194.044 Data Stewardship (VO 2,0) 2022S
and
194.045 Data Stewardship (UE 2,0) 2022S

Andreas Rauber

Tomasz Miksa

Martin Weise

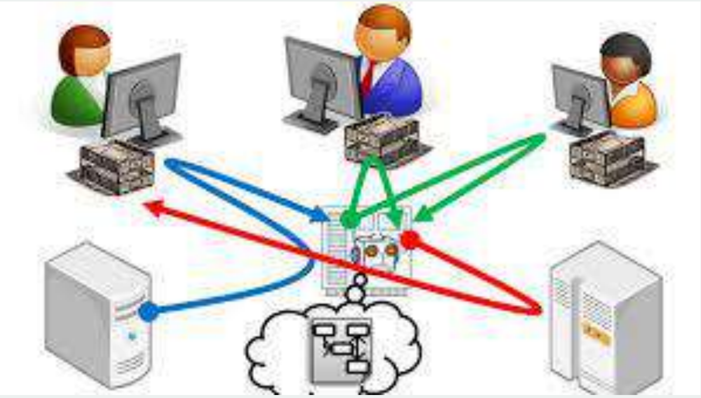
Agenda

- Organisational
  - What is data stewardship?
  - Overview of lecture topics
  - Dates, exams, materials…
  - Exercise overview
  - Selected projects related to this course
  - (Funded) collaboration possibilities

- Introduction to FAIR principles

Data Science

- Data
  - is fuel for research
  - is the result of processes such as
    - capturing, pre-processing, transformation, integration, analysis
- Research Infrastructures
  - [Human Brain Project]
    - neuroscience, computing, and brain-related medicine
  - [Elixir]
    - bioinformatics
  - Large Hadron Collider at CERN
    - 300GB per second of raw data from detectors



Researchers trying to reuse data…

Hanson, Karen; Surkis, Alisa; Yacobucci, Karen: Data Sharing and Management Snafu in 3 Short Acts.
https://doi.org/10.5446/31036

Conversation of two researchers

- Can I see your data?
- It’s on my USB stick
- Can I have it?
- I have in a box and I have moved recently
- Can I have it?
- I forgot to label the boxes…
- (half a year later)
- Thanks, for the USB. However, I cannot read the hexadecimal file on it. How do I open it?
- You need a special program
- What program?
- …

Data Stewardship (DaSt)

Data Stewardship ~ Data Management ~ Digital Curation

Management of data to ensure its

- FAIR
- Reproducibility
- Auditability
- ...

Must be addressed at different levels and is interdisciplinary

- Technical
- Organisational
- Legal

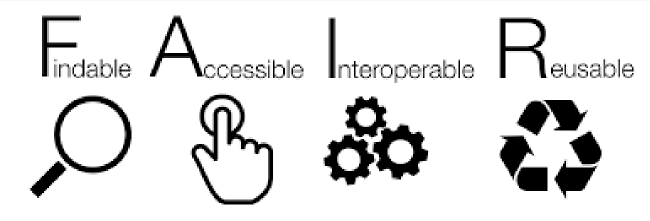
Data Stewardship is needed in all domains

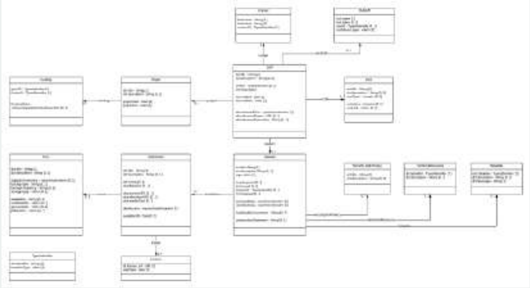
- E.g. Data Stewards must have domain specific knowledge and data management skills

Overview of Lecture Topics

Data Management

- Plans, policies, support services, stewards, etc.
- Legal frameworks, roles and responsibilities, etc.

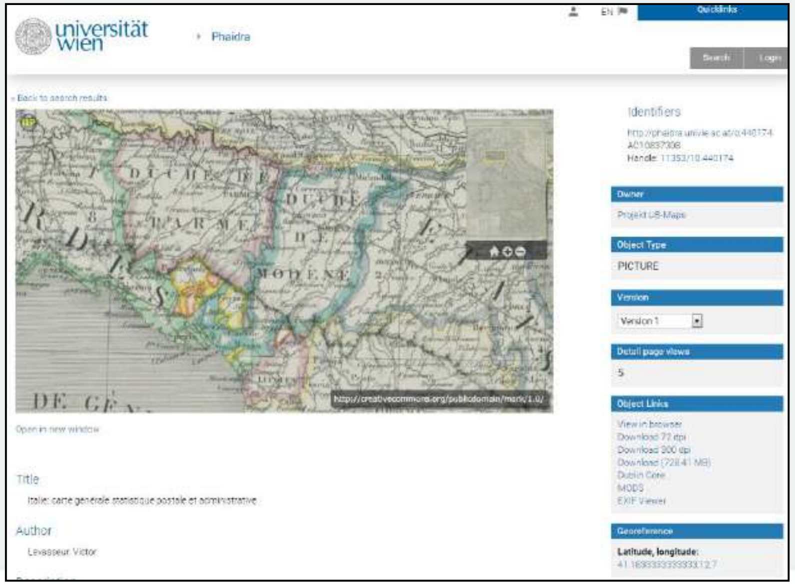




Overview of Lecture Topics

Data repositories

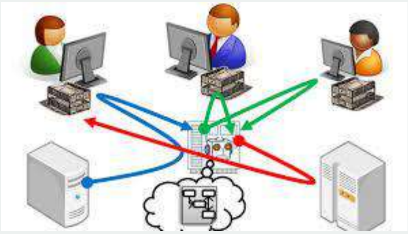
- Software architecture, interfaces, interoperability, data models, standards, ontologies, etc.
- Certification, processes, management, etc.

