Digital Preservation

Data Citation

Stefan Pröll

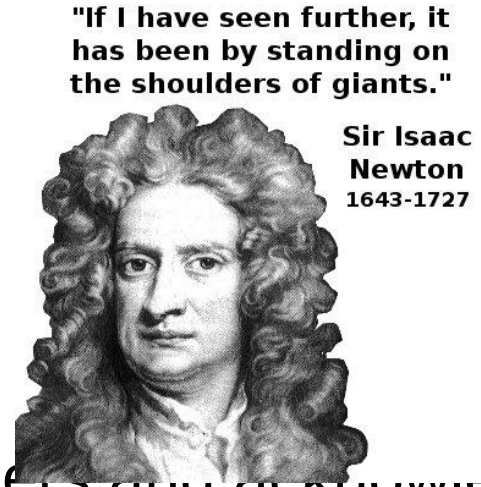
28.04.2014

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



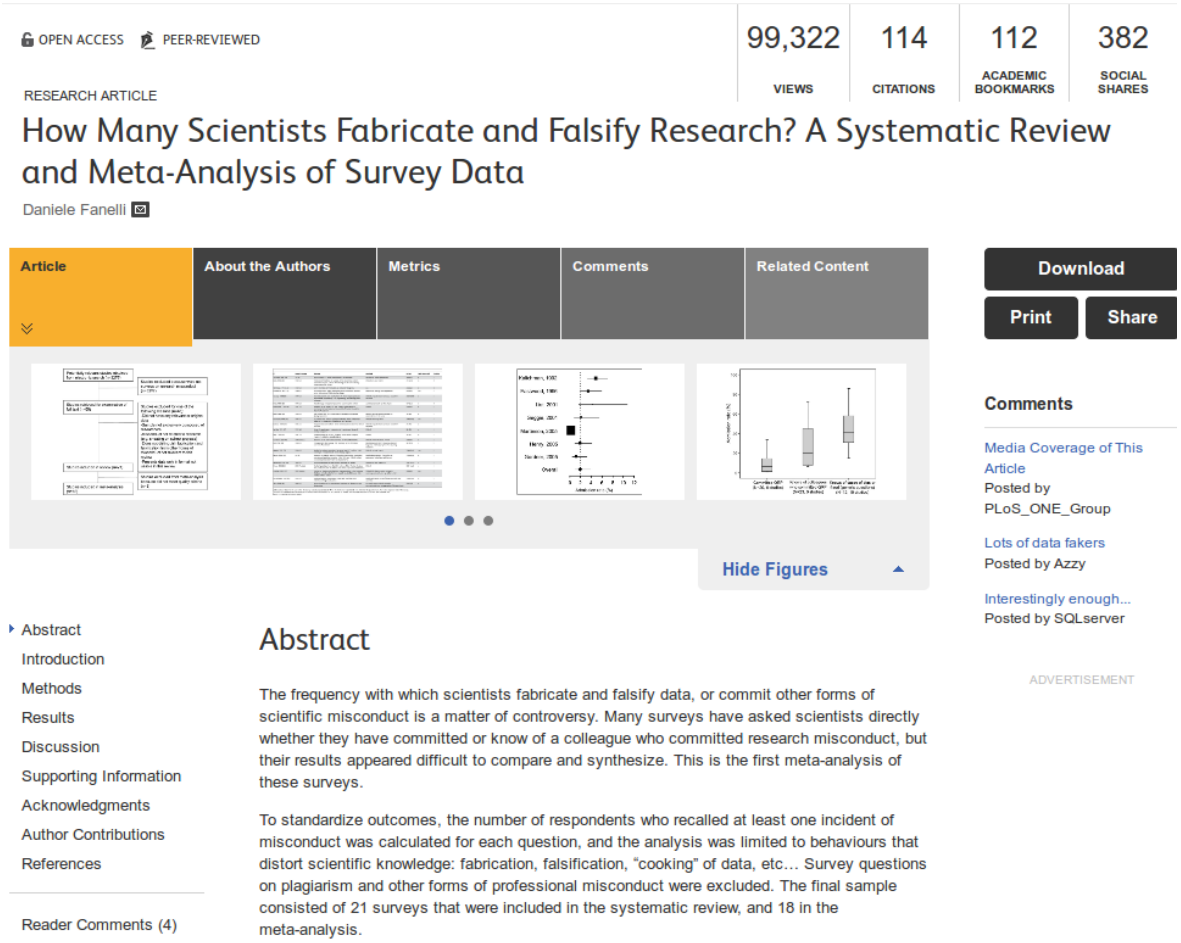
- Giving credit to peers and acknowledge their work

Benefits of Citation

- Reproducibility
- Transparency
- Documentation
- Context
- Identification
- Impact
- Reuse

© Richard Hutten

Citations Help Detecting Scientific Misconduct



OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

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

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

Daniele Fanelli

Article | About the Authors | Metrics | Comments | Related Content

Download | Print | Share

Hide Figures

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)

