Research Data, Libraries, and Croatia

University of Zadar, Croatia, 27 September 2014

Christine L. Borgman
Professor and Presidential Chair in Information Studies
University of California, Los Angeles
and
Visiting Professor KNAW, Digital Archiving and Network Services (DANS)
Netherlands
http://christineborgman.info
Technological advances in mediated communication


Image: Jillian Wallis and Roy Pea
Yugoslavia

Plan for a Yugoslav Scientific-Technical Information Network

December 1972

by B.V. Tell

ARCHIVES
PGI

Serial No. 2873/RMO.RD/DBA
Paris, March 1973

Unesco
Big Data, Little Data, No Data: Scholarship in the Networked World*

• Part I: Data and Scholarship
  – Ch 1: Provocations
  – Ch 2: What Are Data?
  – Ch 3: Data Scholarship
  – Ch 4: Data Diversity

• Part II: Case Studies in Data Scholarship
  – Ch 5: Data Scholarship in the Sciences
  – Ch 6: Data Scholarship in the Social Sciences
  – Ch 7: Data Scholarship in the Humanities

• Part III: Data Policy and Practice
  – Ch 8: Releasing, Sharing, and Reusing Data
  – Ch 9: Credit, Attribution, and Discovery
  – Ch 10: What to Keep and Why

*C. L. Borgman (2015, January) MIT Press
Neelie Kroes, VP European Commission:

To collect, curate, preserve and make available ever-increasing amounts of scientific data, new types of infrastructures will be needed. The potential benefits are enormous but the same is true for the costs. We therefore need to lay the right foundations and the sooner we start the better.

Open access policies

- Australian Research Council
  - Code for the Responsible Conduct of Research
  - Data management plans
- National Science Foundation
  - Data sharing requirements
  - Data management plans
- U.S. Federal policy
  - Open access to publications
  - Open access to data
- European Union
  - European Open Data Challenge
  - OpenAIRE
- Research Councils of the UK
  - Open access publishing
  - Provisions for access to data
Better Data: Better Research

Why manage data?

- Preserve the integrity of the research
- Allow data to be made available for others to use
- Assist researchers to reduce the risk of data loss
- Secure continued access to the value in data

Why connect data?

- Interlink data to people to projects to publications
- Improve the discoverability of data
- Tie data to research achievements
- Provide richer context for data value

Why make data discoverable?

- Enable the demonstration of research excellence
- Allow researchers to build upon existing data, instead of recreating it
- Foster innovation
- Provide the ability to solve big problems across discipline boundaries

Why reuse data?

- Verification of research claims
- New discoveries from existing data
- Integration of sets of data for new analysis
- Re-analysis of expensive, rare or unrepeatable investigations
- Reduction of duplicated effort
EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF SCIENCE AND TECHNOLOGY POLICY
WASHINGTON, D.C. 20502

February 22, 2013

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: John P. Holdren
Director

SUBJECT: Increasing Access to the Results of Federally Funded Scientific Research

1. Policy Principles

The Administration is committed to ensuring that, to the greatest extent and with the fewest constraints possible and consistent with law and the objectives set out below, the direct results of federally funded scientific research are made available to and useful for the public, industry, and the scientific community. Such results include peer-reviewed publications and digital data.

Scientific research supported by the Federal Government catalyzes innovative breakthroughs that drive our economy. The results of that research become the grist for new insights and are assets for progress in areas such as health, energy, the environment, agriculture, and national security.
Chinese science gets mass transformation

Teamwork at centre of Chinese Academy of Sciences reform.