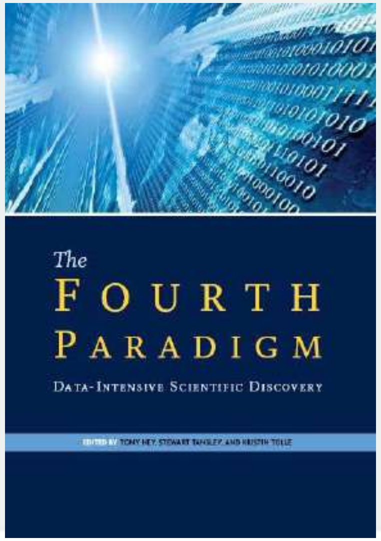


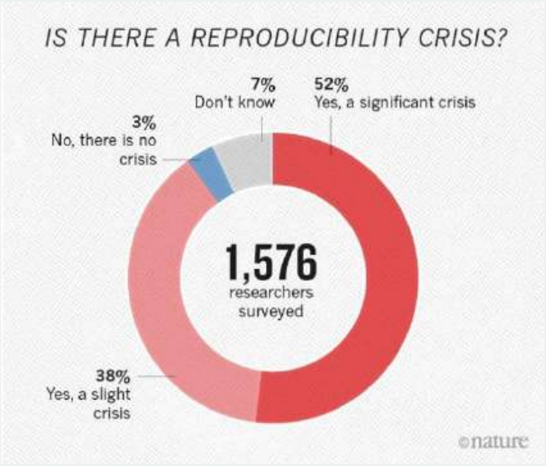
Overview of Lecture Topics

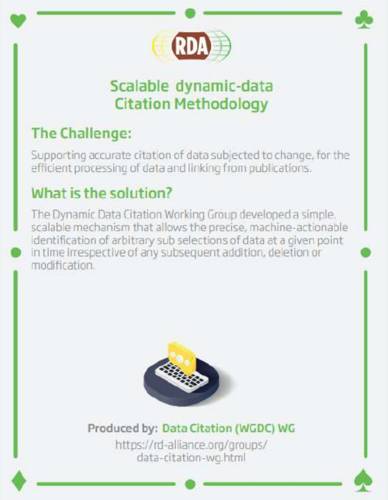
Reproducibility

- Software execution tracing, dependencies, software packaging, etc.
- Data identification and citation, provenance, etc.





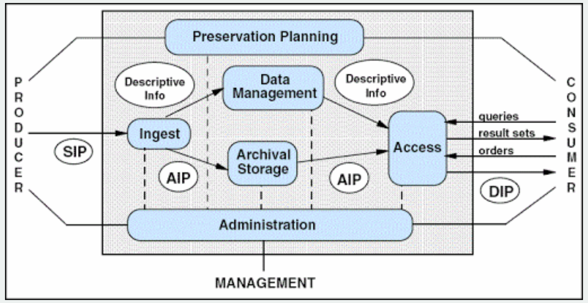




Overview of Lecture Topics

Digital Preservation

- migration, emulation, planning, risk management, etc.



Lecture Goals

As a student/researcher/employee

- Better design and organise your work with data (and code)
  - Select proper tools and software
  - Better design data transformation processes
  - Make aware decisions on experiment/tool/process design with respect to data
- Make your results trustworthy, auditable, reproducible, FAIR…

As a data steward/consultant/architect

- Have a good understanding of all components needed to establish systems/workflows/practices at an organisation

Lecture Dates

07 March 2022 TM: Intro

14 March 2022 AR: DP

21 March 2022 AR: OAIS

28 March 2022 TM: DMPs

4 April 2022 AR: Data Citation

11 April 2022 #Easter

18 April 2022 #Easter

25 April 2022 TM: FAIR

02 May 2022 TM: maDMPs

09 May 2022 TM: Repositories 1

16 May 2022 TM: Repositories 2

23 May 2022 TM: Developing RDM Services + Certification

30 May 2022 VD: Legal aspects and MW: Secure Data Infrastructures

13 June 2022 Exam

Order of specific lectures may change! Always check TUWEL!

Exams

First exam

- 13 June 2022, 12:00 – 14:00

Further exams

- October 2022
- December 2022
- March 2023

PROTIP

- You need to be able to explain 'things', e.g. what they are, why they are needed, what are pros and cons, etc.
- If you just memorise some acronyms then it won't help you much.

Lecture Materials

There is no single book to follow for this lecture!

Slides

- Meant for presenting!
- Contain links to papers/reports/webinars etc.
  - Check them!
- Take notes

DO NOT DEPEND ONLY ON SLIDES WHEN STUDYING FOR THE EXAM!

Lecture Materials

Literature (very basic)

- Barend Mons, Data Stewardship for Open Science: Implementing FAIR Principles, 10.1201/9781315380711
(available online through TU Bibliothek)
- Managing and sharing data by UK Data Archive: http://www.data-archive.ac.uk/media/2894/managingsharing.pdf

Many links provided in the individual lectures – check them!

- https://www.go-fair.org/fair-principles/
- https://www.fairsfair.eu

Guidance for researchers from centres like the one at TUW:

- https://www.tuwien.at/en/research/rti-support/research-data/center-for-rdm

Overview of Exercises

194.045 Data Stewardship (UE 2,0) 2022S

- Practical exercises to get hands-on experience
- Programming skills essential

Exercise 1 (33%)

- Will be published this week in TUWEL
- Deadline ~ end of April

Exercise 2 (67%)

- Deadline ~ end of June
- Groups of 2

Submissions via TUWEL

To pass

- Min 35% of points on each exercise
- Min 50% of point overall to be positive

Late submission by 1 week: 10% points deduction

Overview of Exercises

Just to give an idea of what we did last year! Topics will be different this year!

Exercise 1

- use case - you pick an already existing experiment of yours or create one for this exercise;
- data management - you create a data management plan (DMP) and publish your experiment;
- machine-actionable DMP – you express your DMP using RDA DMP Common Standard.

