Enabling Precise Identification and Citabilityof Dynamic Data

Andreas Rauber

Vienna University of Technology
Favoritenstr. 9-11/188
1040 Vienna, Austria
rauber@ifs.tuwien.ac.at
http://ww.ifs.tuwien.ac.at/~andi

FACULTY OF !NFORMATICS

Outline

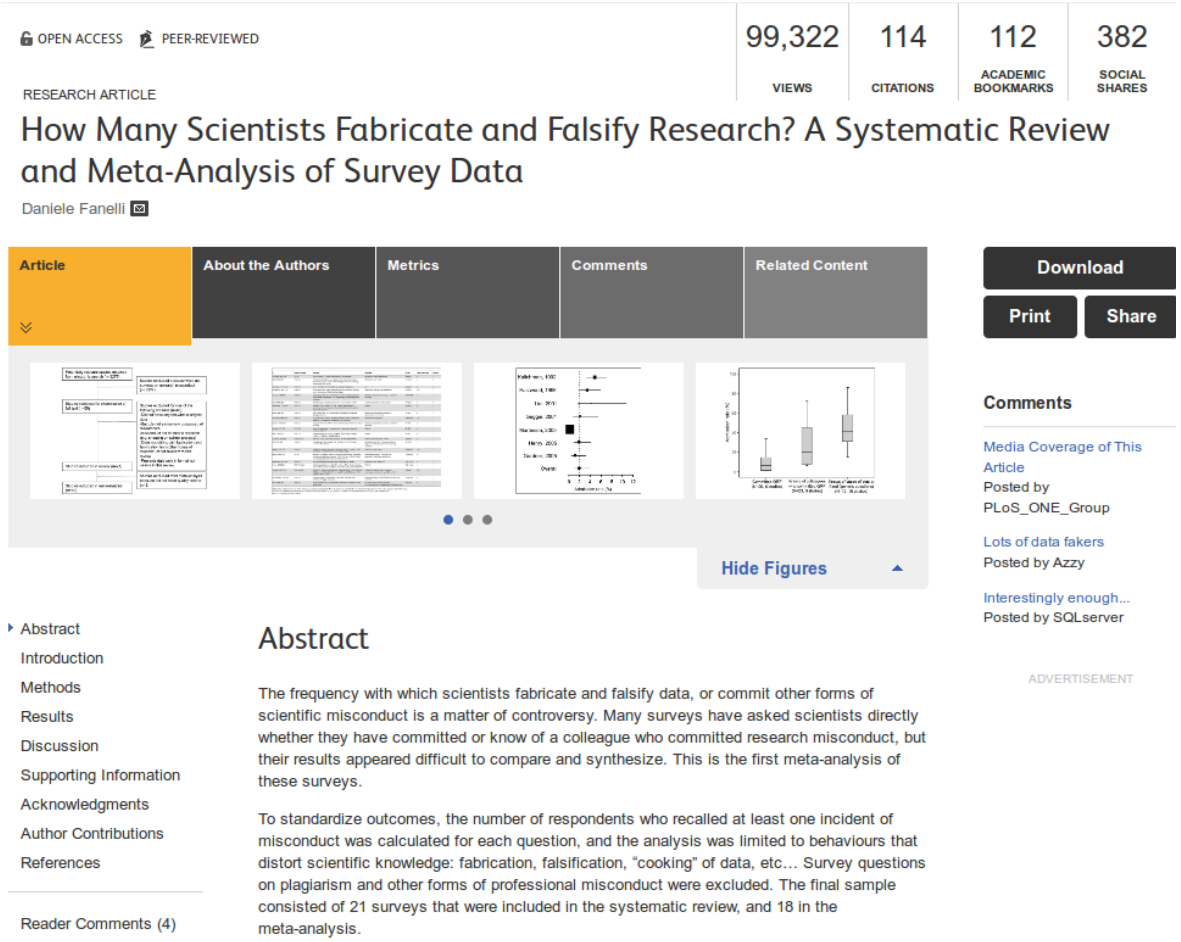
- Why should we want to cite data?
- What identifier system should I use?
- What are the challenges in data identification and citation?
- How should we do it, according to the RDA WG?
- Who is doing it so far, and how?
- Summary

Why to cite data?

- Data is the basis for almost everything
  - eScience, digital humanities,
  - Industry 4.0
  - Driving policies, society, …
- Why should we cite data?
  - Prevent scientific misconduct (“extrinsic”) ?

Prevent Scientific Misconduct

- On average: ~ 2% of scientists admitted they had "fabricated" (made up), "falsified" or "altered" data to "improve the outcome" at least once.
- 34% admitted to other questionable research practices including "failing to present data that contradict one's own previous research" and "dropping observations or data points from analyses based on a gut feeling that they were inaccurate."
- 14% knew someone who had fabricated, falsified or altered data.
- Up to 72% knew someone who had committed other questionable research practices.



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

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.

Comments

Media Coverage of This Article
Posted by PLoS_ONE_Group

Lots of data fakers
Posted by Azzy

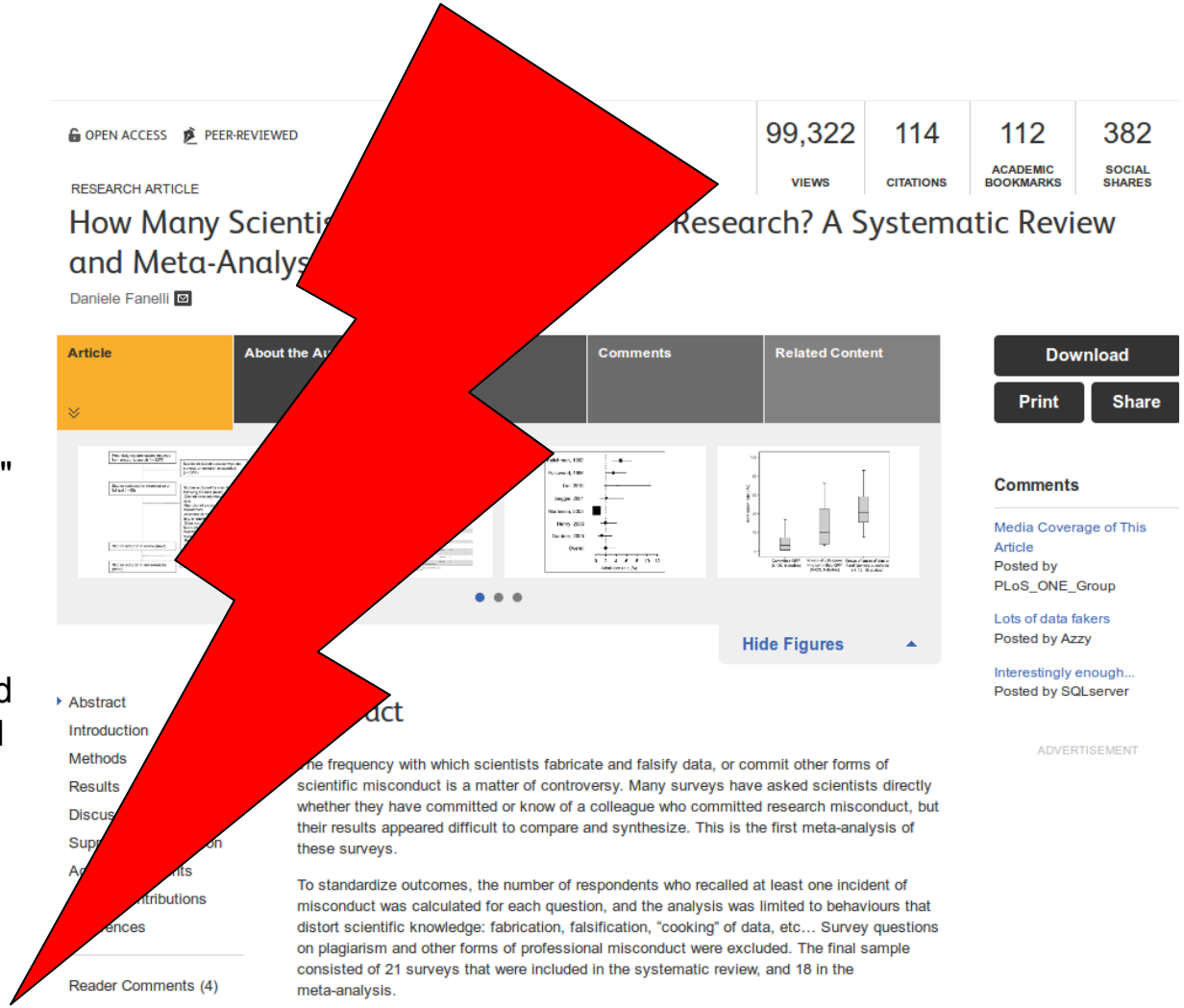
Interestingly enough...
Posted by SQLserver

ADVERTISEMENT

Source: http://www.plosone.org

Prevent Scientific Misconduct

- On average: ~ 2% of scientists admitted they had "fabricated" (made up), "falsified" or "altered" data to "improve the outcome" at least once.
- 34% admitted to other questionable research practices including "failing to present data that contradict one's own previous research" and "dropping observations or data points from analyses based on a gut feeling that they were inaccurate."
- 14% knew someone who had fabricated, falsified or altered data.
- Up to 72% knew someone who had committed other questionable research practices.



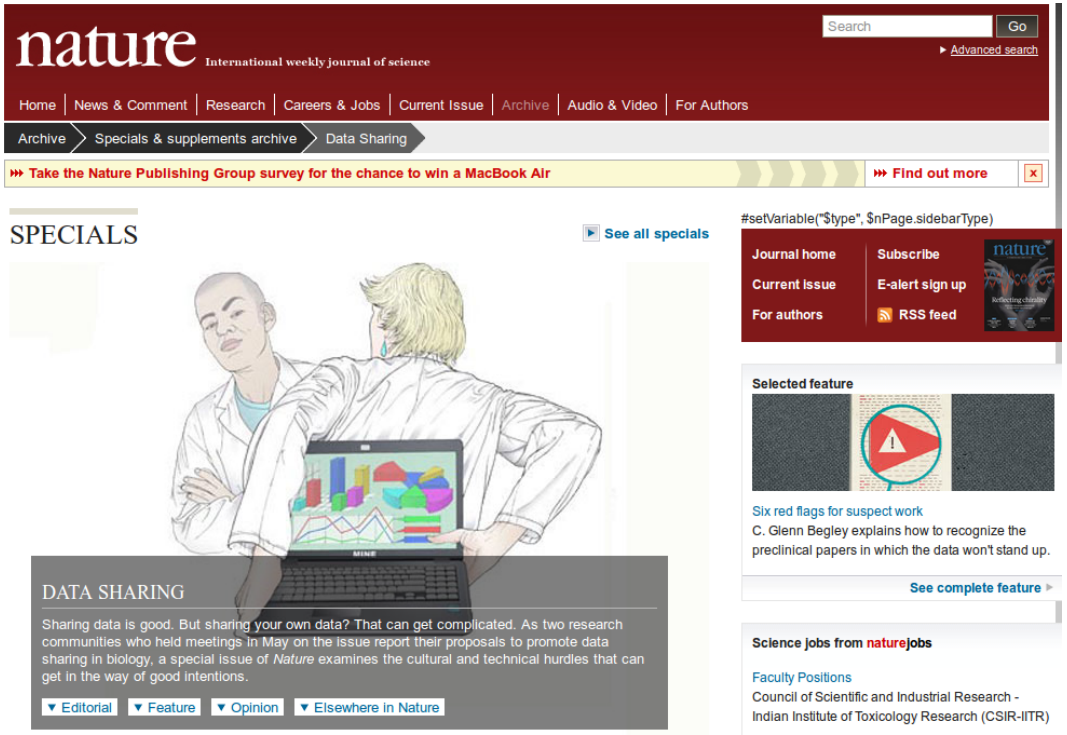
Source: http://www.plosone.org

Why to cite data?

- Data is the basis for almost everything
  - eScience, digital humanities,
  - Industry 4.0
  - Driving policies, society, …
- Why should we cite data?
  - Prevent scientific misconduct (“extrinsic”) ?
  - Give credit (“altruistic”) ?

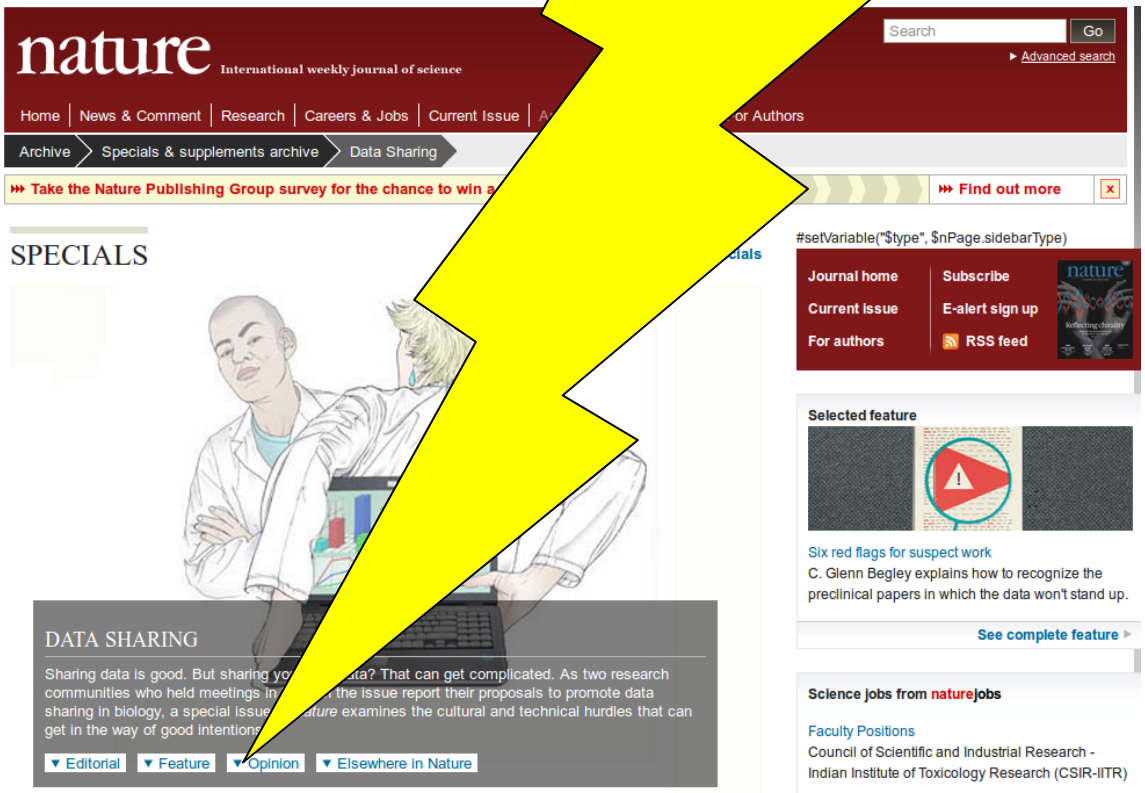
Giving credit

- Prime motivator for sharing data
- Shared data gets cited more frequently
- Citations are the currency of science



Giving credit

- Prime motivator for sharing da
- Shared data gets cited more
- Citations are the currency



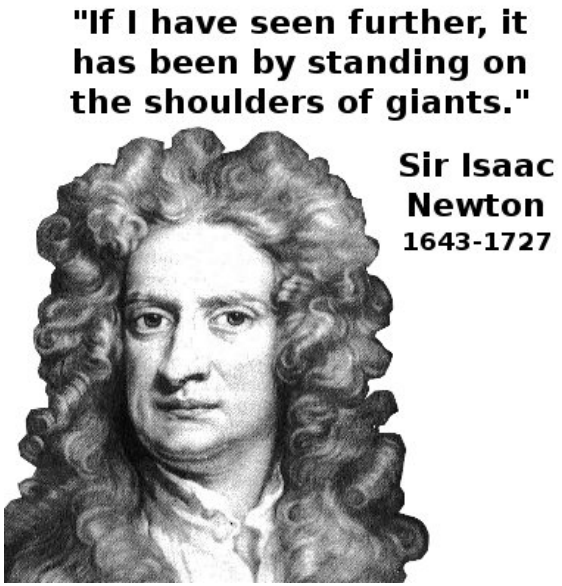
Why to cite data?

- Data is the basis for almost everything
  - eScience, digital humanities,
  - Industry 4.0
  - Driving policies, society, …
- Why should we cite data?
  - Prevent Scientific misconduct (“extrinsic”) ?
  - Give credit (“altruistic”) ?
  - Show solid basis (“egoistic”) ?

Citing to give credit

Why do we cite papers? (“related work”)

- Fundamental basis for own work – foundation!
- No need to prove - it’s been done!
- Speed-up the process, efficiency
- Basis for discourse, scientific work, …



Why to cite data?

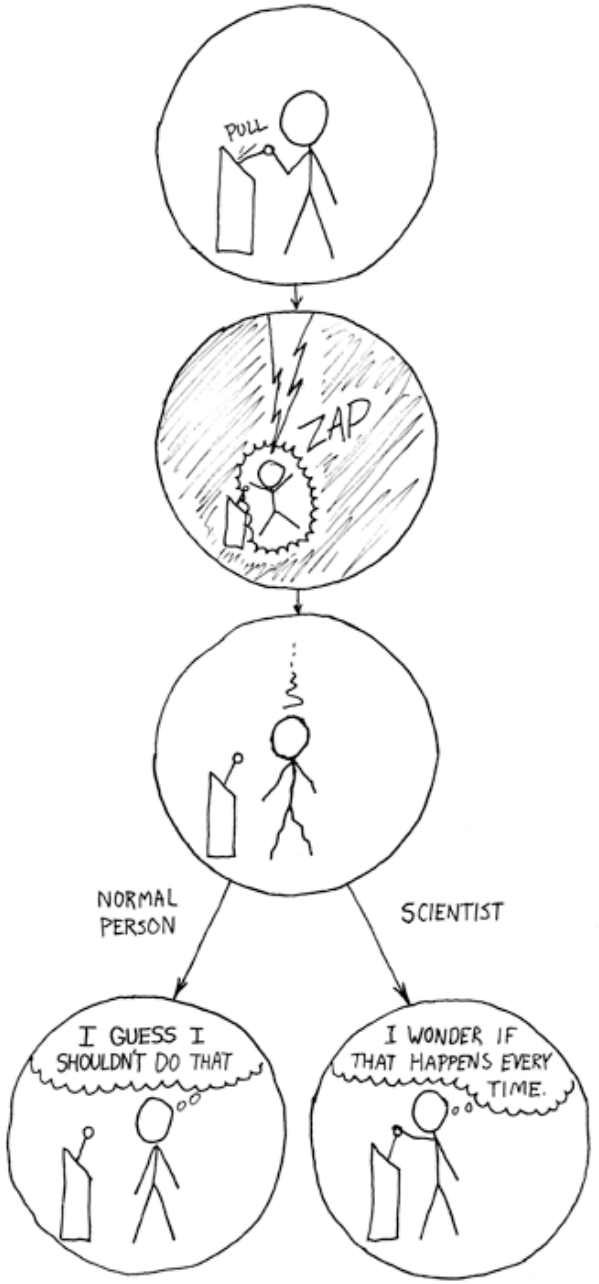
- Data is the basis for almost everything
  - eScience, digital humanities,
  - Industry 4.0
  - Driving policies, society, …

- Why should we cite data?
  - Prevent Scientific misconduct (“extrinsic”) ?
  - Give credit (“altruistic”) ?
  - Show solid basis (“egoistic”) ? ✔
  - Enable reproducibility, re-use (extrinsic + altruistic + egoistic) ?

Reproducibility

- Reproducibility is core to the scientific method
- Focus not on misconduct – but on complexity and the will to produce good work

- Should be easy
  - Get the code, compile, run, …
  - Why is it difficult?



https://xkcd.com/242/

Challenges in Reproducibility

http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0038234

PLOS ONE | Articles | For Authors | About Us | Search | advanced search

OPEN ACCESS PEER-REVIEWED

68,919 VIEWS | 10 CITATIONS | 124 SAVES

RESEARCH ARTICLE

The Effects of FreeSurfer Version, Workstation Type, and Macintosh Operating System Version on Anatomical Volume and Cortical Thickness Measurements

Ed H. B. M. Gronenschild, Petra Habets, Heidi I. L. Jacobs, Ron Mengelers, Nico Rozendaal, Jim van Os, Machteld Marcelis

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

Show Figures

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

Reader Comments (5)

Figures

Abstract

FreeSurfer is a popular software package to measure cortical thickness and volume of neuroanatomical structures. However, little if any is known about measurement reliability across various data processing conditions. Using a set of 30 anatomical T1-weighted 3T MRI scans, we investigated the effects of data processing variables such as FreeSurfer version (v4.3.1, v4.5.0, and v5.0.0), workstation (Macintosh and Hewlett-Packard), and Macintosh operating system version (OSX 10.5 and OSX 10.6). Significant differences were revealed between FreeSurfer version v5.0.0 and the two earlier versions. These differences were on average 8.8±6.6% (range 1.3–64.0%) (volume) and 2.8±1.3% (1.1–7.7%) (cortical thickness). About a factor two smaller differences were detected between Macintosh and Hewlett-Packard workstations and between OSX 10.5 and OSX 10.6. The observed differences are similar in magnitude as effect sizes reported in accuracy evaluations and neurodegenerative studies.

The main conclusion is that in the context of an ongoing study, users are discouraged to update to a new major release of either FreeSurfer or operating system or to switch to a different type of workstation without repeating the analysis; results thus give a quantitative support to successive recommendations stated by FreeSurfer developers over the years. Moreover, in view of the large and significant cross-version differences, it is concluded that formal assessment of the accuracy of FreeSurfer is desirable.

Comments

In praise of prog
Posted by GedR

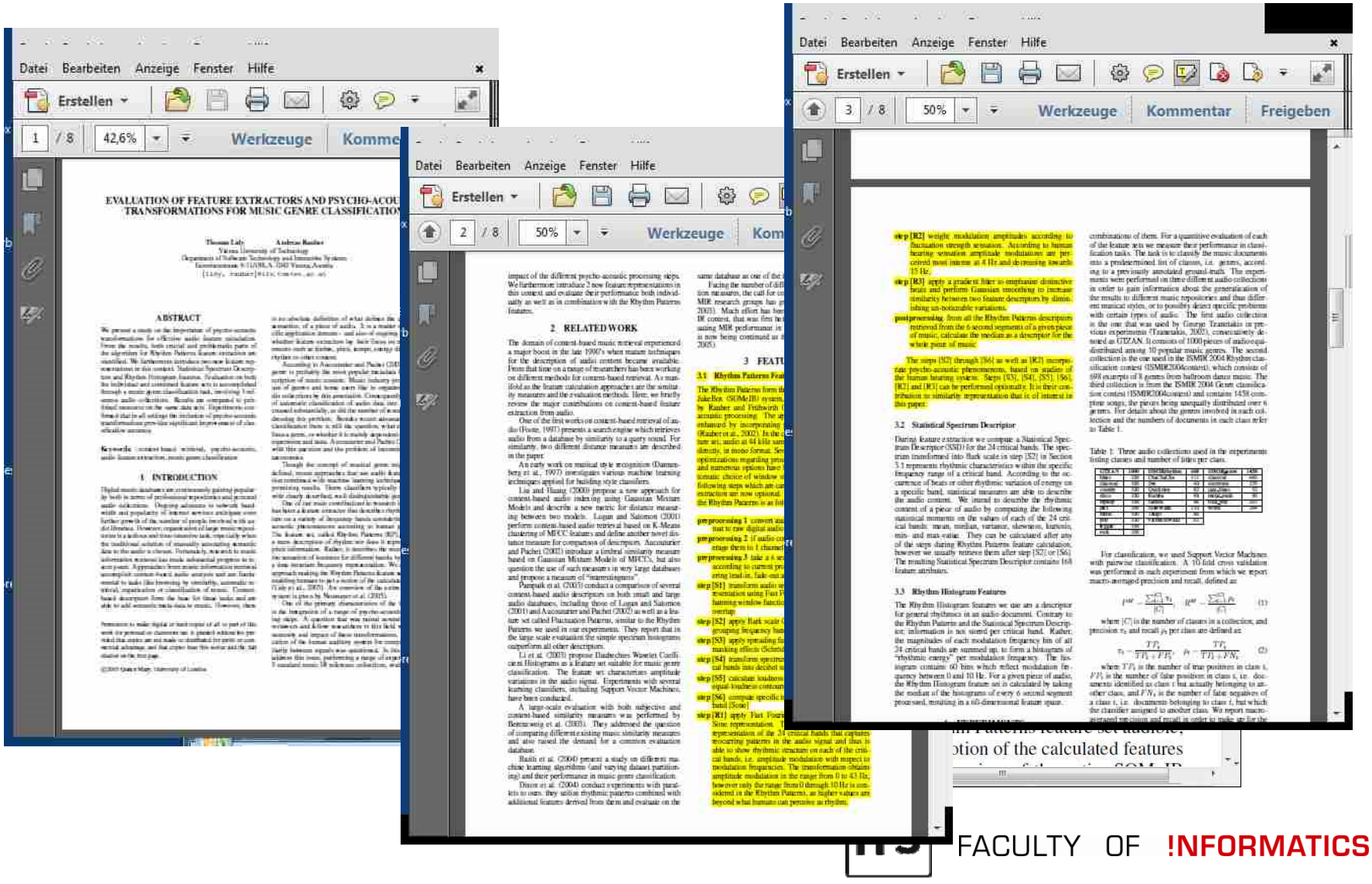
Media Coverage
Article
Posted by
PLoS_ONE_Gro

Comments mac
authors
Posted by EdGr

ifs FACULTY OF !INFORMATICS

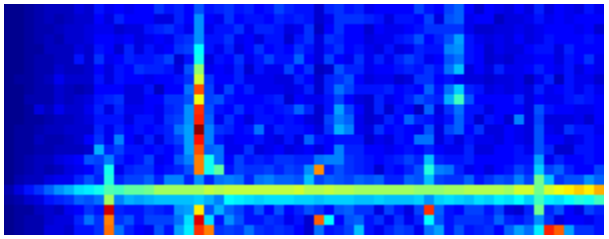
Challenges in Reproducibility

- Excursion: Scientific Processes

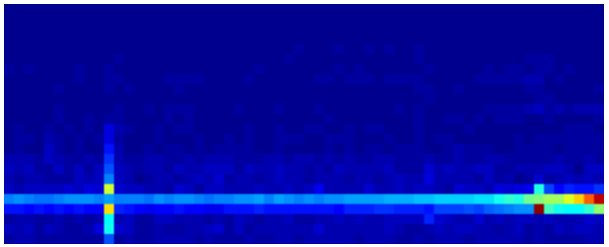


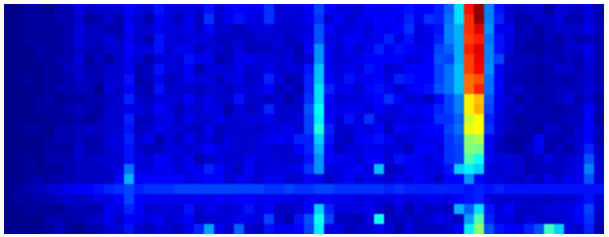
Challenges in Reproducibility

- Excursion: scientific processes

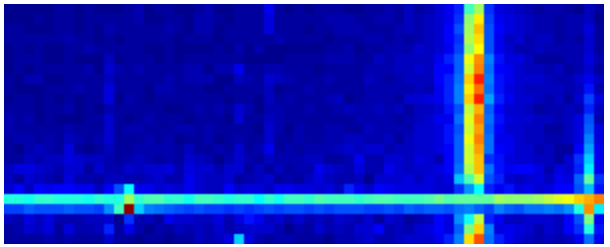


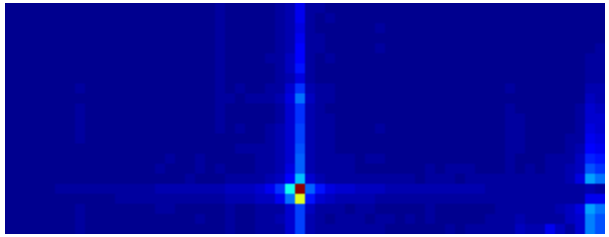
set1_freq440Hz_Am11.0Hz



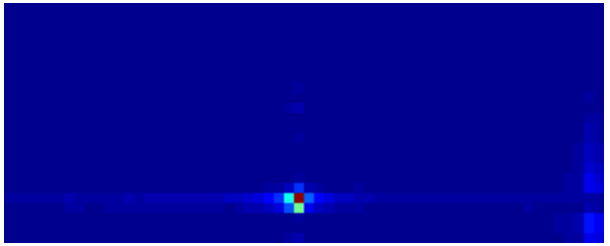


set1_freq440Hz_Am12.0Hz





set1_freq440Hz_Am05.5Hz

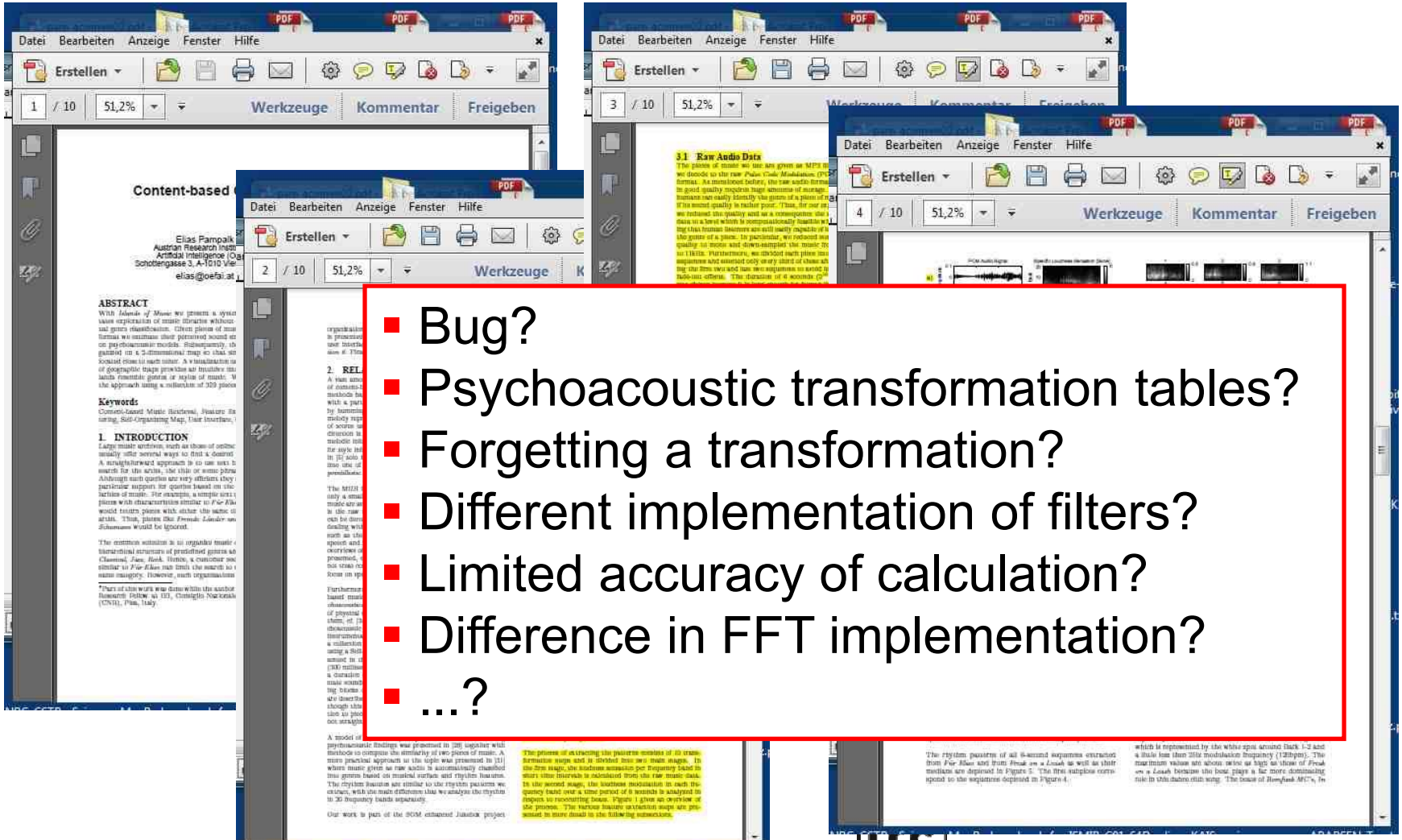


Java

Matlab

Challenges in Reproducibility

- Excursion: Scientific Processes

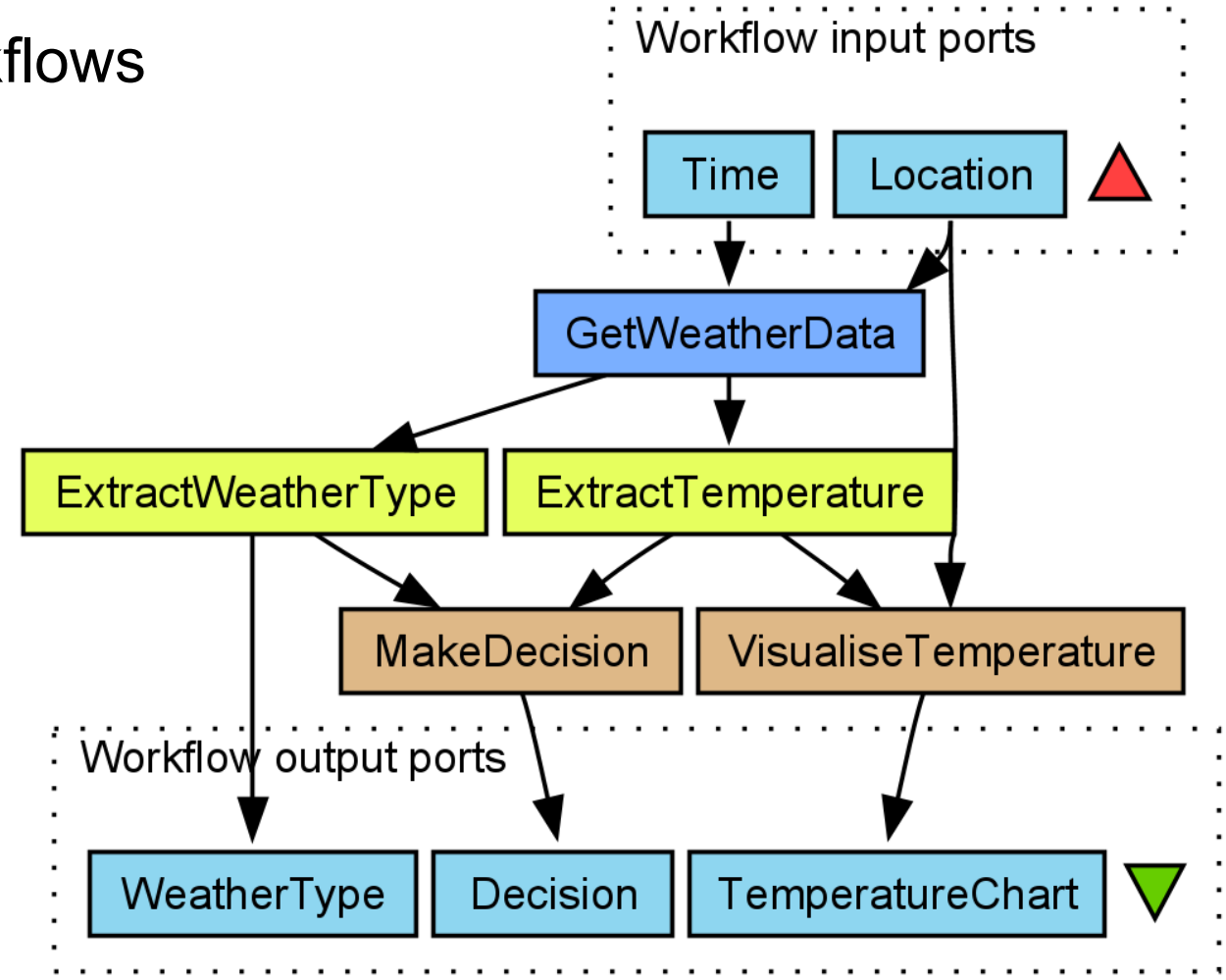


- Bug?
- Psychoacoustic transformation tables?
- Forgetting a transformation?
- Different implementation of filters?
- Limited accuracy of calculation?
- Difference in FFT implementation?
- ...?

