DISCOVERY, ACCESS, AND CITATION OF PUBLISHED RESEARCH DATA

THE DATA CITATION INDEX

NIGEL ROBINSON

29 MAY 2013
OUTLINE

- WHY CREATE THE DATA CITATION INDEX?
- DATA CITATION BEHAVIOUR
- THE CHALLENGES OF METADATA
- THE OUTCOME
DIGITAL EXPANSION

1.8 ZETTABYTES

(1.8 trillion gigabytes) of information will be created and replicated in 2011—and growing fast (it has grown by a factor of 9 in just five years)

DIGITAL SCHOLARSHIP

Digital Scholarship
- Very visible within the literature as a concept
- Articles, projects, university labs all devoted to digital scholarship in various ways

Interested Parties
- Authors/researchers
- Research administrators
- Librarians, data archivists
- Publishers
- Grant funding organizations

Content
- Discipline-specific and multidisciplinary content
- Needs and requirements vary by discipline
- Diverse content formats, with few standards
- Includes collaboration and communications
DEPOSITION OF DATA BY RESEARCHERS

- Publisher website: 24%
- Repository managed by a third party (e.g., domain-...: 36%
- Department or institutional repository: 47%
- Personal website: 51%
- Other: 17%

Q16. Where do you place your non-traditional scholarly output to make it available to others? (n=471)
RESEARCHERS NOT RECEIVING CREDIT

Researchers are not receiving adequate credit for digital scholarship
- Researchers would spend more time producing output if they received credit
- Few are currently receiving adequate credit

Barriers to creating and sharing data:
- Researchers are hesitant to spend time and effort to create and share data because they don’t feel the work is adequately exposed or accredited
- Researchers find it difficult to expose data they have produced because data repositories do not have clear standards or mechanisms in place for doing so

Institutional policies and the lack of standards for sharing digital output is an impediment both to research and sharing

“Lack of knowledge about standards for citation and of proper scholarly recognition and/or evaluation of such materials.” (Researchers, Canada)
BARRIERS TO RESEARCHERS CITING DATA

Researchers agree that data *should* be cited, but there are currently no universally accepted standards for citing data

“Lack of knowledge about standards for citation and of proper scholarly recognition and/or evaluation of such materials.”…

“…cumbersome citation formats including very long internet addresses.”

“Incomplete citation information available (dates and real author names as distinct from aliases)”
BENEFITS OF RESEARCH DATA SHARING

- Advancement of scholarship
- Verification of results
- Promotion of scholar’s work
NIH (2003) Data Sharing Policy that all funding applications of $500,000 or more per year are expected to address data-sharing in their application.

NSF (2011) All funding proposals submitted on or after January 18, 2011, must include a “Data Management Plan” describing how the proposal will conform to NSF policy on the dissemination and sharing of research results.
INTERNATIONAL MANDATE EXTENSION

Aug 2011… “expectation that all our funded researchers should **maximise access to their research data** with as few restrictions as possible. …. submit a **data management and sharing plan** as part of the application process.”

2007… “Researchers are to retain research data and primary materials, **manage storage of research data** and primary materials, maintain confidentiality of research data and primary materials.”
January 14, 2013… “failure to provide the requisite Data Management Plan will result in the application being rejected or terminated.”
IMPACT ON RESEARCH LIBRARIES

Data Management at Harvard

Funding Agencies Guidelines and Policies

National Science Foundation

ARL ASSOCIATION OF RESEARCH LIBRARIES

Transforming Research Libraries

Planning and Visioning

Resources for Planning and Visioning

Envisioning Research Library Futures: A Scenario Thinking Project

E-Science

Guide for Research Libraries: The NSF Data Sharing Policy

Authors: Patricia Hswe and Ann Holt

Overview

In Spring 2010, the National Science Foundation (NSF) announced that it would alter its data sharing policy to require data management plans (DMPs) in future grant proposals to the agency. The Association for Research Libraries has developed...

Creating a Data Management Plan

Managing Your Data

Data Management - Home

Data Management Plans

Creating a Data Management Plan

1. Funding Agency Guidelines: Review your funder’s requirements for data sharing.
VISIBILITY OF RESEARCH DATA

- Grant funding agencies
- Journal publishers
- Data repositories & registration agencies
OBSERVED RESEARCHER PROBLEMS

- Access & discovery
- Citation standards
- Lack of willingness to deposit and cite
- Lack of recognition / credit
WHERE DO WE START?

