Research data in social sciences and humanities

Joachim Schöpfel, University of Lille
Contents

- Terminology and categories
- Data and publications
- Critical issues

- NOT Research data management
Terminology and categories
Data are like cows. If you look them in the face hard enough they generally run away.
(adapted from Dorothy L Sayers)

« Recorded factual material commonly accepted in the scientific community as necessary to validate research findings »

US OMB Circular 110

« Re-usable research results, collected, observed or created for purposes of analysis to produce original results »

University of Edinburgh
Definition mainly by functions (validation, reuse, innovation) and types (and not by nature)

- “Research data refers to information, in particular facts or numbers,
- collected to be examined and considered and as a basis for reasoning, discussion, or calculation.
- In a research context, examples of data include statistics, results of experiments, measurements, observations resulting from fieldwork, survey results, interview recordings and images”

H2020

What is information? Numbers = data?
But what are facts?
Research methods
- Observational data
- Experimental data
- Simulation data
- Derived or compiled data

Input and output
- Primary data (collected)
- Secondary data (produced)
CATEGORIES

- Archived data
- Audiovisual data
- Configuration data
- Databases
- Images
- Network-based data
- Plain text
- Raw data
- Scientific and statistical data formats
- Software applications
- Source code
- Standard office documents
- Structured graphics
- Structured text
- Other

re3data, HUB
CATEGORIES IN SS&H

- Texts
- Tables
- Images, drawings
- Graphs, figures
- Statistics
- Maps
- Photographs
- Databases
- Timelines
- Others

Secondary data in survey

Data in PhD dissertations

Lille studies
Data and publications
DATA PUBLICATION

Enhanced data explanation in supplementary files

Processed Data and data representations

Data published in a research article

Data referenced in research articles and available in repositories

Data in drawers and on researcher harddrives

Data publications, describing available datasets

Raw data and data sets

https://www.elsevier.com/connect/can-data-be-peer-reviewed
Data AND PUBLICATIONS

- Data vehicle
  - Supplementary materials of publication
- Document as data
  - Exploited as primary data source for TDM
- Gateway to data
  - Publication contains links to data, integrated or not in the text
EXAMPLE DISSERTATIONS