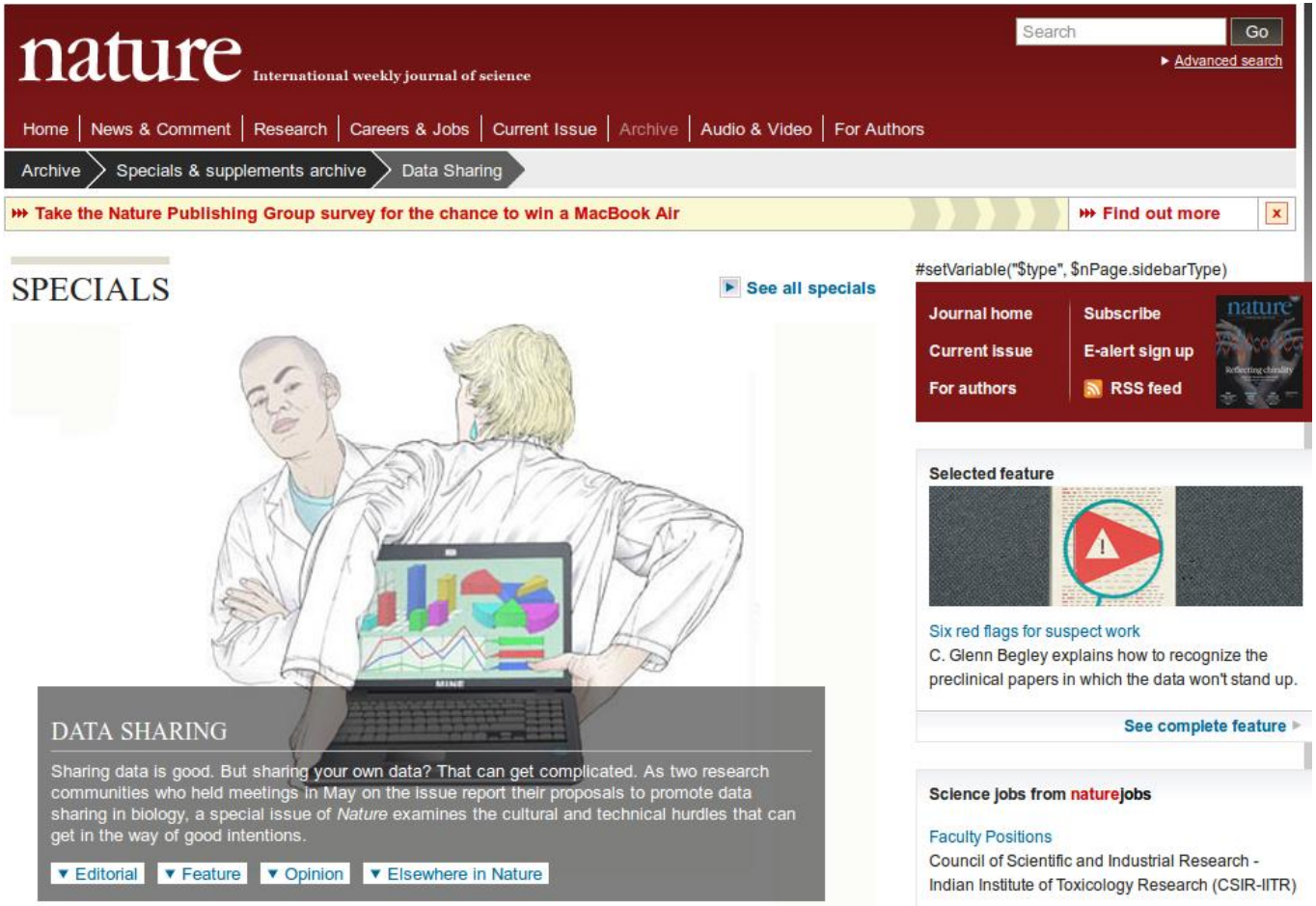
Abstract

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



Paper Publications

- Referencing research papers is well established

ACM DIGITAL LIBRARY

SIGN IN SIGN UP

SEARCH

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

Published in:

communications acm

· 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

1978 Article

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

Tools and Resources

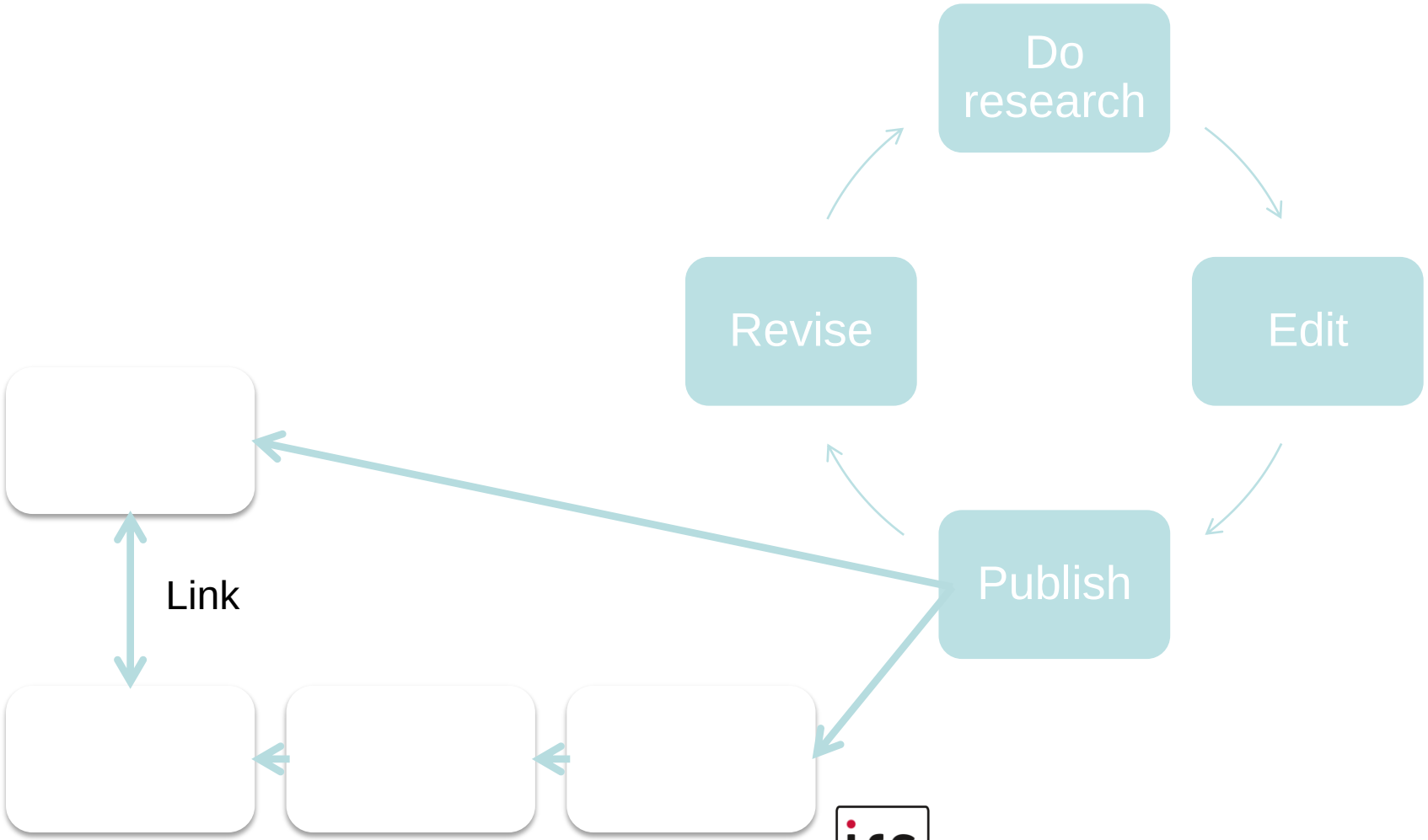
Buy this Article
Request Permissions
TOC Service:
Email RSS
Save to Binder
Export Formats:
BibTeX EndNote ACM Ref
Share:

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

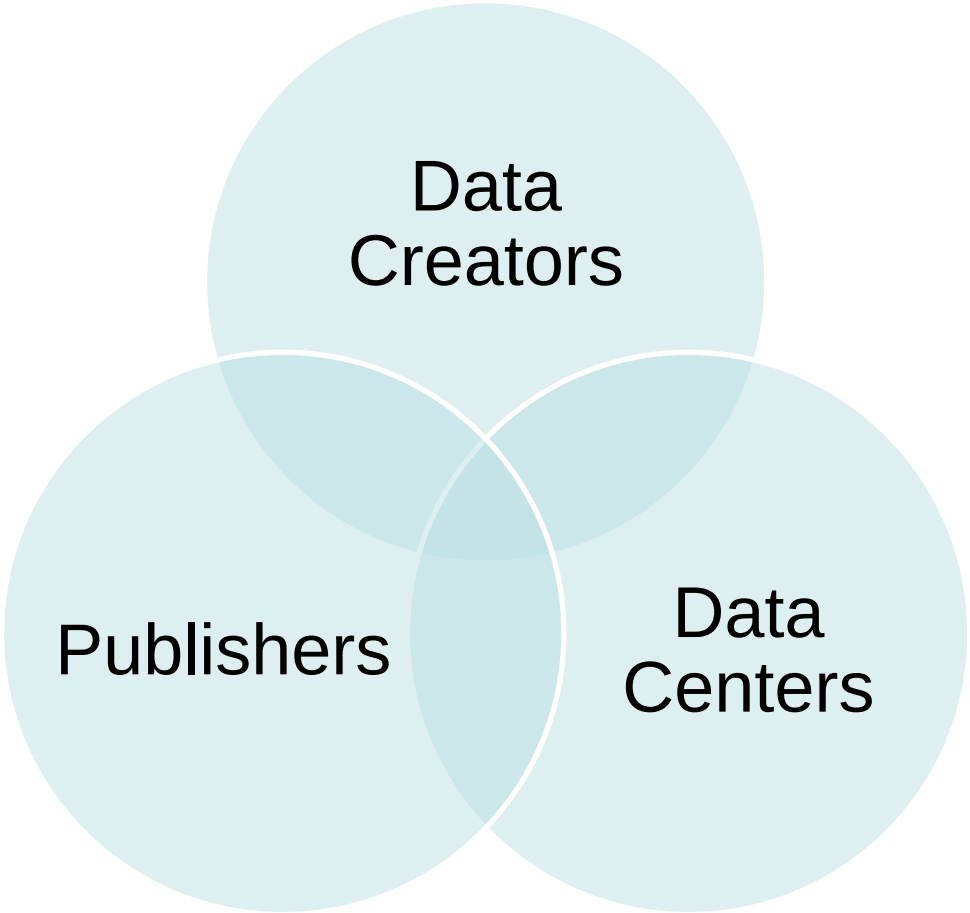
Data Publications

- 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?

Digital Object Life Cycle



Stakeholders



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
- Accepted by researchers
- Citation metrics

Joint Declaration of Data Citation Principles

- 8 Principles created by the Data Citation Synthesis Group
- https://www.force11.org/datacitation
- The Data Citation Principles cover purpose, function and attributes of citations.
- Goal: Encourage communities to develop practices and tools that embody uniform data citation principles.

Joint Declaration of Data Citation Principles (cont'd)

1) **Importance**

   Data should be considered legitimate, citable products of research. Data citations should be accorded the same importance as publications.

2) **Credit and Attribution**

   Data citations should facilitate giving credit and normative and legal attribution to all contributors to the data.

Joint Declaration of Data Citation Principles (cont‘d)

3) **Evidence**

   Whenever and wherever a claim relies upon data, the corresponding data should be cited.

4) **Unique Identification**

   A data citation should include a persistent method for identification that is machine actionable, globally unique, and widely used by a community

Joint Declaration of Data Citation Principles (cont'd)

5) **Access**

Data citations should facilitate access to the data themselves and to such associated metadata, documentation, code, and other materials, as are necessary for both humans and machines to make informed use of the referenced data.

6) **Persistence**

Unique identifiers, and metadata describing the data, and its disposition, should persist - even beyond the lifespan of the data they describe.