- Enable the discovery of data repositories, data studies and data sets in the context of traditional literature
- Help researchers find data sets and studies and track the full impact of their research output
- Provide expanded measurement of researcher and institutional research output and assessment
- Facilitate more accurate and comprehensive bibliometric analyses
As we evaluate repositories for inclusion, some of the things we consider are:

- Editorial Content - ensuring that material is desirable to the research community.
- Persistence and stability of the repository, with a steady flow of new information.
- Thoroughness and detail of descriptive information.
- Links from data to research literature.
DATA REPOSITORIES

- Over 700 repositories identified
## TYPES OF DATA BY DISCIPLINE

<table>
<thead>
<tr>
<th>ART &amp; HUMANITIES</th>
<th>SOCIAL SCIENCES</th>
<th>SCIENCE &amp; TECHNOLOGY</th>
</tr>
</thead>
<tbody>
<tr>
<td>CULTURAL HERITAGE</td>
<td>POLL DATA</td>
<td>MAPS</td>
</tr>
<tr>
<td>LANGUAGE CORPUS</td>
<td>ECONOMIC STATISTICS</td>
<td>ALGORITHMS</td>
</tr>
<tr>
<td>IMAGE COLLECTIONS</td>
<td>LONGITUDINAL DATA</td>
<td>GENOMICS</td>
</tr>
<tr>
<td>RECORDINGS</td>
<td>NATIONAL CENSUS</td>
<td>SKY SURVEYS</td>
</tr>
<tr>
<td></td>
<td>PUBLIC OPINION SURVEYS</td>
<td>ASTROPHYSICS</td>
</tr>
<tr>
<td></td>
<td></td>
<td>REMOTE SENSING</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MUSEUM SPECIMENS</td>
</tr>
</tbody>
</table>
INDEXING A DATA REPOSITORY ON WEB OF KNOWLEDGE

Record Types

- **Repository/Source**: Comprises data studies, data sets and/or microcitations. Stores and provides access to the raw data.

- **Data Study**: Descriptions of studies or experiments with associated data which have been used in the data study. Includes serial or longitudinal studies over time.

- **Data Set**: A single or coherent set of data or a data file provided by the repository, as part of a collection, data study or experiment.

- **Microcitation**: (nanopublication) An assertion about concepts that have been found to be linked by scientific enquiry, and can be uniquely identified and attributed to its author. Made up of three separate parts: a subject, a predicate and an object.
CHALLENGES

- Metadata availability
  - Lack of repository resources
  - Lack of repository expertise

- Metadata quality
  - Metadata inconsistencies

- Data repositories are not static

- Partnerships
DATA CITATION INDEX - METADATA PARTNERSHIPS

DataCite

Repository 1

Repository 2

Repository 3

Data Citation Index

Repository 1

Repository 2

Repository 3

DataCite

Data Citation Index
COLLABORATION BENEFITS

• Any repository providing metadata to the aggregator is included in the Data Citation Index

• Uniform data

• Faster and more frequent updates
DATA CITATION BEHAVIOUR

Current citation style (in full text of article)

III. Data Description

Our first data set is the Bureau of Justice Statistics "Murder Cases in 33 Large Urban Counties." This is a random sample of homicide cases drawn from prosecutors' files. The data set includes information on offender characteristics, victim characteristics and trial outcomes for 2800 murders. The 75 largest counties account for more than half of the murders in the U.S. each year. This data set brings together information on the crime, the offender, the victim, and the sentence. Such information cannot all be linked in other larger data sets such as the Uniform Crime Reporting (UCR) Data or the National Crime Victimization Survey (NCVS). Most crime

Desired/future citation style (as part of cited references)


Search Results within the Data Citation Index present all of the powerful Web of Knowledge options for exploring a body of information.
UniProt Knowledgebase.

Editor(s): Uniprot Consortium
Source: UniProt Knowledgebase
Source URL: http://www.uniprot.org/ (Viewed Date: 13 Dec 2011) Published Year: 2002
Cited References: 0
Abstract: The UniProt Knowledgebase (UniProtKB) is the central hub for the collection of functional information on proteins, with accurate, consistent and rich annotation. In addition to capturing the core data mandatory for each UniProtKB entry (mainly, the amino acid sequence, protein name or description, taxonomic data and citation information), as much annotation information as possible is added. This includes widely accepted biological ontologies, classifications and cross-references, and clear indications of the quality of annotation in the form of evidence attribution of experimental and computational data.