Challenges in Reproducibility

- Workflows

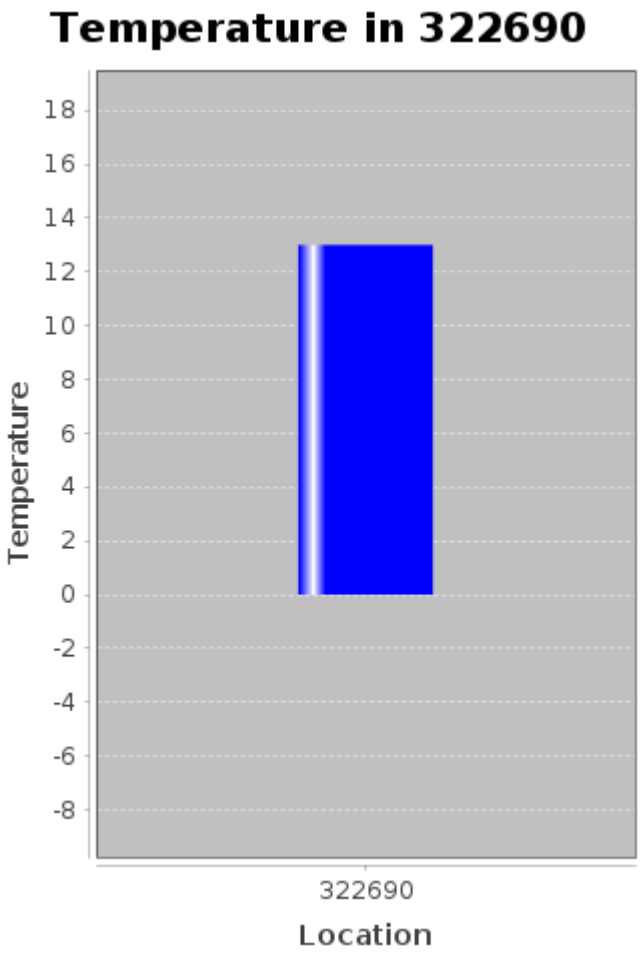
Taverna

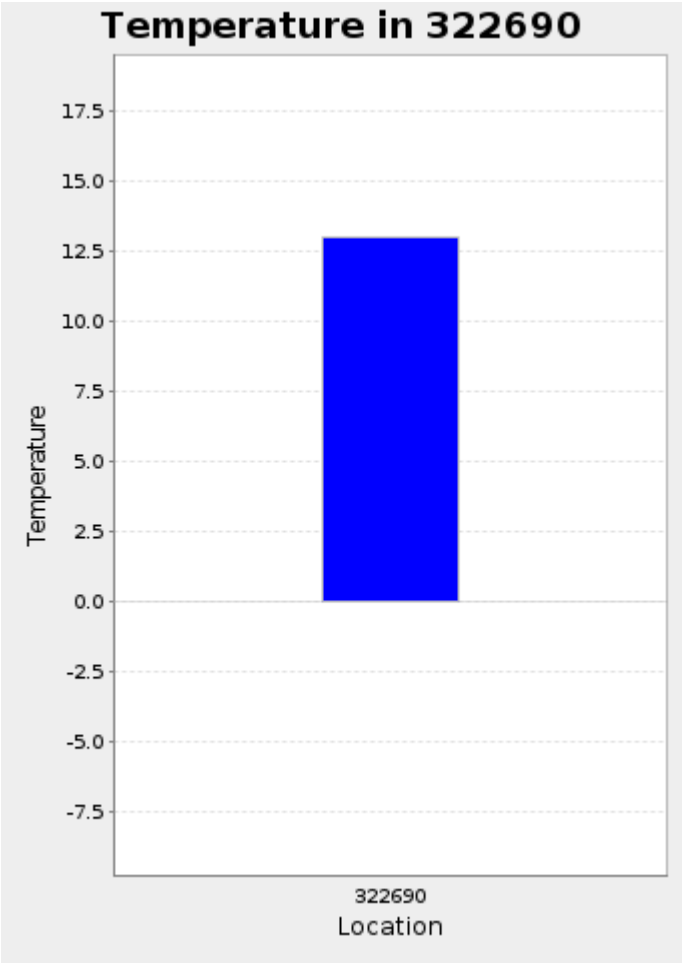


Activiti

Kepler

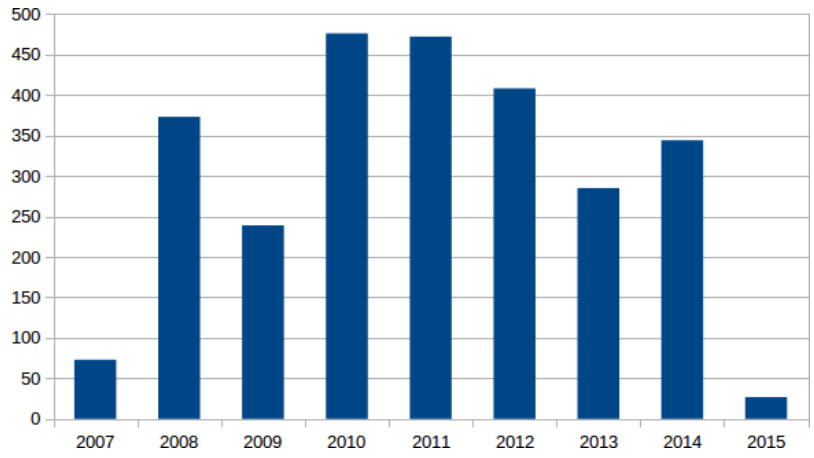
Challenges in Reproducibility





Challenges in Reproducibility

- Large scale quantitative analysis
- Obtain workflows from MyExperiments.org
  - March 2015: almost 2.700 WFs (approx. 300-400/year)
  - Focus on Taverna 2 WFs: 1.443 WFs
  - Published by authors ➔ should be „better quality“
- Try to re-execute the workflows
  - Record data on the reasons for failure along
- Analyse the most common reasons for failures



Workflow Engine	%
Taverna 2	54.7
Taverna 1	20.9
RapidMiner	10.0
Galaxy	2.0
Others	12.4

Challenges in Reproducibility

Re-Execution results

- Majority of workflows fails
- Only 23.6 % are successfully executed
  - No analysis yet on correctness of results…

Processor	# WFs
Original Data Set	1443
- Missing input values	526
- Disabled processors (WSDL services)	180
- Not executable in test environment	6
Final Data Set	731

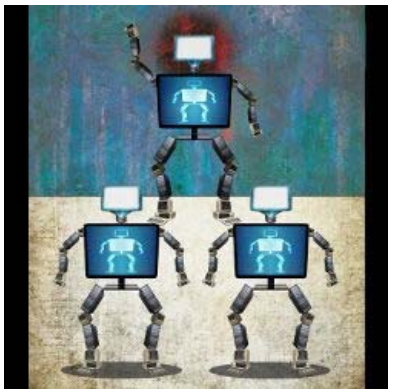
Processor	# WFs	% WFs
Not terminated >48hours	6	0.8
Execution failed	384	52.5
Execution successful	341	46.6

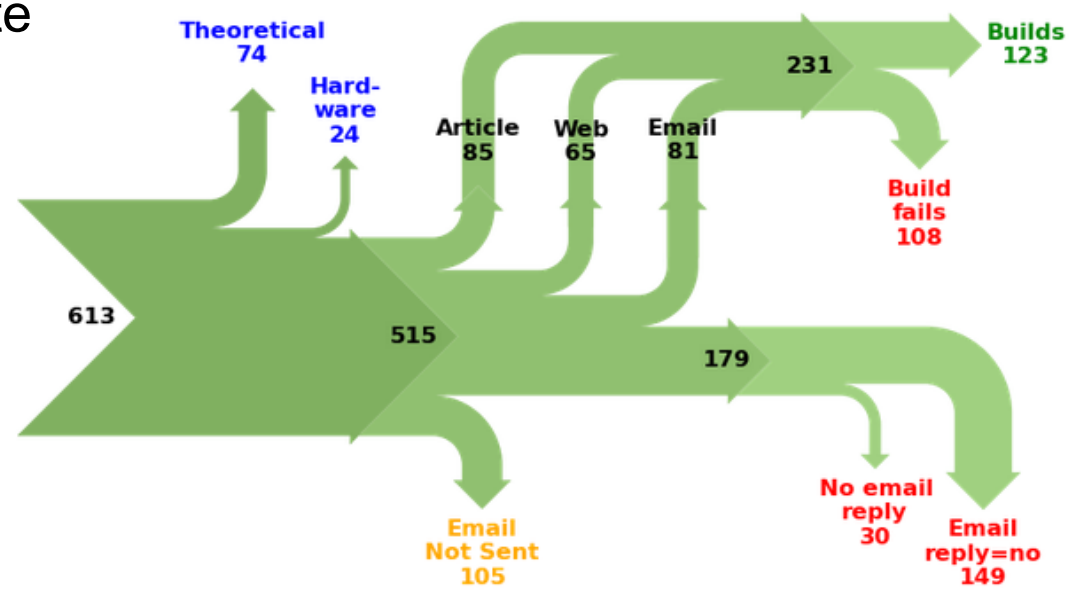
Processor	# WFs
REST unavailable	4
REST unauthenticated	5
Other unauthenticated	40
Missing Resources	14
Tool unavailable	19

Rudolf Mayer, Andreas Rauber, “A Quantitative Study on the Re-executability of Publicly Shared Scientific Workflows”, 11th IEEE Intl. Conference on e-Science, 2015.

Challenges in Reproducibility

- 613 papers in 8 ACM conferences
- Process
  - download paper and classify
  - search for a link to code (paper, web, email twice)
  - download code
  - build and execute

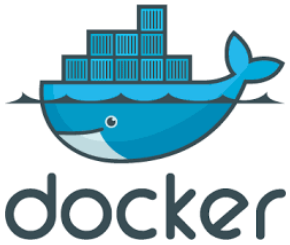




Christian Collberg and Todd Proebsting. “Repeatability in Computer Systems Research,” CACM 59(3):62-69.2016

Reproducibility – solved! (?)

- Provide source code, parameters, data, …
- Wrap it up in a container/virtual machine, …



LXC

…

- Why do we want reproducibility?
- Which levels or reproducibility are there?
- What do we gain by different levels of reproducibility?
- A simple “re-run” is usually not enough
  – otherwise, video would be sufficient….

Types of Reproducibility

- The **PRIMAD Model**[1]: which attributes can we "prime"?
  - Data
    - Parameters
    - Input data
  - Platform
  - Implementation
  - Method
  - Research Objective
  - Actors
- What do we gain by "priming" one or the other?

[1] Juliana Freire, Norbert Fuhr, and Andreas Rauber. Reproducibility of Data-Oriented Experiments in eScience. Dagstuhl Reports, 6(1), 2016.

Types of Reproducibility and Gains

Label	Data		Platform / Stack	Implementation	Method	Research Objective	Actor	Gain
	Parameters	Raw Data						
Repeat	-	-	-	-	-	-		Determinism
Param. Sweep	x	-	-	-	-	-		Robustness / Sensitivity
Generalize	(x)	x	-	-	-	-		Applicability across different settings
Port	-	-	x	-	-	-		Portability across platforms, flexibility
Re-code	-	-	(x)	x	-	-		Correctness of implementation, flexibility, adoption, efficiency
Validate	(x)	(x)	(x)	(x)	x	-		Correctness of hypothesis, validation via different approach
Re-use	-	-	-	-	-	x		Apply code in different settings, Re-purpose
Independent *x* *(orthogonal)*							x	Sufficiency of information, independent verification

Reproducibility Papers

- Aim for reproducibility: for one's own sake – and as Chairs of conference tracks, editor, reviewer, supervisor, …
  - Review of reproducibility of submitted work (material provided)
  - Encouraging reproducibility studies
  - (Messages to stakeholders in Dagstuhl Report)
- Consistency of results, not identity!
- Reproducibility studies and papers
  - Not just re-running code / a virtual machine
  - When is a reproducibility paper worth the effort / worth being published?

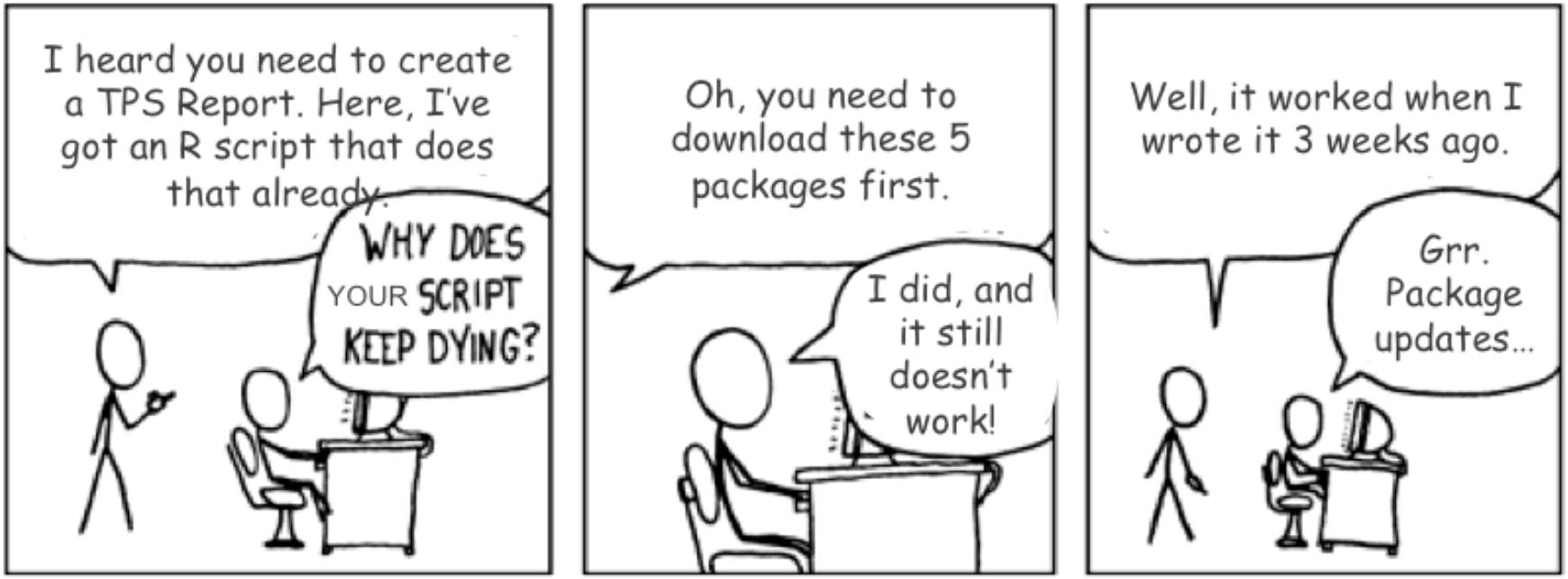
Challenges in Reproducibility

Peer Review and Verification

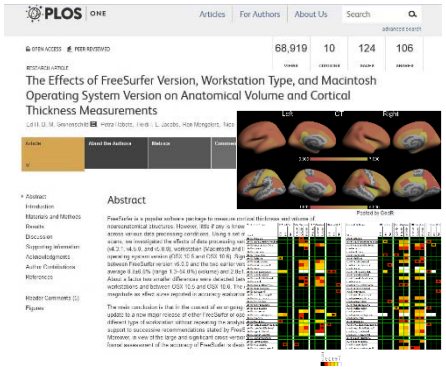
- Peer review is an established process
  - Focused on publications mainly
  - Hardly any data quality reviews
  - Even less reproducibility studies
- Reproducing or replicating experiments is not considered original research
  - No recognition
  - No money
  - A lot of work
- Encourage reproducibility studies
- Needed beyond science!

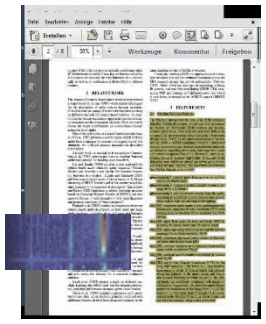
Challenges in Reproducibility

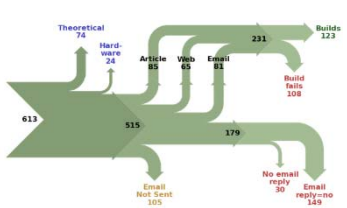
In a nutshell – and another aspect of reproducibility:

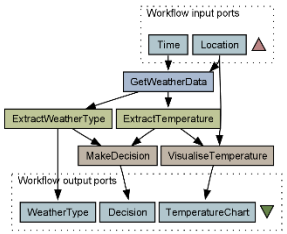


Source: xkcd









ifs

FACULTY OF !INFORMATICS

Data Citation and Reproducibility

Reproducibility

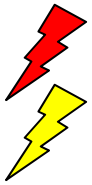
- requires
  - Transparency, requires
    - Documentation, provides
      - Context, requires
        - » **Citation**, requires
          - » **Identification**
- Increases impact
- Increases trust
- Fosters reuse

https://xkcd.com/978/

Why to cite data?

- Data is the basis for almost everything
  - eScience, digital humanities,
  - Industry 4.0
  - Driving policies, society, …

- Why should we cite data?
  - Prevent Scientific misconduct (“extrinsic”) ?
  - Give credit (“altruistic”) ?
  - Show solid basis (“egoistic”) ?
  - Enable reproducibility, re-use (extrinsic + altruistic + egoistic)



Why to cite data?

- Data is the basis for almost everything
  - eScience, digital humanities,
  - Industry 4.0
  - Driving policies, society, …

- Why should we cite data?
  - Prevent Scientific misconduct (“extrinsic”) ?
  - Give credit (“altruistic”) ?
  - Show solid basis (“egoistic”) ?
  - Enable reproducibility, re-use (extrinsic + altruistic + egoistic) ?
  - **Because it’s what you do if you do good work, speeding up the process of scientific discovery, efficiency! (“intrinsic”)**

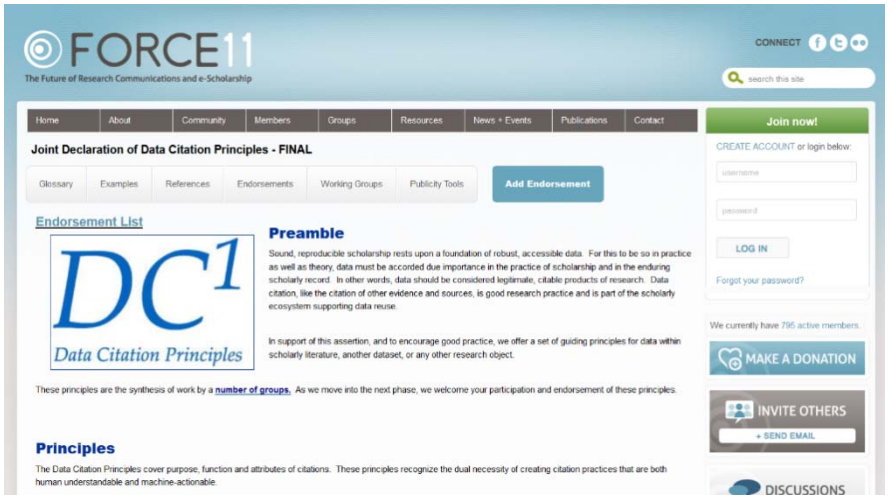
Why to cite data?

- It's what you do! – Lots of benefits
  - Makes live easier because you can build on a solid foundation
  - Speeds up the process because you can re-use existing stuff
  - Helps avoiding / detecting mistakes, improves quality, comparability
  - Reuse increases citations, visibility ("currency")

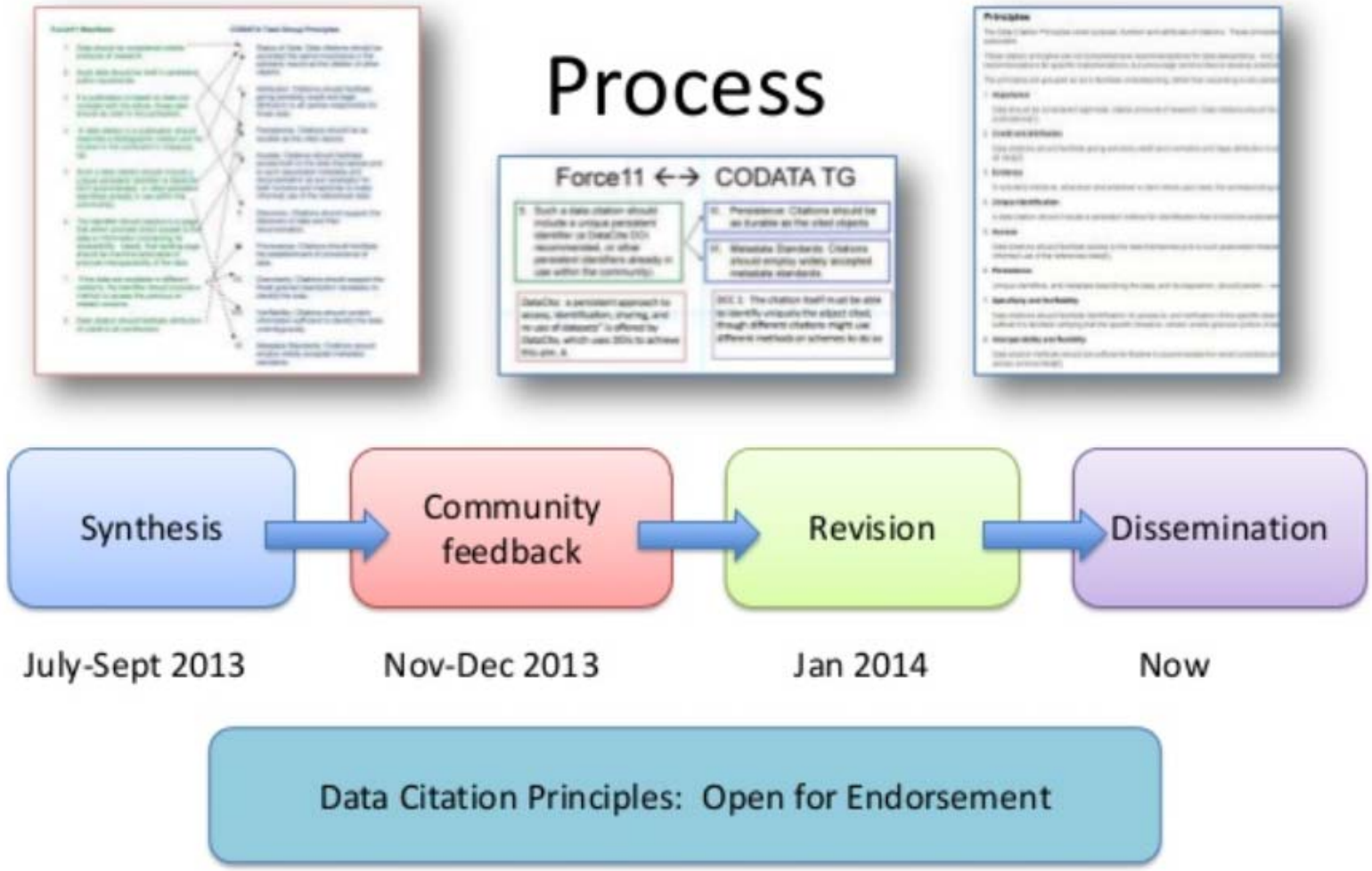
- But:
  - To achieve this it must be easy, straightforward, "automatic"
  - Citing Papers is easy…
  - …what about data?
    (more about this later… first: "we should just do it")

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



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)

Data Citation Principles

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 verification of the specific data that support a claim. Citations or citation metadata should include information about provenance and fixity sufficient to facilitate verifying that the specific time slice, version and/or granular portion of data retrieved subsequently is the same as was originally cited.

Joint Declaration of Data Citation Principles (cont'd)

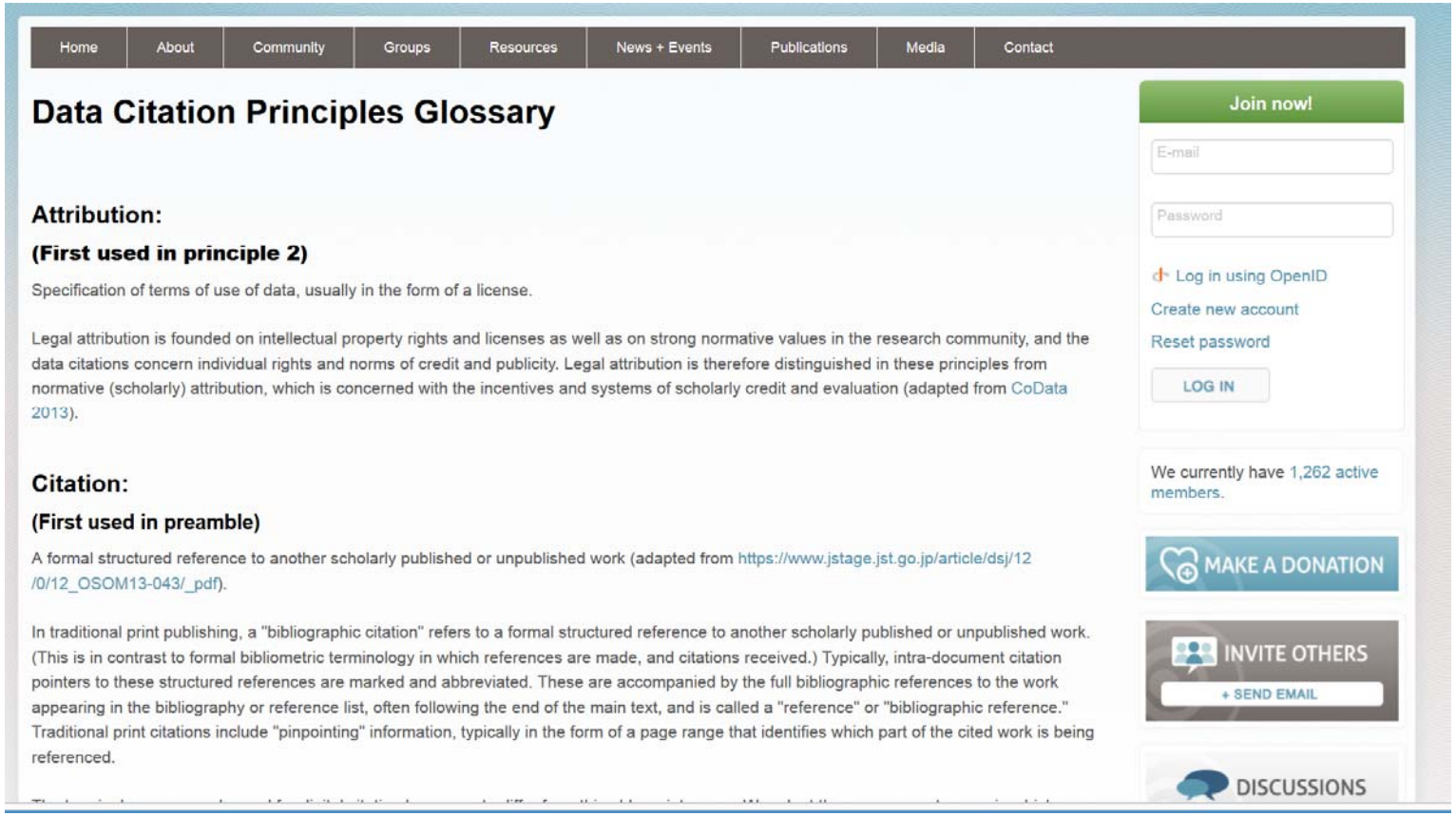
8) Interoperability and flexibility

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.