Joint Declaration of Data Citation Principles (cont'd)

7) **Specificity and Verifiability**

Data citations should facilitate identification of, access to, and verfication of the specific data that support a claim. Citations or citation metadata should include information about provenance and fixity sufficient to facilitate verfiying that the specific timeslice, version and/or granular portion of data retrieved subsequently is the same as was originally cited

Joint Declaration of Data Citation Principles (cont'd)

8) Interoperability and flexibility

Data citation methods should be sufficiently flexible to accommodate the variant practices among communities, but should not differ so much that they compromise interoperability of data citation practices across communities

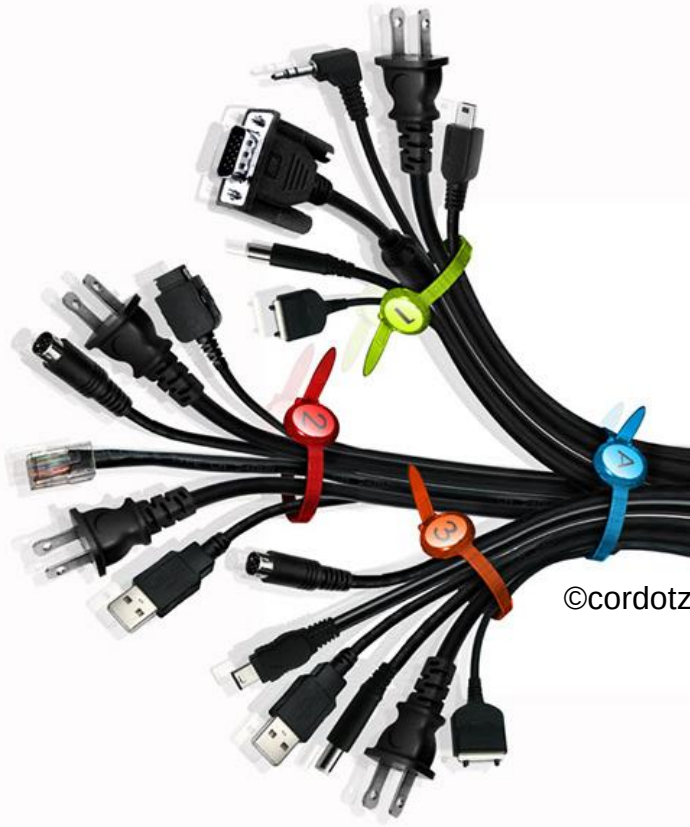
Elements of Data Citation

- 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

Identifiers

- Identifier is a symbol that uniquely identifies an object.
  - Used to identify (digital) objects
  - References the location
  - Provides metadata
  - Can be resolved
  - Several identifier types exist



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

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

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

URLs and Persistency?

- Standard URLs are not forever
  - Describe network locations
  - Not suitable for the long term
  - Link rot: half of the links in publications are not available after 5 to 7 years

Ruslan Eliseev @ 35photo.ru

- Solution: persistent identifiers (PIDs)

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

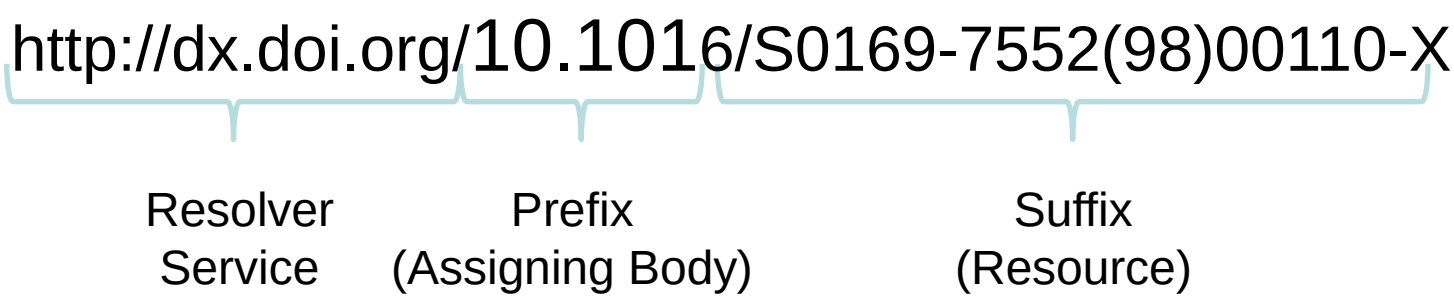
URN

- Main functions of a URN
  - Global scope of names
  - Global uniqueness
  - Persistence
  - Scalability
  - Legacy support
  - Extensibility
  - Independence
  - Resolution

Digital Object Identifier (DOI)

- Digital Identifier of an Object
  - not "Identifier of a Digital Object"
- Identifier scheme administered by the International DOI Foundation (IDF)

- Consists of three parts:



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/

How to Get a DOI

1. Request an account at a DOI registration agency
2. Pay a fee
3. Receive login data and your prefix
4. Establish a DOI suffix
5. Start Citing

DOI: Guidelines

- Suffix must be unique within the prefix
- Suffix is case insensitive
- Use UTF-8 and XML entity encoding
- Use concise suffixes, people have to type them
- Do not use special characters if possible
- No semantics in the suffix

DOI – Facts

- Launched in 2000
- Over 5,000 naming authorities (assigners)
- Over 215,000 DOI name prefixes
- Over 84 million DOI names assigned
- Over one billion DOI resolutions per year
- International Standard: ISO 26324 (May 2012)

[5] http://www.doi.org/factsheets/DOIKeyFacts.html

DOI Registration Agencies

- Are members of the IDF and entitled to assign and maintain DOIs.
- Examples:
  - DataCite
  - CrossRef
  - Bowker
  - CAL
  - Nielsen BookData
  - TIB
  - OPOCE

Helping you to find, access, and reuse data

- Registration Agency for DOIs
- Non-profit membership organization established 2009
- Aims:
  - Establish easier access to research
  - Increase acceptance of research data
  - Support data archiving that will permit results to be verified and re-purposed for future study.