David Cyranoski

23 September 2014

BEIJING
Open Data
Open Data

• A piece of data or content is open if anyone is free to use, reuse, and redistribute it — subject only, at most, to the requirement to attribute and/or share-alike

<table>
<thead>
<tr>
<th>Research Area</th>
<th>Status</th>
<th>Round 1</th>
<th>Round 2</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Application Scientists</strong></td>
<td>Faculty</td>
<td>7</td>
<td>6</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>Staff</td>
<td>5</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Student</td>
<td>3</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td><strong>Technology Researchers</strong></td>
<td>Faculty</td>
<td>4</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Staff</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Student</td>
<td>2</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td></td>
<td>22</td>
<td>21</td>
<td>43</td>
</tr>
</tbody>
</table>

doi:10.1371/journal.pone.0067332.t001
Open Data

• Data that meets the criteria of intelligent openness. Data must be accessible, useable, assessable and intelligible.


*Indian Girl Pitching up Haystack*
Uploaded by The U.S. National Archives on December 9, 2009
Metadata

• Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource.*
  – descriptive
  – structural
  – administrative

*National Information Standards Organization 2004

photo by @kissane
SCHUMACHER, Ernst Friedrich

Small is beautiful: Economics as if people mattered.

http://www.nzdl.org/gsdl/collect/fnl2.2/archives/HASH01a1/51ac3343.dir/p027.gif
that really fouled up that plan.

20 December 1941.
Pretty much activity going on for a change today. The past few days have been fairly easy going for a war. Took on an extra lot of provisions, shells, food and fuel. We are supposed to shift berths today, but I smell something in the air. No liberty to be granted until we tie up at new berth (Manila, Long Beach or ?????). At 5 o'clock this evening we got under way - with sealed orders - for sea. First time since the war. As the PENNSY-MANU steamed passed the NEVADA they gave three cheers which were heartily answered. No one seems to know where we're bound for. Some say Japan, others the West Coast and etc. A crescent shaped moon is shining tonight and a few clouds speckle the sky. First moon I saw since Sunday the 7th. It almost looks friendly. Just heard another rumor we are conveying empty vessels to the West Coast could it be we'll be in California for New Years - I wonder. Everything is relatively quiet so far this evening. Time 7:30 P.M. First war pictures of the attack on Pearl Harbor came out in the papers today.

21 December 1941.
Additional data - shifted into blues this evening. Set the clock ahead 1 hour, so at least we're traveling towards the West Coast. Course 069. Ships rendezvous tomorrow at sunrise in lat. 27-10; long. 56-10. What beats me is the fact we're stationed at Pearl Harbor for nearly two years toward off Japs attacks and after it is all over with the Japs head for home and so do we - that thought moves me to repeat a verse I heard:

22 December 1941.
Two weeks ago today the war started. After gettin underway for sea
Image from a photo album (AL-27) which belonged to Ray Fife, a pioneer aviation mechanic who built a complete plane at age 16. In 1963 he built a reproduction Curtiss Pusher Biplane. This album documents the building of the pusher and also contains photos used to research the project.
Provenance

• Libraries: Origin or source
• Museums: Chain of custody
• Internet: Provenance is information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness.*

*World Wide Web Consortium (W3C) Provenance working group
Open Data

- Openness, flexibility, transparency, legal conformity, protection of intellectual property, formal responsibility, professionalism, interoperability, quality, security, efficiency, accountability, and sustainability.

Discovery and Interpretation

• Identify the form and content
• Identify related objects
• Interpret
• Evaluate
• Open
• Read
• Compute upon
• Reuse
• Combine
• Describe
• Annotate…

Image from Soumitri Varadarajan blog. Iceberg image © Ralph A. Clevenger. Flickr photo
What are data?

Marie Curie’s notebook

NASA Astronomy Picture of the Day

http://www.census.gov/population/cen2000/map02.gif

http://ncl.ucar.edu

http://onlineqda.hud.ac.uk/Intro_QDA/Examples_of_Qualitative_Data.php
A role for self-gravity at multiple length scales in the process of star formation

Alyssa A. Goodman, Erik W. Rosolowsky, Michelle A. Borkin, Jonathan B. Foster, Michael Halle, Jens Kaufmann & Jaime E. Pineda