ETD as data vehicle

ETD as data

ETD as gateway to data
# DATA IN DISSERTATIONS CT’D

<table>
<thead>
<tr>
<th>Domain</th>
<th>Databases</th>
<th>Graphs - figures</th>
<th>Images - drawings</th>
<th>Maps</th>
<th>Others</th>
<th>Photographs</th>
<th>Statistics</th>
<th>Tables</th>
<th>Texts</th>
<th>Timelines</th>
<th>Tous</th>
</tr>
</thead>
<tbody>
<tr>
<td>Archeology</td>
<td>4</td>
<td>2</td>
<td>22</td>
<td>18</td>
<td></td>
<td>11</td>
<td>1</td>
<td>16</td>
<td>15</td>
<td>1</td>
<td>30</td>
</tr>
<tr>
<td>Economical sciences</td>
<td>16</td>
<td>1</td>
<td>5</td>
<td></td>
<td></td>
<td>2</td>
<td>31</td>
<td>36</td>
<td></td>
<td></td>
<td>43</td>
</tr>
<tr>
<td>Educational sciences</td>
<td>8</td>
<td>14</td>
<td>1</td>
<td></td>
<td></td>
<td>5</td>
<td>25</td>
<td>29</td>
<td>1</td>
<td></td>
<td>38</td>
</tr>
<tr>
<td>Foreign languages &amp; literatures</td>
<td>1</td>
<td>1</td>
<td>20</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>21</td>
<td>36</td>
<td>1</td>
<td></td>
<td>46</td>
</tr>
<tr>
<td>French language &amp; literature</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>5</td>
<td>1</td>
<td></td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>Geography</td>
<td>13</td>
<td>7</td>
<td>13</td>
<td></td>
<td></td>
<td>5</td>
<td>3</td>
<td>27</td>
<td>23</td>
<td></td>
<td>33</td>
</tr>
<tr>
<td>History</td>
<td>16</td>
<td>22</td>
<td>39</td>
<td>27</td>
<td></td>
<td>26</td>
<td>14</td>
<td>44</td>
<td>65</td>
<td>12</td>
<td>88</td>
</tr>
<tr>
<td>History of arts</td>
<td>6</td>
<td>17</td>
<td>8</td>
<td>1</td>
<td>8</td>
<td>4</td>
<td>20</td>
<td>1</td>
<td></td>
<td></td>
<td>28</td>
</tr>
<tr>
<td>Information science</td>
<td>2</td>
<td>7</td>
<td>7</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>5</td>
<td>12</td>
<td>20</td>
<td>1</td>
<td>28</td>
</tr>
<tr>
<td>Language science</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>7</td>
<td></td>
<td></td>
<td>7</td>
</tr>
<tr>
<td>Law</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td></td>
<td></td>
<td>4</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td>7</td>
</tr>
<tr>
<td>Management</td>
<td>2</td>
<td>12</td>
<td>10</td>
<td>1</td>
<td></td>
<td>1</td>
<td>7</td>
<td>26</td>
<td>22</td>
<td>2</td>
<td>30</td>
</tr>
<tr>
<td>Others</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>4</td>
<td>1</td>
<td></td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>Philosophy</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>11</td>
<td></td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>Political science</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td></td>
<td></td>
<td>1</td>
<td>6</td>
<td>2</td>
<td></td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>Psychology</td>
<td>2</td>
<td>15</td>
<td>20</td>
<td>1</td>
<td></td>
<td>4</td>
<td>55</td>
<td>65</td>
<td>48</td>
<td></td>
<td>91</td>
</tr>
<tr>
<td>Sociology</td>
<td>2</td>
<td>7</td>
<td>8</td>
<td>4</td>
<td>6</td>
<td>21</td>
<td>28</td>
<td></td>
<td>2</td>
<td></td>
<td>28</td>
</tr>
<tr>
<td><strong>Tous</strong></td>
<td><strong>38</strong></td>
<td><strong>111</strong></td>
<td><strong>175</strong></td>
<td><strong>83</strong></td>
<td><strong>7</strong></td>
<td><strong>65</strong></td>
<td><strong>101</strong></td>
<td><strong>306</strong></td>
<td><strong>376</strong></td>
<td><strong>23</strong></td>
<td><strong>526</strong></td>
</tr>
</tbody>
</table>
DATA IN DISSERTATIONS - ISSUES

Incomplete, inadequate or missing description

Datasets and/or individual data are not or incompletely documented

Research data are presented without any structuration or organisation, often together with other, not reusable material in a kind of information mash-up not suitable for further research

Inadequate organisation

Inadequate format

Data and text are glued together in a PDF file instead of being separated and published in adequate file formats
Data available

(not found)

Data not available

Confidentiality/privacy issues

Or only on demand

No data

“Data sharing is not applicable to this article, as no datasets were generated or analysed during the current study.”
Les données de la recherche en SHS. Une enquête à l’Université de Lille 3 : Rapport final

Hélène Prost 1,2, Joachim Schöpfel 2

1 INIST - Institut de l’information scientifique et technique
2 GERICO - Groupe d’Etudes et de Recherche Interdisciplinaire en Information et Communication

Résumé : Les données de la recherche deviennent l’un des nouveaux défis de la gestion scientifique. Produites dans le cadre des projets de recherche, ces données posent des questions inédites aux laboratoires, bibliothèques et services informatiques des universités : comment conserver ces données, comment les signaler et mettre à disposition d’autres chercheurs, comment faire le lien avec les publications, comment les intégrer dans une politique du libre accès à l’information scientifique ? Avant de se lancer dans un projet de données de la recherche, un établissement doit faire un état des lieux sur le terrain pour mieux connaître les producteurs de ces données, leurs pratiques et besoins dans la gestion des données et leurs outils, mais aussi la nature concrète de ces données. L’Université de Lille 3 a réalisé une étude sur les pratiques, besoins et attentes en matière de gestion des données de la recherche auprès de son personnel scientifique. L’étude est pilotée par le laboratoire GERICO et le SCD de Lille 3. Elle fait partie d’une démarche concertée en faveur de la gestion et du partage des données de la recherche mise en œuvre à partir de 2013, avec plusieurs analyses, séminaires et publications. L’enquête a été préparée avec l’Université Humboldt de Berlin. Le questionnaire contient 22 questions et a été mis en ligne en avril et mai 2015. Il a reçu 270 réponses (taux de réponse 15%). Toutes les disciplines sont représentées, ainsi que toutes les catégories des personnel scientifiques. Les personnes interrogées décrivent un large éventail de données sources. Les corpus (documents textuels) sont de loin la source la plus importante (64%), suivi par les enquêtes et entretiens (47%), observations (41%), expériences (36%) et archives (34%). Quant à la typologie des données
Dissertations and Data: keynote address