Joint Declaration of Data Citation Principles (cont'd)

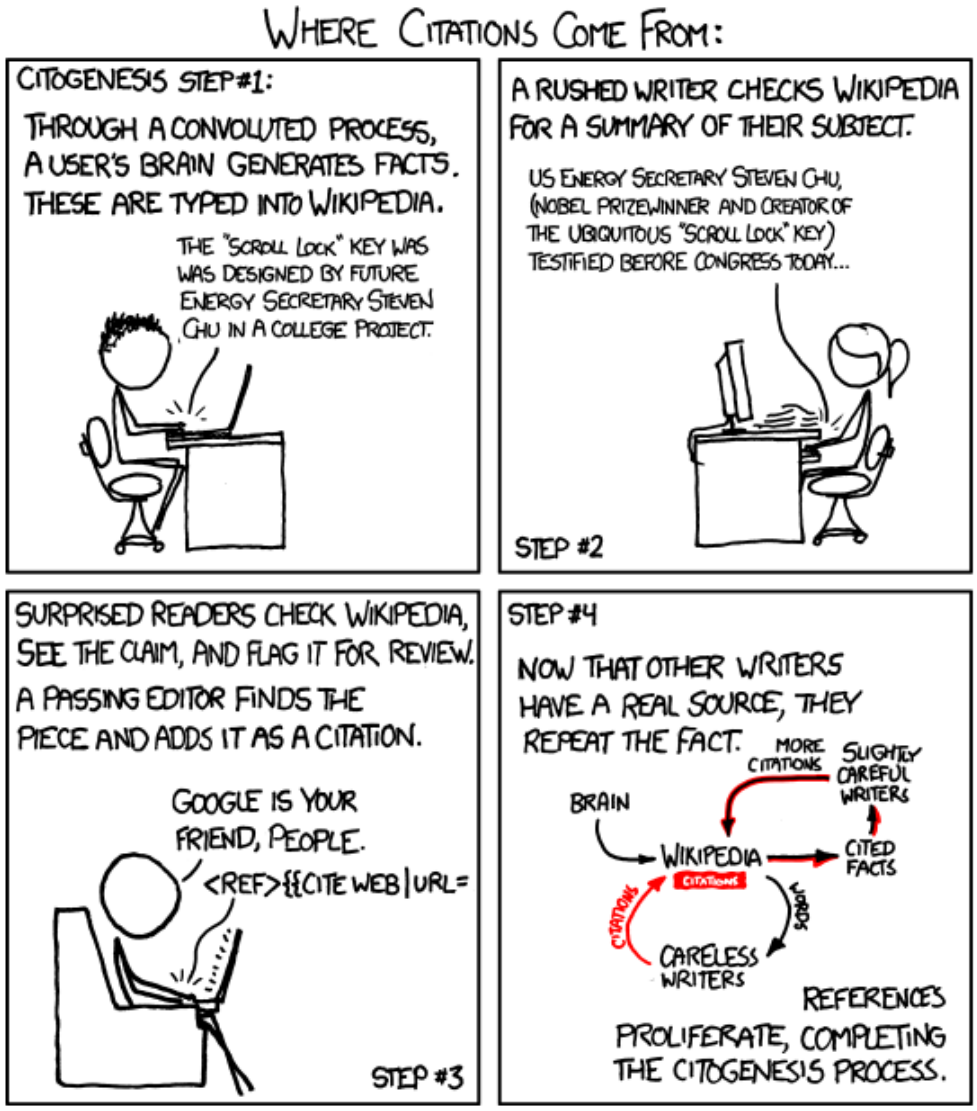
- **Glossary**

  https://www.force11.org/node/4770



Benefits of Citation

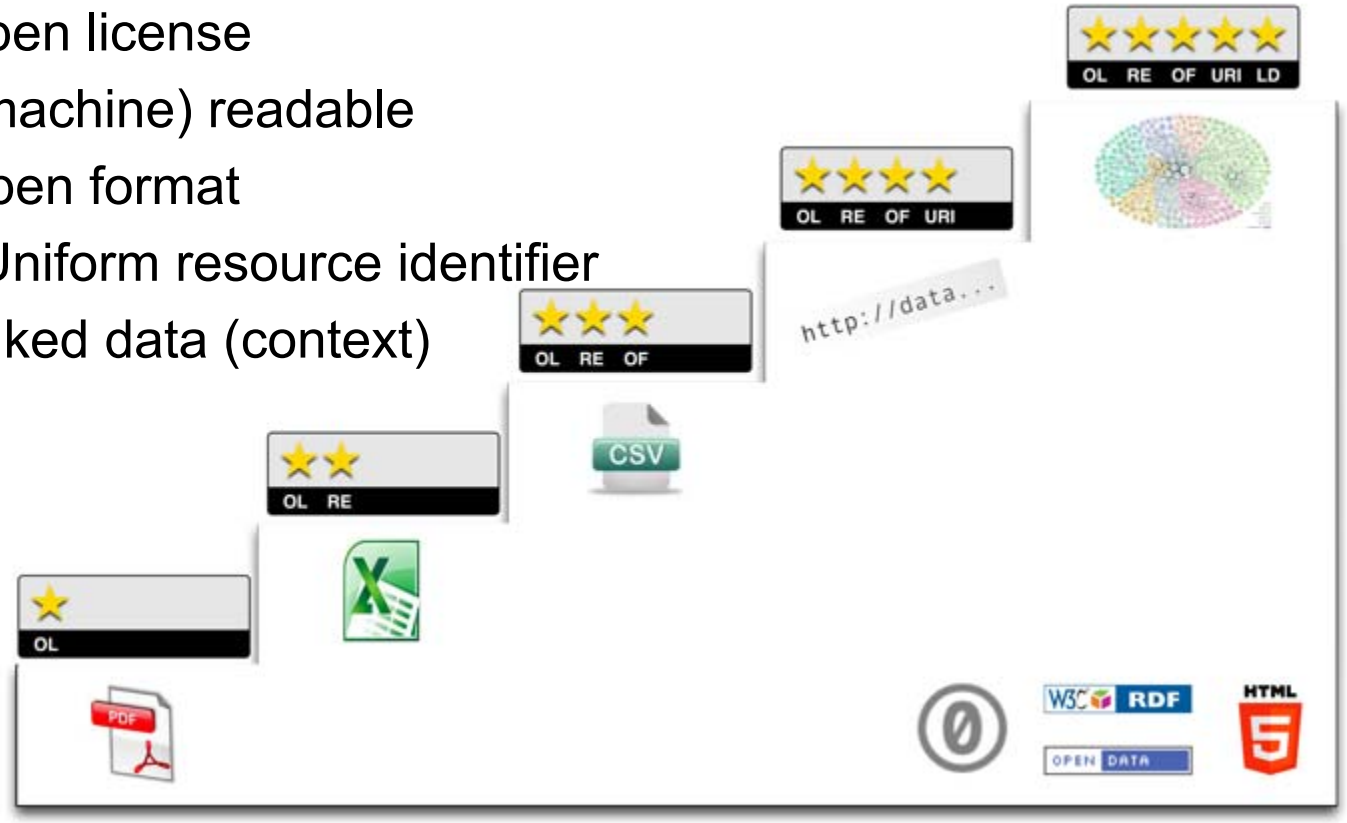
- Identification
- Documentation
- Context
- Impact
- Transparency
- Reproducibility
- Reuse



https://xkcd.com/978/

Excursion: The 5 Star Open Data Model

- 5 levels of data sharing for linked data
- Developed by Tim Berners-Lee
- Levels:
  - OL: open license
  - RE: (machine) readable
  - OF: open format
  - URI: Uniform resource identifier
  - LD: linked data (context)



Excursion: The 5 Star Open Data Model (cont'd)

★	Available on the web (whatever format) *but with an open licence, to be Open Data*
★★	Available as machine-readable structured data (e.g. excel instead of image scan of a table)
★★★	as (2) plus non-proprietary format (e.g. CSV instead of excel)
★★★★	All the above plus, Use open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff
★★★★★	All the above, plus: Link your data to other people's data to provide context

Examples: http://5stardata.info/

Standard 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

- But there is more to it…
  Landing pages – “end credits in movies”

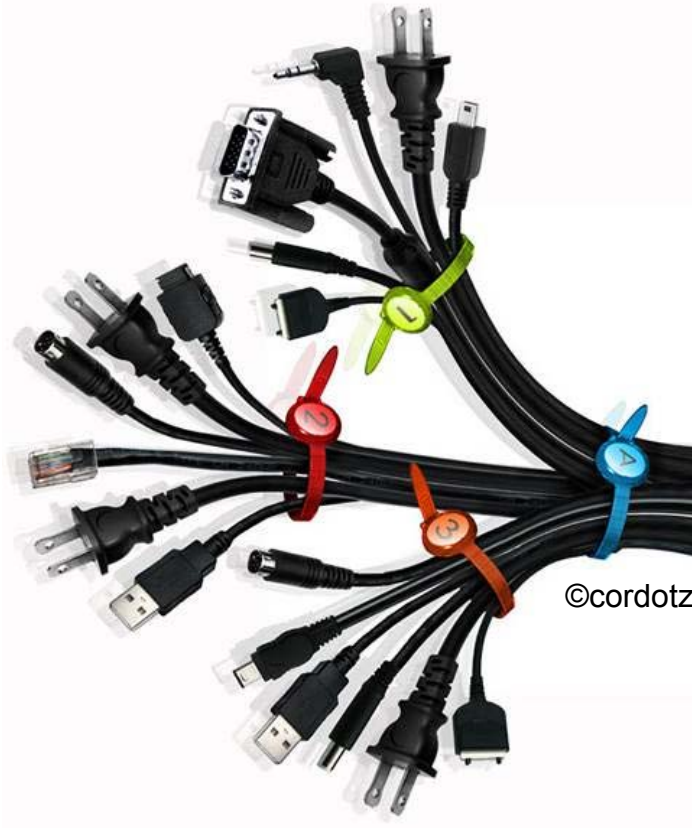
Source: http://www.dcc.ac.uk/resources/how-guides/cite-datasets

Outline

- Why should we want to cite data?
- What identifier system should I use?
- What are the challenges in data identification and citation?
- How should we do it, according to the RDA WG?
- Who is doing it so far, and how?
- Summary

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

Identifiers

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

Identifiers

Unique Identifiers for Digital Objects

- Originally:
  - Uniform Resource Name (URN)
  - Uniform Resource Locator (URL)
  - Uniform Resource Characteristic (URC, metadata, replaced by RDF)
- Uniform Resource Identifier (URI)
  - Encompasses URN and URL
  - Can be resolvable, but need not be
  - Includes ISBN, etc.
- Delegating Methods
  - Handle System
  - Digital Object Identifier (DOI)
  - Persistent URL (PURL)
  - Archival Resource Key (ARK)

Identifiers

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” (precise numbers vary…)

- Solution: persistent identifiers (PIDs)

PURL

- Persistent uniform resource locator
- Developed by Online Computer Library Center in 1995
- Based on HTTP forwarding
  - Only resolution
  - No metadata
- Provides curation and URL resolvers
- Can be hosted on own servers or centrally
- Is free
- Example: http://purl.fdlp.gov/GPO/gpo49354
  - Catalog of U.S. Government Publications

PURL Domains

Logged in as **sproell@sba-research.org** (log out)

PURL Domain Administration

Home | PURLs | Users | Groups | **Domains** | Admin | Help

1) Choose an action to take on domains

Domain administration options.

Create a new domain

2) Create a new domain

Fill in the following information to create a new domain.

Name: Advanced Practitioner Cour

Domain ID: APC2014

Maintainer IDs (one per line): sproell@sba-research.org

Writer IDs (one per line): sproell@sba-research.org

Public? (Applies solely to top-level domains):

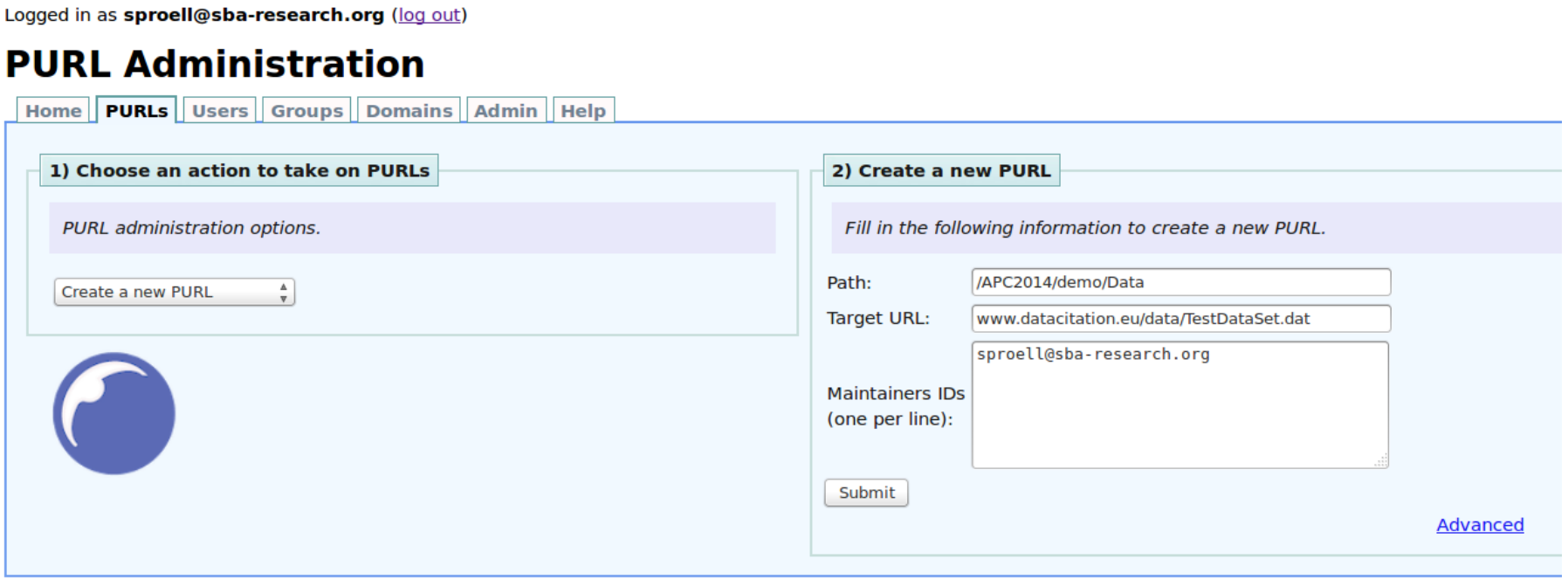
Submit

Create Successful

status: Pending approval

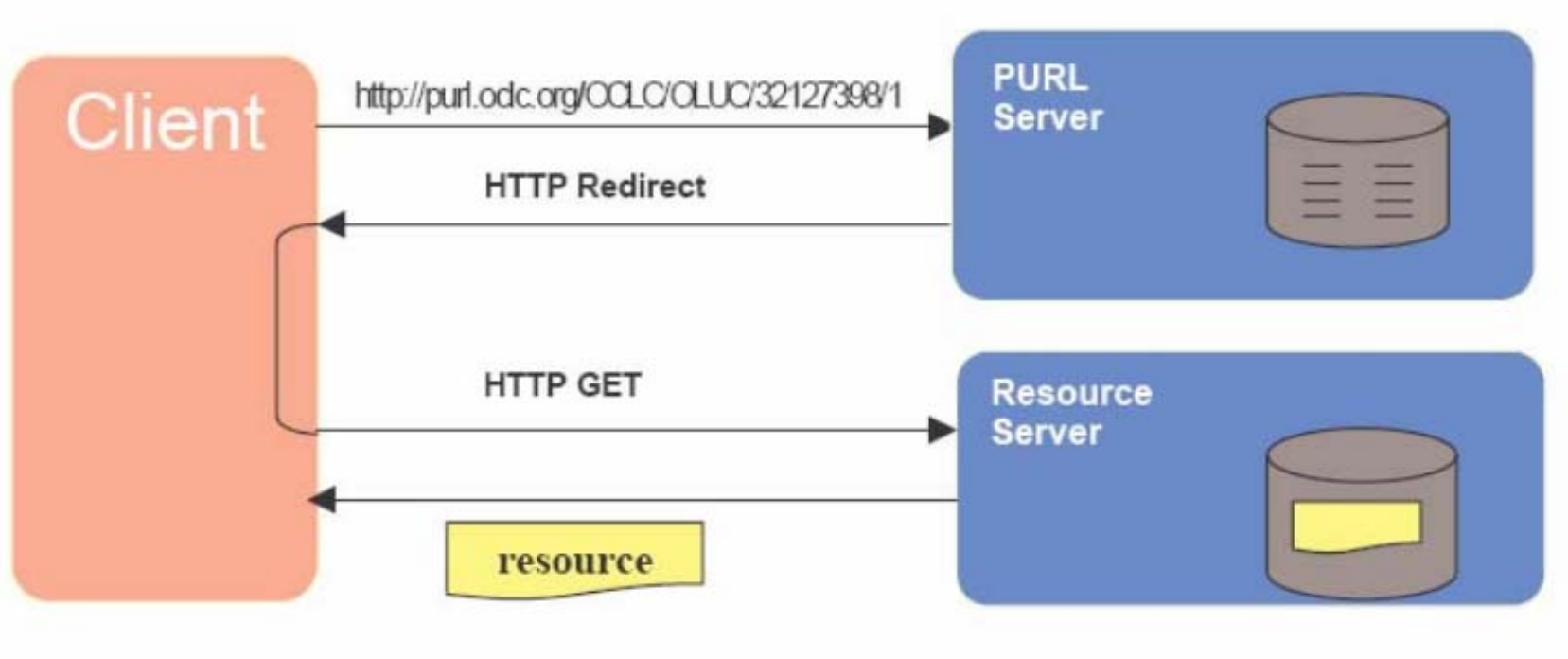
id: /APC2014
name: Advanced Practitioner Course 2014
public: false
maintainers: sproell@sba-research.org
writers: sproell@sba-research.org

PURL Registration



FACULTY OF !NFORMATICS

PURL - Resolution



© MPDL

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

Uniform Resource Identifiers

- Uniform Resource Identifier
  - Name
  - Location (e.g. Web server)
- Combination of
  - namespace identifier (NID) and a
  - namespace specific string (NSS)
- Naming scheme for URNs:
  - urn: <NID> :<NSS>
    - rn:isbn:0451450523
    - urn:ietf:rfc:2648
- Note: Cool URIs don’t change!
  https://www.w3.org/Provider/Style/URI

URN

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

Handle

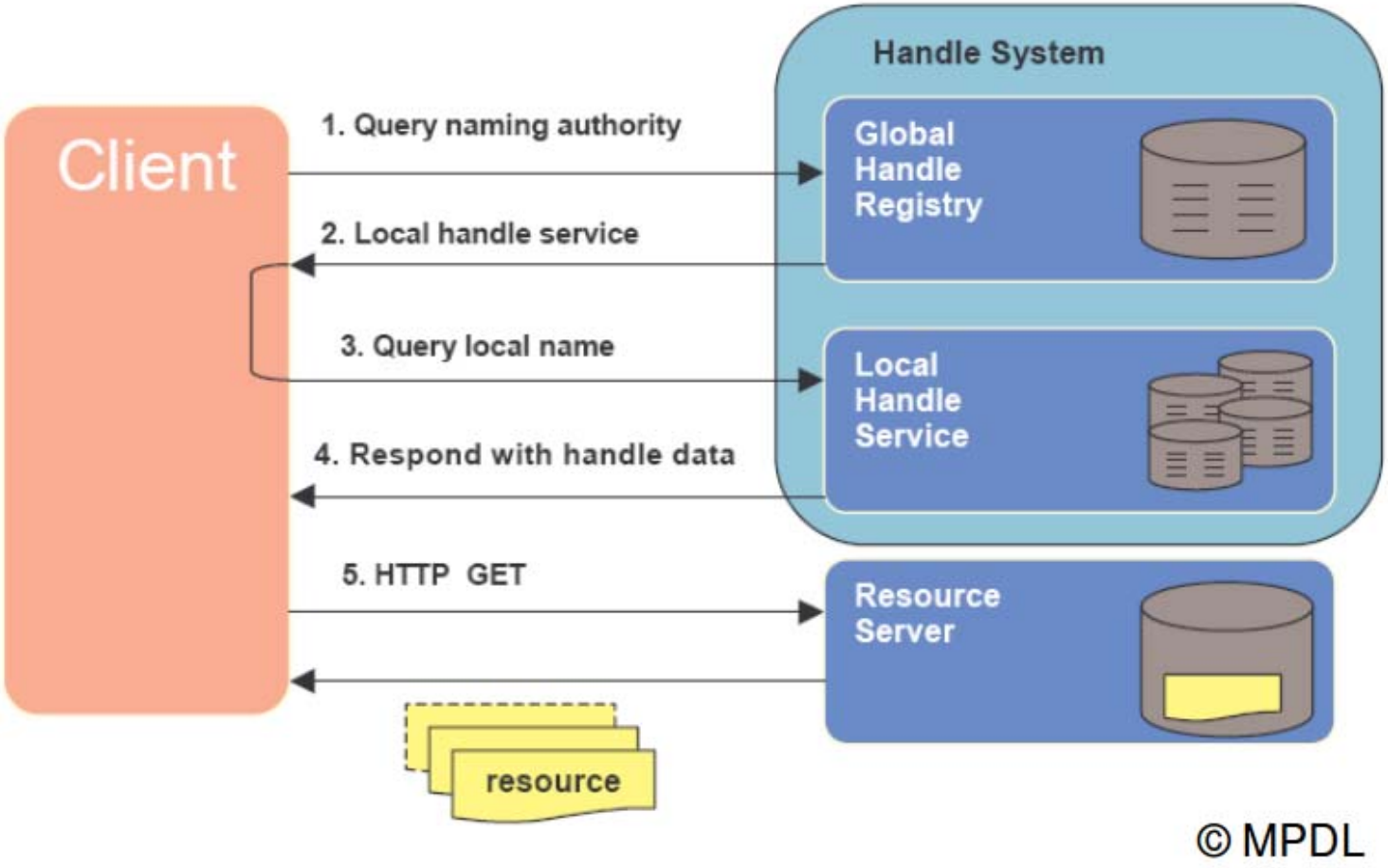
- Distributed persistent naming system
- Conforms to URN framework
- Used by most identifier systems
- Persistent identifier consists of two parts:
  - Naming authority
  - Name (must be unique string to the authority)
- Digital objects on the Internet can be assigned, managed and resolved by handles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

Handle

- Main points
  - Handles are unique and persistent
  - 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 Resolution



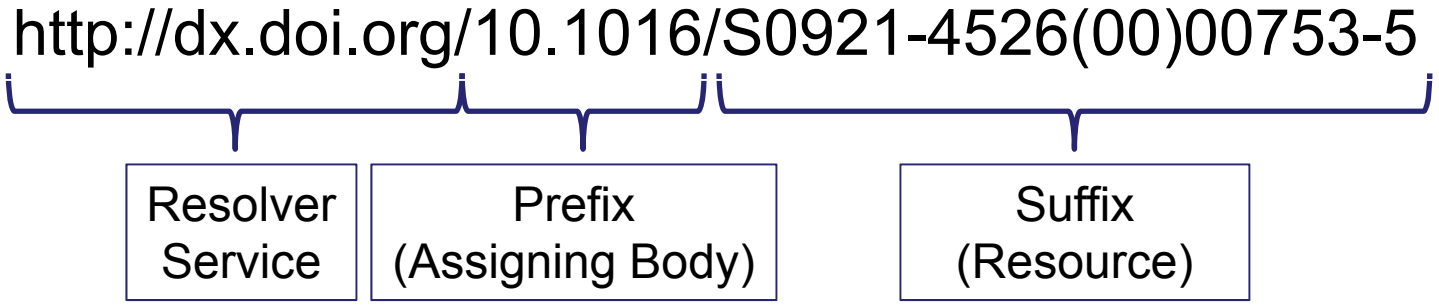
[8] http://www.mpi.nl/DAM-LR/meeting5/Persistent%20Identifiers.pdf

Digital Object Identifier (DOI)

- **Digital Identifier of** an Object
  - not "Identifier of a Digital Object“
  - "click on it and do something"
- Identifier scheme administered by the International DOI Foundation (IDF)
- Relies on the handle concept
- Provides an actionable, interoperable, persistent link

Digital Object Identifier (DOI)

- Consists of three parts:



- Resource can be any entity (thing: physical, digital, or abstract)
- DOI: 10.1594/PANGAEA.724325
- Resolver services lead to landing page
  - http://dx.doi.org/
  - http://dx.doi.org/10.1594/PANGAEA.724325

DOI: Guidelines

- Suffix must be unique within the prefix
- Suffix is case insensitive
- UTF-8
- Recommendations:
  - Use short suffixes, people have to type them
  - Do not use special characters if possible
  - Avoid semantics in the suffix string, as its semantics could change ("**no semantics in an identifier**!")
  - Slightly contradicted in "fragment identifiers" *(personal comment: avoid!)*

DOI

Metadata:

- DOI Kernel Metadata
  https://www.doi.org/doi_handbook/4_Data_Model.html
  - Other identifiers (isbn, issn, …)
  - structural types (e.g. *physical, digital, performance, abstraction)*
  - modes (*audio, visual, tangible, olfactory, tasteable, none*),
  - linkedCreation,
  - linkedParty, date of birth/death, territory, …
  - (several more)
- DOI Data Dictionary
  https://www.doi.org/doi_handbook/schemas/dd/intro.html

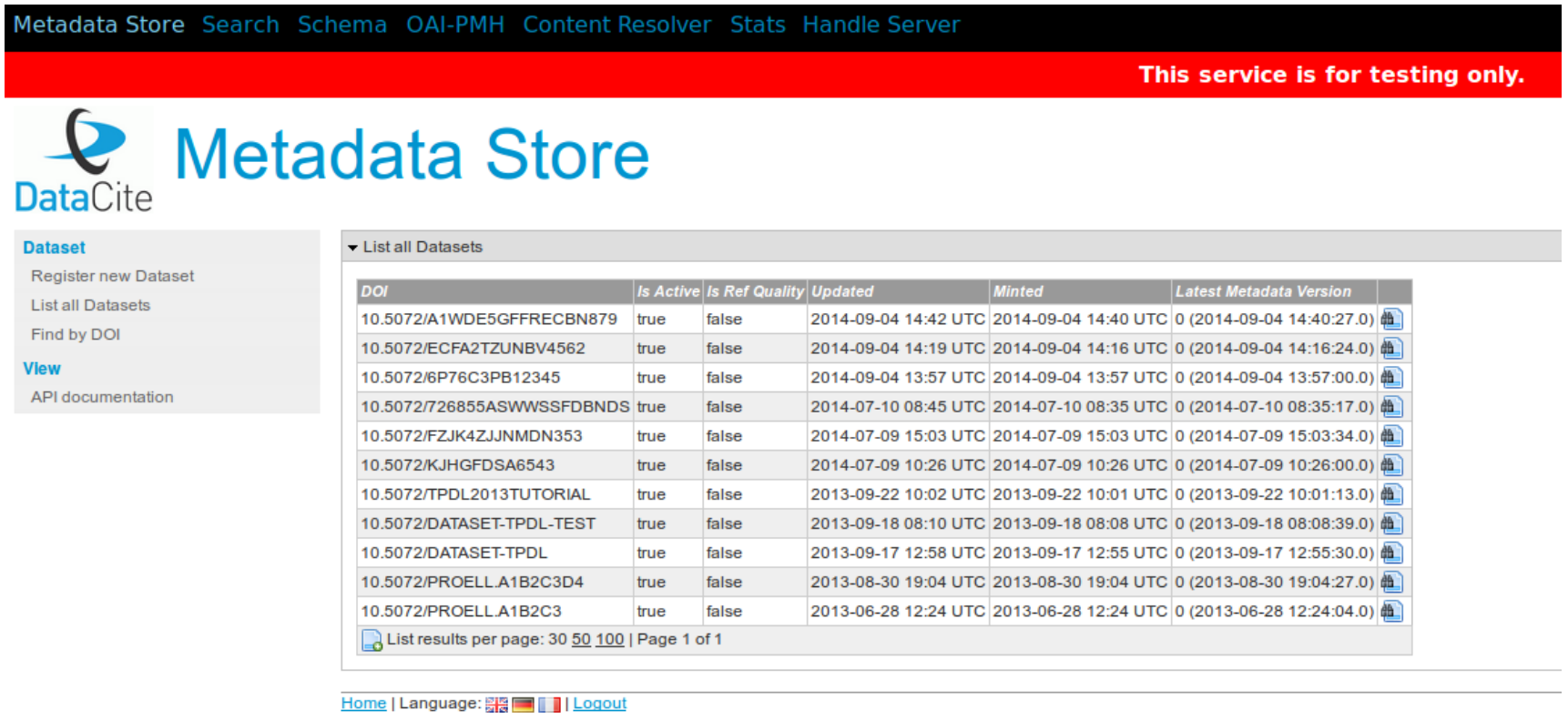
DOI

Example: Formatted Citations:

- curl -LH "Accept: text/x-bibliography; style=**apa**" http://dx.doi.org/10.1126/science.169.3946.635
  - Frank, H. S. (1970). The Structure of Ordinary Water: New data and interpretations are yielding new insights into this fascinating substance. Science, 169(3946), 635–641. doi:10.1126/science.169.3946.635
- curl -LH "Accept: text/x-bibliography; style=**bibtex**" http://dx.doi.org/10.1126/science.169.3946.635
  - @article{Frank_1970, title={The Structure of Ordinary Water: New data and interpretations are yielding new insights into this fascinating substance}, volume={169}, ISSN={1095-9203}, url={http://dx.doi.org/10.1126/science.169.3946.635}, DOI={10.1126/science.169.3946.635}, number={3946}, journal={Science}, publisher={American Association for the Advancement of Science (AAAS)}, author={Frank, H. S.}, year={1970}, month={Aug}, pages={635–641}}

DOI

DataCite Metadata Store



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 (“mint”) a DOI suffix to be linked to your object providing the required metadata
5. Start citing

DOI – Facts

- Launched in 2000
- Over 5,000 naming authorities (assigners)
- Over 20,000 DOI name prefixes
- Over 148 million DOI names assigned
  - Grows 16 % per year!
- Over 5 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
  - TU Wien (starting 2020, via Datacite)

DOI

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/

DOI vs. Handle

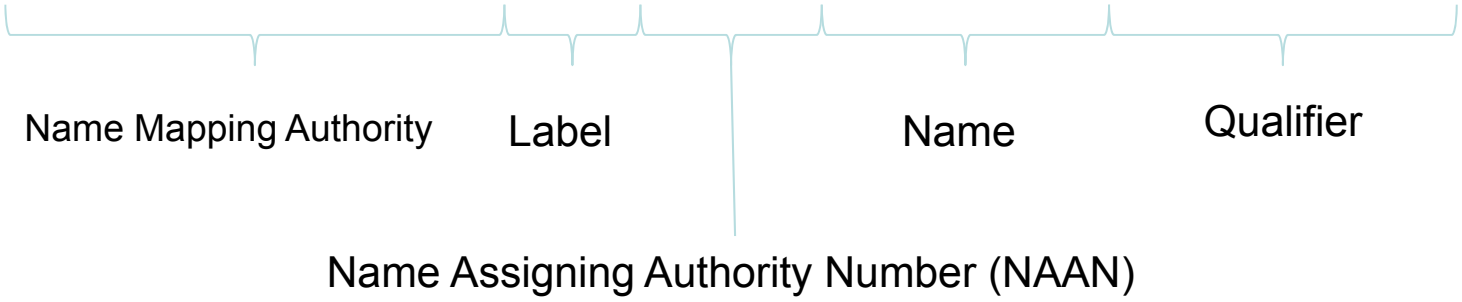
- Handle only provides the resolution service
- DOI uses the Handle System and adds:
  - **Metadata**
  - Consistency of citations
  - Semantic interoperability (data model)
  - Identification of intellectual property entities
- Used by aggregators, impact factor calculation, …

[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:

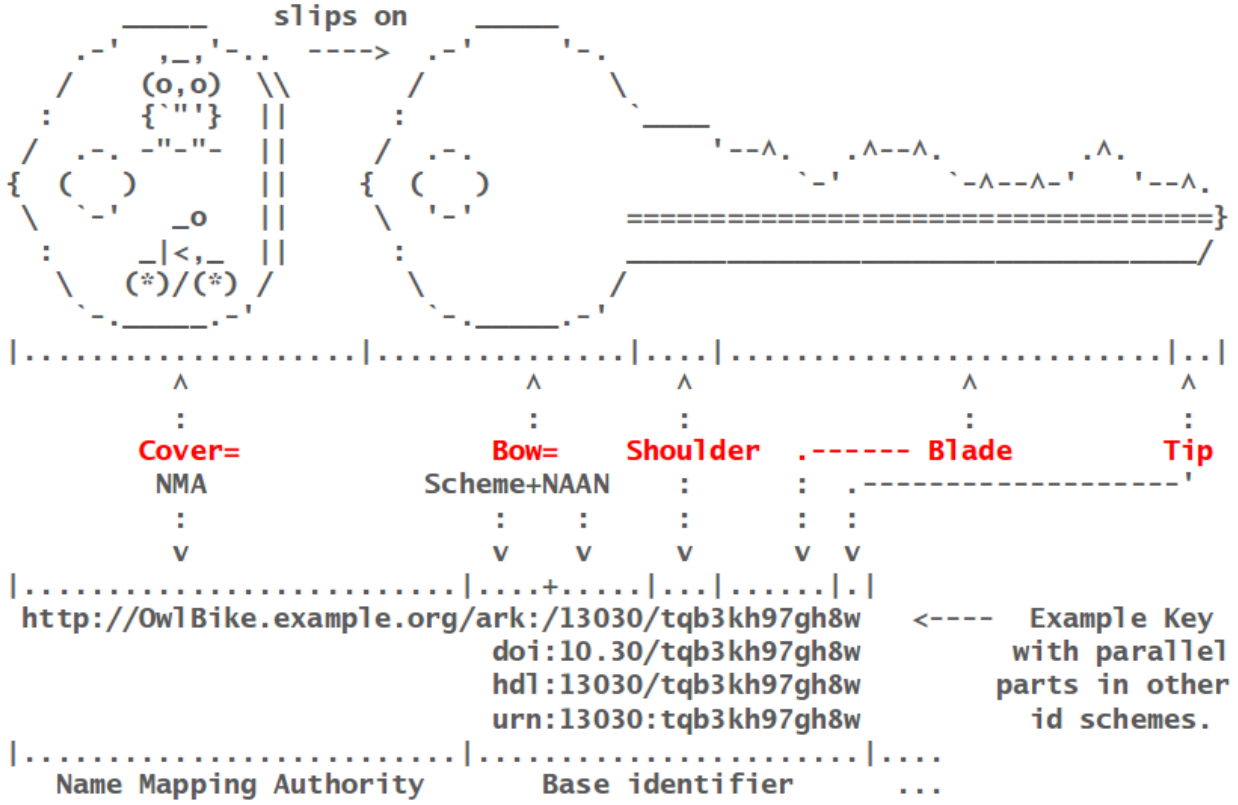
  http://example.org/ark:/13030/654xz321/s3/f8.05v.tiff



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

ARK - Scheme

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



UC3

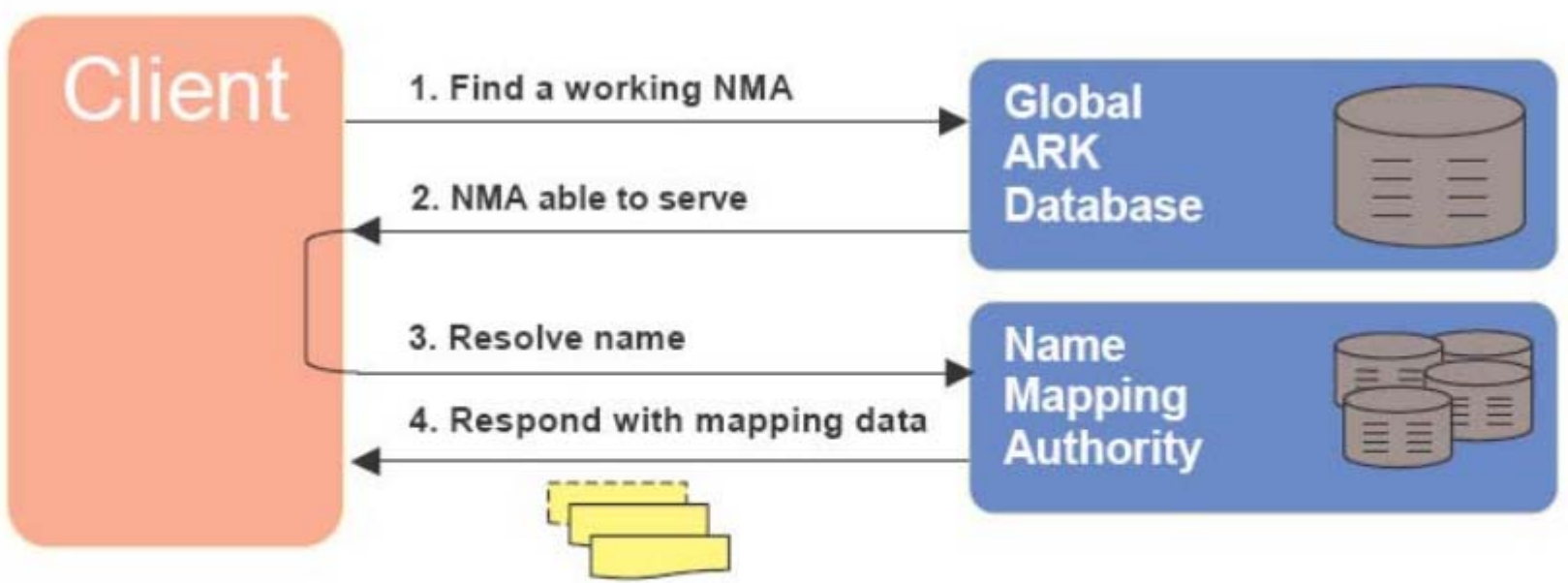
University of California CDL California Digital Library

ifs FACULTY OF !INFORMATICS

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
  - **Metadata** cannot be deleted, stored persistently at resolver!
  - Higher reputation
  - Commercial

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

An Overview of PID Systems

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

ORCID Persistent Identifiers

- ORCID
  - For people (researchers)
  - Resolving name ambiguity
  - Usefull in case of name changes
  - Link research activities and output
  - Examples:
    - https://orcid.org/0000-0002-9272-6225
    - https://orcid.org/0000-0002-4929-7875

Outline

- Why should we want to cite data?
- What identifier system should I use?
- What are the challenges in data identification and citation?
- How should we do it, according to the RDA WG?
- Who is doing it so far, and how?
- Summary

Why to cite data?

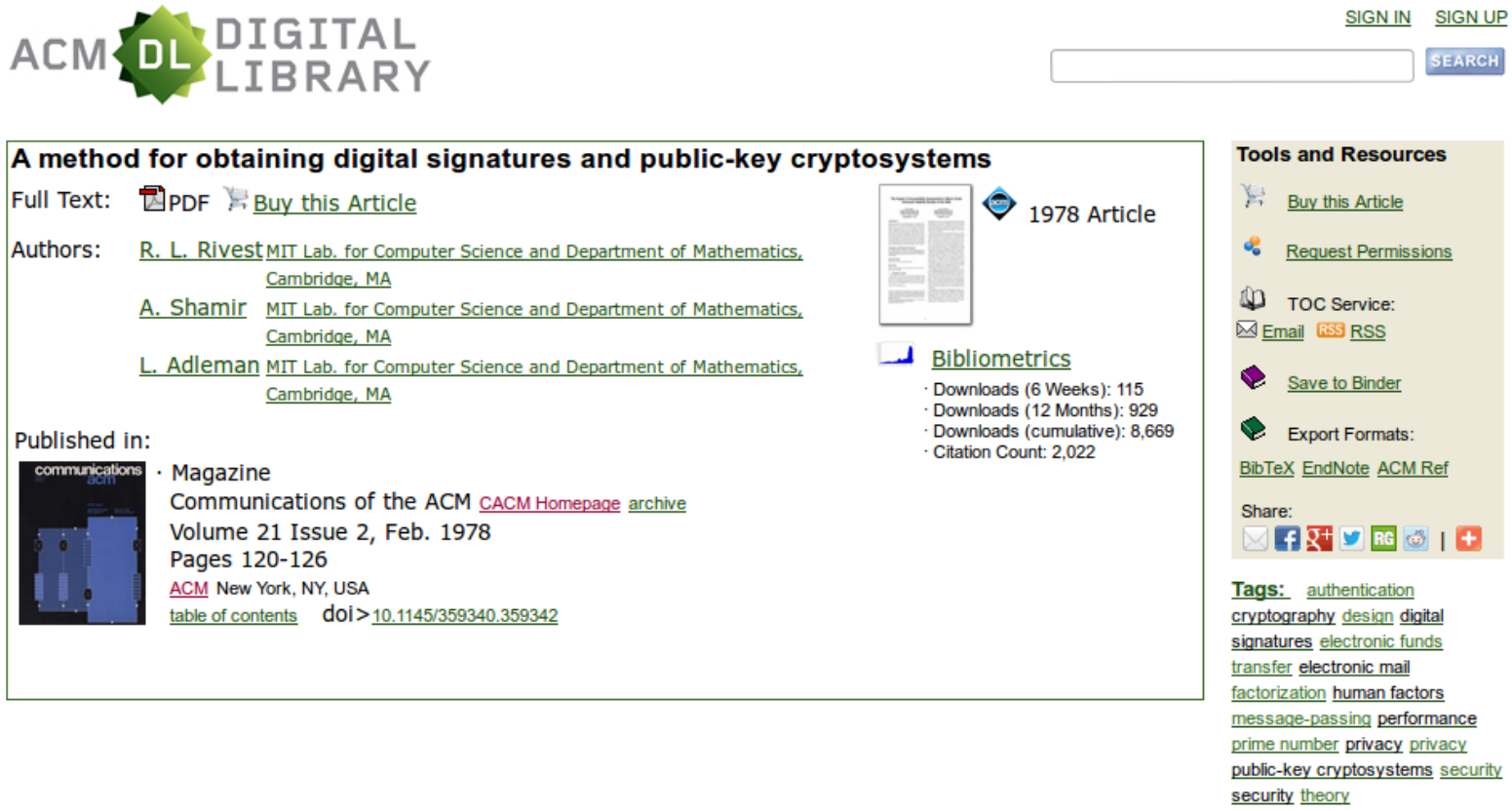
- It's what you do! – Lots of benefits
  - Makes live easier because you can build on a solid foundation
  - Speeds up the process because you can re-use existing stuff
  - Helps avoiding / detecting mistakes, improves quality
  - Reuse increases citations, visibility, currency

Deja-vue

- But:
  - To achieve this it must be easy, straightforward, "automatic"
  - Citing Papers is easy…
  - …what about data?
    (more about this later… first: "we should just do it")

How to cite data?

- Referencing research papers is well established



ACM DL 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:

· 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

Example: Web Page Download

- Example: Web-page download

Natural Language Interfaces: What is the Problem? - A data-driven quantitative analysis

Philipp Cimiano[1] and Michael Minock[2]

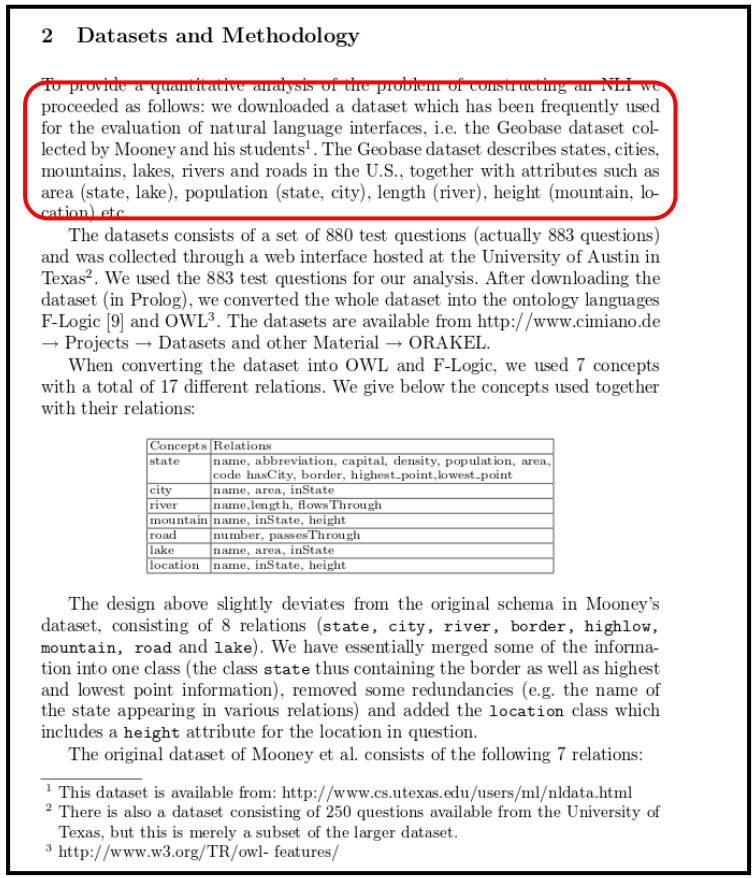
[1]WIS, TU Delft / [2]University of Umea

Abstract. While qualitative analyses of the problems involved in building natural language interfaces (NLIs) have been available, a quantitative grounding in empirical data has been missing. We fill this gap by providing a quantitative analysis on the basis of the Geobase dataset. We hope that this analysis can guide further research in NLIs.

1 Introduction

So far, there has been an impressive amount of research on natural language interfaces (NLIs), i.e. on interfaces allowing users to interact with a certain information system in natural language. While NLIs are not inherently restricted only to the task of answering questions on the basis of a given database or knowledge base, most of the NLIs developed so far have been designed for this purpose. Along these lines, as in most other research on natural language interfaces, we limit ourselves to this restricted view of natural language interfaces essentially as systems providing answers to natural language questions in this paper. Research on NLIs dates back to the 70s and 80s (see [1], [6]) and has yielded increased attention in recent years with a plethora of systems emerging: PRECISE [13], STEP [11], ORAKEL [3], Aqualog [10], GINSENG [2], just to name a few of the very recent systems. What seems missing so far is a description of the problem, in particular a quantitative analysis of the problems inherent in the task of building natural language interfaces. While there have been qualitative analyses of the problems involved in constructing NLIs ([1], [6]), to our knowledge there has been no quantitative analysis grounding the qualitative characteristics of the problem in real data. This is crucial in our view as it can and should guide the development of NLIs in the future, focusing them on the challenging problems. It would also help system developers to focus on a specific phenomenon encountered in NLIs (e.g. resolution of ambiguities) and foster progress in the field by clearly designing and evaluating the solution to a specific phenomenon which would ideally not be specific to one particular approach but reusable across systems. In our view, no real progress can be expected in NLI research only from charts hiding the interesting details and solutions to characteristic problems involved in the task behind top performing precision and recall measures.

The structure of this paper is as follows: in the next Section 2 we describe the dataset we have used to provide a quantitative analysis and describe our methodology. Then, in Section 3 we describe our interesting findings and derive



2 Datasets and Methodology

To provide a quantitative analysis of the problem of constructing an NLI we proceeded as follows: we downloaded a dataset which has been frequently used for the evaluation of natural language interfaces, i.e. the Geobase dataset collected by Mooney and his students[1]. The Geobase dataset describes states, cities, mountains, lakes, rivers and roads in the U.S., together with attributes such as area (state, lake), population (state, city), length (river), height (mountain, location) etc.

The datasets consists of a set of 880 test questions (actually 883 questions) and was collected through a web interface hosted at the University of Austin in Texas[2]. We used the 883 test questions for our analysis. After downloading the dataset (in Prolog), we converted the whole dataset into the ontology languages F-Logic [9] and OWL[3]. The datasets are available from http://www.cimiano.de → Projects → Datasets and other Material → ORAKEL.

When converting the dataset into OWL and F-Logic, we used 7 concepts with a total of 17 different relations. We give below the concepts used together with their relations:

Concepts	Relations
state	name, abbreviation, capital, density, population, area, code hasCity, border, highest_point,lowest_point
city	name, area, inState
river	name,length, flowsThrough
mountain	name, inState, height
road	number, passesThrough
lake	name, area, inState
location	name, inState, height

The design above slightly deviates from the original schema in Mooney's dataset, consisting of 8 relations (`state, city, river, border, highlow, mountain, road` and `lake`). We have essentially merged some of the information into one class (the class `state` thus containing the border as well as highest and lowest point information), removed some redundancies (e.g. the name of the state appearing in various relations) and added the `location` class which includes a `height` attribute for the location in question.

The original dataset of Mooney et al. consists of the following 7 relations:

[1] This dataset is available from: http://www.cs.utexas.edu/users/ml/nldata.html
[2] There is also a dataset consisting of 250 questions available from the University of Texas, but this is merely a subset of the larger dataset.
[3] http://www.w3.org/TR/owl- features/

Example: Web Page Download

To provide a quantitative analysis of the problem of constructing an NLI we proceeded as follows: we downloaded a dataset which has been frequently used for the evaluation of natural language interfaces, i.e. the Geobase dataset collected by Mooney and his students[1]. The Geobase dataset describes states, cities, mountains, lakes, rivers and roads in the U.S., together with attributes such as area (state, lake), population (state, city), length (river), height (mountain, location) etc.

The datasets consists of a set of 880 test questions (actually 883 questions) and was collected through a web interface hosted at the University of Austin in Texas[2]. We used the 883 test questions for our analysis. After downloading the dataset (in Prolog), we converted the whole dataset into the ontology languages F-Logic [9] and OWL[3]. The datasets are available from http://www.cimiano.de → Projects → Datasets and other Material → ORAKEL.

---

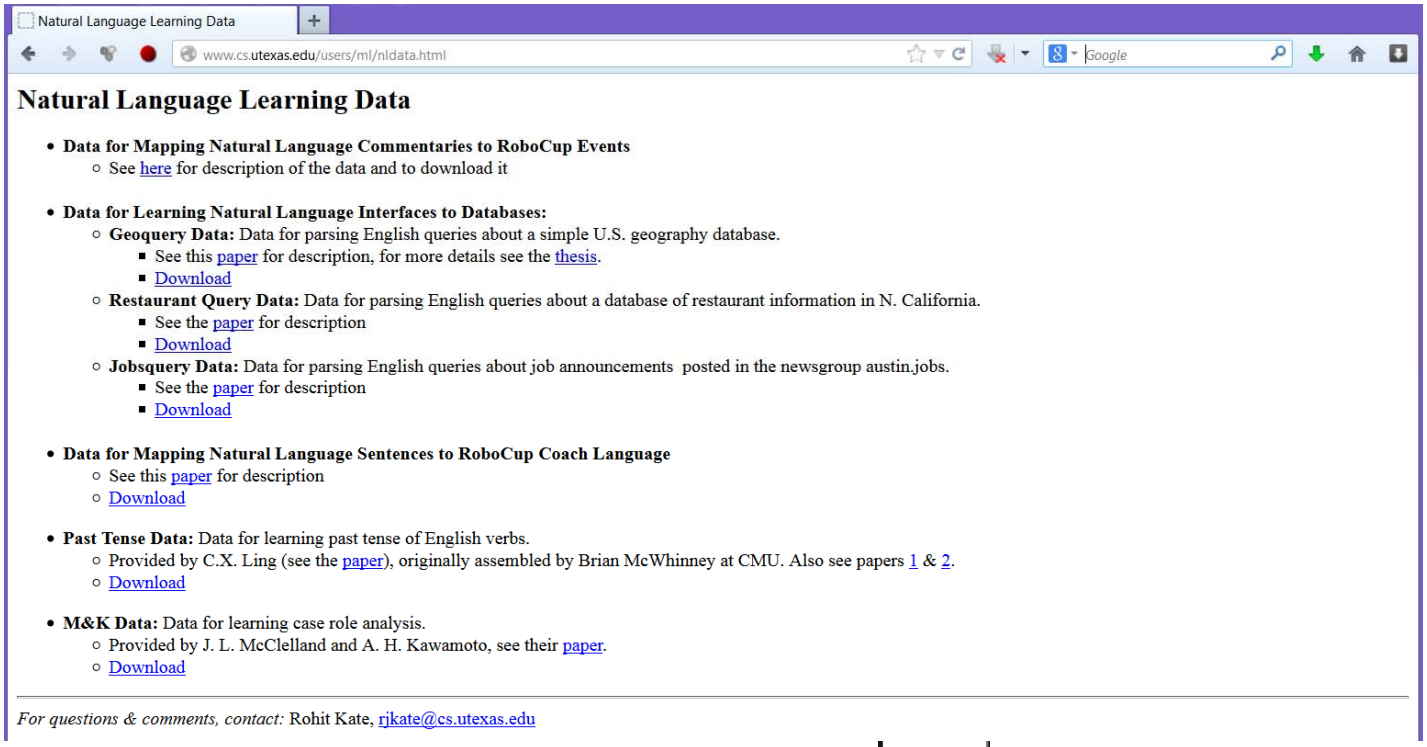
[1] This dataset is available from: http://www.cs.utexas.edu/users/ml/nldata.html

[2] There is also a dataset consisting of 250 questions available from the University of Texas, but this is merely a subset of the larger dataset.

[3] http://www.w3.org/TR/owl- features/

Example: Web Page Download

[1] This dataset is available from: http://www.cs.utexas.edu/users/ml/nldata.html
[2] There is also a dataset consisting of 250 questions available from the University of Texas, but this is merely a subset of the larger dataset.
[3] http://www.w3.org/TR/owl- features/



Example: Sharing Platform

- Example: Data sharing platforms

206 SYSTEMATIC BIOLOGY VOL. 62

of < 1, and are said to be "indecisive" for those branches that cannot be reconstructed. Supermatrices are often characterized by very high levels of missing data, and as such can typically be expected to have PDC values lower than more conventional matrices. The PDC metric has not been broadly applied across a sample of such large combined matrices, but in practice we have observed PDC values between 0.6 and 0.7 to be common for unoptimized matrices compiled from public databases, which frequently contain >70% missing data. Although it is not necessary to have a completely decisive matrix to reconstruct the true tree (because all possible branches in the treespace are not present in all trees), maximizing PDC can nonetheless generally be expected to improve tree resolution in the majority of real-world scenarios.

A related challenge is presented by so-called "rogue taxa"—destabilizing tips that move around the tree enough in replicate searches to cause the collapse of clades and the erosion of branch support values. The instability of a given tip may be compounded by multiple attributes, although perhaps most frequently by (i) data nonintegrity due to sequence misidentification and (ii) very weak phylogenetic signal simply due to a dearth of meaningful phylogenetic information. Identification and removal of rogue taxa is generally expected to help improve the resolution of trees containing such unstable tips (Thorley and Wilkinson 1999; Thomson and Shaffer 2010), and several methods have been presented to accomplish this goal (Thorley and Wilkinson 1999; Thorley and Page 2000; Smith and Dunn 2008). Here, we present and make use of a statistic modified from the *I* score implemented in Mesquite (Maddison and Maddison 2010), which overcomes some of the drawbacks of previously available instability metrics.

In this study, we present results from a supermatrix phylogeny (a PHLAWD megaphylogeny) of Cyperaceae. In this context, we address issues of data integrity as they relate to large combined analyses, and we explore the association between data decisiveness and tree resolution. We compare the results of our tree searches to results from a previously published family level analysis of the Cyperaceae, and we show that maximizing the number of loci and taxa using the methods employed here may in fact yield greater power to resolve phylogenetic trees than more traditional, focused sequencing approaches that rely on smaller numbers of better-sampled loci, even with a highly incomplete alignment.

Materials and Methods

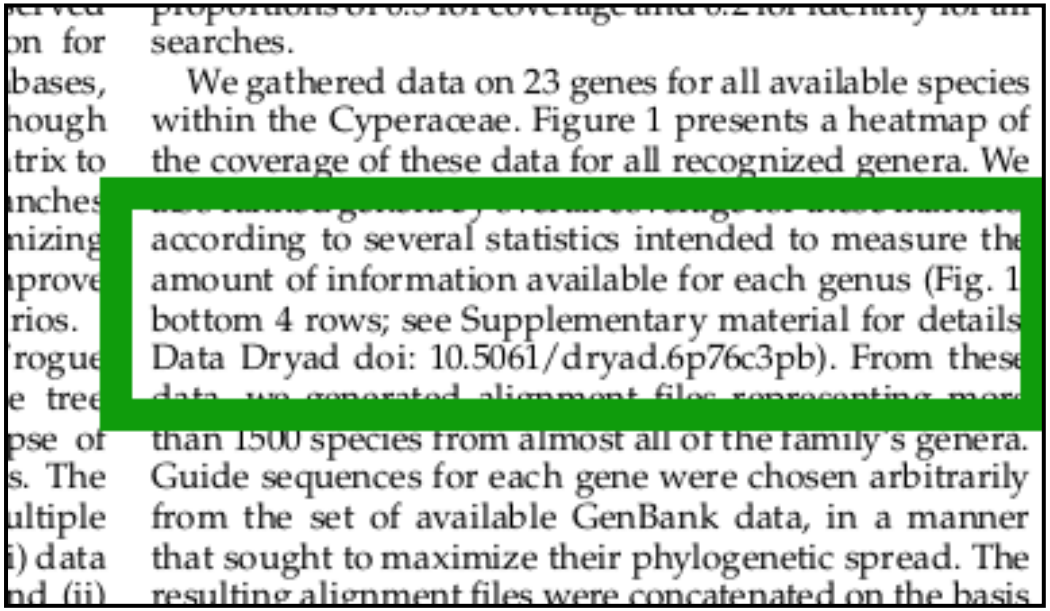
Data Collection and Validation

Data were gathered from GenBank release 185 using the software tool PHLAWD (Smith et al. 2009), which searches the NCBI database for all nucleotide sequences for species within a given taxon that match a given text query. For instance, PHLAWD can be used to yield an alignment of all NCBI sequences matching the query string "internal transcribed spacer" within the NCBI taxon Cyperaceae. Because some nonorthologous sequences may be returned for any given query, PHLAWD requires a set of presupplied guide sequences for each query. Any query results that do not match these guide sequences by some arbitrary minimum amount of coverage and identity are excluded. We used cutoff proportions of 0.3 for coverage and 0.2 for identity for all searches.

We gathered data on 23 genes for all available species within the Cyperaceae. Figure 1 presents a heatmap of the coverage of these data for all recognized genera. We [illegible] according to several statistics intended to measure the amount of information available for each genus (Fig. 1 bottom 4 rows; see Supplementary material for details Data Dryad doi: 10.5061/dryad.6p76c3pb). From these data, we generated alignment files representing more than 1500 species from almost all of the family's genera. Guide sequences for each gene were chosen arbitrarily from the set of available GenBank data, in a manner that sought to maximize their phylogenetic spread. The resulting alignment files were concatenated on the basis of species name using the software phyutility (Smith and Dunn 2008). This approach combines available data from multiple exemplar specimens of each species to make phylogeny estimation at this scale possible, and it has proven effective for estimating phylogeny of large numbers of species when parallel data from multiple markers are not available from a single specimen (Jones et al. 2002; Smith and Donoghue 2008; Smith et al. 2009).

Because data integrity on public databases such as GenBank is uncertain, it is possible to include poor quality or incorrectly identified sequences when data mining these resources. When multiple sequences from different vouchers are concatenated on the basis of species identity, the potential exists for sequences from misidentified species to be concatenated with correctly identified ones, leading to the creation of so-called "chimeric" taxa. The tip nodes that are associated with these mistaken taxa are expected to contain conflicting phylogenetic signal which, at least in the case of bootstrapping, can erode support values for otherwise well-supported clades, because the chimeric tip is either expected to be placed near each of its different constituent species in different replicate searches, or to contain strong enough conflicting signal that it simply causes parts of the tree to collapse. One approach to a posteriori identification of these taxa involves scoring each taxon for a leaf stability index—a statistic that quantifies how much its position changes relative to other taxa in replicate tree searches. Those taxa with the lowest stability are assumed to contain either conflicting phylogenetic signal (especially in the case of chimeric taxa) or such low levels of signal that many placements are equivocal.

Multiple methods of measuring leaf stability have been proposed (Thorley and Wilkinson 1999; Thorley and Page 2000; Smith and Dunn 2008). For this study, we built upon previous work by Maddison and Maddison (2010) in Mesquite version 2.7. They present a statistic,



searches.

We gathered data on 23 genes for all available species within the Cyperaceae. Figure 1 presents a heatmap of the coverage of these data for all recognized genera. We

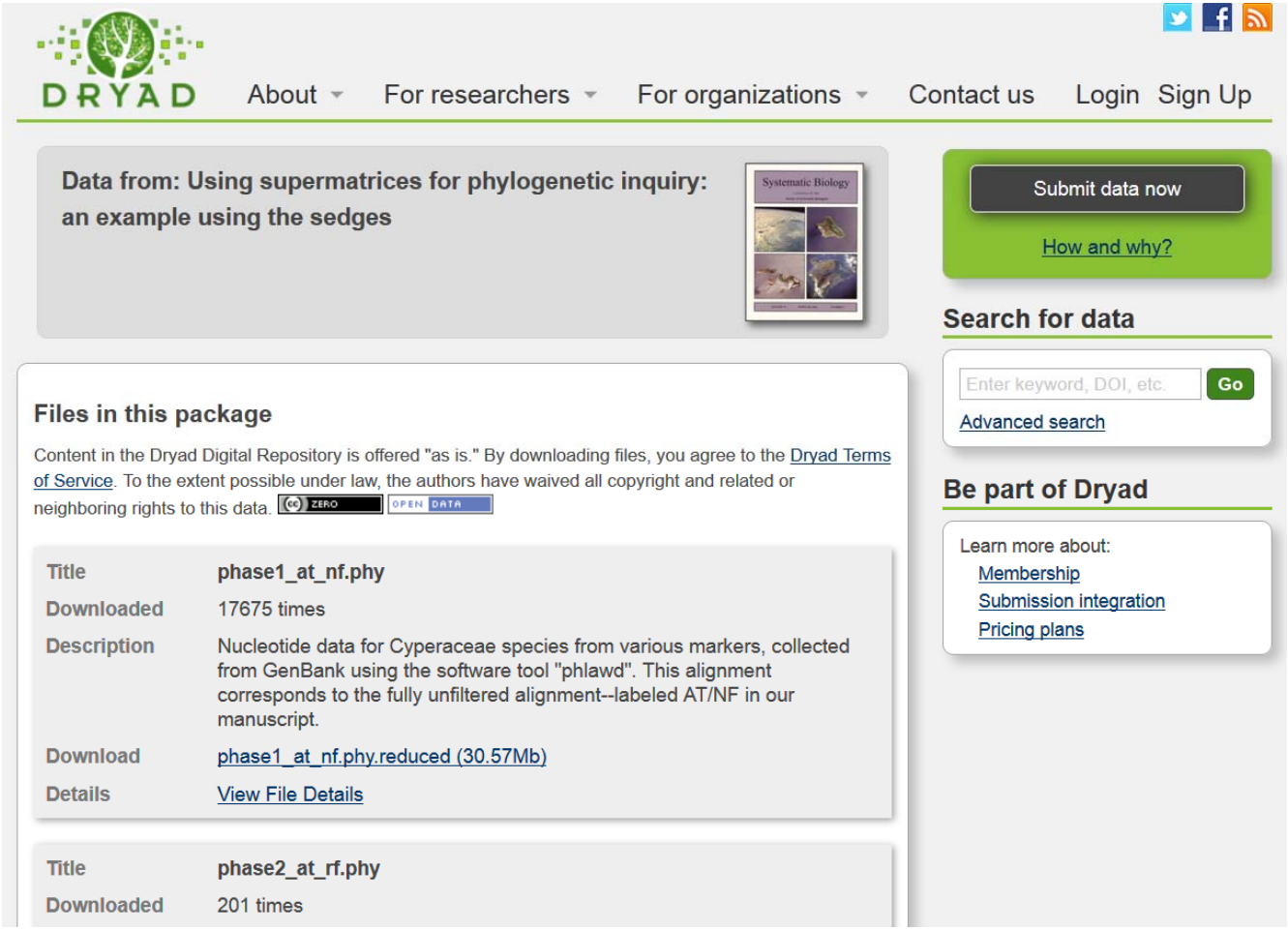
according to several statistics intended to measure the amount of information available for each genus (Fig. 1 bottom 4 rows; see Supplementary material for details Data Dryad doi: 10.5061/dryad.6p76c3pb). From these

than 1500 species from almost all of the family's genera. Guide sequences for each gene were chosen arbitrarily from the set of available GenBank data, in a manner that sought to maximize their phylogenetic spread. The resulting alignment files were concatenated on the basis

http://datadryad.org/resource/doi:10.5061/dryad.6p76c3pb/1

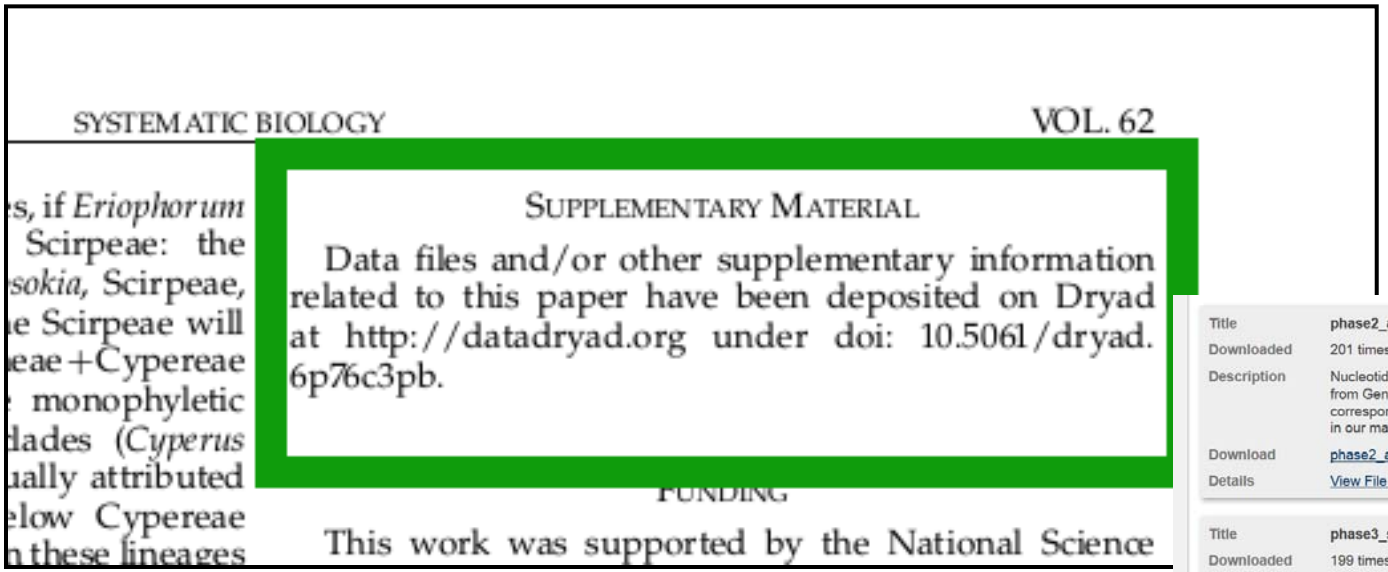
Example: Sharing Platform

- Example: Data sharing platforms



http://datadryad.org/resource/doi:10.5061/dryad.6p76c3pb/1

Example: Sharing Platform



SYSTEMATIC BIOLOGY VOL. 62

s, if *Eriophorum*
Scirpeae: the
sokia, Scirpeae,
e Scirpeae will
eae+Cypereae
monophyletic
lades (*Cyperus*
ually attributed
elow Cypereae
these lineages

SUPPLEMENTARY MATERIAL

Data files and/or other supplementary information related to this paper have been deposited on Dryad at http://datadryad.org under doi: 10.5061/dryad.6p76c3pb.