Self-gravity plays a decisive role in the final stages of star formation, where dense cores (size ~0.1 parsec) inside molecular clouds collapse to form star-plus-disk systems. But self-gravity's role at earlier times (and on larger length scales, such as ~1 parsec) is unclear; some molecular cloud simulations that do not include self-gravity suggest that turbulent fragmentation alone is sufficient to create a mass distribution of dense cores that resembles, and sets, the stellar initial mass function. Here we report a 'dendrogram' (hierarchical tree-diagram) analysis that reveals that self-gravity plays a significant role over the full range of possible scales traced by 12CO observations in the L1448 molecular cloud, but not everywhere in the observed region. In particular, more than 90% of the compact 'pre-stellar cores' traced by peaks of dust emission are projected on the sky within one of the dendrogram's self-gravitating 'leaves'. As these peaks mark the locations of already-forming stars, or of those probably about to form, a self-gravitating cocoon seems a critical condition for their existence. Turbulent fragmentation simulations without self-gravity—even of unmagnetized isothermal material—can yield mass and velocity power spectra very similar to what is observed in clouds like L1448. But a dendrogram of such a simulation shows that nearly all the gas in it (much more than in the observations) appears to be self-gravitating. A potentially significant role for gravity in 'non-self-gravitating' simulations suggests inconsistency in simulation assumptions and output, and that it is necessary to include self-gravity in an realistic simulation of the star-formation process on sub-parsec scales.

Spectral-line mapping shows whole molecular clouds (typically tens to hundreds of parsecs across, and surrounded by atomic gas) to be marginally self-gravitating. When attempts are made to further break down clouds into pieces using 'segmentation' routines, some self-gravitating structures are always found on whatever scale is sampled. But no observational study to date has successfully used one spectral-line data cube to study how the role of self-gravity varies as a function of scale and conditions, within an individual region.

Most past structure identification in molecular clouds has been explicitly non-hierarchical, which makes difficult the quantification of physical conditions on multiple scales using a single data set. Consider, for example, the often-used algorithm CLUMPFIND. In three-dimensional (3D) spectral-line data cubes, CLUMPFIND operates as a watershed segmentation algorithm, identifying local maxima in the position-position-velocity (p-p-v) cube and assigning nearby emission to each local maximum. Figure 1 gives a two-dimensional (2D) view of L1448, our sample star-forming region, and Fig. 2 includes a CLUMPFIND decomposition of it based on 12CO observations. As with any algorithm that does not offer hierarchically nested or overlapping features as an option, significant emission found between prominent clumps is typically either appended to the nearest clump or turned into a small, usually 'pathological', feature needed to encompass all the emission being modelled. When applied to molecular-line

Figure 1 | Near-infrared image of the L1448 star-forming region with contours of molecular emission overlaid. The channels of the colour image correspond to the near-infrared bands J (blue), H (green) and K (red), and the contours of integrated intensity are from 12CO(1-0) emission. Integrated intensity is isospectrally but not quite linearly (see Supplementary Information), related to column density, and it gives a view of all of the molecular gas along lines of sight, regardless of distance or velocity. The region within the yellow box immediately surrounding the protostars has been cropped more deeply in the near-infrared (using Calar Alto) than the remainder of the box (2MASS data only), revealing protostars as well as the scattered starlight known as 'Circumstellar' (see Supplementary Information and Fig. 2a). Asterisks show the locations of the four most prominent embedded young stars or compact stellar systems in the region (see Supplementary Table 1), and yellow circles show the millimetre-dust emission peaks identified as star-forming or 'pre-stellar' cores.

Figure 3 | Schematic illustration of the dendrogram process. Shown is the
Sensor networked science
Sensor network data

Nitrate distribution

[Diagram showing nitrate distribution with a color scale from 2 to 7 mg/l-N]
6. Generally speaking, do you usually think of yourself as a Republican, Democrat, Independent, or what?