Joachim Schöpfel 1, Južnič Primož 2, Hélène Prost 1, Cécile Malleret 3, Ana Češarek 2, Teja Koler-Povh 2

1 GERICICO - Group of Studies and Research Interdisciplinary in Information and Communication
2 University of Ljubljana (SLOVENIA)
3 SCD - Service Commun de la Documentation

Abstract: The keynote provides an overview on the field of research data produced by PhD students, in the context of open science, open access to research results, open-source software development, and the handling of electronic theses and dissertations. The keynote includes recent empirical results and recommendations for good practice and further research. In particular, the paper is based on an assessment of 854 print and electronic dissertations in sciences, social sciences and humanities from the Université de Lille (France) and Ljubljana (Slovenia), submitted between 1987 and 2015, and on a survey on data management with 270 scientists in social sciences and humanities of the University of Lille 3. The keynote starts with an introduction into data-driven science, data lifecycle and data publishing. It then moves on to research data management by PhD students, their practice, their needs and their willingness to disseminate and share their data. After this qualitative analysis of information behaviour, we present the results of a quantitative assessment of research data produced and submitted with dissertations. Special attention is paid to the size of the research data in appendices, to their presentation and link to the text, to their sources and typology, and to their potential for further research. The discussion puts the focus on legal aspects (database protection, intellectual property, privacy, third-party rights) and other barriers to data sharing, reuse and dissemination through open access. Another part adds insight into the potential handling of these data, in the framework of the French and Slovenian dissertation infrastructures. What could be done to valorise these data in a centralized system for electronic theses and dissertations (ETDs)? The topics are formats, metadata (including attribution

http://hal.univ-lille3.fr/hal-01285304
<table>
<thead>
<tr>
<th>Titre du congrès</th>
<th>GL17 International Conference on Grey Literature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audience</td>
<td>Internationale</td>
</tr>
<tr>
<td>Invité</td>
<td>Non</td>
</tr>
<tr>
<td>Comité de lecture</td>
<td>Oui</td>
</tr>
<tr>
<td>Acte</td>
<td>Non</td>
</tr>
<tr>
<td>Vulgarisation</td>
<td>Non</td>
</tr>
<tr>
<td>Langue du document</td>
<td>anglais</td>
</tr>
<tr>
<td>Pays</td>
<td>Pays-Bas</td>
</tr>
<tr>
<td>Domaine</td>
<td>Sciences de l'Homme et Société/Sciences de l'information et de la communication</td>
</tr>
<tr>
<td>Voir aussi</td>
<td><a href="http://dx.doi.org/10.17026/dans-xg6-xnj4">http://dx.doi.org/10.17026/dans-xg6-xnj4</a></td>
</tr>
<tr>
<td>Mots-clés</td>
<td>research data, data repository, institutional repository, open access, open data, electronic theses and dissertations, research data management, Open science</td>
</tr>
</tbody>
</table>
COMMUNICATION (3)

Get exposure and credit for your data: write a data paper for the new peer reviewed, online-only open access Research Data Journal (published by Brill)

For more info: brill.com/yrdj

EASY offers sustainable archiving of research data and access to thousands of datasets.

Search...

Search help

Advanced search  Browse

DISSERTATIONS AND DATA

Overview  Description  Data files (3)

Cites:

Schöpfel, Dr. J. (University of Lille 3)(2015): Dissertations and Data. DANS. http://dx.doi.org/10.17026/dans-xg6-xnj4

2015-12-02 | Schöpfel, Dr. J. (University of Lille 3) | 10.17026/dans-xg6-xnj4

We present the results of a quantitative assessment of research data produced and submitted with dissertations. Special attention is paid to the size of the research data in appendices, to their presentation and link to the text, to their sources and typology, and to their potential for further research. The discussion puts the focus on legal aspects (database protection, intellectual property, privacy, third-party rights) and other barriers to data sharing, reuse and dissemination through open access. Another part adds insight into the potential handling of these data, in the framework of the French and Slovenian dissertation infrastructures. What could be done to valorize these data in a centralized system for electronic theses and dissertations (ETDs)? The topics are formats, metadata (including attribution of unique identifiers), submission/deposit, long-term preservation and dissemination. This part will also draw on experiences from other campuses and make use of results from surveys on data management in the field of electronic theses.
The Geometry of Finite Equilibrium Datasets

