Digital Preservation

Data Citation

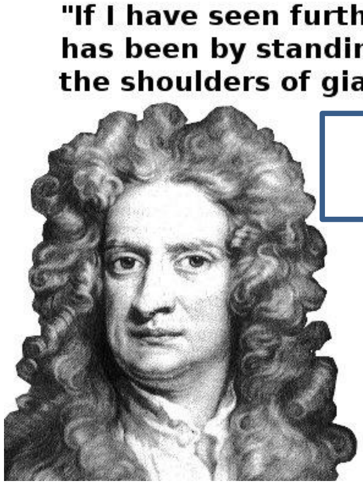
Stefan Pröll

Outline

- Introduction and Motivation
  - Why should we reference?
- Persistent Identifiers
  - Isn‘t a URL enough?
- Citing Datasets
  - Best Practices
- Future Research
  - Dynamic Datasets

Why Should We Cite?

- Science is a collaborative approach



"If I have seen further, it has been by standing on the shoulders of giants."

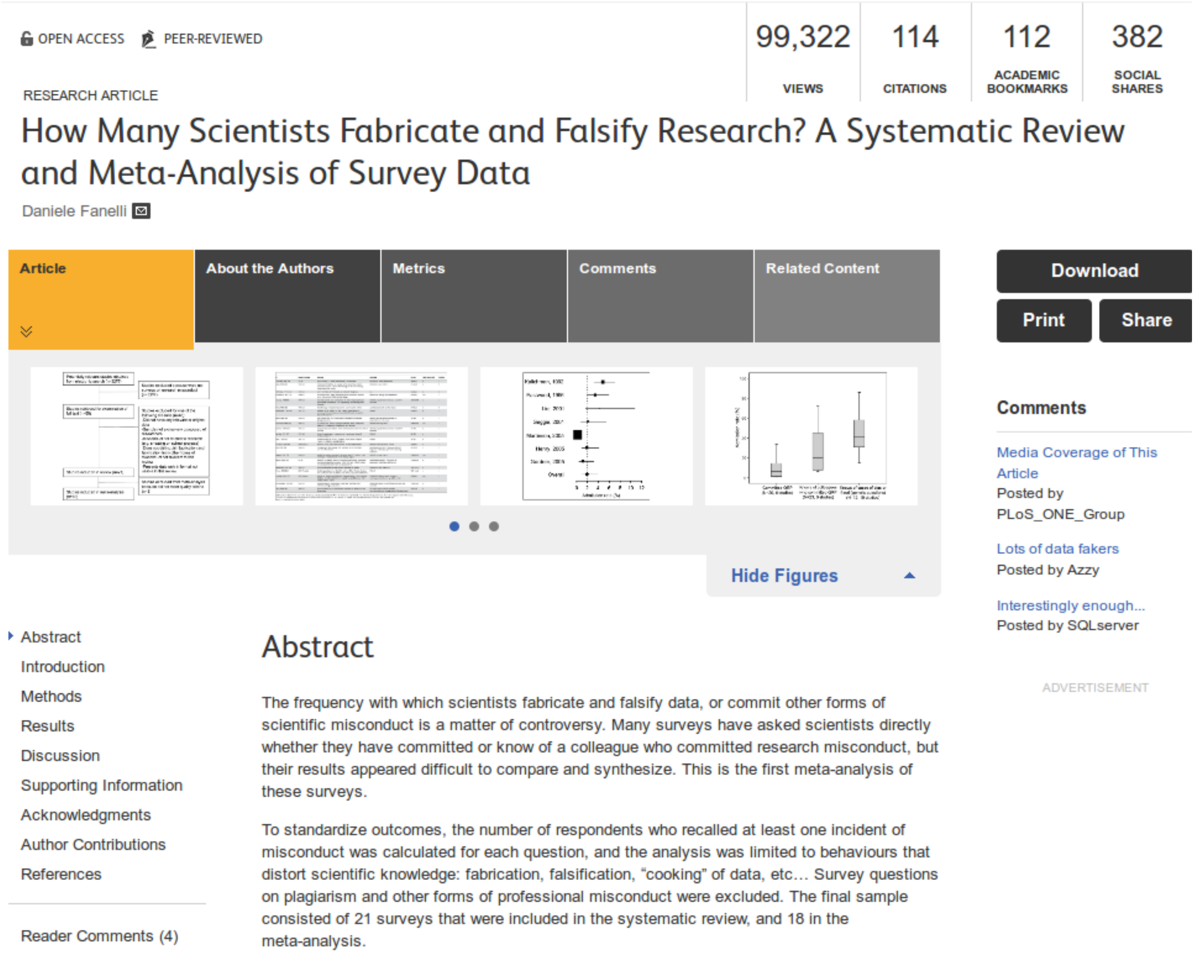
??