[6] http://www.datacite.org/

DataCite

- Open to all not for profit organisations
- Full members: unlimited DOIs
- Provides Meta Data Store
  - Stored metadata about research data
  - Provides metadata schema
  - Provides interface to manage metadata

Example Resolve a DOI

- DOI: 10.1594/PANGAEA.724325
- Resolver services
  - http://dx.doi.org/
  - Content negotiation: http://data.datacite.org/static/index.html
  - http://www.crosscite.org/cn/

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:

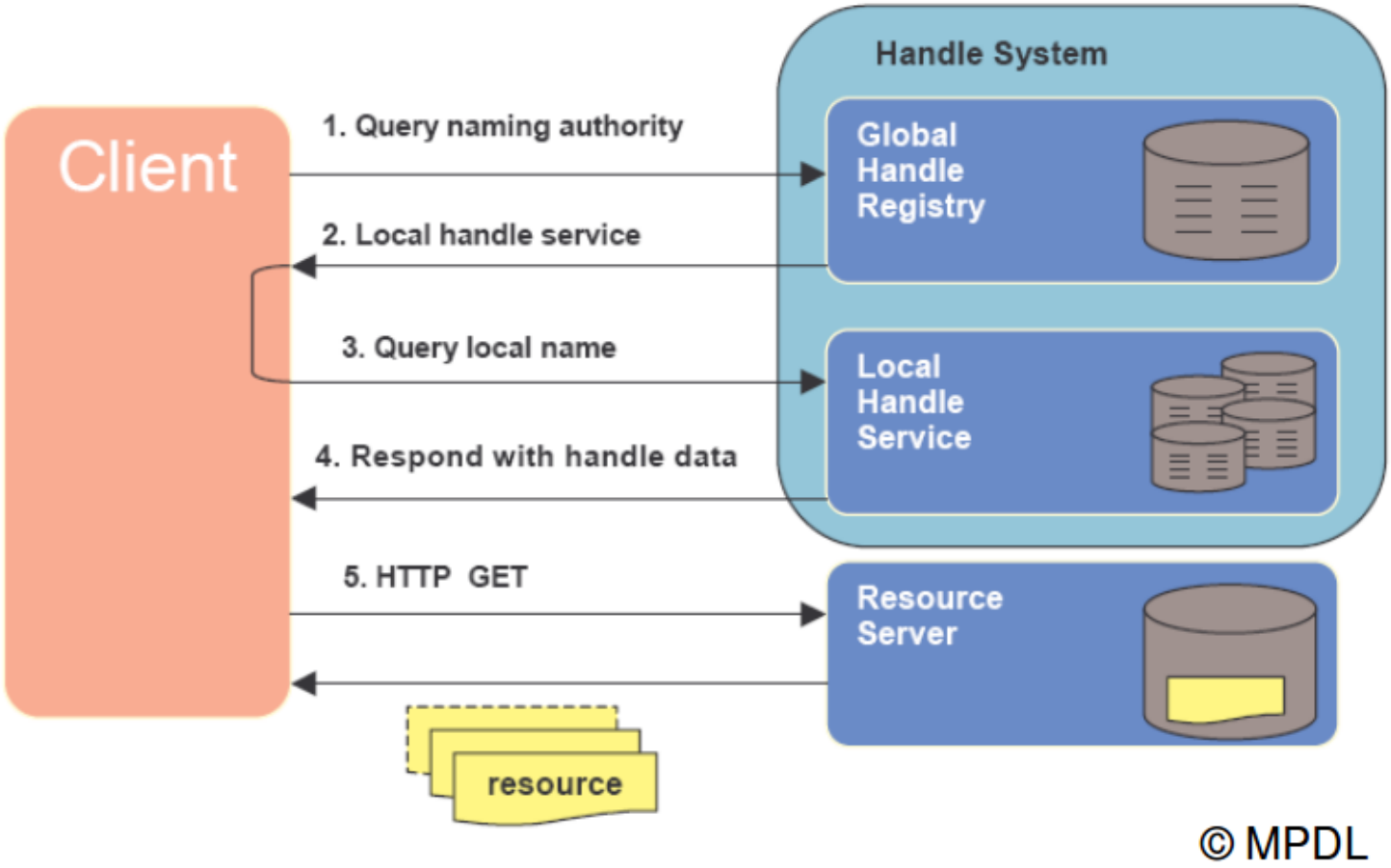
  <Handle Naming Authority> ‚/‘ <Handle Local Name>
- Example:
  - 10.1045/january2013-burns
- Available Services:
  - http://hdl.handle.net

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

[7] http://www.handle.net

Handle Resolution



[8] http://www.mpi.nl/DAM-LR/meeting5/Persistent%20Identifiers.pdf

DOI vs. Handle

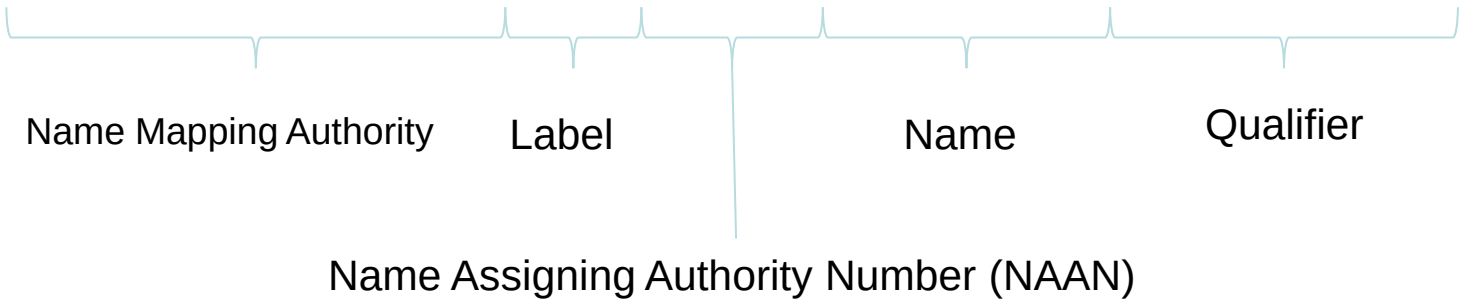
- Handle only provides the resolution service
- DOI uses the Handle System and adds:
  - Persistency of resolutions
  - Consistency of citations
  - Semantically interoperability (data model)
  - Identification of intellectual property entities.