Exercise 2

- Topic #1 SHACL constraints for DCSO
- Topic #2 maDMPs to support reviewers
- Topic #3 RO-Crates and Excel
- Topic #4 Meta-analysis of reproducibility of papers
- Topic #5 Analysis of DMPs from Exercise 1
- Topic #6 Core Trust Seal
- Topic #7 Repository content migration
- Topic #8 GitHub and Sustainability Metrics
- Topic #9 Python client library for TU Data

Selected Data Stewardship Activities and Collaborations

We teach what we learned from research projects and industrial collaborations

- e-Infrastructures Austria
- ROMOR: Research Output Management
- Timbus: Timeless Business Services (EU FP7 IP, via SBA)
- 4C: Collaboration to Clarify the Costs of Curation (FP7 CA, via SBA)
- APARSEN: (EU FP7 NoE, via SBA)
- Scape: Scalable Preservation Systems (EU FP7 IP)
- PLANETS (EU FP6 IP)
- DPE: Digital Preservation Europe (EU FP6 CA)
- …

Center for RDM @ TUW

[Center for Research Data Management]

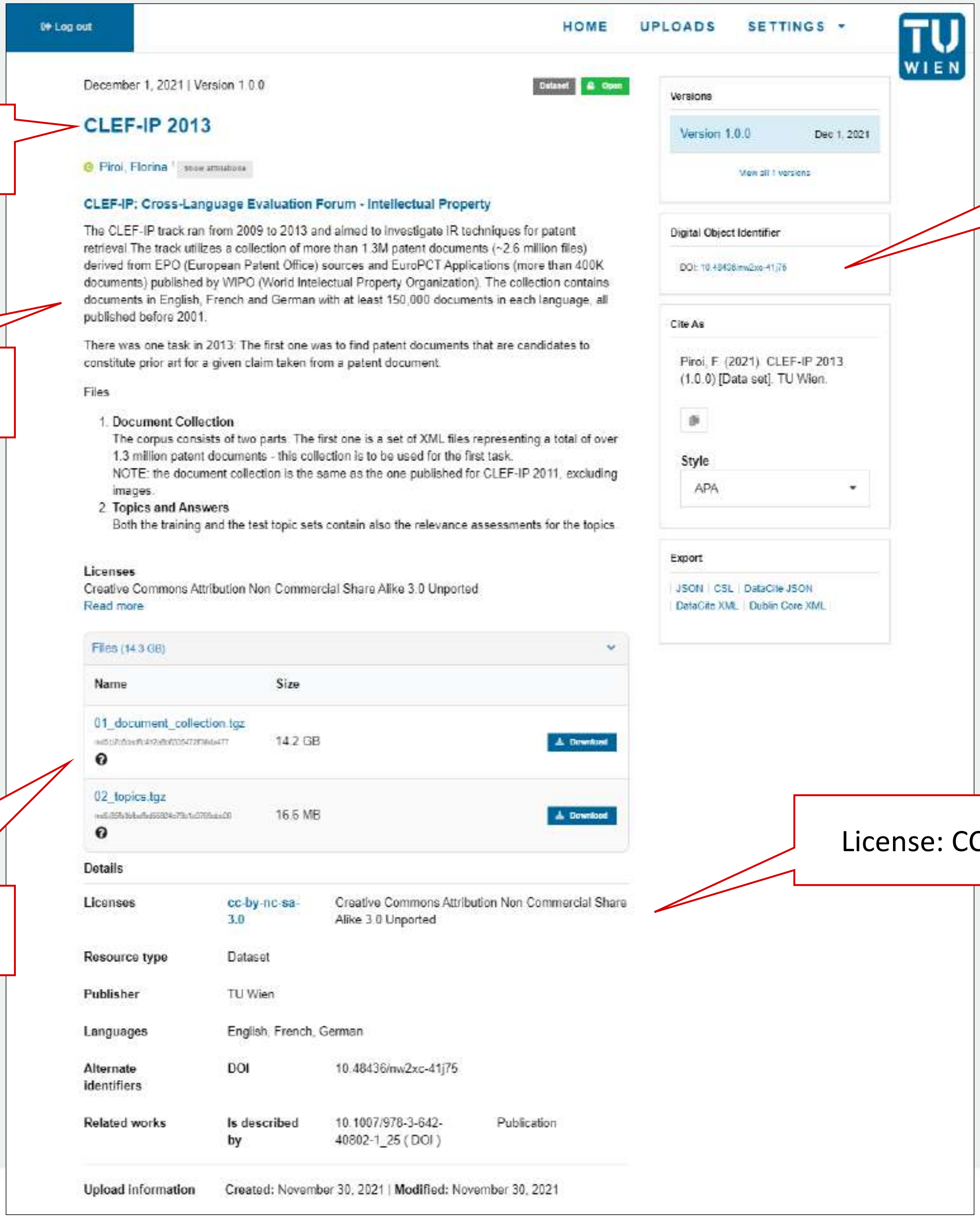
- FAIR Data Austria
  - [TU Data repository]
  - [TU Database repository]
  - [Tools for maDMPs]
  - [FAIR Office Austria]
- Research Data Management (RDM) Policy
- Support to all researchers on RDM

FAIR DATA AUSTRIA

Title of the dataset

Description of the dataset

Files for download


Log out

HOME UPLOADS SETTINGS

TU WIEN

December 1, 2021 | Version 1.0.0

Dataset Open

CLEF-IP 2013

Piroi, Florina

CLEF-IP: Cross-Language Evaluation Forum - Intellectual Property

The CLEF-IP track ran from 2009 to 2013 and aimed to investigate IR techniques for patent retrieval. The track utilizes a collection of more than 1.3M patent documents (~2.6 million files) derived from EPO (European Patent Office) sources and EuroPCT Applications (more than 400K documents) published by WIPO (World Intelectual Property Organization). The collection contains documents in English, French and German with at least 150,000 documents in each language, all published before 2001.

There was one task in 2013: The first one was to find patent documents that are candidates to constitute prior art for a given claim taken from a patent document.

Files

1. Document Collection
   The corpus consists of two parts. The first one is a set of XML files representing a total of over 1.3 million patent documents - this collection is to be used for the first task.
   NOTE: the document collection is the same as the one published for CLEF-IP 2011, excluding images.
2. Topics and Answers
   Both the training and the test topic sets contain also the relevance assessments for the topics.

Licenses
Creative Commons Attribution Non Commercial Share Alike 3.0 Unported
Read more

Files (14.3 GB)

Name	Size	
01_document_collection.tgz	14.2 GB	Download
02_topics.tgz	16.5 MB	Download

Details

Licenses	cc-by-nc-sa-3.0	Creative Commons Attribution Non Commercial Share Alike 3.0 Unported	
Resource type	Dataset		
Publisher	TU Wien		
Languages	English, French, German		
Alternate identifiers	DOI	10.48436/nw2xc-41j75	
Related works	Is described by	10.1007/978-3-642-40802-1_25 (DOI)	Publication

Upload information Created: November 30, 2021 | Modified: November 30, 2021

Versions
Version 1.0.0 Dec 1, 2021
View all 1 versions

Digital Object Identifier
DOI: 10.48436/nw2xc-41j75

Cite As
Piroi, F. (2021). CLEF-IP 2013 (1.0.0) [Data set]. TU Wien.