- Giving credit to peers and acknowledge their work

Benefits of Citation

- Enables reproducibility
- Enhances transparency
- Serves as documentation
- Settles the context
- Allows exact identification of results
- Track the impact of research

Citations Help Detecting Scientific Misconduct



OPEN ACCESS PEER-REVIEWED

99,322	114	112	382
VIEWS	CITATIONS	ACADEMIC BOOKMARKS	SOCIAL SHARES

RESEARCH ARTICLE

How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data

Daniele Fanelli

Comments

Media Coverage of This Article
Posted by PLoS_ONE_Group

Lots of data fakers
Posted by Azzy

Interestingly enough...
Posted by SQLserver

ADVERTISEMENT

- Abstract
- Introduction
- Methods
- Results
- Discussion
- Supporting Information
- Acknowledgments
- Author Contributions
- References

Reader Comments (4)

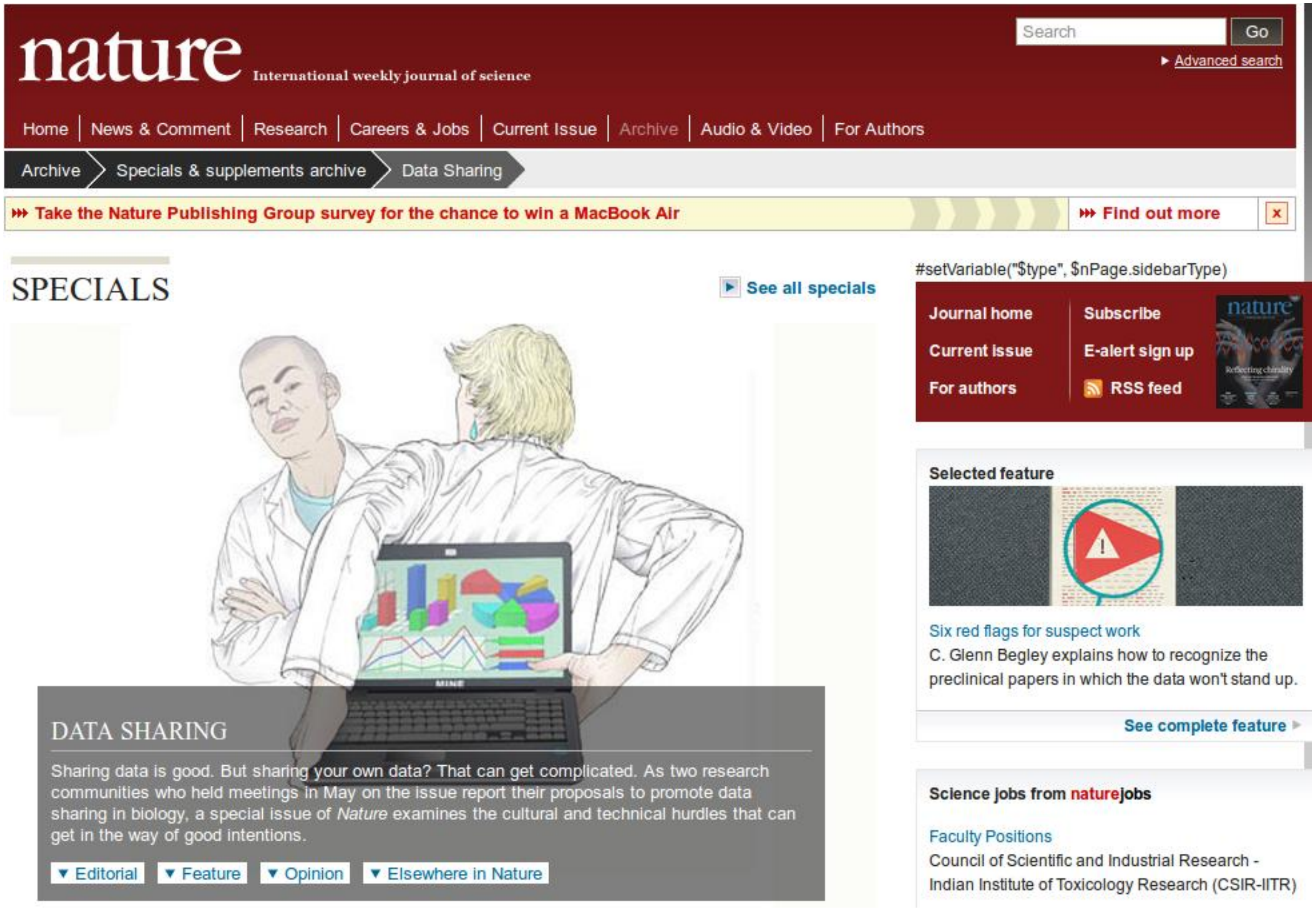
Abstract

The frequency with which scientists fabricate and falsify data, or commit other forms of scientific misconduct is a matter of controversy. Many surveys have asked scientists directly whether they have committed or know of a colleague who committed research misconduct, but their results appeared difficult to compare and synthesize. This is the first meta-analysis of these surveys.

To standardize outcomes, the number of respondents who recalled at least one incident of misconduct was calculated for each question, and the analysis was limited to behaviours that distort scientific knowledge: fabrication, falsification, "cooking" of data, etc… Survey questions on plagiarism and other forms of professional misconduct were excluded. The final sample consisted of 21 surveys that were included in the systematic review, and 18 in the meta-analysis.

Source: http://www.plosone.org

Advance in Research by Sharing



nature International weekly journal of science

Search Go

▸ Advanced search

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Authors

Archive › Specials & supplements archive › Data Sharing

⏵⏵⏵ Take the Nature Publishing Group survey for the chance to win a MacBook Air ⏵⏵⏵ Find out more ×

SPECIALS

▶ See all specials

#setVariable("$type", $nPage.sidebarType)

Journal home | Subscribe
Current issue | E-alert sign up
For authors | RSS feed

DATA SHARING

Sharing data is good. But sharing your own data? That can get complicated. As two research communities who held meetings in May on the issue report their proposals to promote data sharing in biology, a special issue of *Nature* examines the cultural and technical hurdles that can get in the way of good intentions.