FUNDING

This work was supported by the National Science

When using this data, please cite the original publication:

Hinchliff CE, Roalson EH (2012) Using supermatrices for phylogenetic inquiry: an e
using the sedges. Systematic Biology 62(2): 205-219. http://dx.doi.org/10.1093/sys

Additionally, please cite the Dryad data package:

Hinchliff CE, Roalson EH (2012) Data from: Using supermatrices for phylogenetic in
example using the sedges. Dryad Digital Repository. http://dx.doi.org/10.5061
/dryad.6p76c3pb

Cite | Share

Title	phase2_at_rf.phy
Downloaded	201 times
Description	Nucleotide data for Cyperaceae species from various markers, collected from GenBank using the software tool "phlawd". This alignment corresponds to the unscaffolded, rogues-filtered alignment--labeled AT/RF in our manuscript.
Download	phase2_at_rf.phy.reduced (26.56Mb)
Details	View File Details

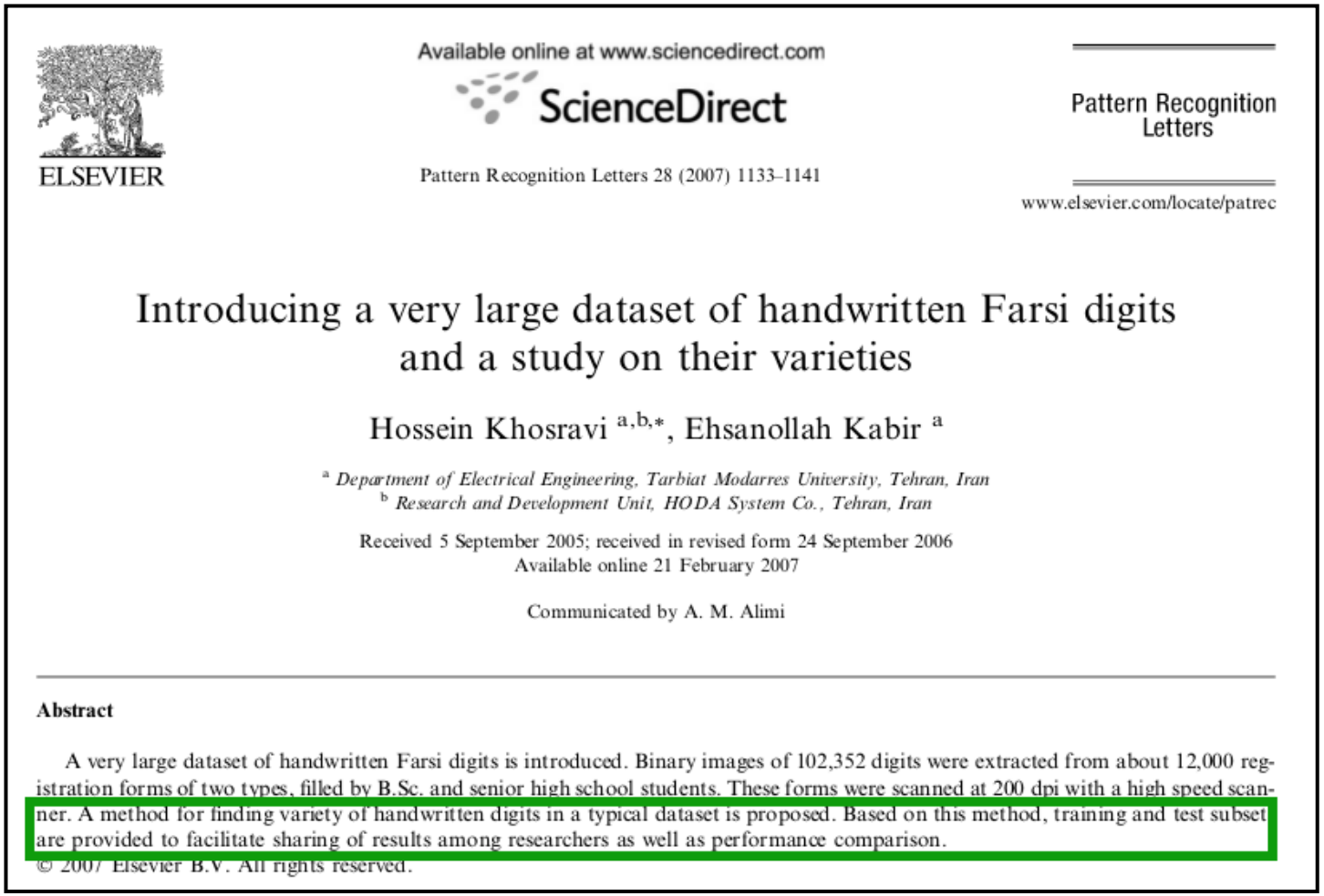
Title	phase3_sc_nf.phy
Downloaded	199 times
Description	Nucleotide data for Cyperaceae species from various markers, collected from GenBank using the software tool "phlawd". This alignment corresponds to the scaffolded alignment with rogues unfiltered--labeled SC/NF in our manuscript.
Download	phase3_sc_nf.phy.reduced (7.766Mb)
Details	View File Details

Title	phase4_sc_rf.phy
Downloaded	206 times
Description	Nucleotide data for Cyperaceae species from various markers, collected from GenBank using the software tool "phlawd". This alignment corresponds to the maximally filtered alignment: scaffolded and having had rogues removed--labeled SC/RF in our manuscript.
Download	phase4_sc_rf.phy.reduced (6.976Mb)
Details	View File Details

http://datadryad.org/resource/doi:10.5061/dryad.6p76c3pb/1

Subset Citation in Papers

- Example: Subsets of data



Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition Letters 28 (2007) 1133–1141
Pattern Recognition Letters
www.elsevier.com/locate/patrec

Introducing a very large dataset of handwritten Farsi digits and a study on their varieties

Hossein Khosravi [a,b,*], Ehsanollah Kabir [a]

[a] Department of Electrical Engineering, Tarbiat Modarres University, Tehran, Iran
[b] Research and Development Unit, HODA System Co., Tehran, Iran

Received 5 September 2005; received in revised form 24 September 2006
Available online 21 February 2007

Communicated by A. M. Alimi

Abstract

A very large dataset of handwritten Farsi digits is introduced. Binary images of 102,352 digits were extracted from about 12,000 registration forms of two types, filled by B.Sc. and senior high school students. These forms were scanned at 200 dpi with a high speed scanner. A method for finding variety of handwritten digits in a typical dataset is proposed. Based on this method, training and test subset are provided to facilitate sharing of results among researchers as well as performance comparison.

Khosravi, Hossein, and Ehsanollah Kabir. "Introducing a very large dataset of handwritten Farsi digits and a study on their varieties." Pattern Recognition Letters 28.10 (2007): 1133-1141.

Subset Citation in Papers

- Example: Subsets of data

Table 1
Some popular digit datasets

Dataset	dpi	Training samples	Test samples	Total samples
CENPARMI	166	4000	2000	6000
CEDAR	300	18,468	2711	21,179
MNIST	Normalized into 20 * 20	60,000	10,000	70,000
USPS	300	7291	2007	9298

The CEDAR[3] digit dataset is available from CEDAR, SUNY[4] at Buffalo. The images were scanned at 300 dpi. The training and test sets contain 18468 and 2711 digits, respectively. The number of samples in both training and test sets differ for each class. Since some images in the test set are poorly segmented, a subset of 2213 well-segmented images are also provided for testing (Liu, 2003).

The MNIST, modified NIST,[5] dataset (LeCun et al., 1995) was extracted from the NIST datasets SD3 and SD7. The training and test sets are composed from both SD3 and SD7. Samples are normalized into 20 * 20 gray-scale images with aspect ratio reserved, and the normalized images are located in a 28 * 28 frame. The dataset is available from LeCun. Number of training and test samples are 60,000 and 10,000 respectively.

At last the USPS digit dataset has 7291 training and 2007 test samples (Hull, 1994). Table 1 lists these datasets briefly.

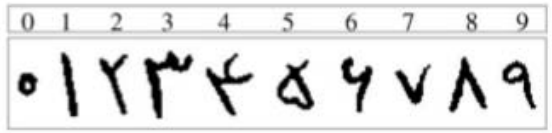


Fig. 1. Sample handwritten Farsi digits.

There were several fields in both types of forms. We used two digit fields from type 1, including *Postal Code* and *National Code*, each of 10 digits length and three digit fields from type 2 including *Record Number*, *Identity Certificate Number* and *Phone Number* that at most have 26 digits, while in average about 20 digits. Both forms are in color. In both types, handwritten texts are in blue or occasionally in black.

3.2. Digit extraction and recognition

To extract the digits, we must find the regions of interest. There were at least two reference marks (squares) in each form (circled in Fig. 2). We first search for these marks using a simple and fast algorithm shown in Fig. 3. If they are not found, the form is rejected. This situation occurs rarely, e.g. when the paper is scanned upside down or the reference square is too noisy. Then, if the reference squares are not in their expected positions, the form is rotated and shifted so that these squares are placed in the

Subset Citation in Papers

- Example: Subsets of data

5. Choosing the training and test sets

To facilitate sharing of results on this dataset between researchers, we provide two distinct datasets for training and test.

From Table 3 it can be seen that the most usual styles are fallen into samples S1, and other varieties are fallen into S2, S3 and S4. So we tried to select most of training samples from S1. To be more accurate we selected from each category a number of samples equal to their proportion in total samples, i.e. 73.47% of training samples were selected from S1, 9.83% from S2 and so on. Then we sat aside training samples and select test samples from the remaining samples, randomly. In this way the training set is a true representation of the whole population, while the test set is selected without any predefined information.

We selected 60,000 samples for training set and 20,000 for test. The remaining samples are also available in another subset (see Appendix A).

Appendix A. Dataset specification and availability

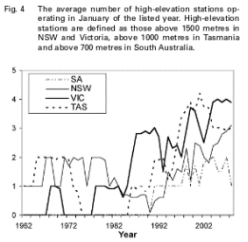
The dataset is available in four separate files, **Total.cdb, Training.cdb, Test.cdb, Remaining.cdb**. The file format is described here with a pseudo code:

```
Skip Header (1024 bytes)
while not End of File
{
    read Start Byte: (1 byte) 0xFF that
    specifies the start of newimage
    read Label: (1 byte) character label
    read Width: (1 byte) character width
    read Height: (1 byte) character height
    read Byte Count: (2 bytes) number of bytes for this character.
    //Runlength coding on each row
    for y = 0 to Height
        while(x < Width)
        {
            read NumOfWhitePixels,
            read NumOfBlackPixels;

        }

}
```

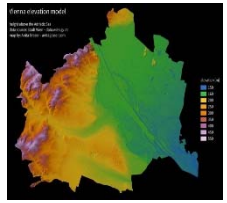
Source codes for reading the dataset files are available in Matlab, C++ and Pascal. To get the dataset please contact kabir@modares.ac.ir, or see the homepage http://www.modares.ac.ir/eng/kabir.

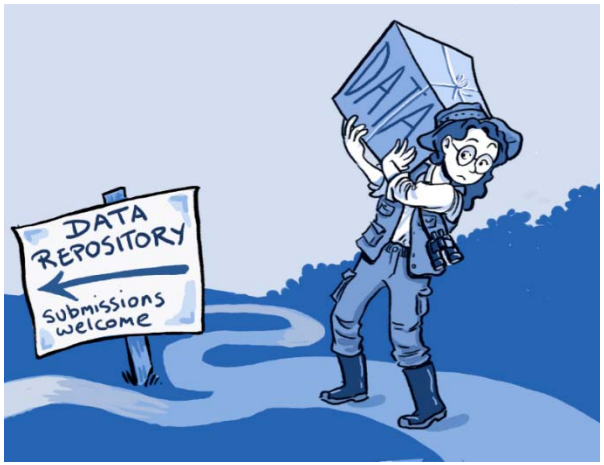
Khosravi, Hossein, and Ehsanollah Kabir. "Introducing a very large dataset of handwritten Farsi digits and a study on their varieties." Pattern Recognition Letters 28.10 (2007): 1133-1141.

Motivation



- Research data is fundamental for science/industry/…
  - Data serves as input for workflows and experiments
  - Data is the source for graphs and visualisations in publications
  - Decisions are based on data
- Data is needed for Reproducibility
  - Repeat experiments
  - Verify / compare results
- Need to provide specific data set
  - Service for data repositories



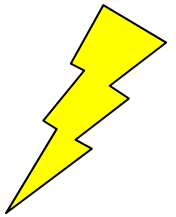


https://commons.wikimedia.org/w/index.php?curid=30978545

1. Put data in data repository,
2. Assign PID (DOI, Ark, URI, …)
3. Make is accessible
   → done!?

Identification of Dynamic Data

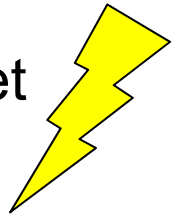
- Usually, datasets have to be static
  - Fixed set of data, no changes:
    no corrections to errors, no new data being added
- But: (research) data is **dynamic**
  - Adding new data, correcting errors, enhancing data quality, …
  - Changes sometimes highly dynamic, at irregular intervals
- Current approaches
  - Identifying entire data stream, without any versioning
  - Using “accessed at” date
  - “Artificial” versioning by identifying batches of data (e.g. annual), aggregating changes into releases (time-delayed!)



- Would like to identify precisely the **data as it existed at a specific point in time**

Granularity of Subsets

- What about the **granularity** of data to be identified?
  - Enormous amounts of CSV data
  - Researchers use specific subsets of data
  - Need to identify precisely the subset used
- Current approaches
  - Storing a copy of subset as used in study -> scalability
  - Citing entire dataset, providing textual description of subset -> imprecise (ambiguity)
  - Storing list of record identifiers in subset -> scalability, not for arbitrary subsets (e.g. when not entire record selected)



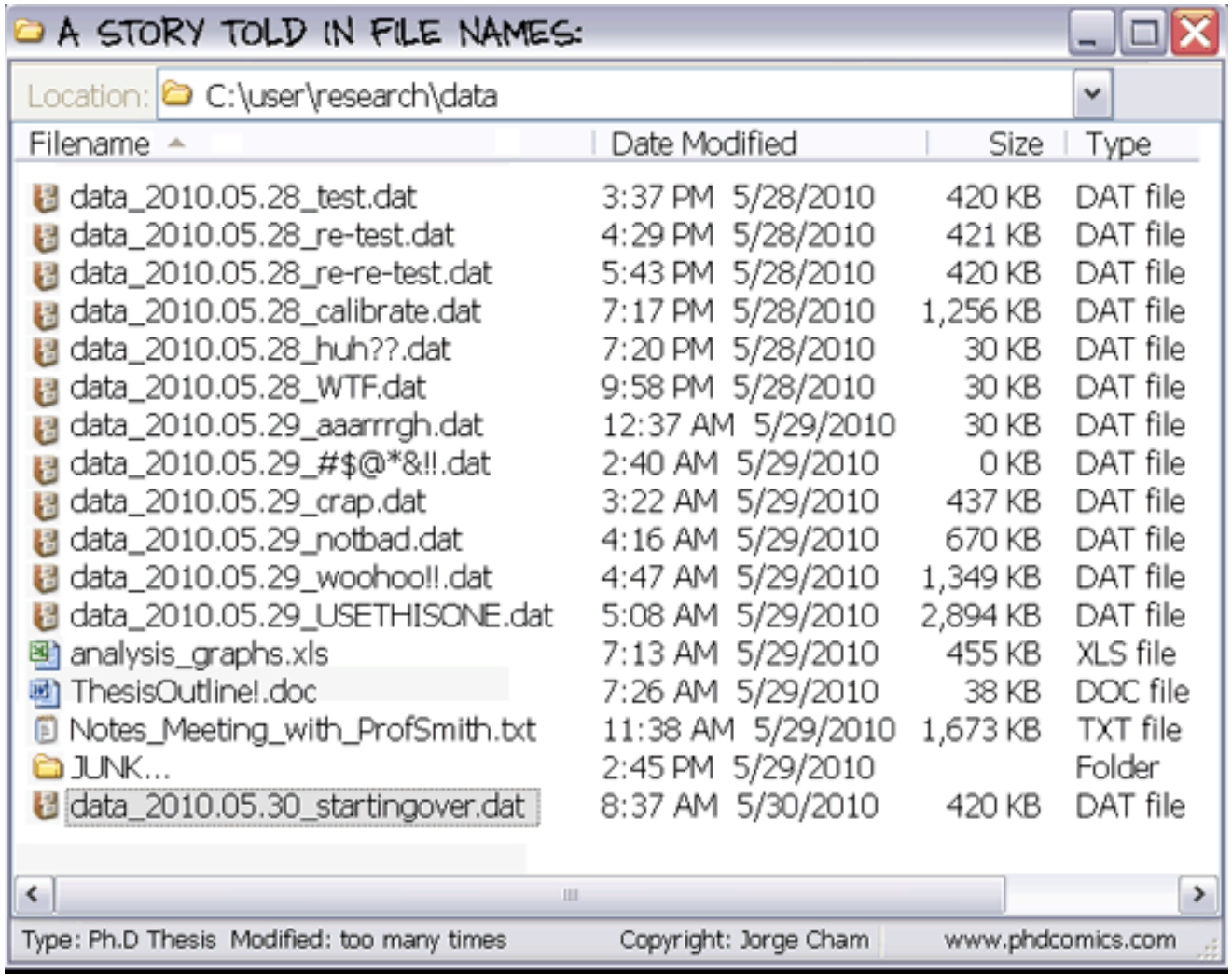
- Would like to be able to identify precisely the **subset of (dynamic) data used** in a process

Data Citation – Requirements

- Dynamic data
  - corrections, additions, …
- Arbitrary subsets of data (granularity)
  - rows/columns, time sequences, …
  - from single number to the entire set
- Stable across technology changes
  - e.g. migration to new database
- Machine-actionable
  - not just machine-readable,
    definitely not just human-readable and interpretable
- Scalable to very large / highly dynamic datasets
  - But: should also work for small and/or static datasets!

What we do NOT want…

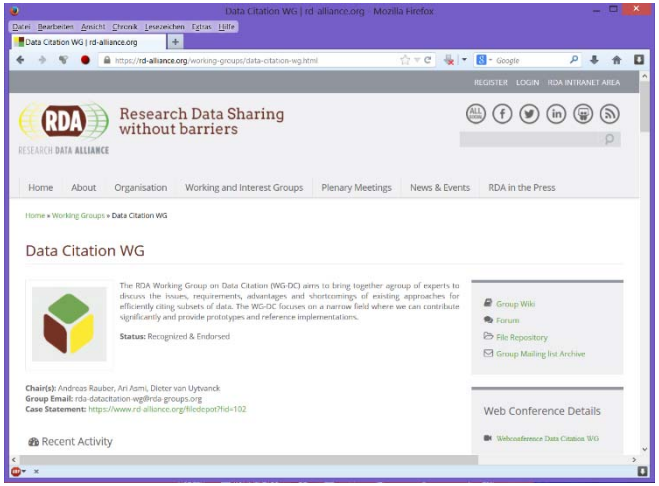
- Common approaches to data management…
  (from PhD Comics: A Story Told in File Names, 28.5.2010)
  Source: http://www.phdcomics.com/comics.php?f=1323



Outline

- Why should we want to cite data?
- What identifier system should I use?
- What are the challenges in data identification and citation?
- How should we do it, according to the RDA WG?
- Who is doing it so far, and how?
- Summary

RDA WG Data Citation



- Research Data Alliance
- WG on **Data Citation: Making Dynamic Data Citeable**
- March 2014 – September 2015
  - Concentrating on the problems of **large, dynamic (changing) datasets**
- Final version presented Sep 2015 at P7 in Paris, France
- Endorsed September 2016 at P8 in Denver, CO
- Since then: supporting adopters

https://www.rd-alliance.org/groups/data-citation-wg.html

RDA WGDC - Solution

- **We have**
  - Data & some means of access („query“)

Dynamic Data Citation

We have: Data + Means-of-access

Dynamic Data Citation:
Cite (dynamic) data dynamically via query!

Dynamic Data Citation

We have: Data + Means-of-access

Dynamic Data Citation:
Cite (dynamic) data dynamically via query!

Steps:

1. Data → versioned (history, with time-stamps)

Dynamic Data Citation

We have: Data + Means-of-access

Dynamic Data Citation:
Cite (dynamic) data dynamically via query!

Steps:

1. Data → versioned (history, with time-stamps)

Researcher creates working-set via some interface:

Dynamic Data Citation

We have: Data + Means-of-access

Dynamic Data Citation:
Cite (dynamic) data dynamically via query!

Steps:

1. Data → versioned (history, with time-stamps)

Researcher creates working-set via some interface:

2. Access → **store & assign PID to "QUERY"**, enhanced with
   - **Time-stamping** for re-execution against versioned DB
   - **Re-writing** for normalization, unique-sort, mapping to history
   - **Hashing** result-set: verifying identity/correctness

   leading to landing page

S. Pröll, A. Rauber. **Scalable Data Citation in Dynamic Large Databases: Model and Reference Implementation.** In IEEE Intl. Conf. on Big Data 2013 (IEEE BigData2013), 2013
http://www.ifs.tuwien.ac.at/~andi/publications/pdf/pro_ieeebigdata13.pdf

Data Citation – Deployment

- Researcher uses workbench to identify subset of data
- Upon executing selection („download“) user gets
  - Data (package, access API, …)
  - PID (e.g. DOI) (Query is time-stamped and stored)
  - Hash value computed over the data for local storage
  - Recommended citation text (e.g. BibTeX)
- PID resolves to landing page
  - Provides detailed metadata, link to parent data set, subset,…
  - Option to retrieve original data OR current version OR changes
- Upon activating PID associated with a data citation
  - Query is re-executed against time-stamped and versioned DB
  - Results as above are returned
- Query store aggregates data usage

Data Citation – Deployment

Note: query string provides excellent provenance information on the data set!

- ubset of data
- er gets
  - Data (package, acce s API, …)
  - PID (e.g. DOI) (Que is time-stamped and stored)
  - Hash value compute over the data for local storage
  - Recommended citati n text (e.g. BibTeX)
- PID resolves to land g page
  - Provides detailed metadata, link to parent data set, subset,…
  - Option to retrieve original data OR current version OR changes
- Upon activating PID associated with a data citation
  - Query is re-executed against time-stamped and versioned DB
  - Results as above are returned
- Query store aggregates data usage

Data Citation – Deployment

- Note: query string provides excellent provenance information on the data set!
- ubset of data
- er gets
  - Data (pac
  - PID (e.g.
  - Hash valu
  - Recommended citat[illegible] text (e.g. BibTeX)
- This is an important advantage over traditional approaches relying on, e.g. storing a list of identifiers/DB dump!!!
- PID resolves to land[illegible]g page
  - Provides detailed metadata, link to parent data set, subset,…
  - Option to retrieve original data OR current version OR changes
- Upon activating PID associated with a data citation
  - Query is re-executed against time-stamped and versioned DB
  - Results as above are returned
- Query store aggregates data usage

Data Citation – Deployment

- Note: query string provides excellent provenance information on the data set!
- ubset of data
- er gets
  - Data (pac
  - PID (e.g.
  - Hash valu
  - Recommended citat text (e.g. BibTeX)
- This is an important advantage over traditional approaches relying on, e.g. storing a list of identifiers/DB dump!!!
- PID resolves
  - Provides det
  - Option to ret
- Identify which parts of the data are used. If data changes, identify which queries (studies) are affected
- Upon activating PID associated w h a data citation
  - Query is re-executed against time-st mped and versioned DB
  - Results as above are returned
- Query store aggregates data usage

Data Citation – Output

- 14 Recommendations grouped into 4 phases:
  - Preparing data and query store
  - Persistently identifying specific data sets
  - Resolving PIDs
  - Upon modifications to the data infrastructure
- 2-page flyer
  https://rd-alliance.org/recommendations-working-group-data-citation-revision-oct-20-2015.html
- More detailed report: Bulletin of IEEE TCDL 2016
  http://www.ieee-tcdl.org/Bulletin/v12n1/papers/IEEE-TCDL-DC-2016_paper_1.pdf

Data Citation – Recommendations

Preparing Data & Query Store

- R1 – Data Versioning
- R2 – Timestamping
- R3 – Query Store

When Resolving a PID

- R11 – Landing Page
- R12 – Machine Actionability

When Data should be persisted

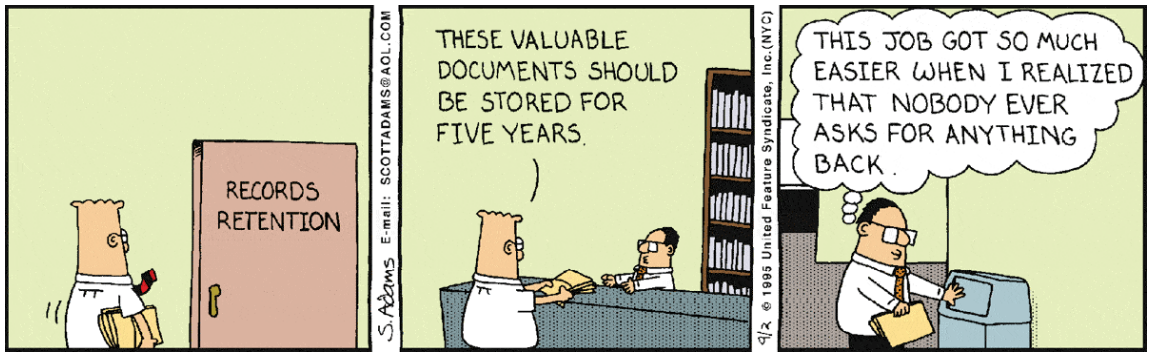
- R4 – Query Uniqueness
- R5 – Stable Sorting
- R6 – Result Set Verification
- R7 – Query Timestamping
- R8 – Query PID
- R9 – Store Query
- R10 – Citation Text

Upon Modifications to the Data Infrastructure

- R13 – Technology Migration
- R14 – Migration Verification

R1: Data Versioning

- **Apply versioning to ensure earlier states of the data can be retrieved**
- Versioning allows tracing the changes
  (static data: no changes – principle still applies)
- No in-place updates or deletes
  - Mark record as deleted, re-insert new record instead of update
  - Keep old versions – only way to be able to “go back”
- Do we really need to keep everything?
  - (“*changes that were never read never existed*”)



Src: http://dilbert.com/strip/1995-09-02

R2: Data Timestamping

- **Ensure that operations on data are timestamped, i.e. any additions, deletions are marked with a timestamp**
- Timestamping is closely related to versioning
- Granularity depends on
  - Change frequency / tracking requirements
    - Per individual operation
    - Batch-operations
    - Grouped in-between read accesses ("*changes that were never read do not matter*")
  - System (data storage, databases)
    - e.g. FAT 2 seconds, NTFS 100 ns, EXT4 1 ns

https://www-03.ibm.com/ibm/history/exhibits/cc/cc_T30.html

R1 & R2: Versioning / Timestamping

Note:

- R1 & R2 are already pretty much standard in many (RDBMS-) research databases
- Different ways to implement, depending on
  - data type / data structure: RDBMS, CSV, XML, LOD, …
  - data volume
  - amount and type of changes
  - number of APIs, flexibility to change them
- Distributed settings:
  - synchronized clocks, or:
  - each node keeps individual, local time
    time-stamps for distributed queries based on local times
    these local times are stored at the query store aggregating the results

Timestamping vs. Semantic Versioning

Why timestamps, why not semantic versioning

- Some prefer to use semantic versioning (minor/major updates that do not / do change behaviour/interface)
  - Advantage: version number indicates relationship btw. versions
  - Disadvantage:
    - Something that was expected to be a not-changing update may turn out to induce changes / side-effects later-on
    - With data, “minor” updates are hard to think of: changing a typo may result in a record being found / not found by a query, encoding changes may break subsequent processing pipelines
    - Different semantics / types of use across different communities
- Recommendation
  - No semantics in identifier (mantra!)
  - Keep identification (version timestamp) and semantics separate
  - Semantic version number in addition to timestamp

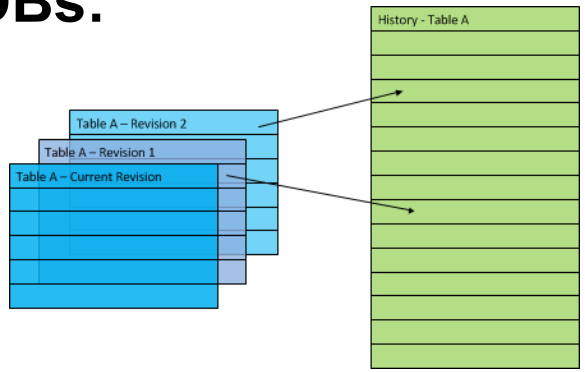
Data Versioning (cont.)