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Strong Democrat</td>
<td>0</td>
<td>2197</td>
<td>143</td>
<td>1271</td>
<td>151</td>
<td>864</td>
<td>227</td>
<td>423</td>
<td>400</td>
<td>370</td>
<td>6,046</td>
</tr>
<tr>
<td>Not very strong Democrat</td>
<td>1</td>
<td>3482</td>
<td>109</td>
<td>1655</td>
<td>89</td>
<td>1,282</td>
<td>321</td>
<td>644</td>
<td>577</td>
<td>597</td>
<td>8,756</td>
</tr>
<tr>
<td>Independent, close to Democrat</td>
<td>2</td>
<td>1788</td>
<td>44</td>
<td>904</td>
<td>51</td>
<td>578</td>
<td>190</td>
<td>341</td>
<td>356</td>
<td>340</td>
<td>4,581</td>
</tr>
<tr>
<td>Independent (Neither, No response)</td>
<td>3</td>
<td>1736</td>
<td>30</td>
<td>855</td>
<td>32</td>
<td>721</td>
<td>205</td>
<td>369</td>
<td>457</td>
<td>477</td>
<td>4,882</td>
</tr>
<tr>
<td>Independent, close to Republican</td>
<td>4</td>
<td>1106</td>
<td>8</td>
<td>743</td>
<td>9</td>
<td>571</td>
<td>158</td>
<td>282</td>
<td>258</td>
<td>244</td>
<td>3,379</td>
</tr>
<tr>
<td>Not very strong Republican</td>
<td>5</td>
<td>2011</td>
<td>8</td>
<td>1259</td>
<td>15</td>
<td>1,170</td>
<td>299</td>
<td>519</td>
<td>500</td>
<td>484</td>
<td>6,265</td>
</tr>
<tr>
<td>Strong Republican</td>
<td>6</td>
<td>1009</td>
<td>8</td>
<td>751</td>
<td>2</td>
<td>662</td>
<td>180</td>
<td>321</td>
<td>307</td>
<td>239</td>
<td>3,479</td>
</tr>
<tr>
<td>Other party, refused to say</td>
<td>7</td>
<td>243</td>
<td>0</td>
<td>75</td>
<td>1</td>
<td>44</td>
<td>17</td>
<td>44</td>
<td>43</td>
<td>63</td>
<td>530</td>
</tr>
<tr>
<td>Don't know</td>
<td>8</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>No answer</td>
<td>9</td>
<td>64</td>
<td>4</td>
<td>29</td>
<td>3</td>
<td>15</td>
<td>9</td>
<td>49</td>
<td>6</td>
<td>9</td>
<td>188</td>
</tr>
</tbody>
</table>

See Appendix D: Recodes, for original question format and method of recoding. See Appendix N for changes across surveys. If planning to perform trend analysis with this variable, please consult GSS Methodological Report 10, 56.

http://dss.princeton.edu/images/gss.gif
Social science data

66. Generally speaking, do you usually vote for the Democratic or Republican candidate?

<table>
<thead>
<tr>
<th>RESPONSE</th>
<th>PUN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strong Democrat</td>
<td></td>
</tr>
<tr>
<td>Not very strong Democrat</td>
<td></td>
</tr>
<tr>
<td>Independent, close to Democrat</td>
<td></td>
</tr>
<tr>
<td>Independent (Neither, No response)</td>
<td></td>
</tr>
<tr>
<td>Independent, close to Republican</td>
<td></td>
</tr>
<tr>
<td>Not very strong Republican</td>
<td></td>
</tr>
<tr>
<td>Strong Republican</td>
<td></td>
</tr>
<tr>
<td>Other party, refused to say</td>
<td></td>
</tr>
<tr>
<td>Don't know</td>
<td></td>
</tr>
<tr>
<td>No answer</td>
<td></td>
</tr>
</tbody>
</table>

See Appendix D: Recodes, for original codes across surveys. If planning to perform correlation or regression, see No. 56.

http://dss.princeton.edu/images/gss.gif
Humanities data

Brick inscribed with the Sutra on Dependent Origination
Gorakhpur district, late 5th century - early 6th century AD. Ashmolean Museum
The CLAROS Data service provides a RESTful interface for the data of the CLAROS Project, and complements the CLAROS Explorer.