▼ Editorial ▼ Feature ▼ Opinion ▼ Elsewhere in Nature

Selected feature

Six red flags for suspect work

C. Glenn Begley explains how to recognize the preclinical papers in which the data won't stand up.

See complete feature ▸

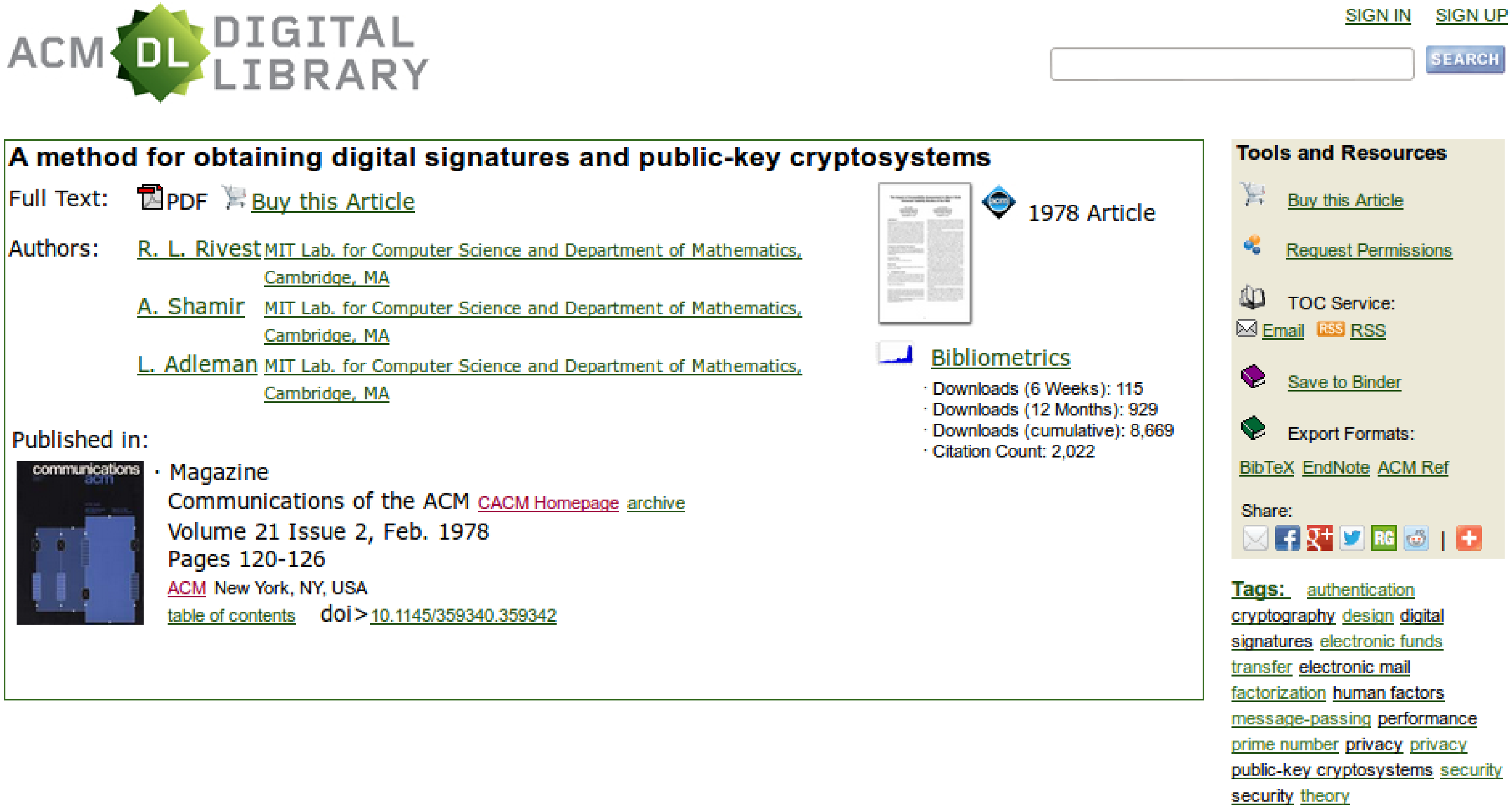
Science jobs from naturejobs

Faculty Positions

Council of Scientific and Industrial Research - Indian Institute of Toxicology Research (CSIR-IITR)

Paper Publications

- Referencing research papers is well established



A method for obtaining digital signatures and public-key cryptosystems

Full Text: PDF Buy this Article

Authors: R. L. Rivest MIT Lab. for Computer Science and Department of Mathematics, Cambridge, MA
A. Shamir MIT Lab. for Computer Science and Department of Mathematics, Cambridge, MA
L. Adleman MIT Lab. for Computer Science and Department of Mathematics, Cambridge, MA

1978 Article

Bibliometrics
· Downloads (6 Weeks): 115
· Downloads (12 Months): 929
· Downloads (cumulative): 8,669
· Citation Count: 2,022

Published in:

· Magazine
Communications of the ACM CACM Homepage archive
Volume 21 Issue 2, Feb. 1978
Pages 120-126
ACM New York, NY, USA
table of contents doi>10.1145/359340.359342

