Why data are not publications: Potential potholes for STM publishers

Christine L. Borgman
Professor and Presidential Chair in Information Studies
University of California, Los Angeles
@scitechprof

Keynote presentation
STM Publishers U.S. Conference
Washington, D.C.
April 22, 2015
PHILOSOPHICAL
TRANSACTIONS:
GIVING SOME
ACCOMPT
OF THE PRESENT
Undertakings, Studies, and Labours
OF THE
INGENIOUS
IN MANY
CONSIDERABLE PARTS
OF THE
WORLD

Vol I.
For Anno 1665, and 1666.

In the SAVOY,
Printed by T. N. for John Martyn at the Bell, a little without Temple-Bar, and James Allestry in Duck-Lane; Printers to the Royal Society.

Theme issue ‘Celebrating 350 years of Philosophical Transactions: life sciences papers’ compiled and edited by Linda Partridge
19 April 2015; volume 370, issue 1666
Data
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
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
Data, Publications, and Scholarship

• Data
• Publications
• Publications vs. Data
• Releasing data
• Keeping data useful

Persistent URL: photography.si.edu/SearchImage.aspx?id=5799
Repository: Smithsonian Institution Archives
Big Data

Volume
Data Size

Data Complexity

Speed of Change

Velocity

Data Sources

Variety

Long tail of data

Volume of data

Number of researchers

Slide: The Institute for Empowering Long Tail Research
Open Data: Free

• 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

Open Data Commons. (2013).

State Library and Archives of Florida, 1922. Flickr commons photo
Open Data: Useful

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

What are data?

Marie Curie’s notebook

Pisa Griffin

Figure 2. Numeric Change in Resident Population for the 50 States, the District of Columbia, and Puerto Rico: 1990 to 2000

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

http://aip.org

hudsonalpha.org

http://onlineqda.hud.ac.uk/Intro_QDA/Examples_of_Qualitative_Data.php

http://ncl.ucar.edu

http://www.census.gov/population/cen2000
Data are representations of observations, objects, or other entities used as evidence of phenomena for the purposes of research or scholarship.

Making useful data

Sloan Digital Sky Survey Telescope, Apache Point, New Mexico

Sensor networks

http://astro.uchicago.edu/~frieman/SDSS-telescope-photos/

http://enl.usc.edu/~jpaek/data/cyclops/bird_nest_2008/figures/nestbox2.jpg
The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

By Chris Anderson  06.23.08
Research process

• Models and theories
• Research questions
• Methods
  – Tools
  – Data sources
  – Practices
  – Infrastructure
  – Domain expertise
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 Kauffmann & Jaime E. Pineda

Self-gravity plays a decisive role in the final stages of star formation, where dense cores (sizes ~0.1 parsecs) 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 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 CO 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 any realistic simulation of the star-formation process on subparsec scales.

Spectral-line mapping shows how 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 clumps 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 spectroscopic 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 modeled. When applied to molecular-line
Center for Embedded Networked Sensing

- NSF Science & Tech Ctr, 2002-2012
- 5 universities, plus partners
- 300 members
- Computer science and engineering
- Science application areas

Slide by Jason Fisher, UC-Merced,
Center for Embedded Networked Sensing (CENS)
Science <-> Data

Engineering researcher: “Temperature is temperature.”

Biologist: “There are hundreds of ways to measure temperature. ‘The temperature is 98’ is low-value compared to, ‘the temperature of the surface, measured by the infrared thermopile, model number XYZ, is 98.’ That means it is measuring a proxy for a temperature, rather than being in contact with a probe, and it is measuring from a distance. The accuracy is plus or minus .05 of a degree. I [also] want to know that it was taken outside versus inside a controlled environment, how long it had been in place, and the last time it was calibrated, which might tell me whether it has drifted.."
The Pisa Griffin Project

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
FIG. 4. ORE Aggregation representing the first stage of the scientific life cycle of a sensor network application in seismology (experiment and deployment planning).