Document Type: Repository
Accession Number: DRUI2012001000682873
Language: English
Funding:

<table>
<thead>
<tr>
<th>Funding Agency</th>
<th>Grant Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>National Institutes of Health</td>
<td>1U41HG006104-01</td>
</tr>
<tr>
<td>European Commission SLING</td>
<td>226673</td>
</tr>
<tr>
<td>NIH GO</td>
<td>2P41HG02273-07</td>
</tr>
<tr>
<td>Swiss Federal Office of Education and</td>
<td></td>
</tr>
<tr>
<td>Science GEN2PHEN</td>
<td>200754</td>
</tr>
<tr>
<td>MICROME</td>
<td>222886-2</td>
</tr>
<tr>
<td>NIH</td>
<td>5R01GM080646-04</td>
</tr>
<tr>
<td>NIH</td>
<td>3R01GM080646-04S2</td>
</tr>
<tr>
<td>NIH</td>
<td>1R01GM080720-01</td>
</tr>
</tbody>
</table>

Funding information for the Repository is presented when available.
Data Citation Index℠

UniProt Knowledgebase.

Editor(s): UniProt Consortium
Source: UniProt Knowledgebase
Source URL: http://www.uniprot.org/
Cited References: 0

Abstract: The UniProt Knowledgebase offers accurate, consistent and rich annotation of protein sequences and amino acid sequence. Protein name or other identifiers can be added. This includes widely used columns of quality annotation in the format.

Document Type: Repository
Accession Number: DRC1DATA2011
Language: English
Funding:

Funding Agency
National Institutes of Health
European Commission SLING
NIH GO
Swiss Federal Office of Education and Science
GEN2PHEN

MICROME
222862-2
NIH
5R01GM080646-04
NIH
3R01GM080646-04S2

From protein sequences to 3D-structures and beyond: the example of the UniProt Knowledgebase

Hinz, Ursula. From protein sequences to 3D-structures and beyond: the example of the UniProt Knowledgebase. CELLULAR AND MOLECULAR LIFE SCIENCES, APR 2010.


Times Cited: 11

Create Citation Alert
This article has been cited 11 times in Web of Knowledge.


MacDonald, Justin A. Intrinsically Disordered N-Terminus of Calponin Homology-Associated Smooth Muscle Protein (CHASMP) Interacts with the Calponin Homology Domain to Enable Tropomyosin Binding. BIOCHEMISTRY, APR 3 2012.


[view all 11 citing articles]

Cited References: 0

Additional information
How to cite this Resource
Suggest a correction
<table>
<thead>
<tr>
<th>Study Title</th>
<th>Data Type</th>
<th>Link to External Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>USIA/DIVO</td>
<td>Poll # USIA1957-XX10.</td>
<td>Data study</td>
</tr>
<tr>
<td>USIA Poll # 1987-187072: American Image.</td>
<td>Data study</td>
<td>Link to External Source</td>
</tr>
<tr>
<td>Harvard/Health Canada Poll # 2003-SARS: SARS.</td>
<td>Data study</td>
<td>Link to External Source</td>
</tr>
<tr>
<td>USIA Poll # 1200140: El Salvador's Economic Situation/Politics/Osama bin Laden.</td>
<td>Data study</td>
<td>Link to External Source</td>
</tr>
</tbody>
</table>
Data Citation Index℠

Welcome

Dataset Abstract

Title: USIA Poll # 1957-XX10: War/Science/Foreign Policy [GMUSIA1957-XX10]

Survey Sponsor: United States Information Agency [USIA]

Survey Firm: DIVO-Institut fur Wirtschaftsforschung, Sozialforschung und Angewandte Mathematik

Study Date: November, 1957

Sample: West German adults—21 years old and over

Sample Size: 813

Variables: 187

Abstract: Feelings toward various countries (4); USSR favorability (2); US favorability (2); feelings toward political leaders (4); atomic energy (1); West Germany aligning with east or west (1); side in US-Russia war (1); West German war involvement (1); America preventing war (1); Soviet Russia preventing war (1); US vs. USSR as stronger power (2); stronger scientifically developed country (3); stronger military power (2); stronger atomic power (1); ensuring security (1); NATO effectiveness (2); increasing military for modern weapons (2); American forces in West.

