site and tools are terrible, but the data is important"
- "sound international reputation"
- "highest quality with excellent links to other databases"
What is missing?

- “For me, **long-tail research datasets** are most important, and I would therefore like to see more repositories supporting these (i.e. institutional ones)”
- “Every discipline in the natural sciences needs to be able to **access its heritage data** (those in analogue forms) which cannot presently be accessed electronically (observations are still, and only, in their virgin forms on paper, forms, books, photographs, unreadable early mag tapes, etc.).”
- “Some types of data **still lack community standards** that would allow creation of open data resources.”
- “the open publication of **source codes or scripts** of simulation programs or analytical / numerical solutions is extremely important”
Conclusions & recommendations

Based on the findings of the survey, we have made the following recommendations to the Belmont Forum:

- that funders should make open data archiving mandatory, to take into account the main motivators revealed by the survey,
- to strengthen support and training activities,
- to further facilitate interoperability between data infrastructures, and
- to support the long-term sustainability of archives and data infrastructures.
Thank you for your attention!

www.bfe-inf.org

Acknowledgements:
• To my co-authors Birgit Gemeinholzer, University of Giessen, and Andrew Treloar, ANDS
• Slides 2-3 by Belmont Forum colleagues
• All members of the Open Data WG

Birgit Schmidt, University of Göttingen, bschmidt@sub.uni-goettingen.de