FIG. 5. ORE Aggregation representing the second stage of the scientific life cycle of a sensor network application in seismology (data collection).
Random walk

http://www2.ess.ucla.edu/~jewitt/oort2-random.html
Publications <-> Data

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

Publications <-> Data

- Article 1
- Article 2
- Article 3
- Article 4
- Article n
- Dataset time 1
- Dataset time 2
- Observation time 1
- Visualization time 3
- Community collection 1
- Repository 1
Publications <-> Data

• 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...

“Publishing” data

• To make public
• To release, post, or share some unit of data
• To link a publication with a dataset
• To peer review some unit of data
• To curate, preserve, and steward data
  – Build coherent collections
  – Sustain scientific value
  – Make discoverable and accessible

ROSC Community Group Charter

Jun Zhao | Posted on: April 17, 2013

Research Object for Scholarly Communication (ROSC) Community Group Charter

Goal

The primary goal of the Community Group is to provide a platform for scholars, librarians, publishers, archivists and policy makers to exchange requirements and expectations for supporting a new form of scholarly communication, i.e. making the actual research assets available as first-class objects to enable better reuse and reproduce of research results and knowledge. These research assets, including data used and generated in an investigation, methods used for producing the data, as well as people and organisations involved in the study, are what we call Research Objects.

It is within the scope of the Community Group to articulate these requirements, gather use cases, survey related work, and analyse and propose various best practices and guidance towards establishing community-wide practice of facilitating the publishing and sharing of the so-called research objects. This Community Group does NOT intend to produce a new specification to achieve the above goals, but may collaborate on forming requirements for other related specifications and vocabularies produced within or outside W3C groups.

Deliverables
Achieving human and machine accessibility of cited data in scholarly publications

Joan Starr¹, Eleni Castro², Mercè Crosas², Michel Dumontier³, Robert R. Downs⁴, Ruth Duerr⁵, Laurel Haak⁶, Melissa Haendel⁷, Ivan Herman⁸, Simon Hodson⁹, Joe Hourclé¹⁰, John Ernest Kratz¹, Jennifer Lin¹¹, Lars Holm Nielsen¹², Amy Nurnberger¹³, Stefan Pröll¹⁴, Andreas Rauber¹⁵, Simone Sacchi¹³, Arthur P. Smith¹⁶, Michael Taylor¹⁷, Tim Clark¹⁸

¹ California Digital Library, University of California, Office of the President, Oakland, CA, USA
² Institute of Quantitative Social Science, Harvard University, Cambridge, MA, USA
³ Stanford Center for Biomedical Informatics Research, School of Medicine, Stanford University, Palo Alto, CA, USA
⁴ Center for International Earth Science Information Network (CIESIN), Columbia University, Palisades, NY, USA
⁵ University of Colorado at Boulder, National Snow and Ice Data Center, Boulder, CO, USA
⁶ ORCID, Inc., Bethesda, MD, USA
⁷ Oregon Health and Science University, Portland, OR, USA
⁸ W3C/CWI, Amsterdam, The Netherlands
⁹ CODATA (ICSU Committee on Data for Science and Technology), Paris, FR
¹⁰ Solar Data Analysis Center, NASA Goddard Space Flight Center, Greenbelt, MD, USA
¹¹ Public Library of Science, San Francisco, CA, USA
¹² European Organization for Nuclear Research (CERN), Geneva, Switzerland
¹³ Columbia University Libraries/Information Services, New York, NY, USA
¹⁴ SBA Research, Vienna, AT
¹⁵ Institute of Software Technology and Interactive Systems, Vienna University of Technology / TU Wien,
Precondition:

Researchers share data
Ways to release data

• Centralized data production
  – Top down investments in data
  – Pooled data resources for the community

• Decentralized data production
  – Bottom up investments in data
  – Local data resources pooled later