Semantic Versioning

- Semantic versions are “only” **assertions on states of the data at certain points in time**, eg
  - Data may be transient / still undergoing changes, whereas after a certain points in time it has reached a state where no further changes are expected
  - Certain states of data may not be intended for permanent retention, whereas others may have guarantees of availability over time
- Assertions specified as tags associated to queries, e.g.
  - Query *“Select * FROM <table> WHERE timestamp_added < ts1 and ts_deleted >ts1”* may carry the assertions *“status: not expected to change”* and *“availability: 7 years” (preferably from controlled vocabularies)*
- *Subset queries are “nested queries” on such “stable versions”*

R1 & R2: Versioning / Timestamping

Implementation options for e.g. relational DBs:

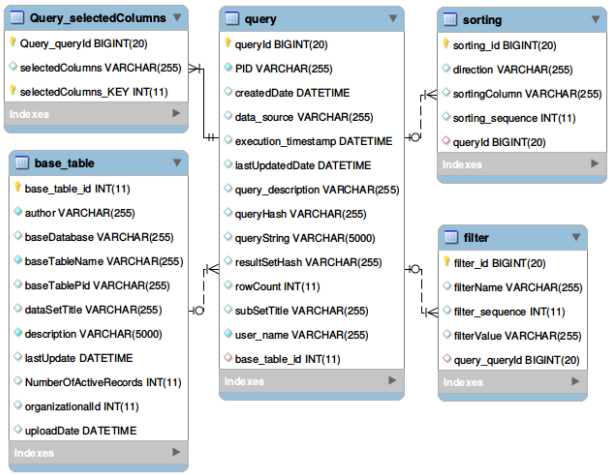


- History Table
  - Utilizes full history table
  - Also inserts reflected in history table
  - Doubles storage space, no API adaptions
- Integrated
  - Extend original tables by temporal metadata
  - Expand primary key by timestamp/version column
  - Minimal storage footprint, changes to all APIs
- Hybrid
  - Utilize history table for deleted record versions with metadata
  - Original table reflects latest version only
  - Minimal storage footprint, some API change, expensive query re-writes

- Solution to be adopted depends on trade-off
  - Storage Demand
  - Query Complexity
  - Software/API adaption

R3: Query Store

- **Provide means for storing queries and the associated metadata in order to re-execute them.**
- Approach is based upon queries.
  - Therefore we need to preserve the queries
    - Original and re-written (**R4, R5**), potentially migrated (**R13**)
  - Query parameters and system settings
  - Execution metadata
  - Hash keys (multiple, if re-written) (**R4, R6**)
  - **Persistent identifier(s) (R8)**
  - Citation text (**R10**) …
- Comparatively small,
  even for high query volumes



R4: Query Uniqueness

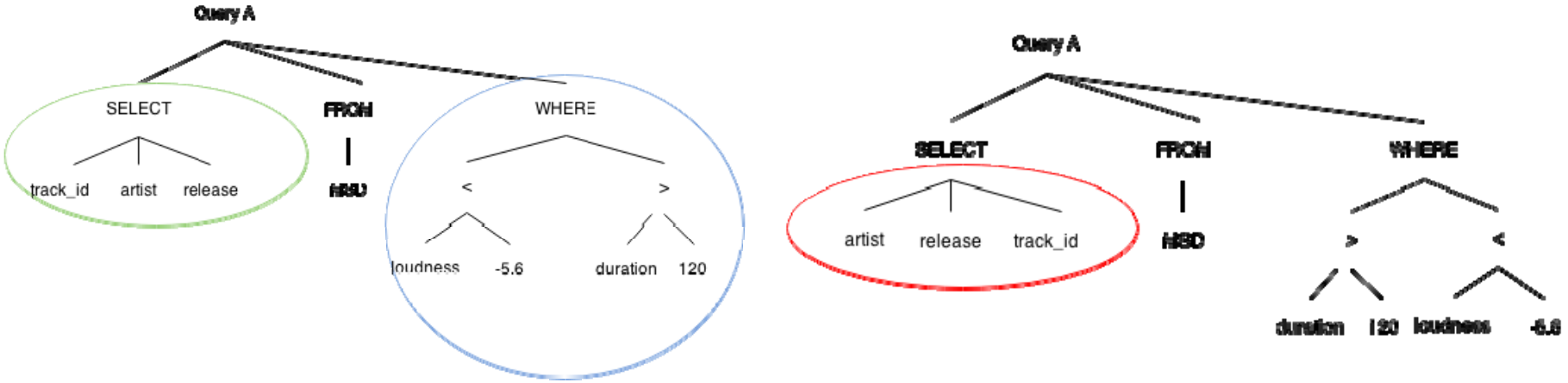
- **Re-write the query to a normalized form so that identical queries can be detected.**
  Compute checksum of the normalized query to efficiently detect identical queries
- Detecting identical queries can be challenging
  - Query semantics can be expressed in different ways
  - Different queries can deliver identical results
  - Interfaces can be used for maintaining a stable query structure
- Best effort, no perfect solution
- Usually not a problem if queries generated via standardized interfaces, e.g. workbench – optional!
- Worst case: two PIDs for semantically equivalent queries

R4: Query Uniqueness

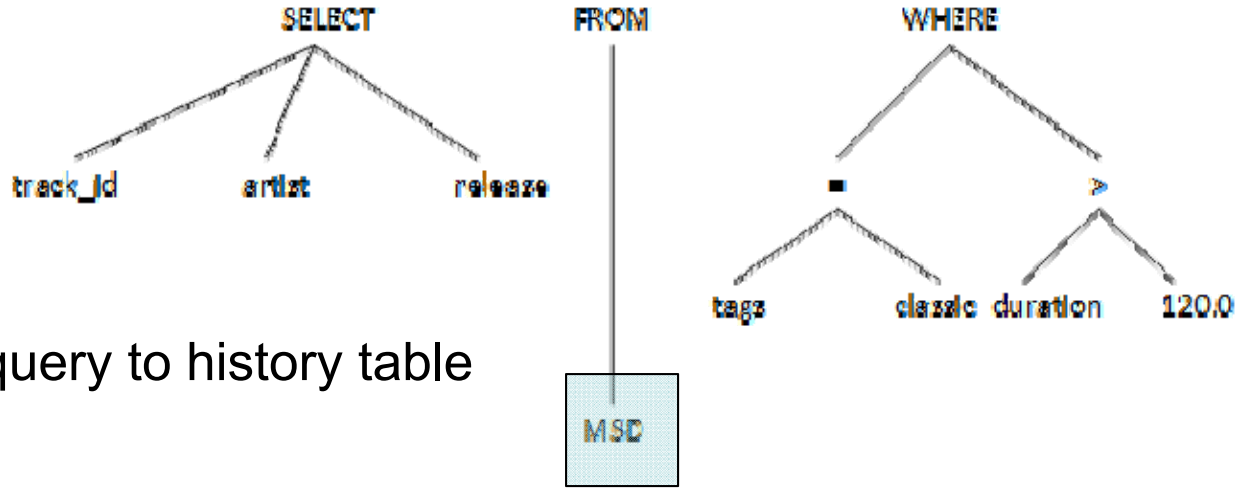
- Query re-writing needed to
  - **Standardization/Normalization** of query to help with identifying semantically identical queries
    - upper/lower case spelling, sorting of filter parameters, …
  - Re-write to **adapt to versioning approach** chosen (versioning in operational tables, separate history table, …), e.g. **identify last change to result set touched** upon (i.e. select including elements marked deleted, check most recent timestamp, to determine correct PID assignment)
  - **Add timestamp (t-Δt)** to any select statement in query
  - **Apply unique sort** to any table touched upon in query prior to query to ensure unique sort (see **R5**)

R4: Query Uniqueness

- Normalizing queries to detect identical queries
  - WHERE clause sorted
  - Calculate query string hash
  - Identify semantically identical queries
  - → non-identical queries: columns in different order



R4: Query Uniqueness

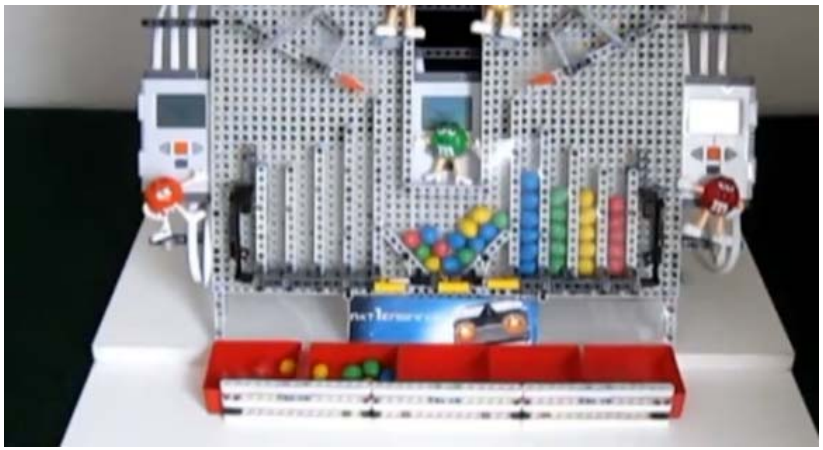


- Adapt query to history table

```
SELECT results.track_id, results.artist, results.release
        FROM MSD AS results JOIN (
                SELECT track_id, max(timestamp) AS latestTimestamp
                FROM MSD
                WHERE timestamp <= (SELECT @queryExecutionTimestamp)
                AND (track_id NOT IN
                        (SELECT track_id FROM MSD AS deletedRecords
                                WHERE deletedRecords.status_mark = 'deleted'
                                AND (deletedRecords.timestamp < @queryExecutionTimestamp))
                        )
        GROUP BY track_id
) AS version ON results.track_id = version.track_id AND results.timestamp = version.latestTimestamp

WHERE
    results.tags = 'classic'  AND results.duration> 120
ORDER BY results.track_id;
```

R5: Stable Sorting

- **Ensure that the sorting of the records in the data set is unambiguous and reproducible**
- The sequence of the results in the result set may not be fixed, but data processing results may depend on sequence
  - Many databases are set based
  - The storage system may use non-deterministic features
- If this needs to be addressed, apply default sort (on id) prior to any user-defined sort
- Optional!



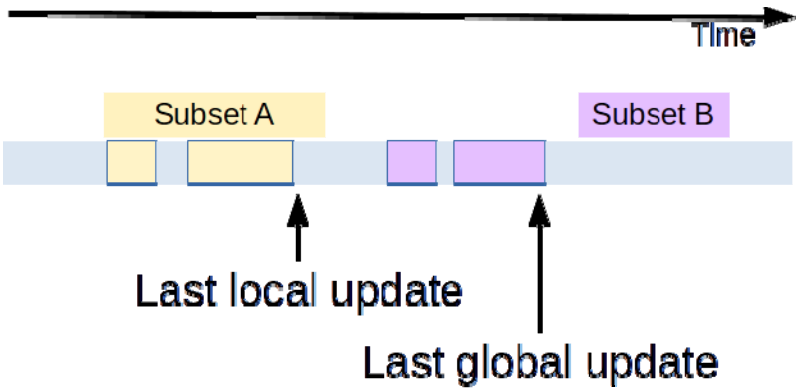
http://www.geek.com/

R6: Result Set Verification

- **Compute fixity information (also referred to as checksum or hash key) of the query result set to enable verification of the correctness of a result upon re-execution.**
- Correctness:
  - No record has changed within a data subset
  - All records which have been in the original data set are also in the re-generated data set
- Compute a hash key
  - Allows to compare the completeness of results
  - For extremely large result sets:
    potentially limit hash input data,
    e.g. only row headers + record id's

R7: Query Timestamping

- **Assign a timestamp to the query based on the last update to the entire database (or the last update to the selection of data affected by the query or the query execution time).**
- Allows to map the execution of a query to a state of the database
  - Execution time: default solution, simple, potentially privacy concerns?
  - Last global update: simple, **recommended**
  - Last update to affected subset: complex to implement
- All equivalent in functionality! (transparent to user)



R8: Query PID

- **Assign a new PID to the query if either the query is new or if the result set returned from an earlier identical query is different due to changes in the data. Otherwise, return the existing PID of the earlier query to the user.**
- **Existing PID**: Identical query (semantics) with identical result set, i.e. no change to any element touched upon by query since first processing of the query
- **New PID**: whenever query semantics is not absolutely identical
  (irrespective of result set being potentially identical!)

R8: Query PID

- Note:
  - Identical result set alone does not mean that the query semantics is identical
  - Will assign different PIDs to capture query semantics
  - Need to normalize query to allow comparison
- Process:
  - Re-write query to adapt to versioning system, stable sorting, …
  - Determine query hash
  - Execute user query and determine result set hash
  - Check query store for queries with identical query hash
    - If found, check for identical result set hash
- 2 PIDs: (compare e.g. paper in journal)
  - precise subset of (static) data, as an excerpt of
  - a larger, dynamically evolving data stream

R9: Store the Query

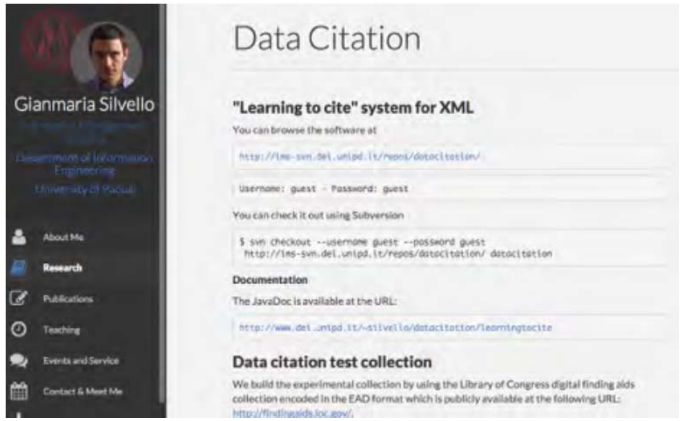
- **Store query and metadata (e.g. PID, original and normalised query, query and result set checksum, timestamp, superset PID, data set description, and other) in the query store.**
  - Query store is central infrastructure
  - Stores query details for long term
  - Provides information even when the data should be gone
  - Responsible for re-execution
  - Holds data for landing pages
  - Stores sensitive information
- Not necessarily ALL queries (staging area)

R10: Create Citation Texts

- **Generate citation texts in the format prevalent in the designated community for lowering the barrier for citing and sharing the data.**
  Include the PID in the citation text snippet.
- Researchers are “lazy”/efficient
  - Support citing by allow them to copy and paste citations for data
  - Citations contain text including PIDs and timestamps
  - Adapted for each community
- **2 PIDs!**
  - Superset: the “database” and it’s holder (repository, data center)
  - Subset: based on the query
  - Accumulate credits for subset and (dynamic) data collection/holder

Suggested citation text:	Stefan Proell (2015) "Austria Facts" created at 2015-10-07 10:51:55.0, PID [ark:12345/qmZi2wO2vv]. Subset of CIA: "The CIA WorldFactbook", PID [ark:12345/cLfH9FjxnA]

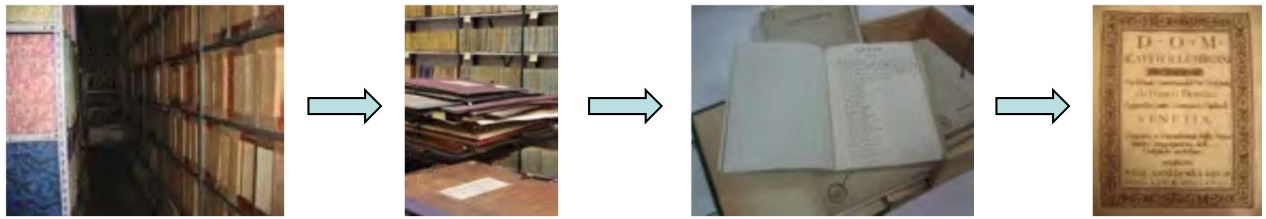
R10: Automated Citation Texts

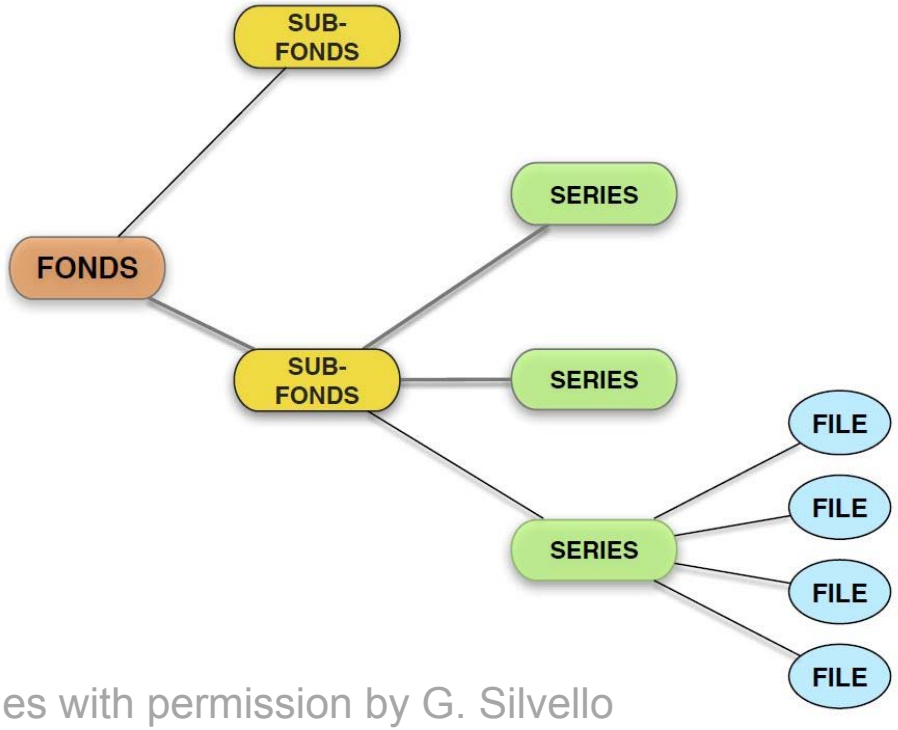


- Can be created automatically
  - relatively simple for relational
  - more complex for hierarchical/XML
- Learning to Cite:
  - Gianmaria Silvello. Learning to Cite Framework: How to Automatically Construct Citations for Hierarchical Data. Journal of the Association for Information Science and Technology (JASIST), Volume 68 issue 6, pp. 1505-1524, June 2017.
  - http://www.dei.unipd.it/~silvello/datacitation

R10: Automated Citation Texts

- EAD: Encoded Archival Description






```
<ead>
    <eadheader>
        […]
    </eadheader>
    <archdesc level="fonds">
        […]
        <did>[…]</did>
            <dsc level="fonds">
                […]
                <c01 level="sub-fonds">
                    […]
                </c01>
                <c01 level="sub-fonds">
                    […]
                    <c02 level="series">
                        […]
                    </c02>
                    <c02 level="series">
                        […]
                    </c02>
                    <c02 level="series">
                        […]
                        <c03 level="file">
                            […]
                        </c03>
                        <c03 level="file">
                            […]
                        </c03>
                        <c03 level="file">
                            […]
                        </c03>
                        <c03 level="file">
                            […]
                        </c03>
                    </c02>
                </c01>
            </dsc>
    </archdesc>
</ead>
```

Slides with permission by G. Silvello

R10: Automated Citation Texts

- A human-readable citation:

Correspondence, 1951-1956,

"The Elements of Legal Theory" (unpublished). Books, box 135. Part II:

Writings (1905-1984), box 129-152. Huntington Cairns Papers.

Manuscript Division, Library of Congress.

http://hdl.loc.gov/loc.mss/eadmss.ms001024

R10: Automated Citation Texts

- A human-readable citation:

Citable unit

```
Correspondence, 1951-1956
```

Contextual Information (from ancestors of the citable unit)

```
"The Elements of Legal Theory" (unpublished). Books, box 135. Part II:

Writings (1905-1984), box 129-152. Huntington Cairns Papers.

Manuscript Division, Library of Congress.
```

```
http://hdl.loc.gov/loc.mss/eadmss.ms001024
```

(Persistent) Unique identifier of the EAD file

R10: Automated Citation Texts

- A machine-readable citation:
  - Conjunction of XML paths

```
/ead/eadheader/eadid && /ead/eadheader/filedesc/publicationstmt/publisher && /ead/
archdesc/did/unittitle && /ead/archdesc/dsc/c01[10]/did/unittitle && /ead/archdesc/
dsc/c01[10]/did/unittitle/unitdate && /ead/archdesc/dsc/c01[10]/did/container/@type
&& /ead/archdesc/dsc/c01[10]/did/container &&  /ead/archdesc/dsc/c01[10]/c02/did/
container/@type && /ead/archdesc/dsc/c01[10]/c02/did/container && /ead/archdesc/dsc/
c01[10]/c02/did/unittitle &&  /ead/archdesc/dsc/c01[10]/c02/c03[4]/did/unittitle
&& /ead/archdesc/dsc/c01[10]/c02/c03[4]/did/container/@type && /ead/archdesc/dsc/
c01[10]/c02/c03[4]/did/container &&  /ead/archdesc/dsc/c01[10]/c02/c03[4]/c04[2]/
did/unittitle && /ead/archdesc/dsc/c01[10]/c02/c03[4]/c04[2]/c05[1]/did/unittitle
```

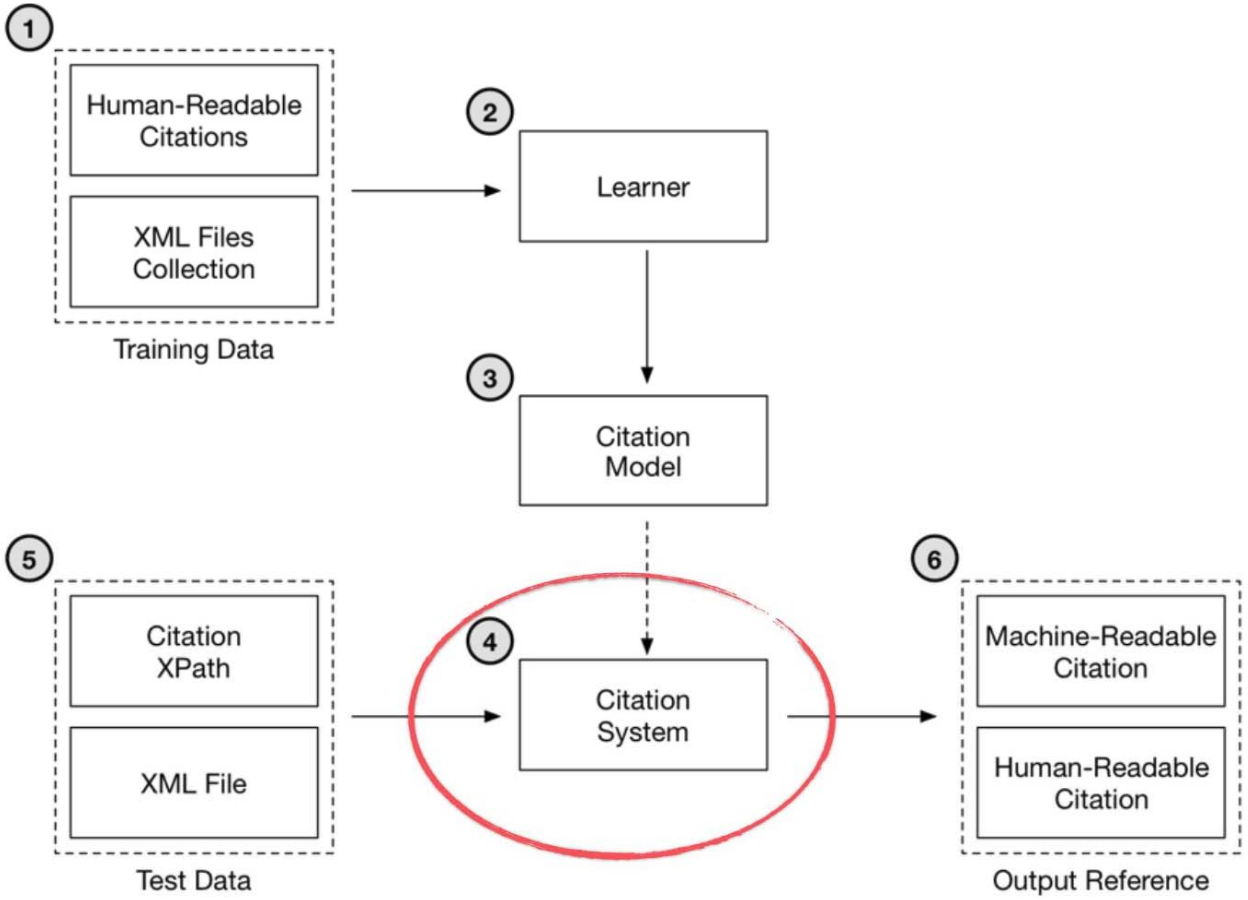
R10: Automated Citation Texts

- Mapping machine-readable to human-readable:

Human-Readable Citation	Machine-Readable Citation
http://hdl.loc.gov/loc.mss/eadmss.ms001024	/ead/eadheader/eadid
Manuscript Division, Library of Congress	/ead/eadheader/filedesc/publicationstmt/publisher
Huntington Cairns Papers	/ead/archdesc/did/unittitle
Part II: Writings	/ead/archdesc/dsc/c01[10]/did/unittitle
1905-1984	/ead/archdesc/dsc/c01[10]/did/unittitle/unitdate
box	/ead/archdesc/dsc/c01[10]/did/container/@type
129-152	/ead/archdesc/dsc/c01[10]/did/container
By Cairns	/ead/archdesc/dsc/c01[10]/c02[1]/did/unittitle
box	/ead/archdesc/dsc/c01[10]/c02[1]/did/container/@type
129	/ead/archdesc/dsc/c01[10]/c02[1]/did/container/
Books	/ead/archdesc/dsc/c01[10]/c02[1]/c03[4]/did/unittitle
box	/ead/archdesc/dsc/c01[10]/c02[1]/c03[4]/did/container/@type
135	/ead/archdesc/dsc/c01[10]/c02[1]/c03[4]/did/container
"The Elements of Legal Theory" (unpublished)	/ead/archdesc/dsc/c01[10]/c02[1]/c03[4]/c04[2]/did/unittitle
Correspondence, 1951-1956	/ead/archdesc/dsc/c01[10]/c02[1]/c03[4]/c04[2]/c05[1]/did/unittitle

R10: Automated Citation Texts

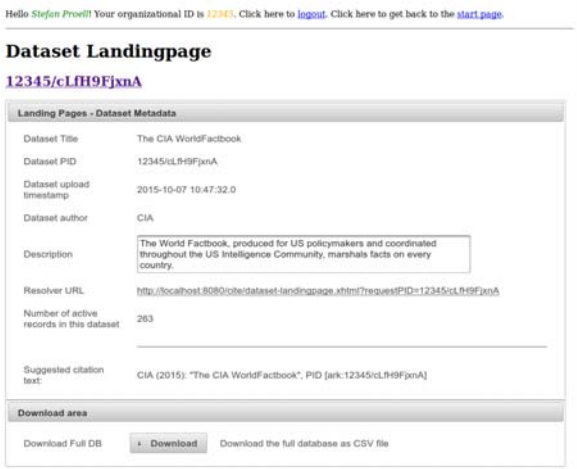
- Learning citation models



Slides with permission by G. Silvello

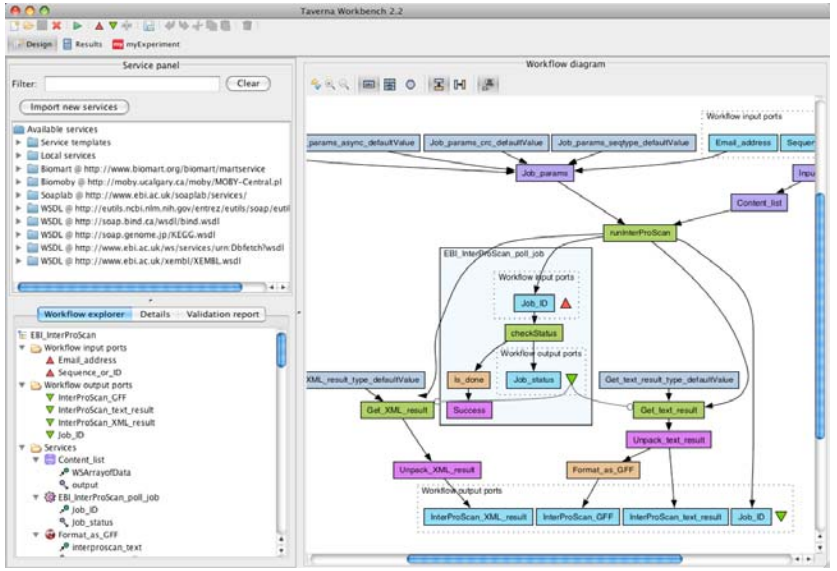
R11: Landing Page

- **Make the PIDs resolve to a human readable landing page that provides the data (via query re-execution) and metadata, including a link to the superset (PID of the data source) and citation text snippet.**
  - Data sets and subsets uniquely identifiable by their PID, which resolves to a human readable landing page.
  - Landing page reachable by a unique URL, presented in a Web browser
  - Not all information needs to be provided on landing page
    (e.g. query strings frequently not relevant / potential security threat)



R12: Machine Actionability

- **Provide an API / machine actionable landing page to access metadata and data via query re-execution.**
  - Experiments are increasingly automated
  - Machines most likely to consume data citations
  - Allows machines to resolve PIDs, access metadata and data
  - Note: does NOT imply full / automatic access to data!
    - Authentication
    - Load analysis
  - Handshake, content negotiation, …
  - Allows automatic meta-studies, monitoring, …



R13: Technology Migration

- **When data is migrated to a new representation (e.g. new database system, a new schema or a completely different technology), migrate also the queries and associated fixity information.**
  - Technology evolves and data may be moved to a new technology stack
  - Query languages change
- Migration required
  - Migrate data and the queries (both are with the data center!)
  - Adapt versioning, re-compute query hash-keys
  - Maybe decide to keep “original” queries in the provenace trace
- Note: such data migrations constitute major projects, usually happen rarely – require all APIs to be adapted, …

R13: Technology Migration

- **Consider e.g. Schema Modification Operators (SMOs)**
  - CREATE TABLE R
  - DROP TABLE R
  - RENAME TABLE R
  - COPY TABLE R INTO S
  - PARTITION TABLE R INTO S with *cond*, T with !*cond*
  - DECOMPOSE TABLE R INTO S(A,B) T(A,C)
  - JOIN TABLE R,S INTO T WHERE cond
  - ADD COLUMN C [as const | func(A)] INTO R
  - DROP COLUMN C FROM R
  - RENAME COLUMN A IN R TO B
- How will the queries need to be re-written to address them?

R14: Migration Verification

- **Verify successful data and query migration, ensuring that queries can be re-executed correctly.**
- Sanity check: After migration is done, verify that the data can still be retrieved correctly
- Use query and result set hashes in the query store to verify results
- If hash function is incompatible/cannot be computed on new system as hash input data sequence cannot be obtained, pairwise comparison of subset elements
  - May constitute new PID / data subset in this case, as subsequent processes will not be able to use it as input if result set presentation has changed, breaks processes

RDA Recommendations - Summary

- Building blocks of supporting dynamic data citation:
  - Uniquely identifiable data records
  - Versioned data, marking changes as insertion/deletion
  - Time stamps of data insertion / deletions
  - “Query language” for constructing subsets
- Add modules:
  - Persistent query store: queries and the timestamp (either: <when issued> or <of last change to data>)
  - Query rewriting module
  - PID assignment for queries that enables access
- Stable across data source migrations (e.g. diff. DBMS), scalable, machine-actionable

RDA Recommendations - Summary

- ***Benefits***
  - Allows **identifying, retrieving and citing the precise data subset** with minimal storage overhead by only storing the versioned data and the queries used for extracting it
  - Allows retrieving the data both **as it existed** at a given point in time as well as the **current view** on it, by re-executing the same query with the stored or current timestamp
  - It allows to cite even an **empty set**!
  - The query stored for identifying data subsets provides valuable **provenance data**
  - Query store collects **information on data usage**, offering a basis for data management decisions
  - **Metadata** such as checksums support the verification of the correctness and **authenticity** of data sets retrieved
  - The same principles work for **all types of data**

RDA Recommendations - Summary

- ***Some considerations and questions***
  - May data be deleted?
    Yes, of course, given appropriate policies. Queries may then not be re-executable against the original timestamp anymore
  - Does the system need to store every query?
    No, only data sets that should be persisted for citation and later re-use need to be stored.
  - Can I obtain only the most recent data set?
    Queries can be re-executed with the original timestamp or with the current timestamp or any other timestamp desired.
  - Which PID system should be used?
    Any PID system can, in principle, be applied according to the institutional policy.