Tools and Resources

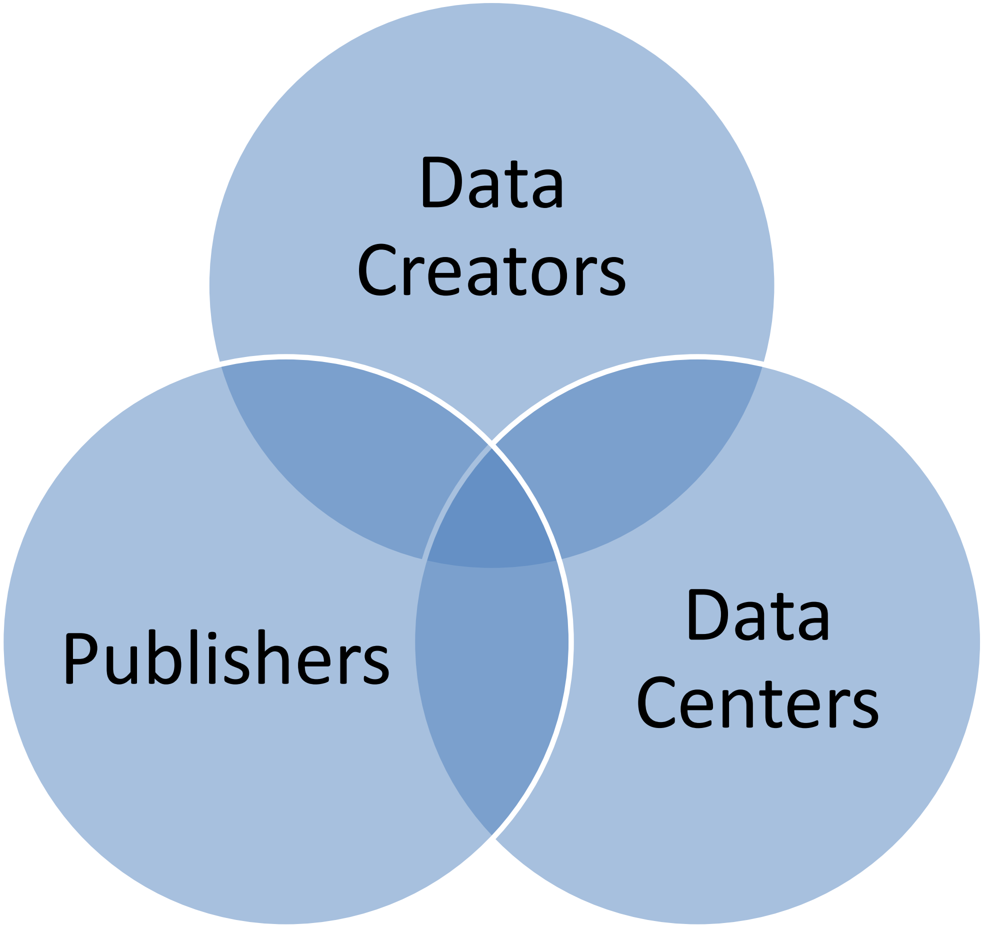
Buy this Article
Request Permissions
TOC Service:
Email RSS RSS
Save to Binder
Export Formats:
BibTeX EndNote ACM Ref
Share:

Tags: authentication cryptography design digital signatures electronic funds transfer electronic mail factorization human factors message-passing performance prime number privacy privacy public-key cryptosystems security security theory

Data Publications

- Data is an essential part of research
  - Majority of papers is based upon research data
  - Needed for validation and reproduction of experiments
- Challenges
  - Encourage researchers to share
  - Different data formats
  - Potentially large storage size
  - Who maintains it?

Stakeholders



Data Center Policies

Criteria for Assessing Value of Data

- Relevance to mission
- Scientific value
- Uniqueness
- Potential for redistribution
- Non-Replicability
- Costs
- Documentation
- ….