Style
APA

Export
JSON | CSL | DataCite JSON | DataCite XML | Dublin Core XML

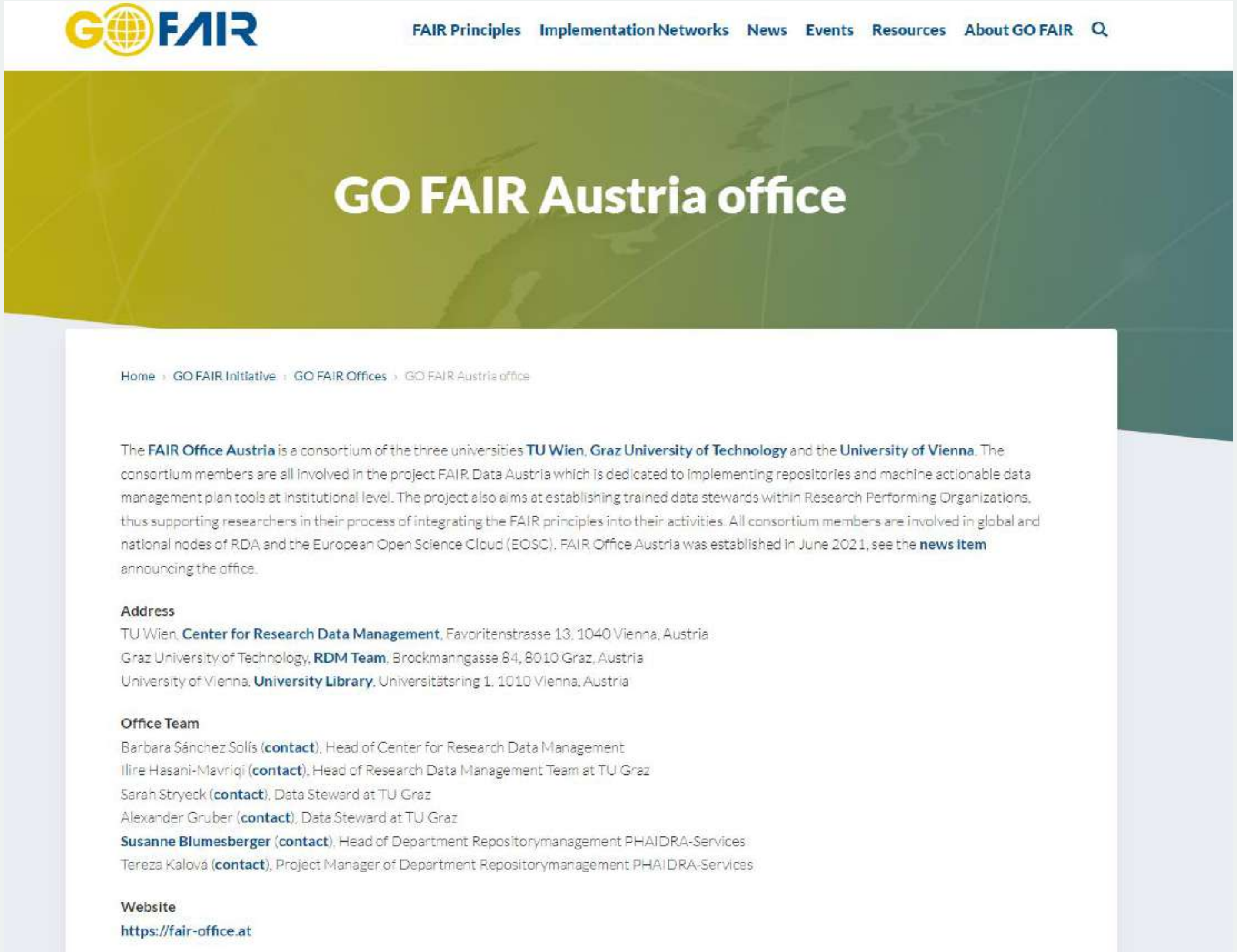
Persistent identifier (DOI):
10.48436/nw2xc-41j75

License: CC-BY-NC-SA-3.0

https://researchdata.tuwien.ac.at

https://fair-office.at/

GO-FAIR



https://www.go-fair.org/go-fair-initiative/go-fair-offices/go-fair-austria-office/