Outline

- Why should we want to cite data?
- What identifier system should I use?
- What are the challenges in data identification and citation?
- How should we do it, according to the RDA WG?
- Who is doing it so far, and how?
- Summary

Reference Implementations & Standards

Reference Implementations

- MySQL – Stefan Pröll
- CSV via MySQL – Stefan Pröll
- CSV-files via GIT – Kristof Meixner
- NoSQL via MongoDB in CKAN – Florian Wörister

Standards

- ISO 690:2010, Information and documentation — Guidelines for bibliographic references and citations to information resources, p.43, Sec. 16.13.5, 4th ed. 2018-09-11 – Juha Hakala
- ESIP: Data Citation Guidelines for Earth Science Data Version 2 (draft) – Mark Parsons

Key Adopters so far…

- Electronic Health Records at the Univ. of Washington in St. Louis – Leslie McIntosh
- Vermont Monitoring Cooperative – James Duncan
- Virtual Atomic and Molecular Data Center (VAMDC) – Carlo Maria Zwölf
- Climate Change Centre Austria (CCCA)
- Open Earth Observation - Earth Observation Data Center (EODC) – Wolfgang Wagner, Bernhard Gößwein
- IPSL (CNRS) – Sebastian Denville

Ongoing Adoption Projects

- Ocean Networks Canada
- Deep Carbon Observatory
- MyHealth MyData
- Smart Data Platform at NICT
- Dendro Data Repository

Pilots / Adopters

- Series of Webinars presenting implementations
  - Recordings, slides, supporting papers
  - https://www.rd-alliance.org/group/data-citation-wg/webconference/webconference-data-citation-wg.html

  - Implementing of the RDA Data Citation Recommendations by the **Climate Change Centre Austria (CCCA) for a repository of NetCDF files**
  - Implementing the RDA Data Citation Recommendations for **Long-Tail Research Data / CSV files**
  - Implementing the RDA Data Citation Recommendations in the **Distributed Infrastructure of the Virtual and Atomic Molecular Data Center (VAMDC)**
  - Implementation of Dynamic Data Citation at the **Vermont Monitoring Cooperative**
  - Adoption of the RDA Data Citation of Evolving Data Recommendation to **Electronic Health Records**

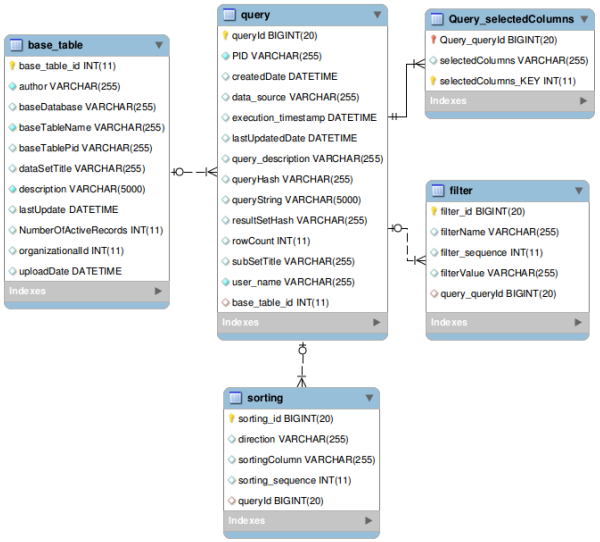
Reference Implementation for CSV Data (and SQL)

Stefan Pröll, SBA
Christoph Meixner, TU Wien

research data sharing without barriers
rd-alliance.org

Large Scale Research Settings

- RDA recommendations implemented in data infrastructures
- Required adaptions
  - Introduce versioning, if not already in place
  - Capture sub-setting process (queries)
  - Implement dedicated query store to store queries
  - A bit of additional functionality (query re-writing, hash functions, …)
- Done! ?
  - “Big data”, database driven
  - Well-defined interfaces
  - Trained experts available
  - “*Complex, only for professional research infrastructures*” ?

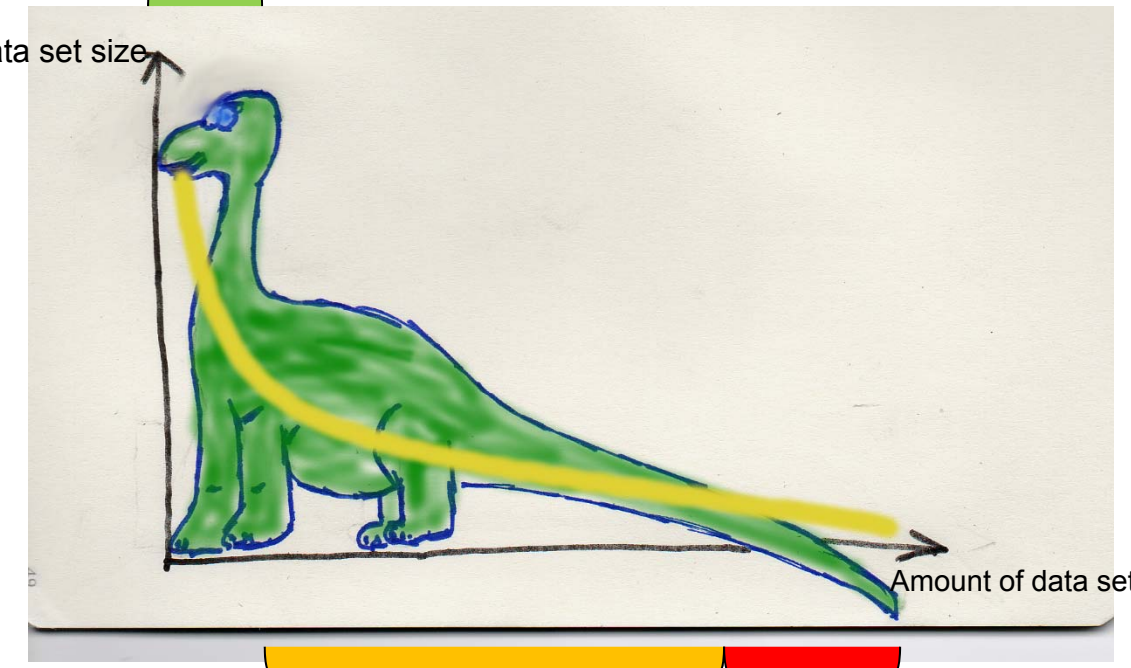


Slide by Stefan Pröll

FACULTY OF !INFORMATICS

Long Tail Research Data

Big data,
well organized,
often used and cited



Less well organized,
non-standardised
no dedicated infrastructure

"Dark data"

Slide by Stefan Pröll

Prototype Implementations

- Solution for small-scale data
  - CSV files, no “expensive” infrastructure, low overhead

2 Reference implementations :

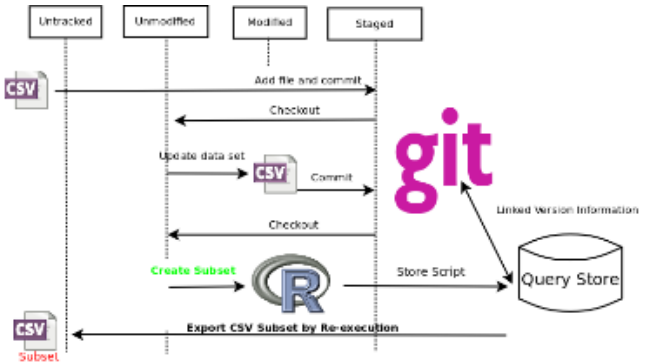
- **Git** based Prototypes: widely used versioning system
  - A) Using separate folders
  - B) Using branches
- **MySQL** based Prototype:
  - C) Migrates CSV data into relational database

- Data backend responsible for versioning data sets
- Subsets are created with scripts or queries via API or Web Interface
- Transparent to user: always CSV

Git-Based Reference Implementation

Git Implementation 1

- Upload CSV files to Git repository (versioning)
- Subsets created via scripting language (e.g. R)
  - Select rows/columns, sort, returns CSV + metadata file
  - Metadata file with script parameters stored in Git
  - (Scripts stored in Git as well)
- PID assigned to metadata file
  - Use Git to retrieve proper data set version and re-execute script on retrieved file

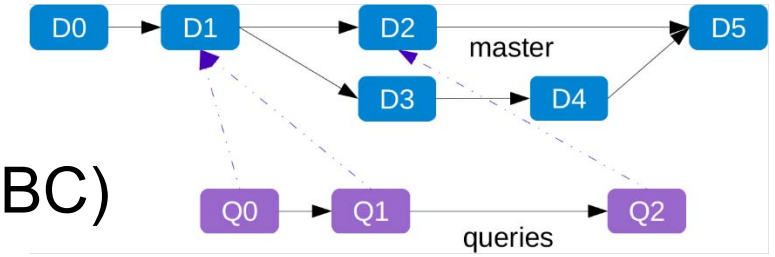



```
# PID=1234/abcdefgh
# Repository_Path=/media/Data/Git-Repository
# Execution_Time=2015-09-30:11:07:09
# Subset Tool=R scripting front-end version 3.2.2 (2015-08-14)
# Subset_Tool_Path=/usr/bin/Rscript
# Input_Script_Path=supercomputing/top5-script.r
# Input_Script_Hash=bef5d...d7861:supercomputing/top5-script.r
# Dataset Path=supercomputing/supercomputer.csv
# Dataset_Commit_Hash=acaed...4cf9c:supercomputer.csv
# Output_Path=/tmp/supercomputer-top5.csv

# Original execution:
#   /usr/bin/Rscript supercomputing/top5-script.r \
#   /media/Data/Git-repository/supercomputing/supercomputer.csv
#   /tmp/supercomputer-top5.csv

# Recommended re-execution
# Retrieve script
git --git-dir=/media/Data/Git-Repository/.git/ \
show bef5d...d7861:supercomputing/top5-script.r \
> /tmp/reproduced-datasets/top5-script.r
# Retrieve data set
git --git-dir=/media/Data/Git-Repository/.git/ \
show 47bed...b9792:supercomputing/supercomputer.csv \
> /tmp/reproduced-datasets/supercomputer.csv
# Rexecute
/usr/bin/Rscript supercomputing/top5-script.r \
/tmp/reproduced-datasets/supercomputer.csv \
/tmp/reproduced-datasets/supercomputer-top5.csv
```

Git-Based Reference Implementation

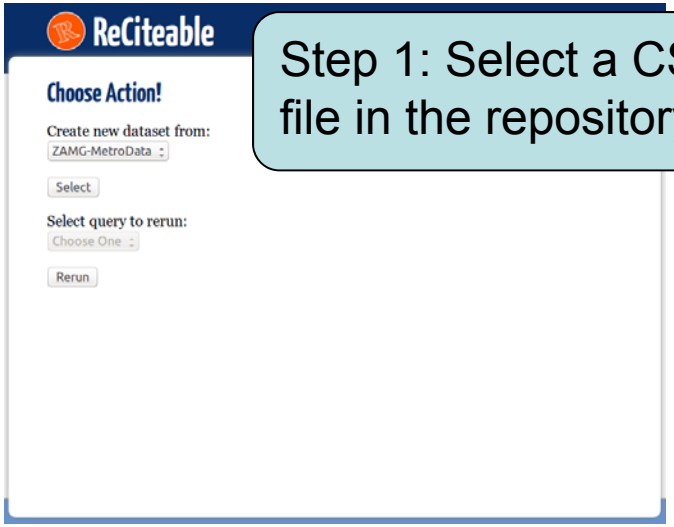
Git Implementation 2

- Addresses issues
  - common commit history, branching data
- Using Git branching model:
  Orphaned branches for queries and data
  - Keeps commit history clean
  - Allows merging of data files
- Web interface for queries (CSV2JDBC)
- Use commit hash for identification
  - Assigned PID hashed with SHA1
  - Use hash of PID as filename (ensure permissible characters)

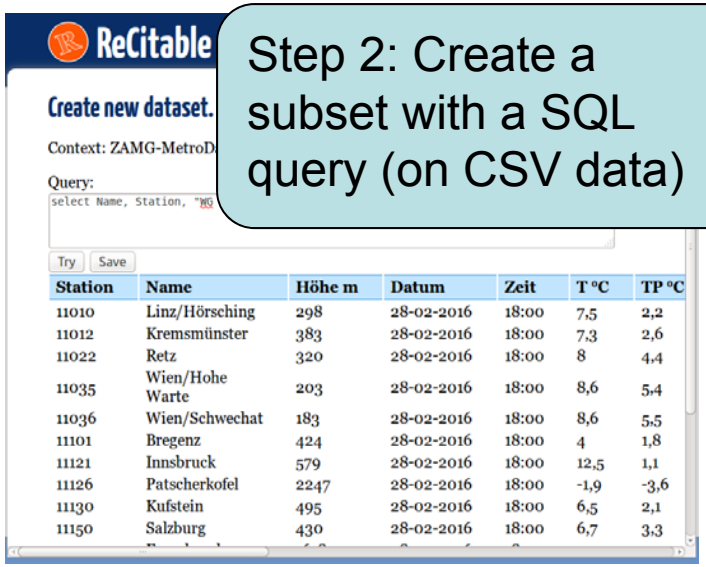


Git-Based Reference Implementation

- Prototype: https://github.com/Mercynary/recitable

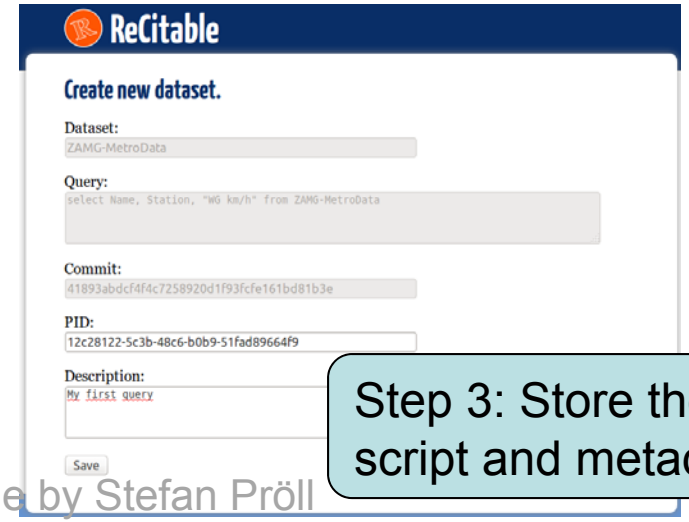


Step 1: Select a CSV file in the repository

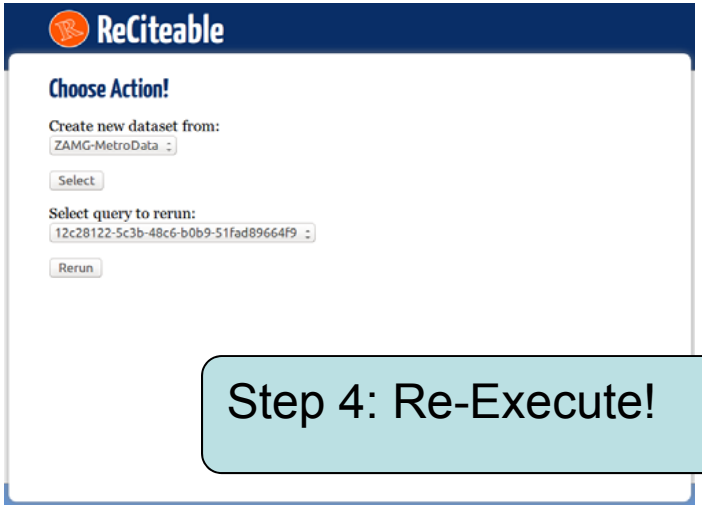


Station	Name	Höhe m	Datum	Zeit	T °C	TP °C
11010	Linz/Hörsching	298	28-02-2016	18:00	7,5	2,2
11012	Kremsmünster	383	28-02-2016	18:00	7,3	2,6
11022	Retz	320	28-02-2016	18:00	8	4,4
11035	Wien/Hohe Warte	203	28-02-2016	18:00	8,6	5,4
11036	Wien/Schwechat	183	28-02-2016	18:00	8,6	5,5
11101	Bregenz	424	28-02-2016	18:00	4	1,8
11121	Innsbruck	579	28-02-2016	18:00	12,5	1,1
11126	Patscherkofel	2247	28-02-2016	18:00	-1,9	-3,6
11130	Kufstein	495	28-02-2016	18:00	6,5	2,1
11150	Salzburg	430	28-02-2016	18:00	6,7	3,3

Step 2: Create a subset with a SQL query (on CSV data)



Step 3: Store the query script and metadata



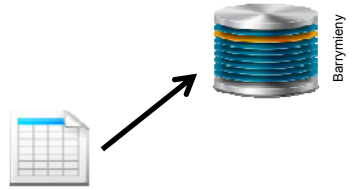
Step 4: Re-Execute!

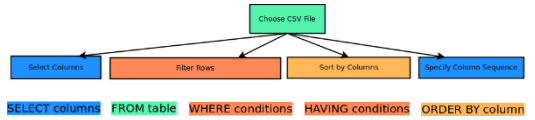
Slide by Stefan Pröll

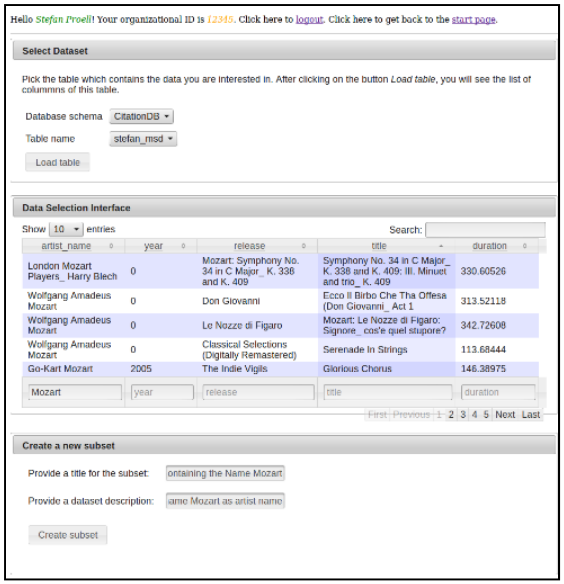
MySQL-Based Reference Implementation

MySQL Prototype

- Data upload
  - User uploads a CSV file into the system
- Data migration from CSV file into RDBMS
  - Generate table structure
  - Add metadata columns (versioning)
  - Add indices (performance)
- Dynamic data
  - Insert, update and delete records
  - Events are recorded with a timestamp
- Subset creation
  - User selects columns, filters and sorts records in web interface
  - System traces the selection process
  - Exports CSV

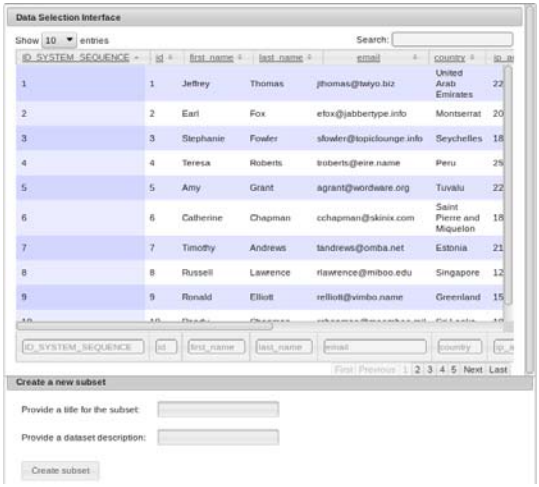






MySQL-Based Reference Implementation

- Source at Github:
  - https://github.com/datascience/RDA-WGDC-CSV-Data-Citation-Prototype
- Videos:
  - Login: https://youtu.be/EnraIwbQfM0
  - Upload: https://youtu.be/xJruifX9E2U
  - Subset: https://www.youtube.com/watch?v=it4sC5vYiZQ
  - Resolver: https://youtu.be/FHsvjsUMiiY
  - Update: https://youtu.be/cMZ0xoZHUyI



CSV Reference Implementations

- Stefan Pröll, Christoph Meixner, Andreas Rauber
  Precise Data Identification Services for Long Tail Research Data.
  Proceedings of the intl. Conference on Preservation of Digital Objects (iPRES2016), Oct. 3-6 2016, Bern, Switzerland.
- Source at Github:
  https://github.com/datascience/RDA-WGDC-CSV-Data-Citation-Prototype
- Videos:
  - Login: https://youtu.be/EnraIwbQfM0
  - Upload: https://youtu.be/xJruifX9E2U
  - Subset: https://www.youtube.com/watch?v=it4sC5vYiZQ
  - Resolver: https://youtu.be/FHsvjsUMiiY
  - Update: https://youtu.be/cMZ0xoZHUyI


Precise Data Identification Services for Long Tail Research Data

WG Data Citation Pilot
CBMI @ WUSTL

Cynthia Hudson Vitale, Leslie McIntosh, Snehil Gupta
Washington University in St.Luis

research data sharing without barriers
rd-alliance.org

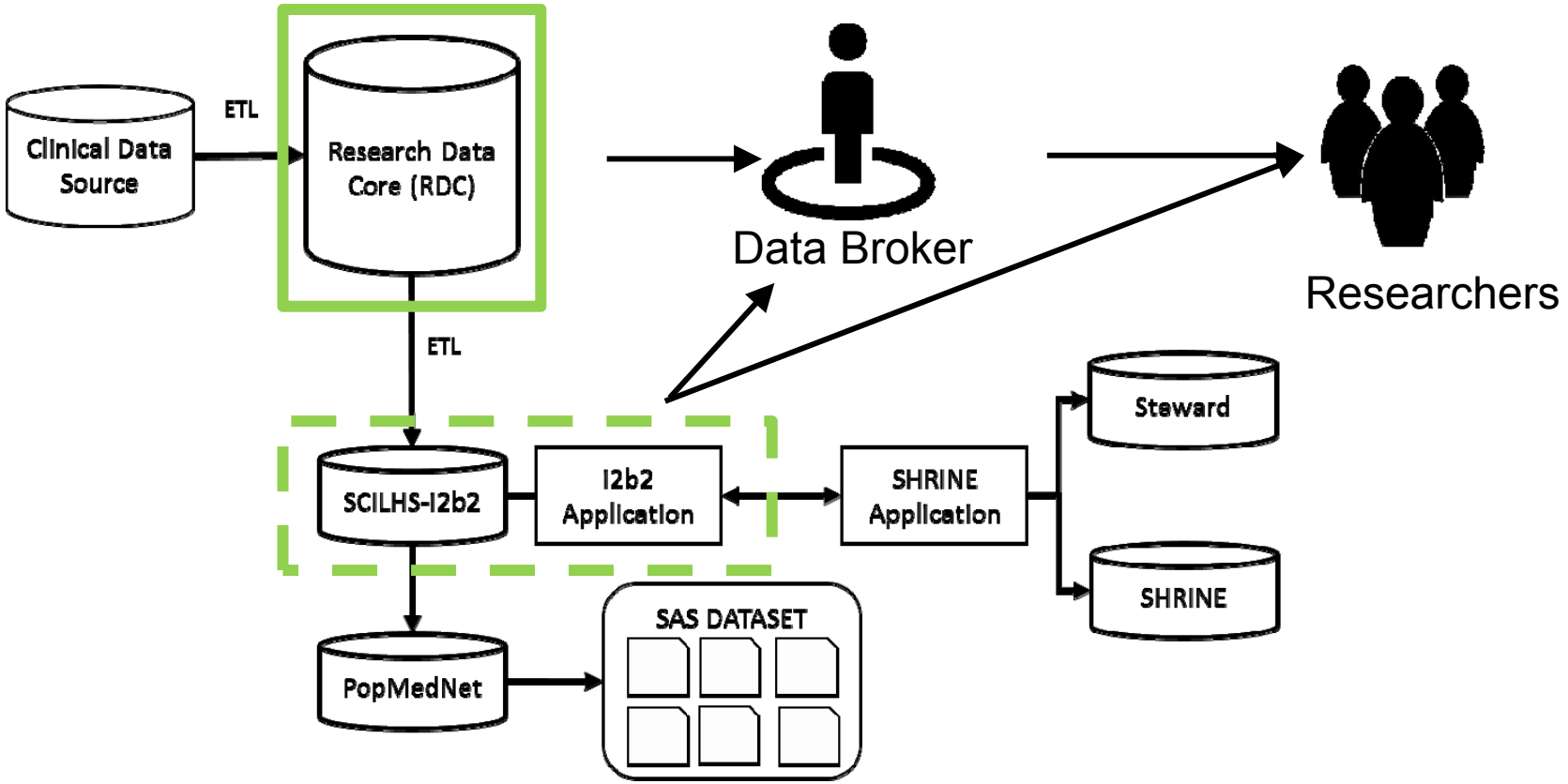
Biomedical Adoption Project Goals

- Implement RDA Data Citation WG recommendation to local Washington U i2b2
- Engage other i2b2 community adoptees
- Contribute source code back to i2b2 community

- Repository
  https://github.com/CBMIWU/Research_Reproducibility
- Slides
  http://bit.ly/2cnWorU
- Bibliography
  https://www.zotero.org/groups/biomedical_informatics_resrepro

Slide by Leslie McIntosh

RDA-MacArthur Grant Focus



Slide by Leslie McIntosh

R1 and R2 Implementation

1

PostgreSQL Extension
“temporal_tables”

2

triggers

RDC.table

c1	c2	c3

sys_period

3

RDC.hist_table*

c1	c2	c3	sys_period

*stores history of
data changes

Return on Investment (ROI) - Estimated

- 20 hours to complete 1 study
- $150/hr (unsubsidized)
- $3000 per study
- 115 research studies per year
- **14 replication studies**

Slide by Leslie McIntosh

CBMI Research Reproducibility Resources

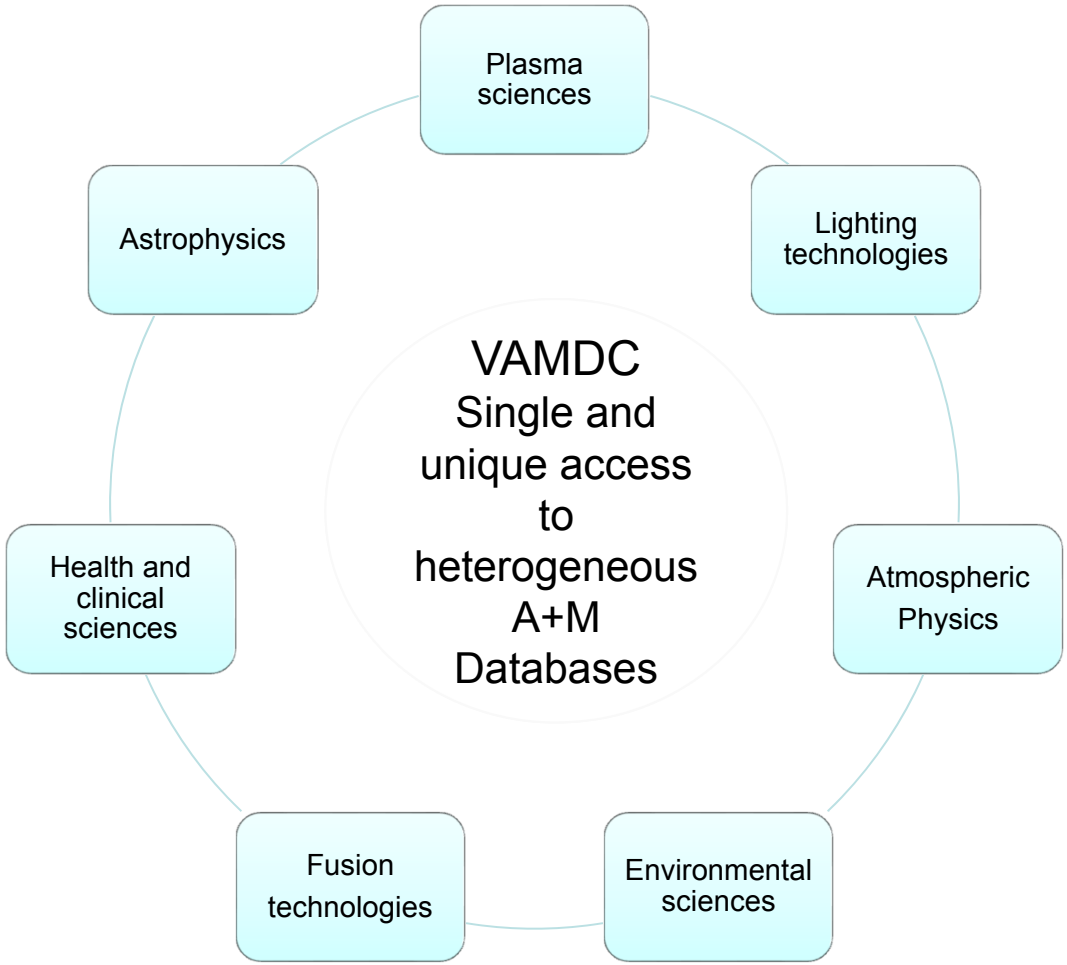
- Repository
  https://github.com/CBMIWU/Research_Reproducibility
- Slides
  http://bit.ly/2cnWorU
- Bibliography
  https://www.zotero.org/groups/biomedical_informatics_resrepro

From RDA Data Citation Recommendations to new paradigms for citing data from VAMDC

C.M. Zwölf and VAMDC Consortium
carlo-maria.zwolf@obspm.fr

research data sharing without barriers
rd-alliance.org

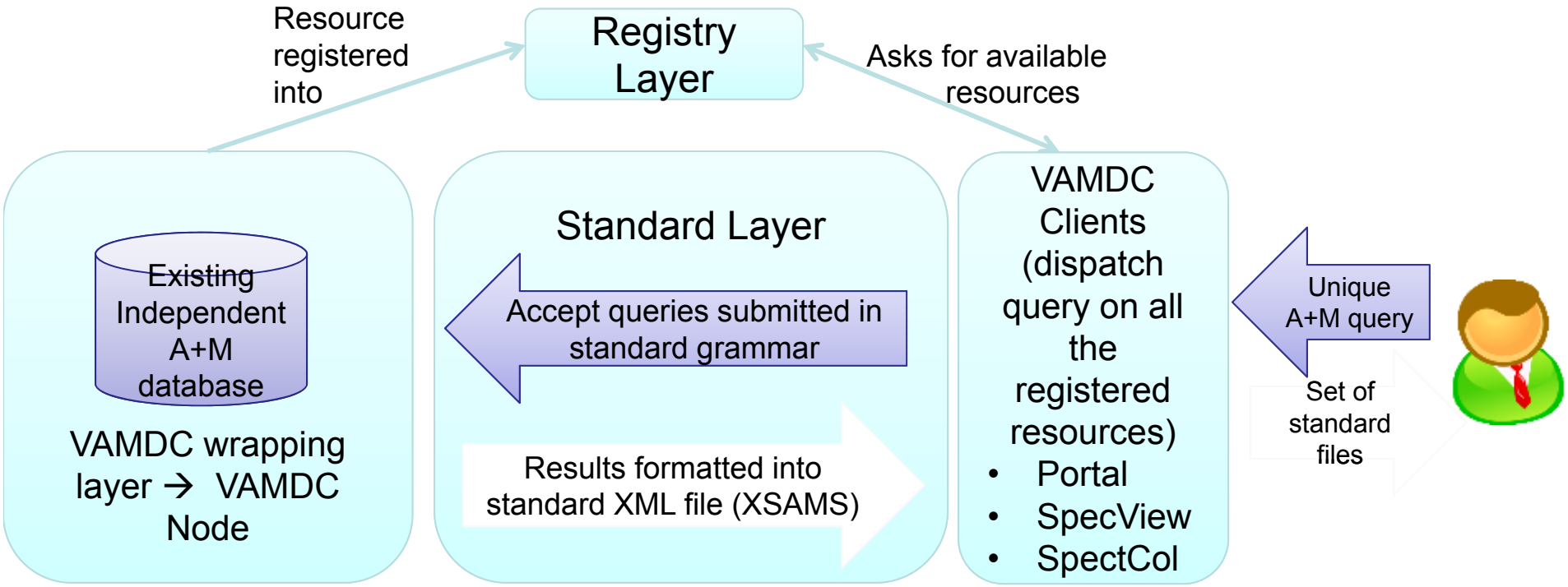
The Virtual Atomic and Molecular Data Centre



- Federates 29 heterogeneous databases http://portal.vamdc.org/
- The "V" of VAMDC stands for Virtual in the sense that the e-infrastructure does not contain data. The infrastructure is a wrapping for exposing in a unified way a set of heterogeneous databases.
- The consortium is politically organized around a Memorandum of understanding (15 international members have signed the MoU, 1 November 2014)
- High quality scientific data come from different Physical/Chemical Communities
- Provides data producers with a large dissemination platform
- Remove bottleneck between data-producers and wide body of users

Slide by Carlo Maria Zwölf

VAMDC Infrastructure Technical Architecture



- VAMDC is agnostic about the local data storage strategy on each node.
- Each node implements the access/query/result protocols.
- There is no central management system.
- Decisions about technical evolutions are made by consensus in Consortium.