[11] http://www.doi.org/factsheets/DOIHandle.html

Archival Resource Key (ARK)

- URLs with long-term support
- Maintained by California Digital Library
- Identify objects of any type (digital, physical, people, vocabulary terms, art...)
- Schema:

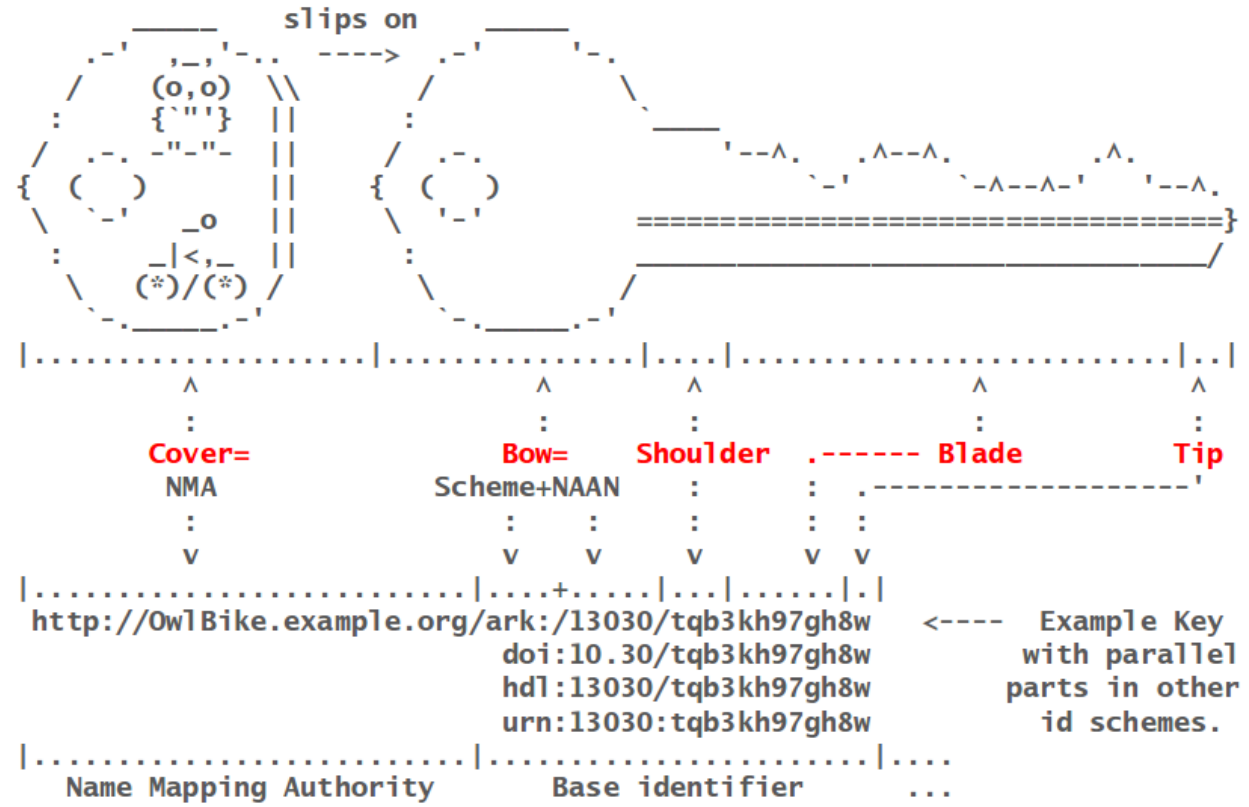
  http://example.org/ark:/13030/654xz321/s3/f8.05v.tiff