Research Data Alliance

Removing barriers in data sharing

Open community

- 9,600 members from 137 countries

Working and Interest Groups

- DMP Common Standards WG
- Dynamic Data Citation WG
- …

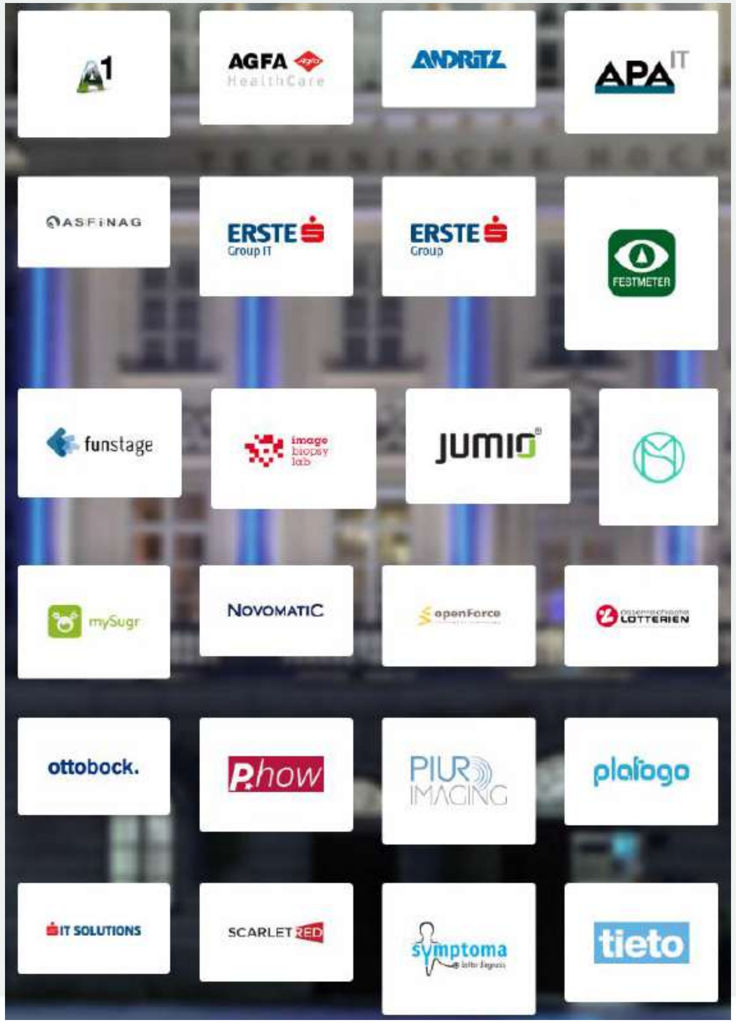
RDA Recommendations

https://www.rd-alliance.org

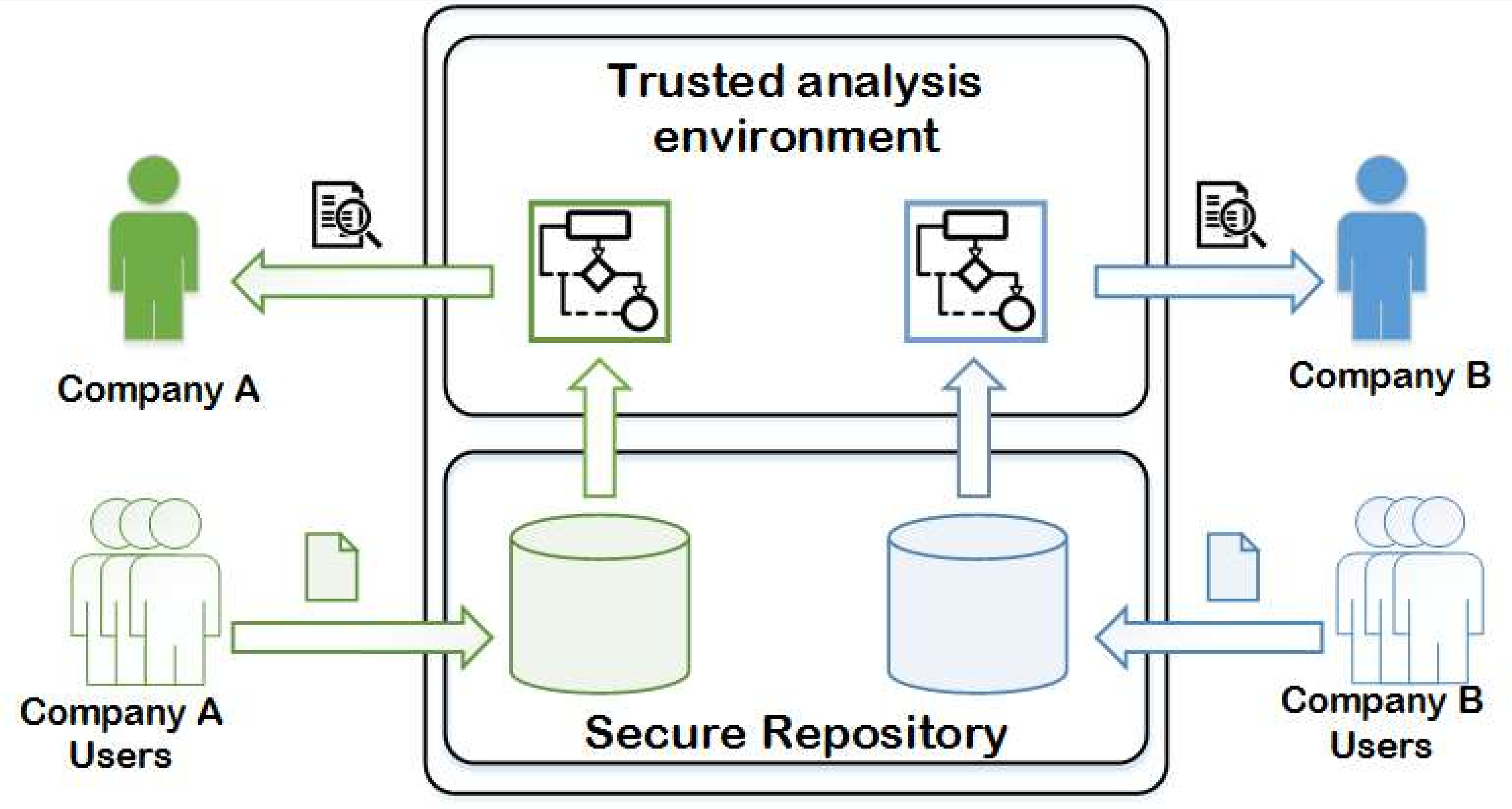
Innovationslehrgang Data Science und Deep Learning

