Seeking data: A use (user) case

Kathleen Gregory, DANS
Paul Groth, University of Amsterdam

GO Inter Kick-Off
01 October 2019, Köln
Data search as a use case

- Data needs and discovery strategies of users
- Possible approaches to data search
What data are needed and why?

Data needs are diverse and sometimes span categories.
Data as an information need

Diverse disciplines, diverse sources, diverse scales...

Integrating diverse data is challenging.

<table>
<thead>
<tr>
<th>Users in this community...</th>
<th>Need this type of data</th>
<th>Needs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Astronomy</td>
<td>Data from sky surveys, telescopes, archives, repositories, data catalogues, and virtual observatory systems</td>
<td>Foreground – new questions of old data</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Background – baselines, instrument calibration, physical properties, model inputs; data integration</td>
</tr>
<tr>
<td>Earth &amp; Environmental Sciences</td>
<td>Plant, animal, water, weather, solar observations; soil analyses; rock thin-section and satellite images; maps; geographic, demographic and census data; continuously collected and transmitted data; data at temporal/spatial scales; raw and summarized data</td>
<td>Foreground – new questions of old data, meta-analyses</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Background – calibration, context, baselines, reference, model inputs, verification, comparison, environmental planning, policy- and decision making, education; instrument monitoring; data integration</td>
</tr>
<tr>
<td>Biomedicine</td>
<td>Images; complete fMRI studies; pathology results; patient observations and demographics; population-level disease data; behavioural data</td>
<td>Foreground - disease/disorder research, new visualizations</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Background - evaluations, 3-D anatomical pictures; preparing research outputs; education; patient care</td>
</tr>
<tr>
<td>Field Archaeology</td>
<td>Field notebooks, photographs, artefacts; stratigraphic baselines; data at temporal/spatial scales</td>
<td>Background – comparison; data integration/triangulation; training, dissertations</td>
</tr>
<tr>
<td>Social Sciences</td>
<td>Survey data (often only one question is of interest); long-running datasets/surveys; interviews; archival documents, images, videos</td>
<td>Foreground – re-interpret datasets; new questions, comparative research; Background – comparison, preparations; training, dissertations</td>
</tr>
</tbody>
</table>
Data as an information need

...diverse data types

How do people discover data?

• Discovery is challenging
Why is it challenging to find data?

- Data are not accessible: 27%
- Data are in many different places: 20%
- Online tools are inadequate: 15%
- Don't know where or how to look for data: 14%
- Data are not digital: 14%
- Don't have necessary personal network: 8%
- Other: 3%

Multiple responses possible. Percents are percent of responses to question (n = 4080).
How do people discover data?

- Discovery is challenging
- Via accustomed channels – literature, search engines and personal networks
How frequently do you use the following to find data?

<table>
<thead>
<tr>
<th>Method</th>
<th>Researchers</th>
<th>Support prof</th>
</tr>
</thead>
<tbody>
<tr>
<td>Literature</td>
<td>75%</td>
<td>19%</td>
</tr>
<tr>
<td>Search engine</td>
<td>56%</td>
<td>15%</td>
</tr>
<tr>
<td>Domain repos</td>
<td>45%</td>
<td>11%</td>
</tr>
<tr>
<td>Gov source</td>
<td>36%</td>
<td>16%</td>
</tr>
<tr>
<td>Network</td>
<td>31%</td>
<td>16%</td>
</tr>
<tr>
<td>Data search engine</td>
<td>38%</td>
<td>25%</td>
</tr>
<tr>
<td>Multidisc repos</td>
<td>40%</td>
<td>31%</td>
</tr>
<tr>
<td>Prof association</td>
<td>19%</td>
<td>53%</td>
</tr>
<tr>
<td>Support prof</td>
<td>12%</td>
<td>35%</td>
</tr>
<tr>
<td>Code repos</td>
<td>9%</td>
<td>65%</td>
</tr>
<tr>
<td>Commercial</td>
<td>30%</td>
<td>66%</td>
</tr>
</tbody>
</table>
Do you use the following to discover, access or make sense of data?

<table>
<thead>
<tr>
<th>Method</th>
<th>Discovering</th>
<th>Accessing</th>
<th>Sensemaking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attending conferences</td>
<td>31%</td>
<td>16%</td>
<td>19%</td>
</tr>
<tr>
<td>Conversations with personal networks</td>
<td>27%</td>
<td>24%</td>
<td>34%</td>
</tr>
<tr>
<td>Disciplinary mailing list or forum</td>
<td>18%</td>
<td>14%</td>
<td>12%</td>
</tr>
<tr>
<td>Contacting data creator</td>
<td>9%</td>
<td>19%</td>
<td>27%</td>
</tr>
<tr>
<td>New collaborations with data creators</td>
<td>15%</td>
<td>17%</td>
<td>19%</td>
</tr>
</tbody>
</table>

How respondents make use of social connections in discovering data (n=3311), accessing data (n=3589) and making sense of data (n=3031). Percents are percent responses for each option; multiple responses possible.
Searching for data in another discipline

**Different sources for different data**

For astronomical data, we have used mainly online virtual observatories. The general data (for teaching purposes), we've used digital/institutional repositories *(Respondent ID 305)*
Searching for data in another discipline

**Success**

*Wasting hours on the internet (although you didn’t ask about success)* (Respondent ID 1484)

*Just today I have been playing with Google's new database search tools - their dataset search, and public data access. Each led me to interesting data on topics where my grasp is ....errr... limited* (Respondent ID 328)
Possible approaches to data search
Spreadsheet Events

Data search

Data from: Political Power in Boston, Massachusetts and Charleston, South Carolina, 1828-1843

DOI link
https://doi.org/10.3886/ICPSR08653.v1

Dataset provided by
Inter-university Consortium for Political and Social Research

Authors
Pease, Jane H.; Pease, William H.

License
https://www.icpsr.umich.edu/icpsrweb/icpsr/studies/8653/terms

Available download formats from providers
ASCII

Time period covered
1828 - 1843

Area covered
Data search – is it just a regular search engine?

Probably not – see


Using Structured Metadata from Data Providers

When Google’s search engine processes a Web page with schema.org/Dataset mark-up, it understands that there is dataset metadata there and processes that structured metadata to create
Constructive Data Search

Title: Population growth of European cities

<table>
<thead>
<tr>
<th>City</th>
<th>Country</th>
<th>Population growth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Helsinki</td>
<td>Finland</td>
<td>0.12%</td>
</tr>
<tr>
<td>Oslo</td>
<td>Norway</td>
<td>3.00%</td>
</tr>
<tr>
<td>Stavanger</td>
<td>Norway</td>
<td>0.50%</td>
</tr>
</tbody>
</table>

(c) Column population assistance.

(b) Row population assistance.

### The Search Process from Different Technical Perspectives

![Diagram](image)

**Fig. 2** An abstract view of the search process, comprising of querying, query processing, data handling and results presentation, alongside approaches to each step by different related communities.

Conclusions

- Cataloging information needs
- Thinking about data search in different contexts
- Data search as a social process
- Lots of work that can be used but needs to be tied together and tailored
- The need for benchmarks
Questions & Discussion

Kathleen Gregory
kathleen.gregory@dans.knaw.nl

Paul Groth
p.groth@uva.nl
References