[9] https://wiki.ucop.edu/display/Curation/ARK

ARK - Scheme

Locksmith jargon: shoulder, blade, tip, bow, cover



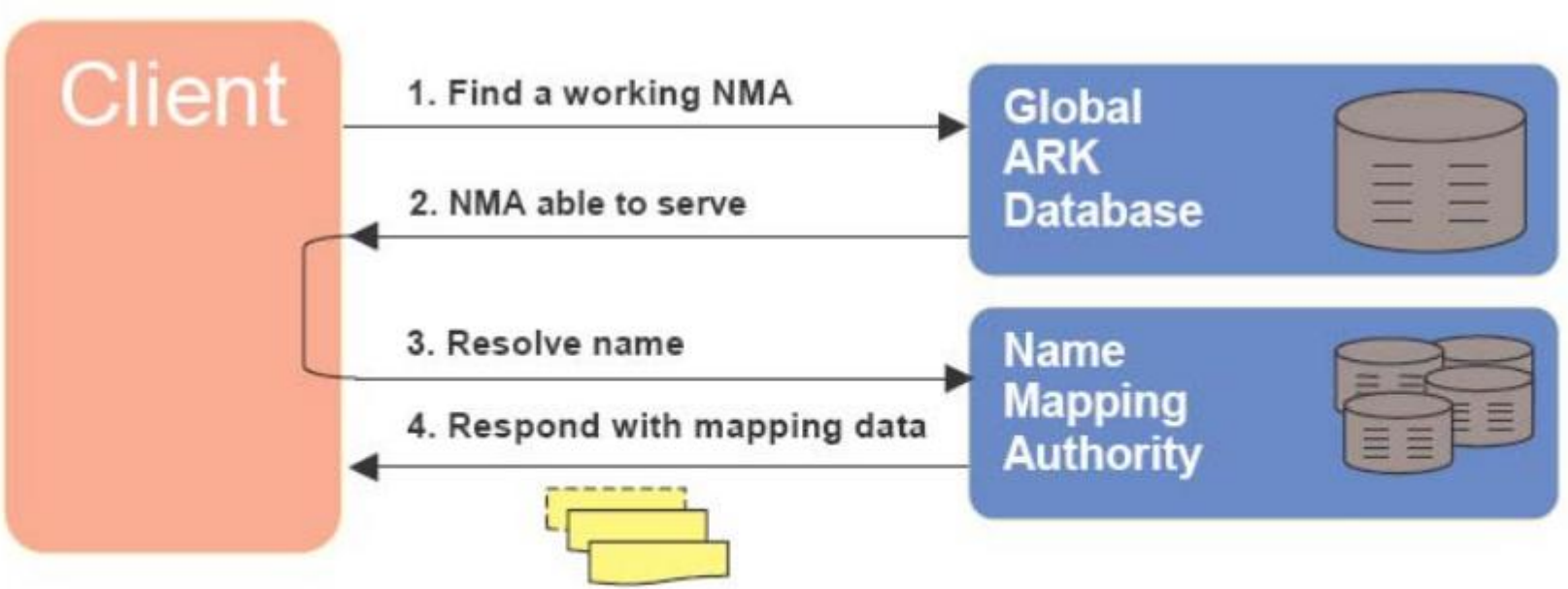
UC3

University of California CDL California Digital Library

ARK

- Currently there are 183 NAANs
  - Universities
  - Libraries
  - Google
  - http://www.cdlib.org/services/uc3/naan_registry.txt
- Any institution can obtain a NAAN by contacting CDL
- ARK can be self hosted
- ARKs are free

ARK



© MPDL

[10] http://www.mpi.nl/DAM-LR/meeting5/Persistent%20Identifiers.pdf

ARK

- Integrated services: example
  - http://texashistory.unt.edu/ark:/67531/metapth123456
  - http://texashistory.unt.edu/ark:/67531/metapth123456/?
  - http://texashistory.unt.edu/ark:/67531/metapth123456/??

ARK vs. DOI

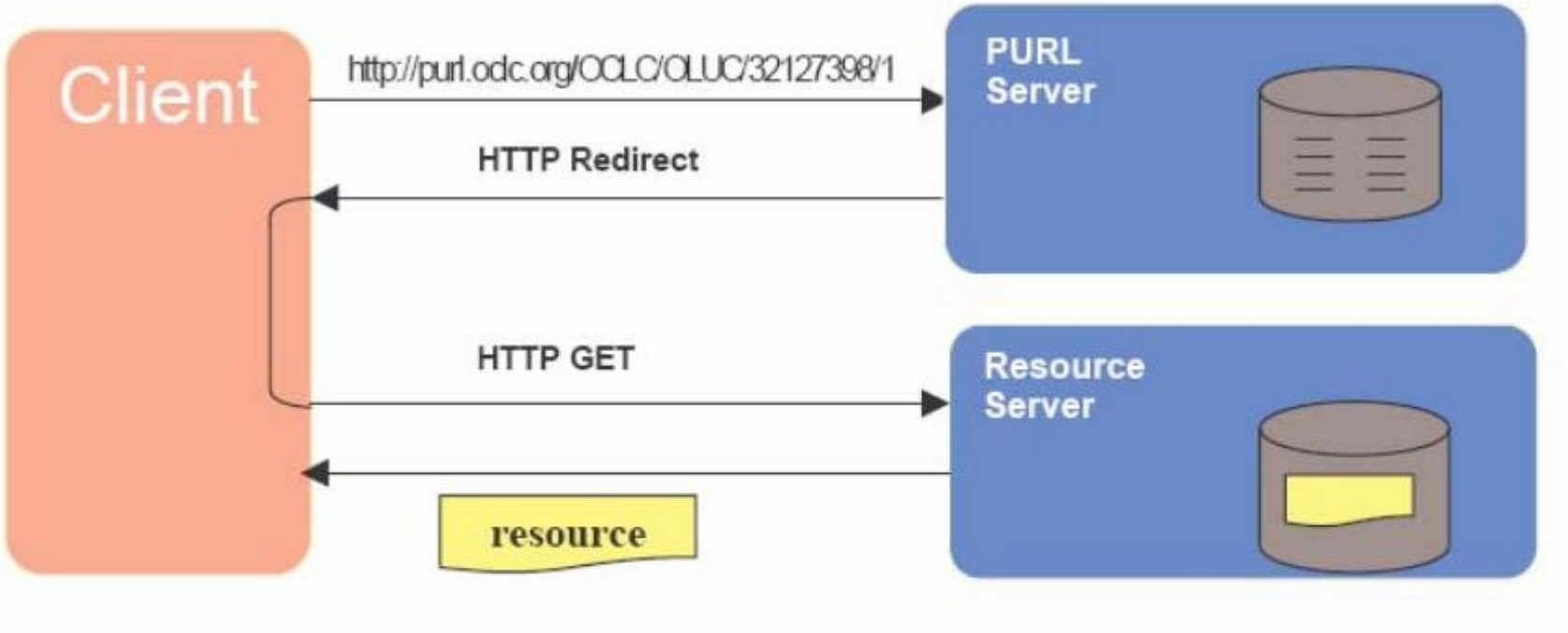
- ARK
  - Subset facilities
  - Can be deleted
  - Good for early stage of live cycle
  - Free
- DOI
  - Can not be deleted
  - Higher reputation
  - Commercial

[12] http://n2t.net/ezid/home/understanding

PURL

- Persistent uniform resource locator
- Based on HTTP forwarding
  - Only resolution
  - No metadata
- Provides URL curation and URL resolvers
- Can be hosted on own servers or centrally
- Is free