➢ It is both technical and political challenging to implement the WG recommendations.

Slide by Carlo Maria Zwölf

Let us implement the recommendation!!

Implementation will be an overlay to the standard / output layer, thus independent from any specific data-node

Tagging versions of data

Two layers mechanisms

1 → Fine grained granularity:
Evolution of XSAMS output standard for tracking data modifications

2 → Coarse grained granularity:
At each data modification to a given data node, the version of the Data-Node changes

With the second mechanism we know that something changed : in other words, we know that the result of an identical query may be different from one version to the other. The detail of which data changed is accessible using the first mechanisms.

Slide by Carlo Maria Zwölf

Let us implement the recommendation!!

Implementation will be an overlay to the standard / output layer, thus independent from any specific data-node

Tagging versions of data	Query Store
Two layers mechanisms	
1 → Fine grained granularity: Evolution of XSAMS output standard for tracking data modifications	Is built over the versioning of Data
2 → Coarse grained granularity: At each data modification to a given data node, the version of the Data-Node changes	Is plugged over the existing VAMDC data-extraction mechanisms
	Due to the distributed VAMDC architecture, the Query Store architecture is similar to a log-service

With the second mechanism we know that something changed : in other words, we know that the result of an identical query may be different from one version to the other. The detail of which data changed is accessible using the first mechanisms

Slide by Carlo Maria Zwölf

Data-Versioning: Overview of the fine grained mechanisms

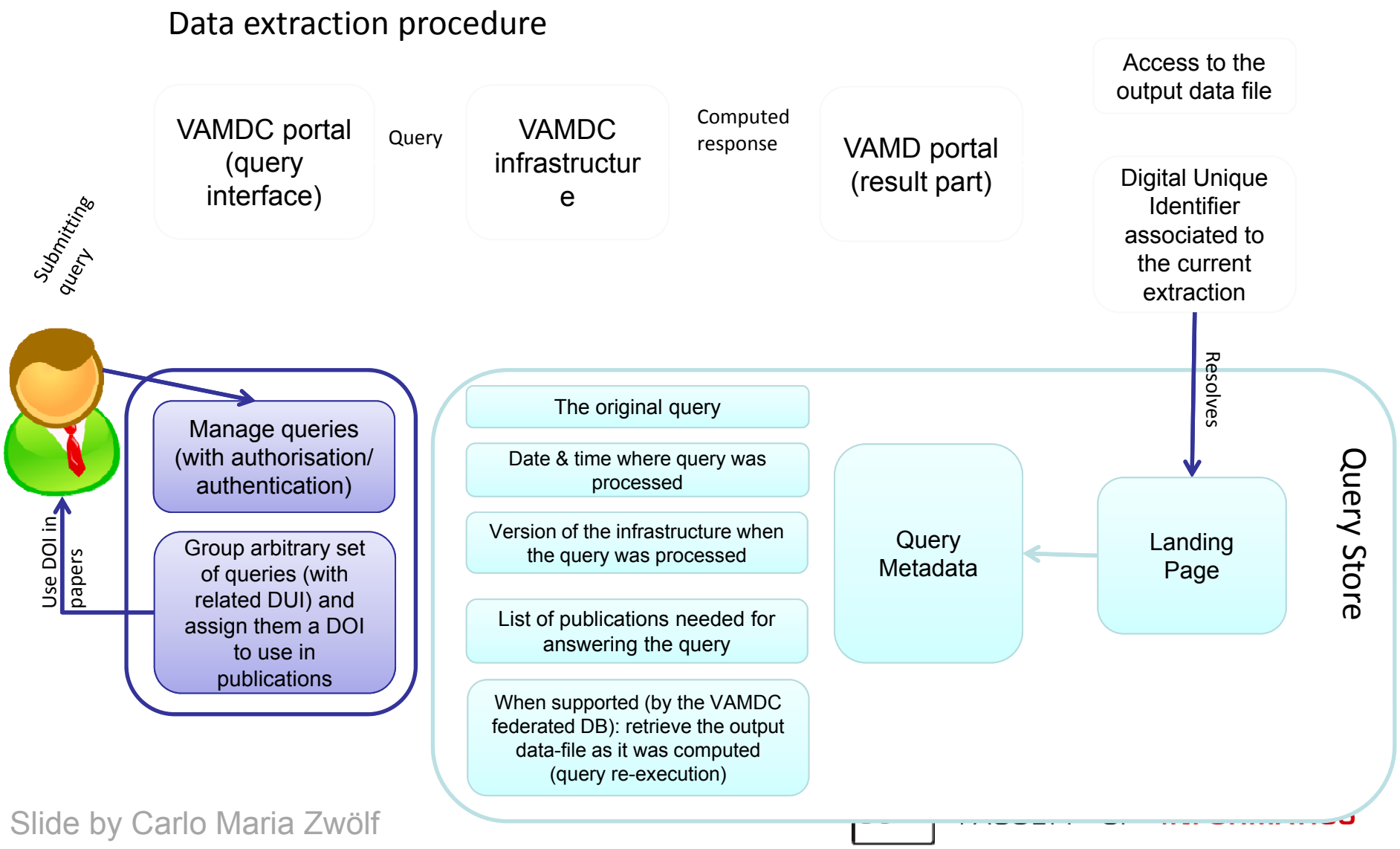
This approach has several advantages:

- It solves the data tagging granularity problem
- It is independent from what is considered a dataset
- The new files are compliant with old libraries & processing programs
  - We add a new feature, an overlay to the existing structure
  - We induce a structuration, without changing the structure

New model for datasets citation and extraction reproducibility in VAMDC, C.M. Zwölf, N. Moreau, M.-L. Dubernet, *J. Mol. Spectrosc. (2016)*, http://dx.doi.org/10.1016/j.jms.2016.04.009 Arxiv version: https://arxiv.org/abs/1606.00405

Let us focus on the query store:

Sketching the functioning – From the final-user point of view:



Slide by Carlo Maria Zwölf

Let us focus on the query store:

The difficulty we have to cope with:

- Handle a query store in a distributed environment (RDA did not design it for these configurations)
- Integrate the query store with the existing VAMDC infrastructure

The implementation of the query store is the goal of a jointly collaboration between VAMDC and RDA-Europe

- Development started during spring 2016
- Final product released during 2017

Collaboration with Elsevier for embedding the VAMDC query store into the pages displaying the digital version of papers.
Designing technical solution for

- Paper / data linking at the paper submission (for authors)
- Paper / data linking at the paper display (for readers)

Climate Change Centre Austria (CCCA)

Chris Schubert
chris.Schubert@ccca.ac.at

research data sharing without barriers
rd-alliance.org

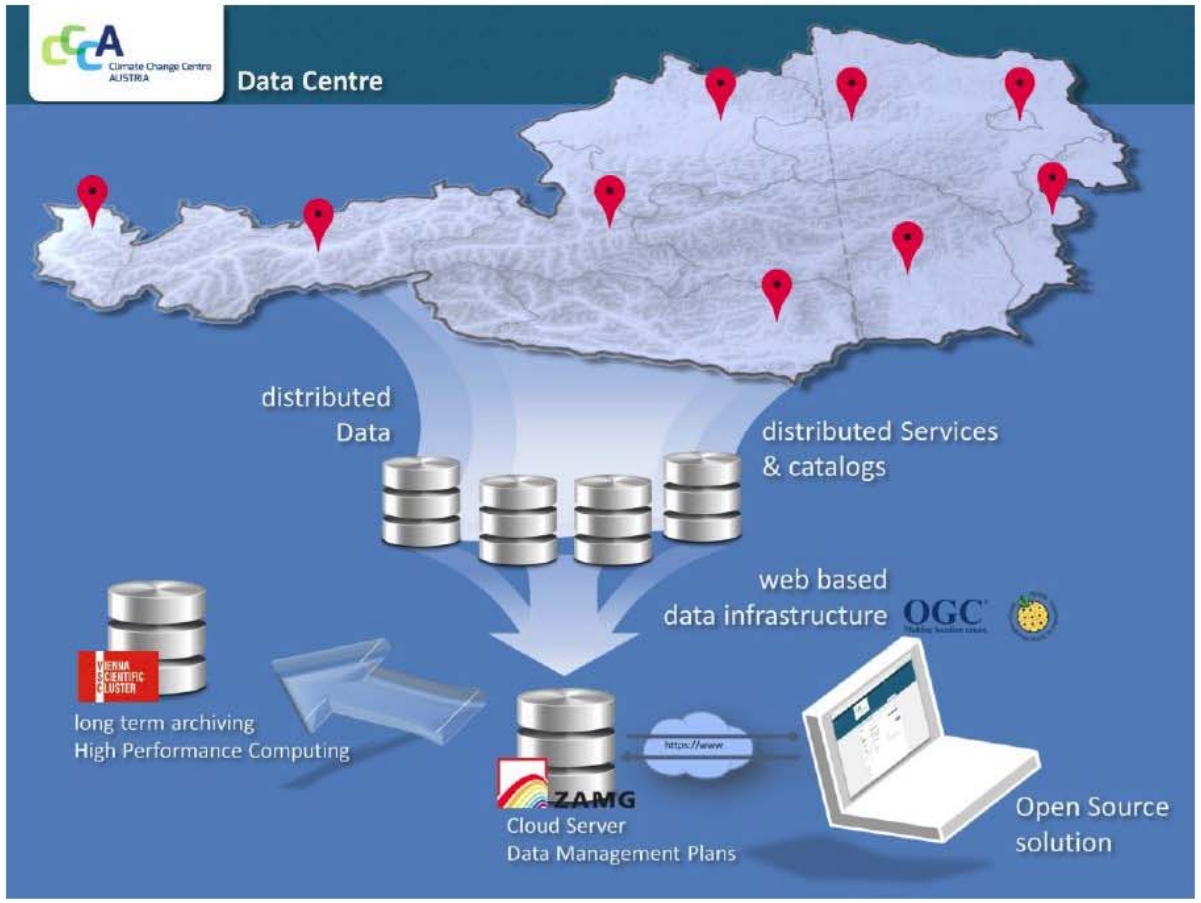
CCCA

Climate Change Centre Austria

- Climate research network for sustained, high-quality Austrian climate research.
- 28 members (11 universities, 13 non-university institutions, 4 supporting members)
- Structure: Coordination Office (Vienna, BOKU), Service Centre (Univ. Graz), Data Centre (ZAMG, Vienna)
- Service available at http://data.ccca.ac.at

Slide by Chris Schubert

CCCA

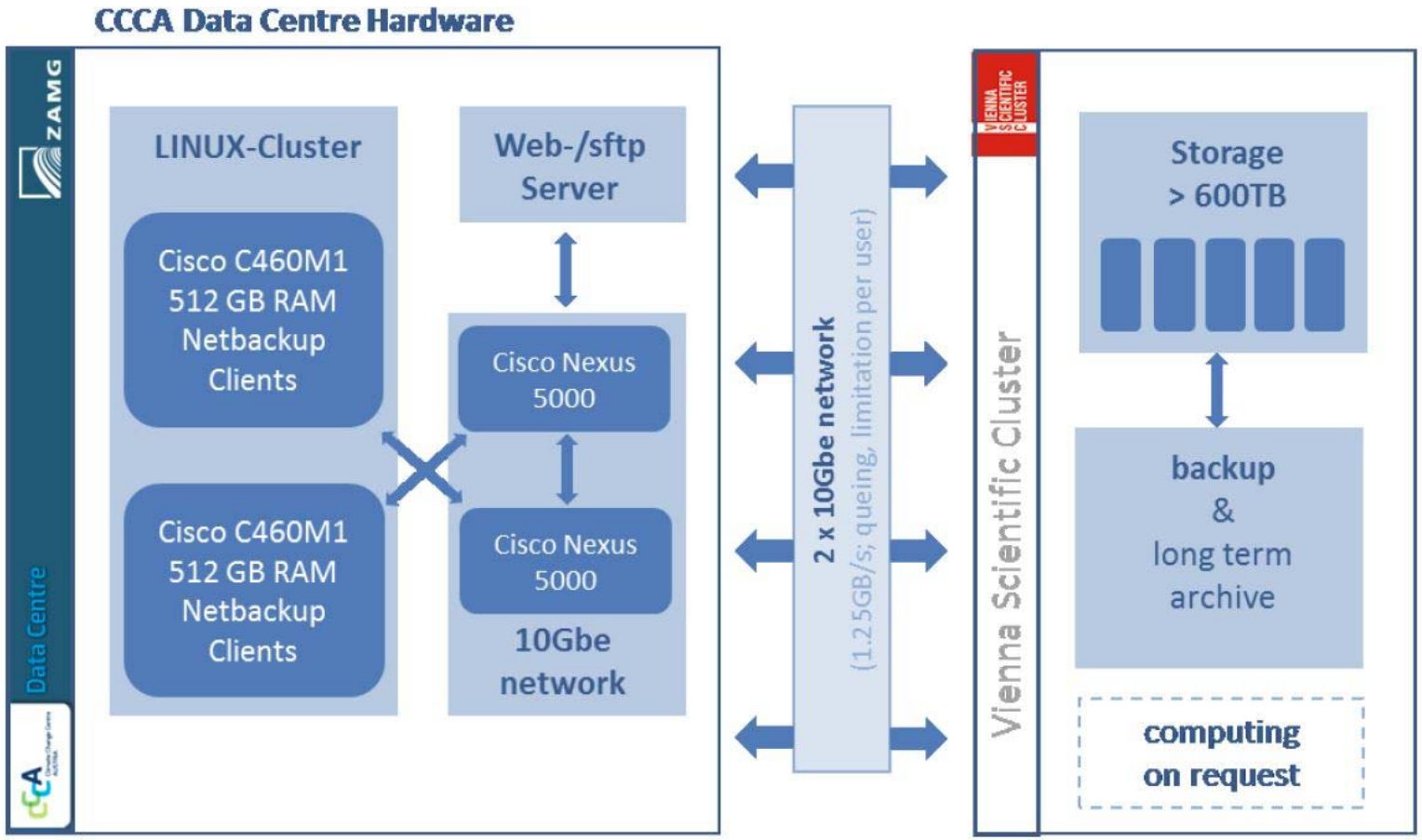


CCCA Data Centre

- **provision** of climate-relevant information, data, algorithms, reports
- **interoperable interfaces** to international portals, standards, legislation (e.g. INSPIRE)
- conception for **long term archiving** of research data & repositories
- capacity building, consultancy and **support** for **data sharing**

Slide by Chris Schubert

CCCA


CCCA Data Centre Hardware
ZAMG
Data Centre
LINUX-Cluster
Cisco C460M1
512 GB RAM
Netbackup
Clients
Cisco C460M1
512 GB RAM
Netbackup
Clients
Web-/sftp
Server
Cisco Nexus
5000
Cisco Nexus
5000
10Gbe
network
2 x 10Gbe network
(1.25GB/s; queing, limitation per user)
Vienna Scientific Cluster
Storage
> 600TB
backup
&
long term
archive
computing
on request

CCCA

… a data portal among many others?

FEATURE No. 4 & 5

- handle® Service implemented to serve persistent identifier (PID) -> fundamental for DataCitation

hdl.handle.net/20.500.11756/7b9374de

Cite this resource:

Using this data set or resource, you should cite this data set according to the given copyright conditions with following citation rules.

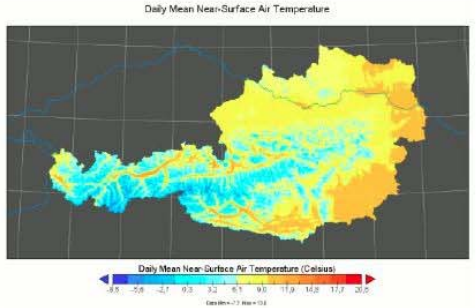
> Hiebl et al. (2016). cdd-1961-2011-annual (Ver. 1). Retrieved from CCCA Data Centre: https://hdl.handle.net/20.500.11756/fa338331. Access Date: February 22, 2017

Your Publication


Scaled distribution mapping: a bias correction method that preserves raw climate model projected changes

formal
Data Citation

Your Data

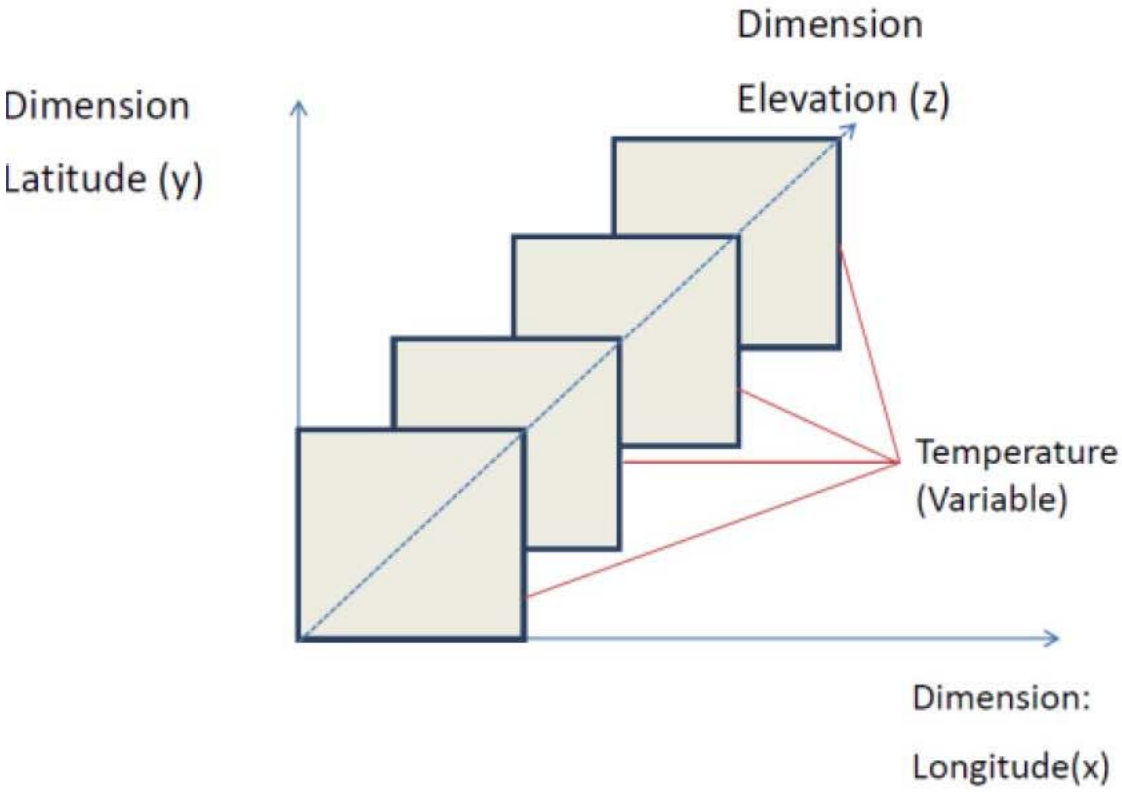


Slide by Chris Schubert

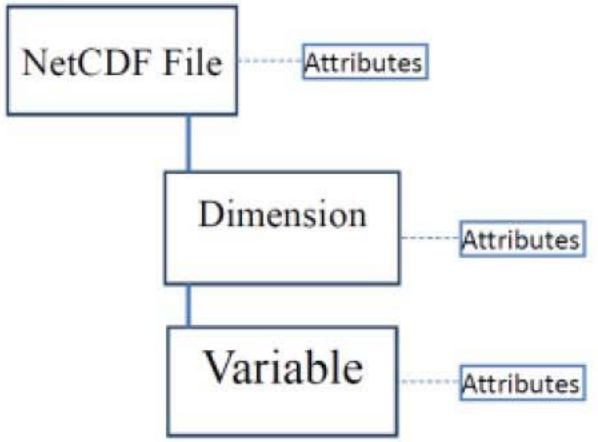
FACULTY OF !INFORMATICS

CCCA

- NetCDF Files:



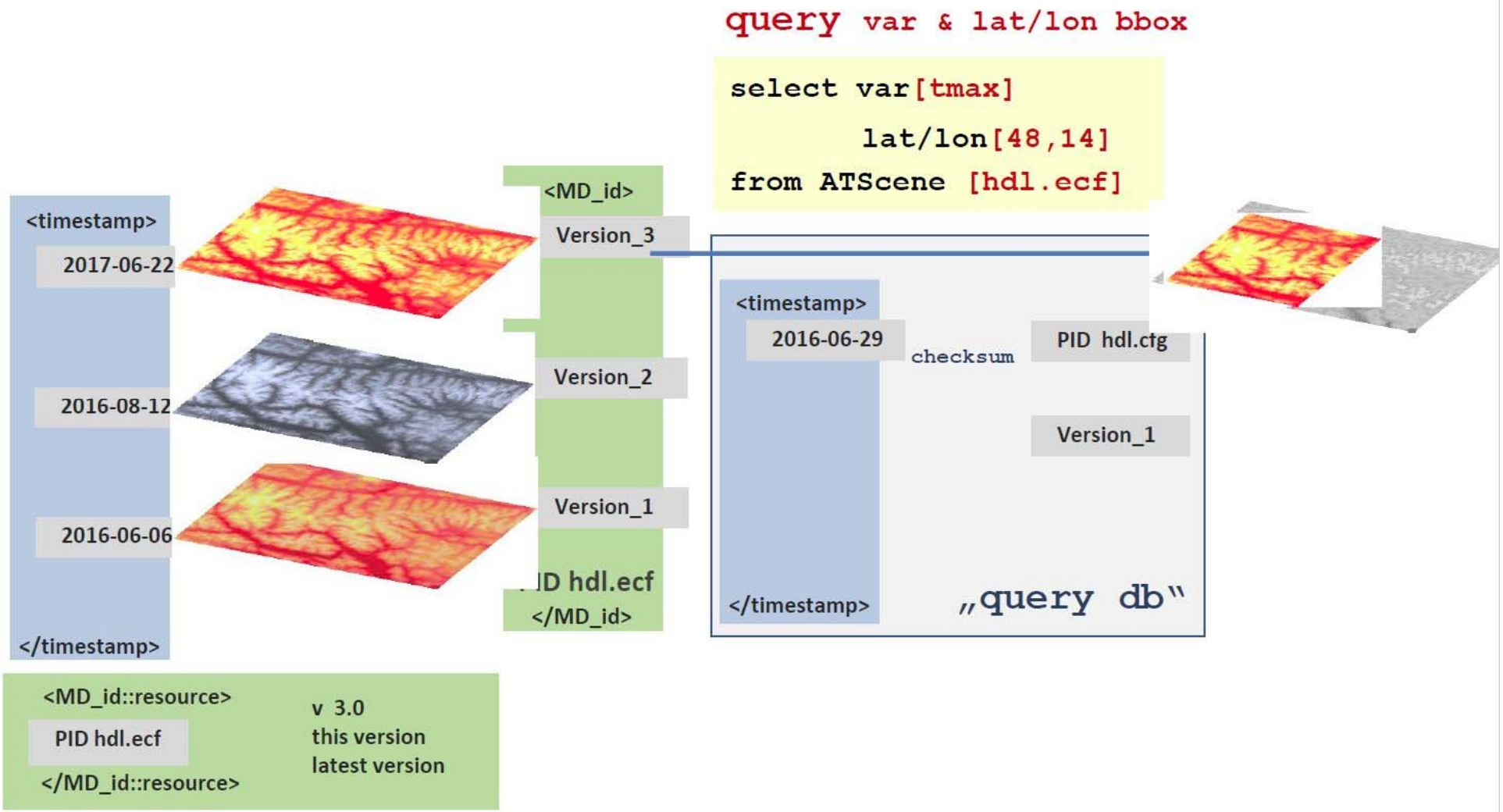
modified and based on UCAR Unidata, www.unidata.ucar.edu/



Slide by Chris Schubert

FACULTY OF !INFORMATICS

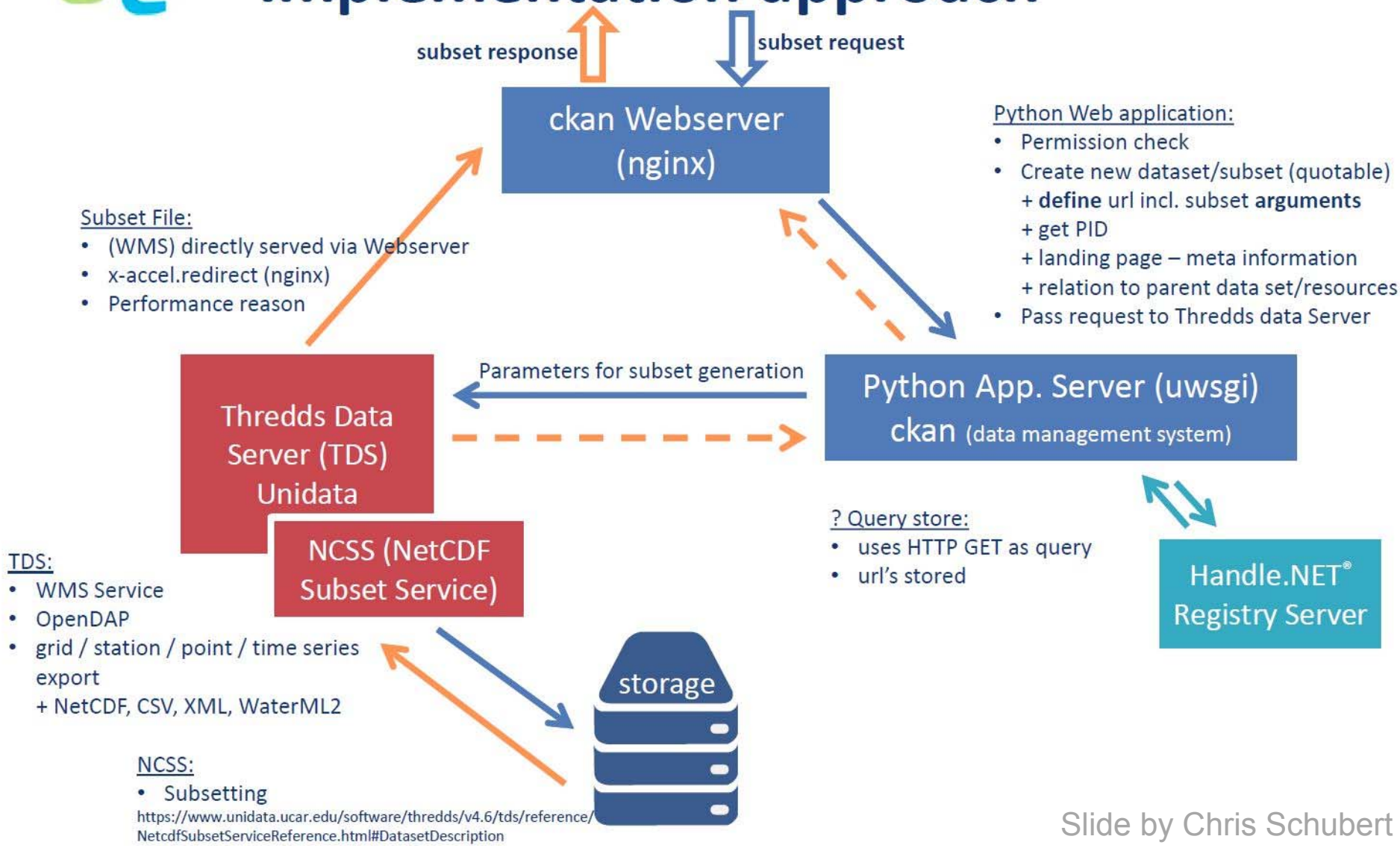
CCCA



Slide by Chris Schubert

CCCA

Implementation approach



Slide by Chris Schubert

CCCA

NCSS (NetCDF Subset Service)

- for subsetting datasets
- uses **HTTP GET** as query in following scheme:
  http://{host}/{context}/{service}/{**dataset**}[/**dataset.html | {?query}**]

Subsetting parameter used:

- **var** - names of our layer
- **north, south, east, west** - for the geographical extend, the bounding box
- **time_start, time_end, time_duration** - for time extend, limited only on 5 years interval
- **accept** - specify the returned format

query store (uwsgi) ckan

All "http get" stored as url in our ckan data store

PID:
hdl.handle.net/20.500.11756/93887ecf

https://data.ccca.ac.at/tds_proxy/ncss/**1dba52b2-4fd0-4fa1-a3ac-cfb0b94a7670**?**north**=47.73168822550699**&west**=9.021605998277664&**accept**=netCDF&**var**=tas**&east**=12.031859904527664**&south**=46.77724203092812

Outline

- Why should we want to cite data?
- What are the challenges in data identification and citation?
- How should we do it, according to the RDA WG?
- Who is doing it so far, and how?
- Summary

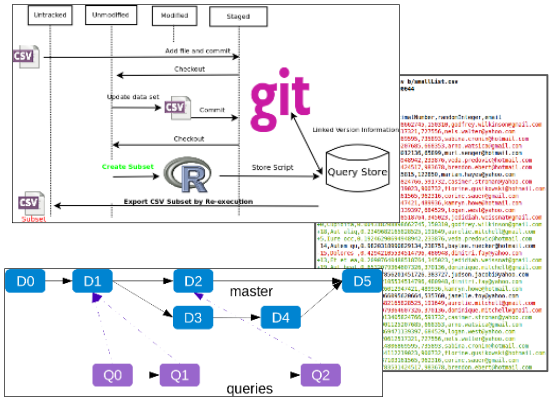
Summary

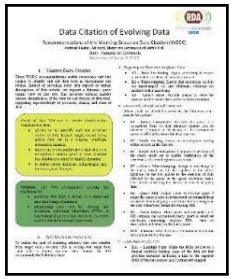
- Data citation essential for **solid** and **efficient** science (but not just for science!)
- It is more than just giving credit
- Human-readable and machine-actionable
- RDA recommendations
  - Time-stamp and version data if it is evolving
  - Provide PIDs to arbitrary subsets via selection mechanism (“query”) (rather than statically assigned PIDs to pre-defined subsets)
- 2 PIDs:
  - for evolving intellectual object
  - for precise, static subset

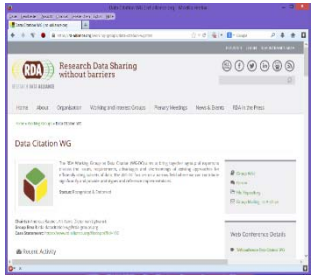
Benefits

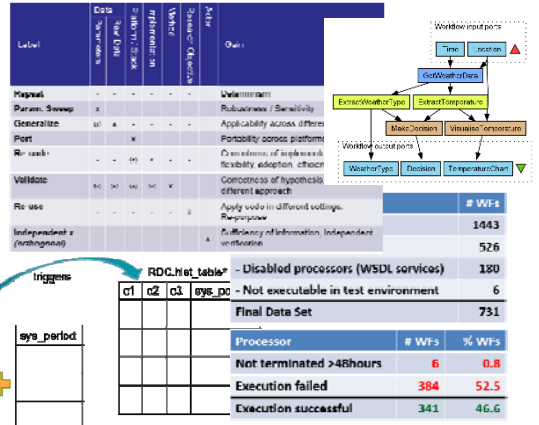
- **Precisely identify any arbitrary subset of data**
- Principles applicable to all types of data
- Straightforward to implement in most settings
- Optimizations for high-volume / very dynamic data possible
- Transparent for the analyst / data scientist
- Reduces documentation effort for analysts / data scientist
- Reduces data management complexity for data centre
- Increases traceability of results, **trust**

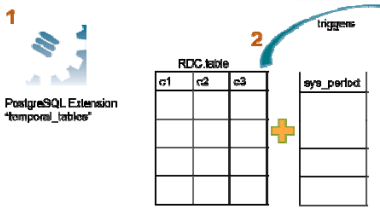
Thank you!











Thanks!

https://rd-alliance.org/working-groups/data-citation-wg.html