This service provides metadata about archaeology and art in machine-readable formats such as RDF, JSON and KML. The data for the CLAROS project are modelled using the Erlangen OWL-DL 1.0 implementation of the CIDOC Conceptual Reference Model using the http://purl.org/net/crm-owl# namespace.

You can use the Objects and People views to start exploring, or try your hand at a SPARQL query.

Hosted by the University of Oxford's e-research centre, QeRC
This interface was built using Fuseki, humfrey, and a tweaked frontend, all open-source software.

http://data.clarosnet.org/
The aim of this project is to perform a comparative study of three artworks (bronze casts of Islamic provenance), to discover evidence of similarities and to get new insight on their origin.

Probably produced within the Islamic Mediterranean in the eleventh century, the Griffin has incised on its body a long inscription in Arabic expressing good wishes. Captured by the Pisans, it underwent an extraordinary transformation: for centuries it was a terrifying, sound-producing guardian figure on top of the roof of Pisa Cathedral. The present project is focused on the Griffin but also includes alongside it other bronze animal sculptures such as a Lion and a Falcon. It is hoped that the interdisciplinary study of the Griffin will shed light on the significance of such objects in a global Mediterranean culture.

Videos

The Pisa Griffin: an introduction

http://vcg.isti.cnr.it/griffin/

Arte islamica, ippogrifo, XI sec 03, own work
Data are representations of observations, objects, or other entities used as evidence of phenomena for the purposes of research or scholarship.

Description and Classification

The OBO Foundry is a collaborative experiment involving developers of science-based ontologies in creating a suite of orthogonal interoperable reference ontologies in the biomedical sciences followed by other relevant efforts in this domain.

In addition to a listing of OBO ontologies, this site also provides a statement of the current state of OBO ontology development. We welcome feedback and encourage participation.

Click any column header to sort the table by that column. The link to the table of contents is the small link at the very top left of the page.
Precondition:

Researchers share data
Rewards

- Publications
- Publications
- Publications
- Publications
- Publications
- Grants
- Awards and honors
- Teaching
- Service
- Data

http://blog.startfreshtoday.com/Portals/170402/images/improve-credit-score1.jpg
Incentives

• Publications that report the research vs.
• Data that are reusable by others

Image: Alyssa Goodman, Harvard Astronomy
Functions of Scholarly Publications

- **Legitimization**
  - Authority, quality
  - Priority, trustworthiness
- **Dissemination**
  - Awareness
  - Diffusion
  - Publicity
- **Access, preservation, curation**
  - Availability
  - Discovery
  - Retrieval
  - Persistence

Responsibility

Publications are arguments made by authors, and data are the evidence used to support the arguments.

Responsibility

• Publications
  – Independent units
  – Authorship is negotiated

• Data
  – Compound objects
  – Ownership is rarely clear
  – Attribution
    • Long term responsibility: Investigators
    • Expertise for interpretation: Data collectors and analysts
Attribution of data

• Legal responsibility
  – Licensed data
  – Specific attribution required

• Scholarly credit: contributorship
  – “Author” of data
  – Contributor of data to this publication
  – Colleague who shared data
  – Software developer
  – Data collector
  – Instrument builder
  – Data curator
  – Data manager
  – Data scientist
  – Field site staff
  – Data calibration
  – Data analysis, visualization
  – Funding source
  – Data repository
  – Lab director
  – Principal investigator
  – University research office
  – Research subjects
  – Research workers, e.g., citizen science...
Knowledge Infrastructures:
Intellectual Frameworks and Research Challenges

Report of a workshop sponsored by the National Science Foundation and the Sloan Foundation
University of Michigan School of Information, 25-28 May 2012
Australian National Data Service

Our Vision: More Australian researchers reusing research data more often

ANDS is enabling the transformation of:

<table>
<thead>
<tr>
<th>Data that are:</th>
<th>to</th>
<th>Structured Collections that are:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unmanaged</td>
<td></td>
<td>Managed</td>
</tr>
<tr>
<td>Disconnected</td>
<td></td>
<td>Connected</td>
</tr>
<tr>
<td>Invisible</td>
<td></td>
<td>Findable</td>
</tr>
<tr>
<td>Single-use</td>
<td></td>
<td>Reusable</td>
</tr>
</tbody>
</table>

Australian Research Data Commons
ANDS is building the Australian Research Data Commons: a cohesive collection of research resources from all research institutions, to make better use of Australia's research data outputs.