Subject Area: Social Sciences - Other Topics

Associated Records: [View All]

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Data study</th>
<th>Link to External Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>USIA/DIVO</td>
<td>Poll # USIA1957-XX10.</td>
<td>Data study</td>
</tr>
<tr>
<td>USIA Poll # 1987-187072: American Image.</td>
<td>Data study</td>
<td>Link to External Source</td>
</tr>
<tr>
<td>Harvard/Health Canada Poll # 2003-SARS, SARS.</td>
<td>Data study</td>
<td>Link to External Source</td>
</tr>
<tr>
<td>USIA Poll # 1200140, El Salvador’s Economic Situation/Politics/Osama bin Laden.</td>
<td>Data study</td>
<td>Link to External Source</td>
</tr>
</tbody>
</table>

Link out directly to the original item, in this case a Data Study.
GSE12195: Mutations of multiple genes deregulate the NF-kB pathway in diffuse large B cell lymphoma.

From Repository: Gene Expression Omnibus
Authors: Pasqualucci, Laura; Basso, Katia
Published Year: 2009
Cited References: 0

Abstract: Diffuse large B-cell lymphoma (DLBCL), the most common form of lymphoma in adulthood, comprises distinct subtypes including germinal center B cell-like (GCB) and activated B cell-like (ABC) DLBCL. Gene expression of its most aggressive subtype, ABC-DLBCL, is associated with constitutive activation of the NF-kB transcription factor in a fraction of cases. It remains unclear whether NF-kB activation in these tumors represents an intrinsic program for pathogenic events. Here we show that >50% of ABC-DLBCL and a smaller fraction of GCB-DLBCL carry somatic inactivating mutations of the tumor suppressor genes TRAF2 and CARD11, which encode a ubiquitin-modifying enzyme involved in regulation of NF-kB responses, in the context of the patients displaying biallelic inactivation by mutations and/or deletions, suggesting a tumor suppressor role of TRAF2 and CARD11 in producing molecules with significantly enhanced ability to activate NF-kB. Thus, the rest of DLBCL is caused by genetic lesions affecting multiple genes, whose loss or activation may promote lymphoma by deregulated NF-kB responses. We show that most ABC-DLBCL and a smaller fraction of GCB-DLBCL display genetic pathway alterations with A20 representing the most frequently mutated gene.

Document Type: Data study
Data Type: Expression profiling by array
Accession Number: DRR100700000275875
Language: English
Author Keywords: Phenotypic characterization of human DLBCL
Addresses:
1. Columbia University Institute for Cancer Genetics, 1130 St. Nicholas Ave., New York, 10032, USA.
2. Columbia University Institute for Cancer Genetics, 150 St. Nicholas Ave., New York, 10032, USA.
E-mail Address: lp1741@columbia.edu; kdb451@columbia.edu

Web of Science Category: Biochemistry & Molecular Biology; Genetics & Heredity
Subject Area: Biochemistry & Molecular Biology; Genetics & Heredity

<table>
<thead>
<tr>
<th>Taxonomic Data:</th>
<th>Organism Classifier</th>
<th>Organism Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUPER TAXA</td>
<td>Animals, Chordata, Vertebrata, Mammalia, Primates</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>TAXA NOTES</td>
<td>Animals, Chordates, Humans, Mammals, Primates, Vertebrates</td>
<td>Homo sapiens</td>
</tr>
</tbody>
</table>

Miscellaneous: Transcription; missense mutation; Gene Expression Profiling; phenotype: Tumor; B-Cell lymphoma; Germinal Center; genomics; Molecular Genetics.

Associated Records: [View All]
- GSM476306: Lymphoblastoid B cell line_JARC 304. Data set [Link to External Source]
- GSM476291: Follicular lymphoma_FL.E.15. Data set [Link to External Source]
- GSM476279: Follicular lymphoma_FL.E.03. Data set [Link to External Source]
OUTCOMES - DATA CITATION INDEX

• Discovery of data most important to scholarly research
• Data linked to published research literature
• Measures of data use and reuse
• New metrics for digital scholarship
Thank you

Nigel Robinson
nigel.robinson@thomsonreuters.com