http://www.dcc.ac.uk/resources/how-guides/appraise-select-data

Data Citation Requirements

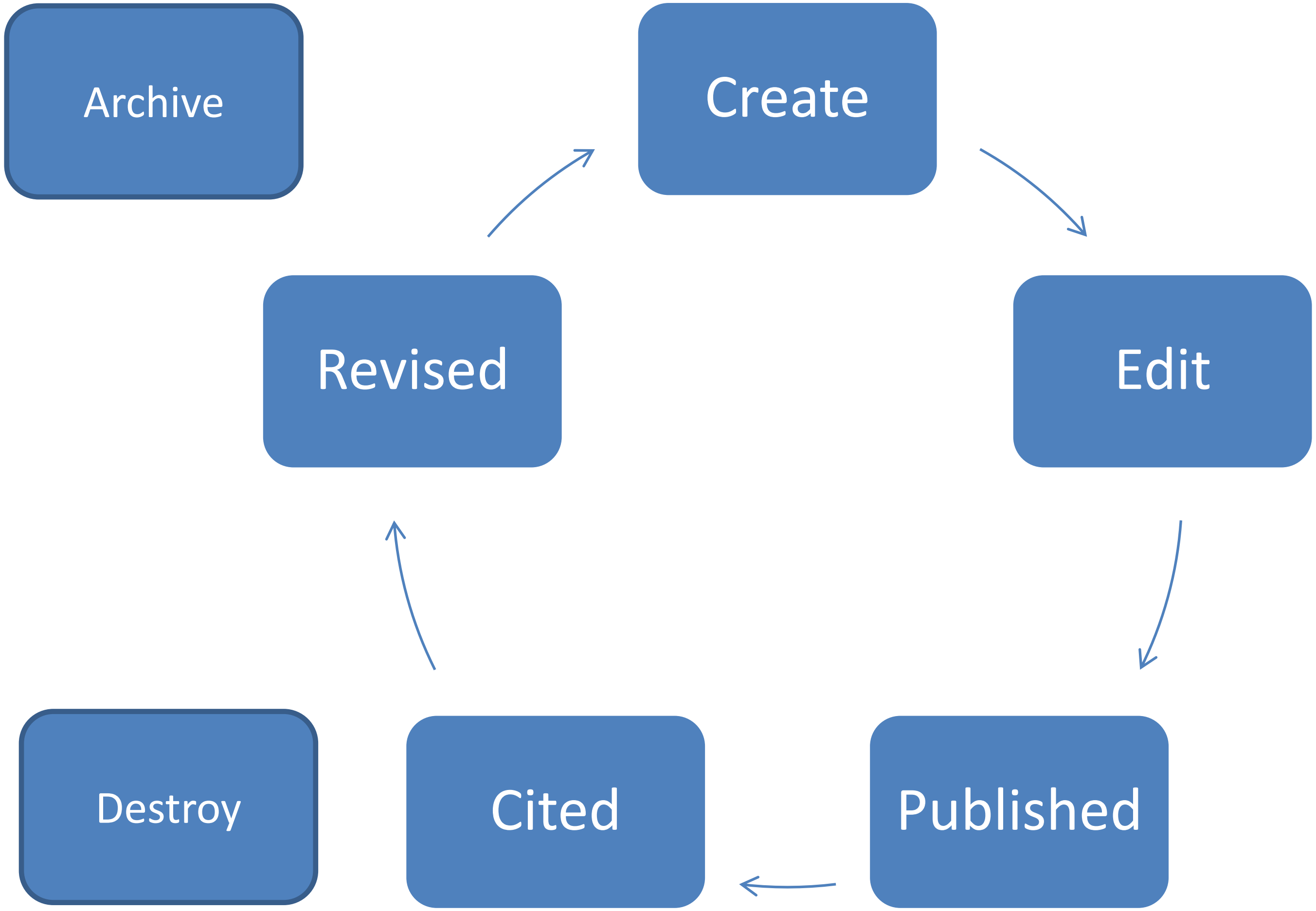
- Unique identification
- Identify subsets and complete dataset
- Machine readable metadata
- Human readable metadata
- Citation metrics

Elements of Data Citation

- Classical bibliographic details:
  - Author, date, edition
  - Publisher, version
- Specific details:
  - Feature name, resource type
  - Unique numeric fingerprint (hash)
  - Persistent identifier
  - Location

Source: http://www.dcc.ac.uk/resources/how-guides/cite-datasets

Digital Object Life Cycle



Archive

Create

Edit

Published

Cited

Revised

Destroy

Identifiers

- Identifier is a symbol that uniquely identifies a digital object. Can be dependent on the context.
- Are URLs identifiers?
- Are URLs persistent?
  - Can URLs be mapped to the digital object life cycle?