PURL - Resolution



© MPDL

http://www.mpi.nl/DAM-LR/meeting5/Persistent%20Identifiers.pdf

An Overview of PIDs

	DOI	ARK	PURL
Actionable	✓	✓	✗
Metadata included	✓ ✗	✓ ✓	✗ ✓
Self hosting	✓	✓	✗
Centralized	✗	✓	✗
Subsets	✓	✓	✓
Opacity			
Community Acceptance	✓ ✗	✓ ✓	✓ ✓
Free	✓	✗	✗
Commercial			

Other Persistent Identifiers

- ORCID
  - For researchers
- OpenURL
  - Includes metadata in URLs
  - Not opaque

Current Challenges

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

Dynamic Data Citation

- So far citable datasets have to be static
- Research data is dynamic
  - How to reference subsets in a dynamic environment?
  - How to create specific subsets?
- Granularity?
  - Assigning PIDs to every record does not scale

Example Music Workflow

- Ground truth and metadata stored in a database
- Training data can be changed
- Measure the effect on the classification results and compare versions
- So far:
  - Each version of the ground truth stored in a separate archive
  - Each archive referenced

Example Music Workflow

Idea: Reference the query, not the result set.

PID

Query

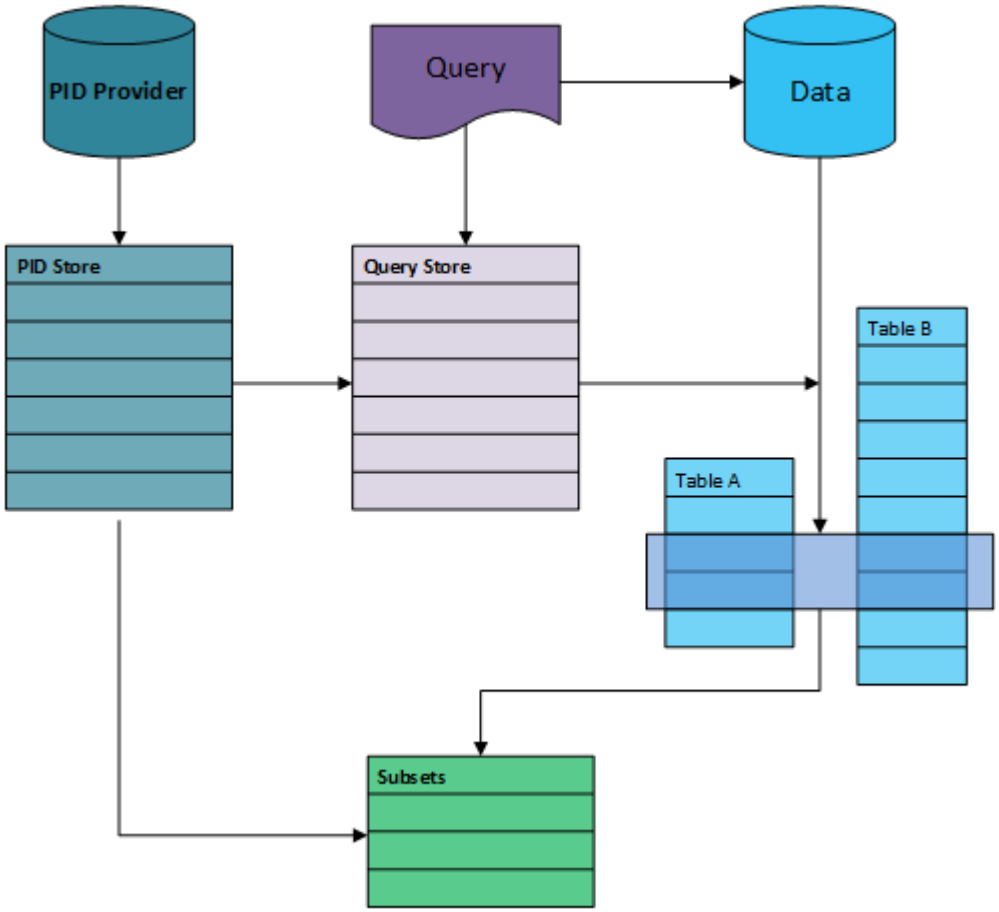
Music File	Genre	Version
Bach	Chamber music	3
Beatles	Poprock	3
Bach	Baroque music	2
Beatles	Pop	2
Bach	Classic	1
Beatles	Rock	1

Result

Dynamic Data Citation

- Example: Relational Database Systems
  - Produce subsets by using SQL
  - Instead of storing the complete result set, store only the query with timing information
  - Store all versions of the records with timestamps
  - Trace inserts, updates and deletes
  - Assign PID to query and timestamp their issuing
  - Re-execute query against the

Citing Dynamic Data in Databases



Dynamic Data Citation

- Steps for creating and citing dynamic data:

1. Record timestamps for all data operations
2. Maintain a history
3. Store queries
4. Create hash of the resultset
5. Assign PID to query
6. Re-execute adapted query
7. Verify correctness

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

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

Literature and Links

- http://www.dlib.org/dlib/march07/altman/03altman.html
- http://www.dcc.ac.uk/resources/how-guides/cite-datasets
- http://www.dlib.org/dlib/january11/starr/01starr.html
- http://dx.doi.org/10.1109%2F2.901164
- http://www.doi.org/factsheets/DOIKeyFacts.html
- http://www.datacite.org
- http://www.handle.net
- http://www.mpi.nl/DAM-LR/meeting5/Persistent%20Identifiers.pdf
- https://wiki.ucop.edu/display/Curation/ARK
- http://www.doi.org/factsheets/DOIHandle.html
- http://n2t.net/ezid/home/understanding
- http://sagecite.knowledgeblog.org/2011/07/28/why-do-we-need-datacitation

Further Pointers

- http://www.ariadne.ac.uk/issue56/tonkin
- http://ands.org.au/guides/persistent-identifiers-working.html
- http://hdl.handle.net/
- http://dx.doi.org/
- http://www.dcc.ac.uk/resources/how-guides/appraise-select-data

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