• Post on websites

• Share privately upon request…
The Sloan Digital Sky Survey has created the most detailed three-dimensional maps of the Universe ever made, with deep multi-color images of one third of the sky, and spectra for more than three million astronomical objects. Learn and explore all phases and surveys—past, present, and future—of the SDSS.
ICT Diffusion and Distribution Dataset, 1990-2007 (ICPSR 23562)

Principal Investigator(s): Howard, Philip N., University of Washington; Busch, Laura, University of Washington; Cohen, Spencer, University of Washington

Summary:
This dataset covers the years 1990 through 2007 and contains two types of indicators for the global distribution of information, communication and technology (ICT) resources. The data includes gini coefficients for the distribution of Internet access within countries, and a technology diffusion index that weights the distribution of broadband subscribers, personal computers, mobile phones, Internet users, and international Internet bandwidth by economic output. The data are secondary source data... (more info)

Access Notes
- These data are available only to users at ICPSR member institutions. Because you are not logged in, we cannot verify that you will be able to download these data.

Dataset(s)
- Dataset - Download All Files (4.9 MB)
- Documentation: Codebook.pdf
- Data: SAS, SPSS, Stata, ASCII, Delimited

Study Description

Citation

Export Citation:
- RIS (generic format for RefWorks, EndNote, etc.)
- EndNote XML (EndNote X4.0.1 or higher)

Funding
This study was funded by:
- Peoples and Practices Group (IIS-0713074)
- National Science Foundation (IIS-0713074)

Scope of Study
Subject Terms: communications systems, computer use, information dissemination, information systems, Internet, technology
If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology

Jillian C. Wallis, Elizabeth Rolando, Christine L. Borgman

Published: July 23, 2013 • DOI: 10.1371/journal.pone.0067332

Abstract

Research on practices to share and reuse data will inform the design of infrastructure to support data collection, management, and discovery in the long tail of science and technology. These are research domains in which data tend to be local in character, minimally structured, and of short-term interest.
Lack of incentives to release data

- Labor to document data
- Benefits to unknown others
- Competition
- Control
- Confidentiality...
Lack of incentives to reuse data

• Identify useful data
  – Documentation
  – Interpretation
  – Software

• Cleaning

• Trust

• Credit

• Licensing...

Keeping Data Useful

Flickr Commons Photo: Women working in the Pinion Department at Bulova Watch, Southern Methodist University Libraries; Creator: Richie, Robert Yarnall (1908-1984), 1937
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
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
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)

British Library, provenance record: Bestiary - caption: 'Owl mobbed by smaller birds'
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

Image from Soumitri Varadarajan blog. Iceberg image © Ralph A. Clevenger. Flickr photo
Data Curation and Stewardship

- Services and tools
- Data management planning
- Selection and appraisal
- Metadata, provenance
- Migration
- Economics
- Infrastructure

http://www.librarygirl.net/2013/08/putting-your-best-foot-forward-tl.html
# Economics of the Knowledge Commons

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

Proportion of Repositories by Continent - Worldwide

This chart is based on the number of repositories in each Continent. However, some organisations have two or more repositories - over 20 in some cases - and this arguably skews the results.

For a different viewpoint, please see the equivalent chart for Repository Organisations, in which each organisation only counts once, regardless of how many repositories it hosts.

For further data, please see the corresponding table of repositories sorted by country.
Conclusions

• Data
  – Representations used as evidence
  – One person’s signal is another’s noise
• Publications
  – Peer-reviewed documents
  – Networks of research objects
• Publications vs. data
  – Authorship vs. attribution, credit
  – Relationships are many to many
• Releasing data
  – “Publishing data” is an oxymoron
  – Lack of incentives to release or reuse
• Keeping data useful
  – Invest in repositories, infrastructure, expertise
  – Value propositions vary
Acknowledgements

UCLA Data Practices team

• Peter Darch, Milena Golshan, Irene Pasquetto, Ashley Sands, Sharon Traweek, Camille Mathieu

• Former members: Rebekah Cummings, David Fearon, Ariel Hernandez, Elaine Levia, Jaklyn Nunga, Rachel Mandel, 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