Data Stewardship was a module

http://www.ifs.tuwien.ac.at/idsdl/



FFG WellFort



HEARTBALANCE

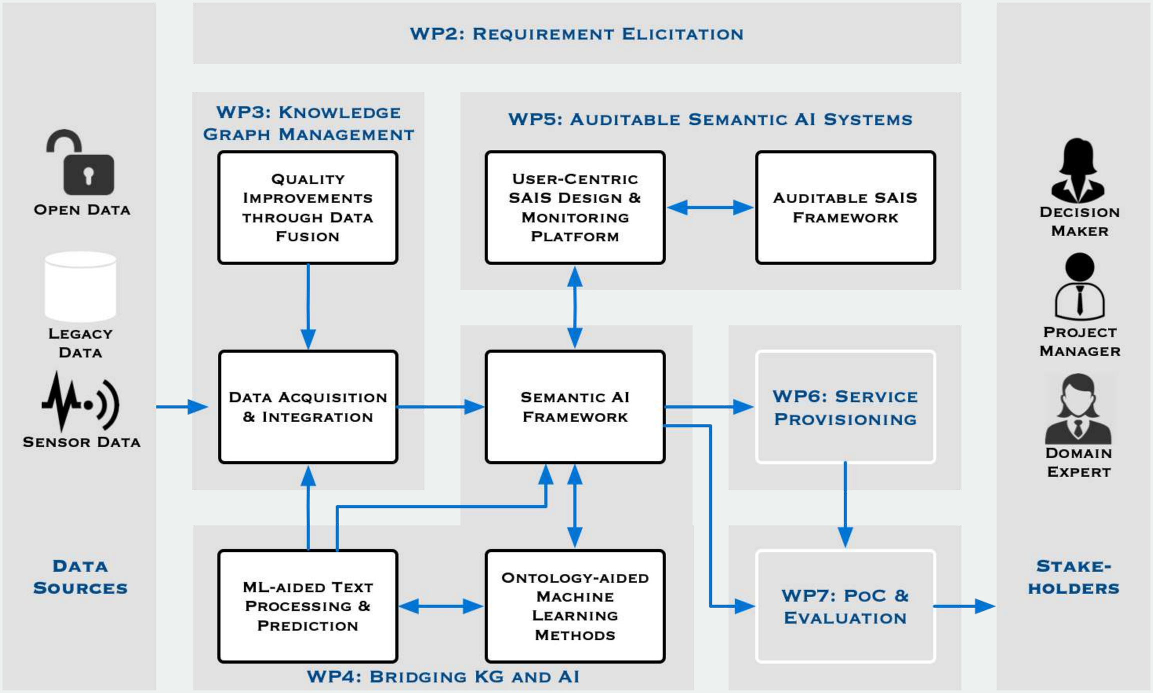
FFG
Österreichische Forschungsförderungsgesellschaft