Research Data Australia
Research Data Australia, ANDS' flagship service, provides a comprehensive window into the Australian Research Data Commons. It is an Internet-based discovery service designed to provide rich connections between data, projects, researchers and institutions, and promote visibility of Australian research data collections in search engines.

Search for research data on researchdata.ands.org.au
Reuse across place and time

- Reuse by investigator
- Reuse by collaborators
- Reuse by colleagues
- Reuse by unaffiliated others
- Reuse at later times
  - Months
  - Years
  - Decades
  - Centuries

http://chandra.harvard.edu/photo/2013/kepler/kepler_525.jpg
# Economics of the Knowledge Commons

<table>
<thead>
<tr>
<th>Subtractability / Rivalry</th>
<th>Low</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Exclusion</strong></td>
<td><strong>Public Goods</strong></td>
<td><strong>Common-pool resources</strong></td>
</tr>
<tr>
<td>Difficult</td>
<td>General knowledge</td>
<td>Libraries</td>
</tr>
<tr>
<td></td>
<td>Public domain data</td>
<td>Data archives</td>
</tr>
<tr>
<td>Easy</td>
<td><strong>Toll or Club Goods</strong></td>
<td><strong>Private Goods</strong></td>
</tr>
<tr>
<td></td>
<td>Subscription journals</td>
<td>Printed books</td>
</tr>
<tr>
<td></td>
<td>Subscription data</td>
<td>Raw or competitive data</td>
</tr>
</tbody>
</table>

Adapted from C. Hess & E. Ostrom (Eds.), *Understanding knowledge as a commons: from theory to practice*. Cambridge, Mass.: MIT Press.
Conclusions

• Stakeholders
• Economics
• Sustainability
• Knowledge infrastructures
Stakeholders in digital stewardship

- Scholars
- Students
- Readers / public at large
- Universities
- Funding agencies
- Repositories
- Governments
- Private enterprise
- Libraries
- Archives
- Museums...
Economics of knowledge commons

- Private goods
- Common-pool resources
- Toll or club goods
- Public goods
Sustainability

• What to keep
• Why to keep them
• How to keep them
• How long to keep them
• Who will govern them
• How to manage free riders
• What kinds of expertise are required
Knowledge Infrastructures

• Technical fabric
  – Data archives
  – Metadata, provenance, classification, collections
  – Tools, services, support

• Social fabric
  – Data curation workforce
  – Digital stewardship workforce

• Trust fabric
  – Memory institutions
  – Stewardship
Sewing with the Needle in the Haystack

• Add value to data early and often
  – Metadata
  – Provenance
• Invest in tools and services for interpretation
• Anticipate reuses
• Support serendipity
Acknowledgements

UCLA Data Practices team

- Peter Darch, Milena Golshan, Irene Pasquetto, Ashley Sands, Sharon Traweek

- Former members: Rebekah Cummings, David Fearon, Ariel Hernandez, Elaine Levia, Jaklyn Nunga, Matthew Mayernik, Alberto Pepe, Kalpana Shankar, Katie Shilton, Jillian Wallis, Laura Wynholds, Kan Zhang

- Research funding: National Science Foundation, Alfred P. Sloan Foundation, Microsoft Research, DANS-Netherlands

- University of Oxford: Balliol College, Oliver Smithies Fellowship, Oxford Internet Institute, Oxford eResearch Center, Bodleian Library