Yves Balasko (yves@balasko.com) and Mich Ty_three

No 09-07, Discussion Papers from University of Copenhagen, Department of Economics

Abstract: We investigate the geometry of finite datasets defined by equilibrium prices, income distributions, and total resources. We show that the equilibrium condition imposes no restrictions if total resources are collinear, a property that is robust to small perturbations. We also show that the set of equilibrium datasets is pathconnected when the equilibrium condition does impose restrictions on datasets, as for example when total resources are widely non-collinear.

Keywords: equilibrium manifold; rationalizability; pathconnectedness (search for similar items in EconPapers)
JEL-codes: D31 D31 (search for similar items in EconPapers)
Date: 2009-03
References: View references in EconPapers View complete reference list from CitEc
Citations: View citations in EconPapers (1) Track citations by RSS feed

Downloads: (external link)

Related works:
Journal Article: The geometry of finite equilibrium datasets (2009) EconLabs
This item may be available elsewhere in EconPapers: Search for items with the same title.

Export reference: BibTeX RIS (EndNote, ProCite, RefMan) HTML/Text

Persistent link: http://EconPapers.repec.org/RePEc:kud:kuiedp:0907

Access Statistics for this paper

More papers in Discussion Papers from University of Copenhagen, Department of Economics Oster Farimagsgade 5, Building 26, DK-1353 Copenhagen K., Denmark. Contact Information at EconPapers Series data maintained by Thomas Hoffmann (thho@eb.dk).
OTHER PUBLICATIONS

Books – OpenEdition

Data journals (later)

http://books.openedition.org/ksp/244
PUBLICATIONS AS DATA

- TDM of research publications
  - The case of ETDs
  - The future: writing dissertations as usual?
- Legal issues
- Technical issues
- Impact on publications?
  - Structure
  - Content
  - Format
Critical issues
GENERAL ISSUES

- Separation text/data
- Preferred formats cf. DANS
- Metadata (generic, specific)
- Persistent identifiers (DOI, ORCID)
- Altmetrics (DOI) and usage (low?)
- Backup, storage, preservation, sharing, reusage
  - Quality of data repositories (Data Seal of Approval...)

DARIAH-EU
Impact of disciplines
- Typology
- Or rather, profiles
- Related to methodology and instruments (ex survey data...)

Evaluation: a strong need for a standard and generic approach
- Metadata
- Identifiers (DOI, handle...)

Preservation and sharing: disciplinary and generic repositories
- Cf. re3data.org
- Cf. HAL, Figshare
Research data are evaluated as research output
Mix between primary and secondary data
Contrary to publication, data quality or volume are NOT evaluated...

...but the data management
- Existence of a DMP
- Description and identification
- Conservation
- Sharing

Especially by funders
LEGAL ISSUES

❖ Intellectual property?
  ❖ Career strategy
  ❖ Publications

❖ Data base protection (sui generis)?
❖ Third party rights?

❖ Confidentiality?
  ❖ Private company information
  ❖ Corporate secrets

❖ Privacy issues?
POLITICAL ISSUES (1)
POLITICAL ISSUES (2)

Netherlands’ EU presidency on 4 and 5 April 2016. It is a living document reflecting the present state of open science evolution. Based on the input of all participating experts and stakeholders as well as outcomes of preceding international meetings and reports, a multi-actor approach was formulated to reach two important pan-European goals for 2020:

1. **Full open access for all scientific publications**
   This requires leadership and can be accelerated through new publishing models and compliance with standards set.

2. **A fundamentally new approach towards optimal reuse of research data**
   Data sharing and stewardship is the default approach for all publicly funded research. This requires definitions, standards and infrastructures.

3. **New assessment, reward and evaluation systems**
   New systems that really deal with the core of knowledge creation and account for the impact of scientific research on science and society at large, including the economy, and incentivise citizen science.

4. **Alignment of policies and exchange of best practices**
   Practices, activities and policies should be aligned and best practices and information should be shared. It will increase clarity and comparability for all parties concerned and help to achieve joint and concerted actions. This should be accompanied by regular monitoring-based stocktaking.

**Twelve action items with concrete actions to be taken**

Twelve action items have been included in this Call for Action. They all contribute to
REFERENCES