Sources: http://ands.org.au/guides/persistent-identifiers-working.html

Existing Unique Identifier Models

- Traditional Mechanisms
  - International Standard Serial Number (ISSN)
    - Unique eight-digit number
    - Identifiers periodical publications
    - Can be encoded as URN
  - International Standard Book Number (ISBN)
    - Unique commercial book identifier barcode
    - 13 (since 2007) or 10 digits with checksum
    - ISBN-10: 3836217155
    - ISBN-13: 978-3836217156

Unique Identifiers for Digital Objects

- Locator Based Mechanisms
  - Uniform Resource Identifier (URI)
  - Uniform Resource Locator (URL)
  - Uniform Resouce Name (URN)
  - National Bibliographic Numbers (NBNs)
- Delegating Methods
  - DOI
  - The Handle System
  - Digital Object Identifier
  - Persistent URL (PURL)
  - Archival Resource Key (ARK)

URL

- Uniform Resource Locator
- Particular form of URIs
- Addressing documents
- Can only be used to locate resources
- Depend on DNS information
- URLs are not persistent
- Problems
  - Links may break
  - No mechanism to handle broken links

URN

- Uniform Resource Name
- Combination of namespace identifier (NID) and a namespace specific string (NSS)
- Naming scheme for URNs:
- urn: <NID> :<NSS>
- Example: urn:isbn:0451450523

URN (2)

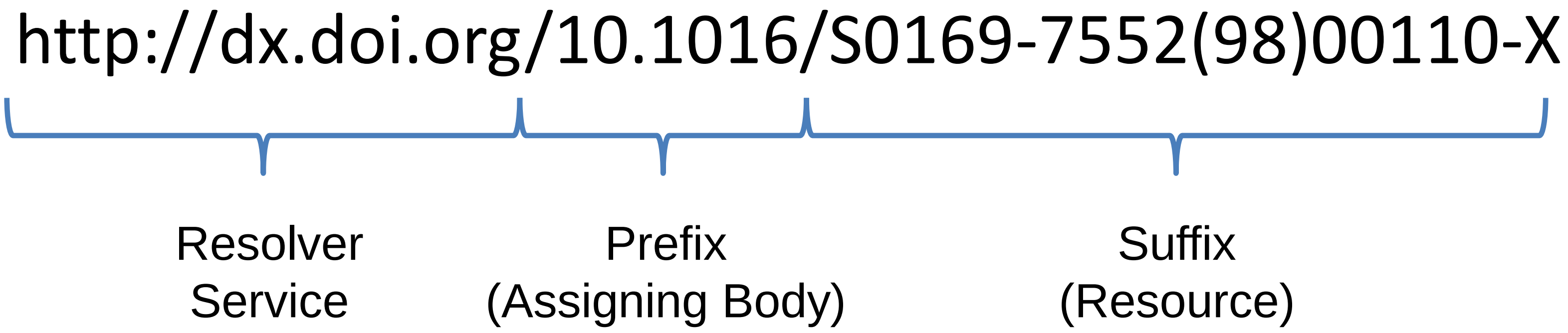
- Main functions of a URN
  - Global scope of names
  - Global uniqueness
  - Persistence
  - Scalability
  - Legacy support
  - Extensibility
  - Independence
  - Resolution
- Persistence of identifiers
- Possibility to resolve them

Persistency?

- Standard URLs are not forever
  - Network locations
  - Not suitable for the long term
  - Link rot: half of the links in publications are not available after 5 to 7 years
- Persistent identifiers need to be maintained
  - Proactively
  - Throughout the whole lifecycle
- Can a URL be persistent?

Digital Object Identifier (DOI)

- Identifier scheme administered by the International DOI Foundation
- Consists of three parts:



http://dx.doi.org/10.1016/S0169-7552(98)00110-X

Resolver Service

Prefix (Assigning Body)

Suffix (Resource)

DOI

- Publisher (organizations) register and get a unique ID (Prefix)
- Resource gets an ID (Suffix) which is unique within the prefix
- Resolver services maintain the link between the endpoint (e.g. URL) and the resource
  - http://dx.doi.org/

Handle

- Distributed persistent naming system
- Conforms to URN framework
- Used by DOI (Digital Object Identifier) system
- Persistent identifier consists of two parts:
  - Naming authority
  - Name (must be unique string to the authority)
- Digital objects on the Internet can be assigned, managed and resolved by handles.
- Resolved by global handle service

Handle (2)

- Main points
  - Handles are unique and persistent
  - Handle system supports internationalization
  - Operations on handle system have to be authorized
- Syntax:

  <Handle Naming Authority> ‚/‘ <Handle Local Name>
- Example:
  - 10.1045/january2013-burns
- Available Services:
  - http://hdl.handle.net