TU WIEN Informatics

FFG OBARIS

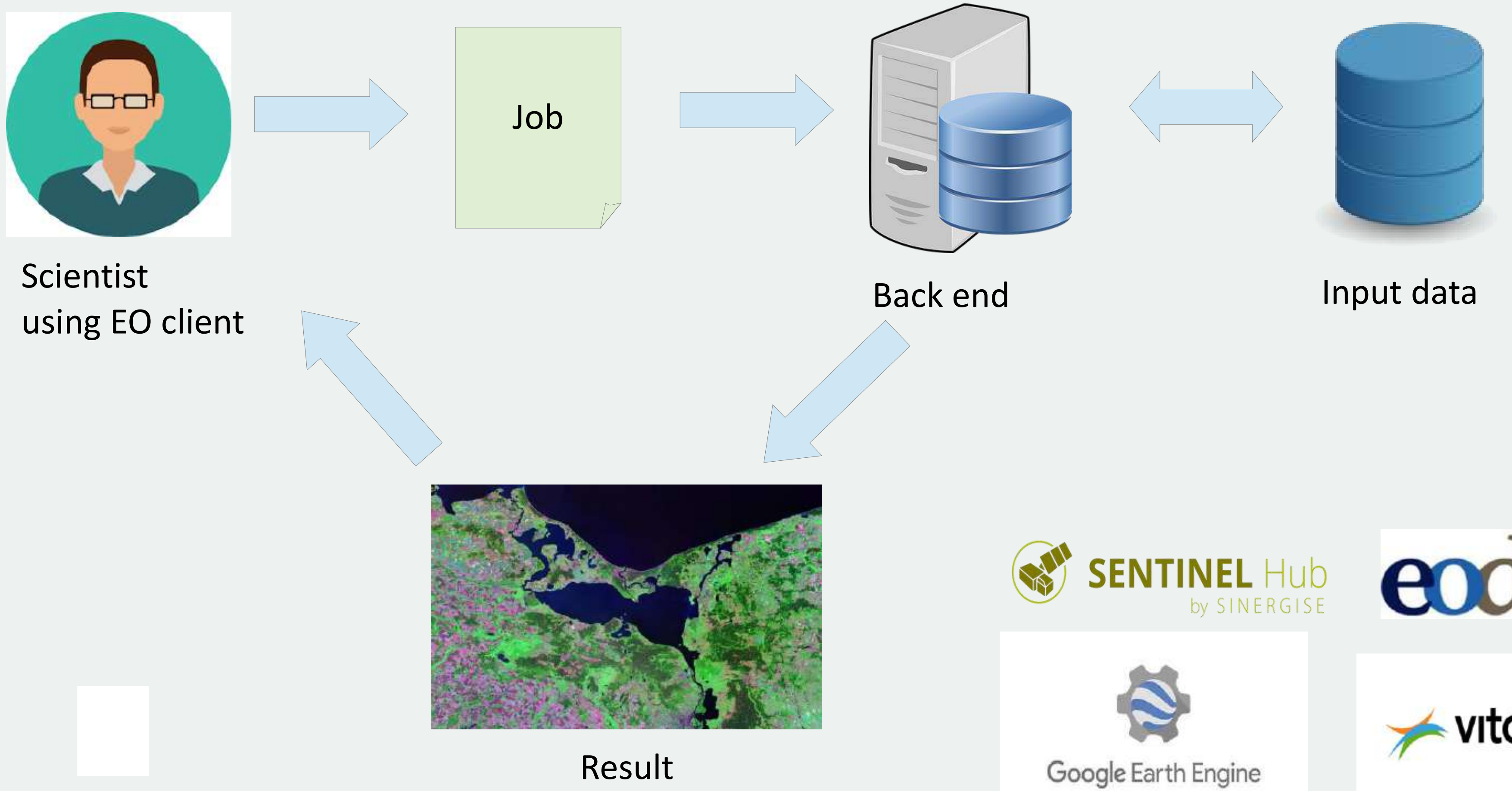
Auditable Semantic AI Systems

Use case: *Predictive analytics of water quality*



http://www.obaris.org

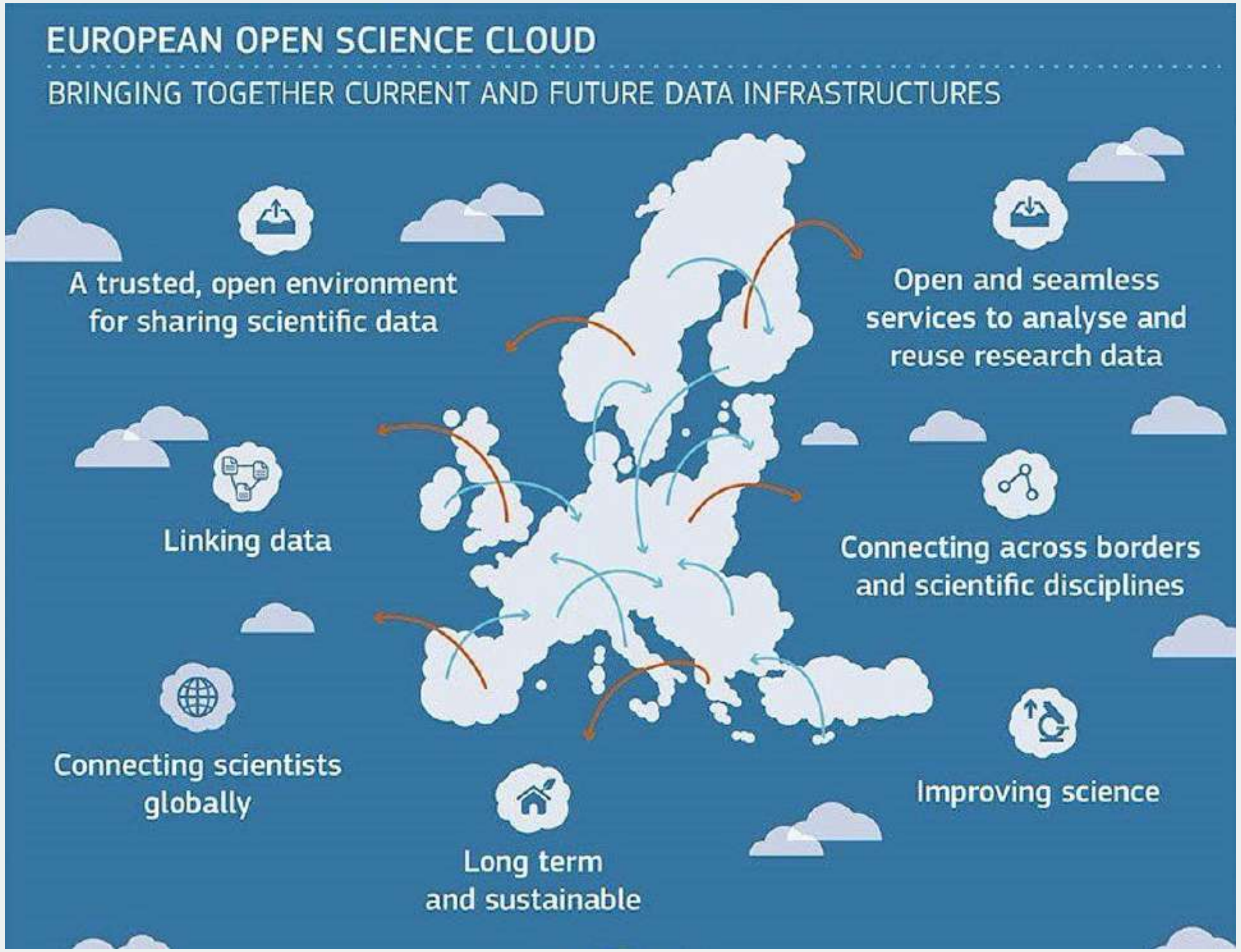
EU openEO



SENTINEL Hub by SINERGISE
eodc
Google Earth Engine
vito

https://openeo.org

European Open Science Cloud (EOSC)



EOSC-Life

EOSCsecretariat.eu
Setup and management of the EOSC Secretariat supporting the EOSC Governance

Funding opportunity

FEMTech Funding opportunity

- Funding by FFG (Austrian Research Funding Agency)
- Target group: **female students**
- Duration: 2-6 months, **paid**
- Can be used to work on a thesis or project (Praktikum)

In collaboration with SBA Research

- https://www.sba-research.org

More info

- https://www.ffg.at/femtech-praktika

First come, first served! Application has just opened!

Interested?

➔ Contact me (tmiksa@sba-research.org)

If you're looking for a topic…

We offer

- Master thesis topics
  - “success stories”
- Interdisciplinary projects, e.g. with other faculties

Focus

- Data management, versioning, reproducibility, auditability…

Various application domains

No synthetic problems

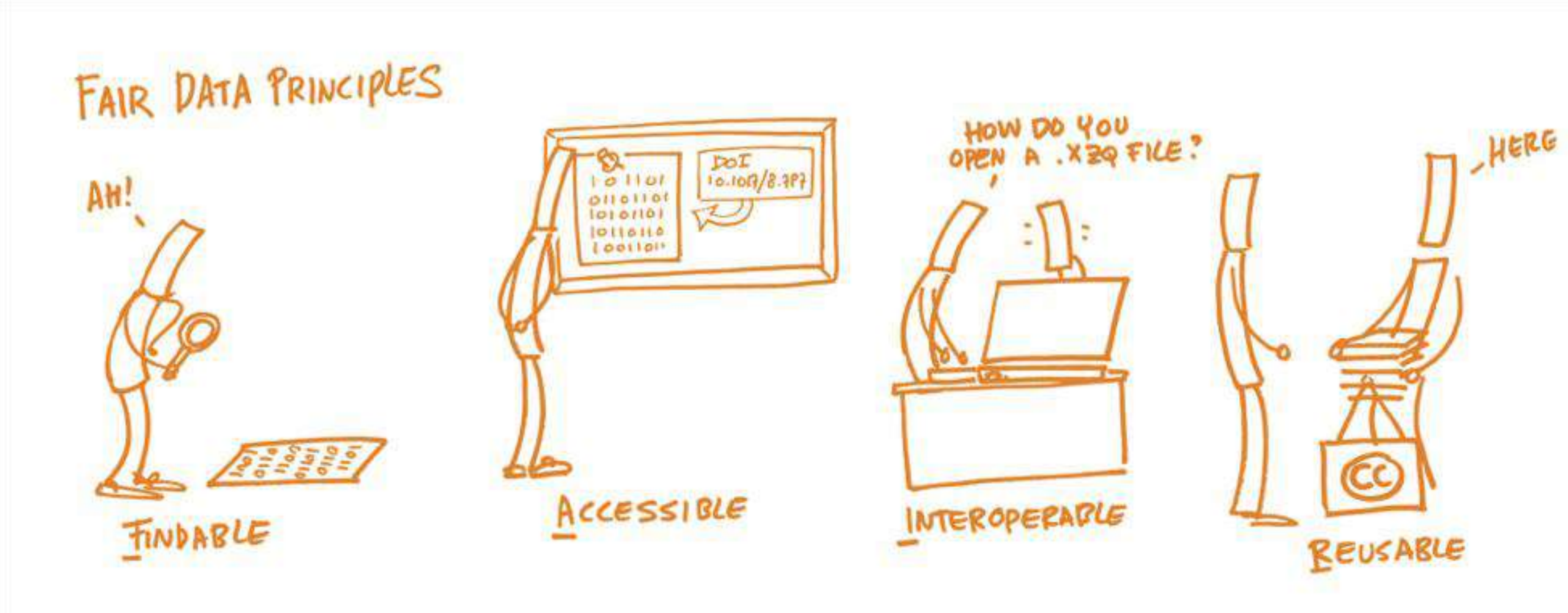
Usually the topics are not offered in TISS.