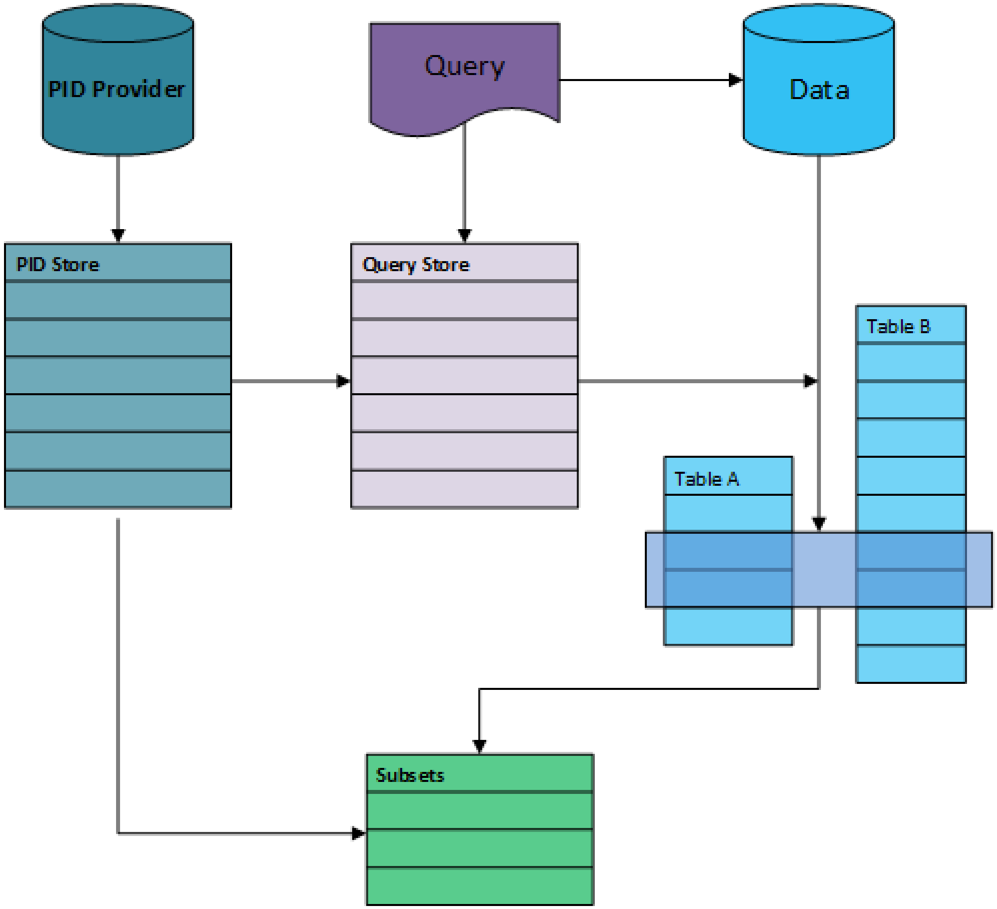
Current Challenges

- Granularity
  - How to define subsets?
  - Should individual data records be cited?
  - Assign each database row a DOI?
- Identify people (authors)?
- Dynamic data
  - How to treat evolving data?

Referencing Dynamic Subsets

- What is needed:
  - Uniquely identifiable data records
  - Time stamps of data
  - Versioned data, considering markings of deleted, altered or inserted data records
  - Precise query language for constructing subsets
  - Persistent query store that keeps queries and the timestamp of their issuing
  - An identification mechanism for queries, that enables access

Citing Dynamic Data in Databases



Referencing Dynamic Subsets

- High Level Requirements
  - Dynamic data
    - Queries need to be stored
    - Temporal data and queries
  - Assemble subsets
  - Scalability is enabled
  - Implementation is transparent
  - Machine actionable

Literature and Links

- http://www.dcc.ac.uk/resources/how-guides/appraise-select-data
- http://www.dcc.ac.uk/resources/how-guides/cite-datasets
- http://dl.acm.org/citation.cfm?doid=602421.602422
- http://ands.org.au/guides/persistent-identifiers-working.html
- http://hdl.handle.net/
- http://dx.doi.org/

Thank you for your attention.

sproell@sba-research.org