Just come and talk to us!

FAIR Principles

Introduction

FAIR Principles (very simplified :))



FAIR vs fair

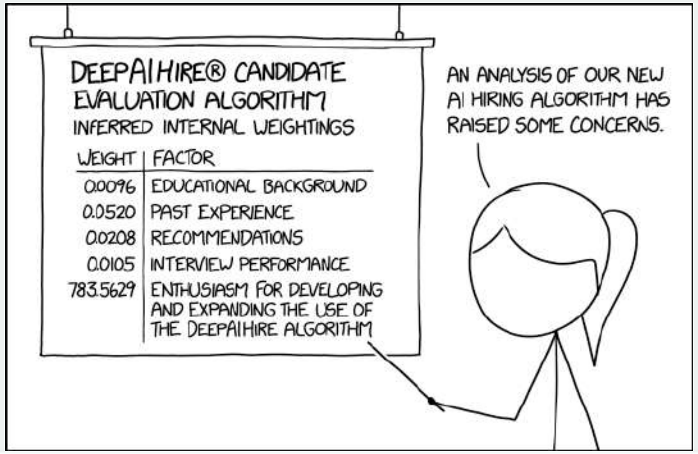
FAIR principles =/= Algorithmic fairness

To be FAIR

- To apply/use FAIR principles
- Focus on how data is managed, etc.

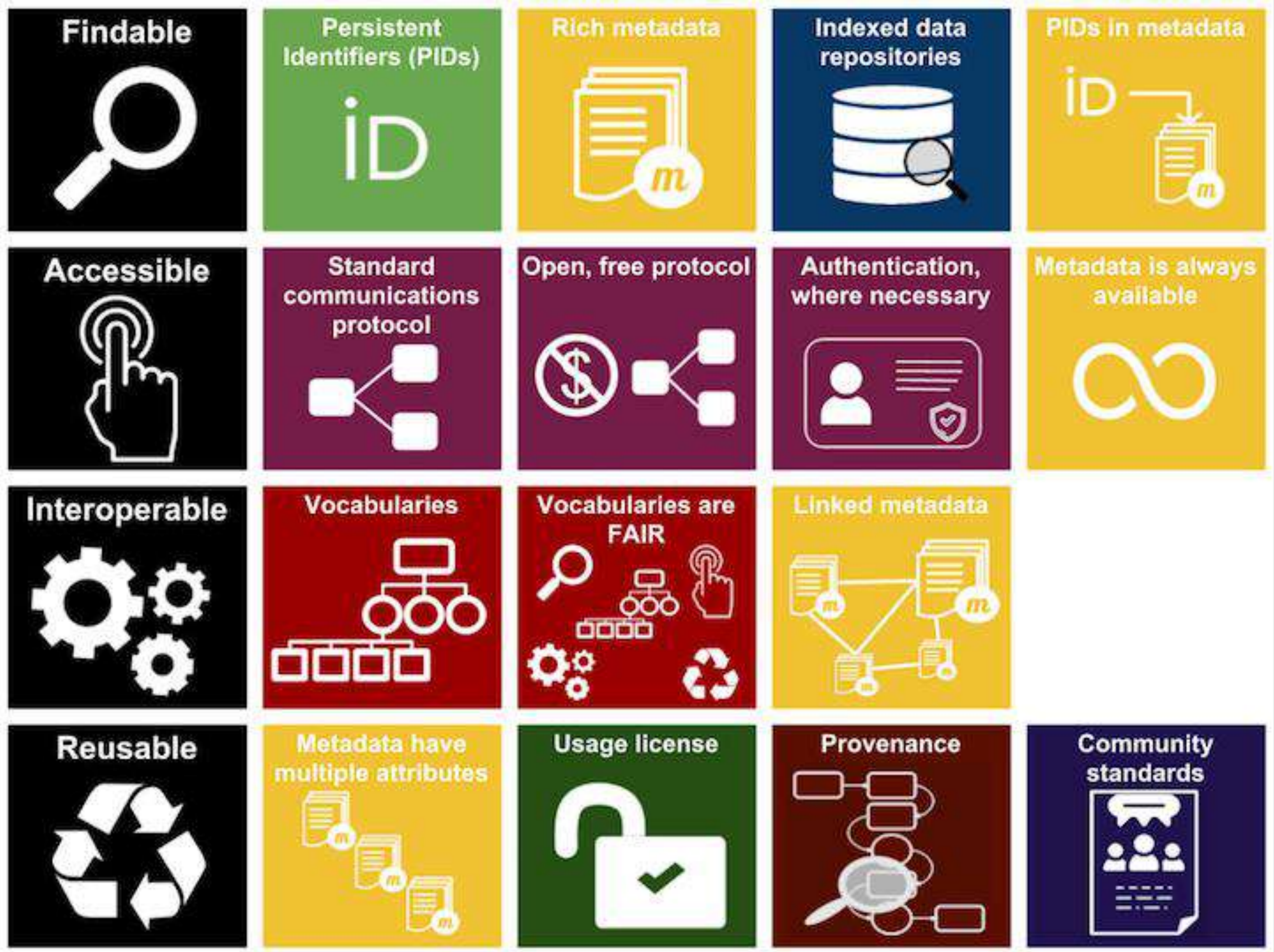
To be fair

- Evade bias
- Focus on design and implementation



https://xkcd.com/2237/

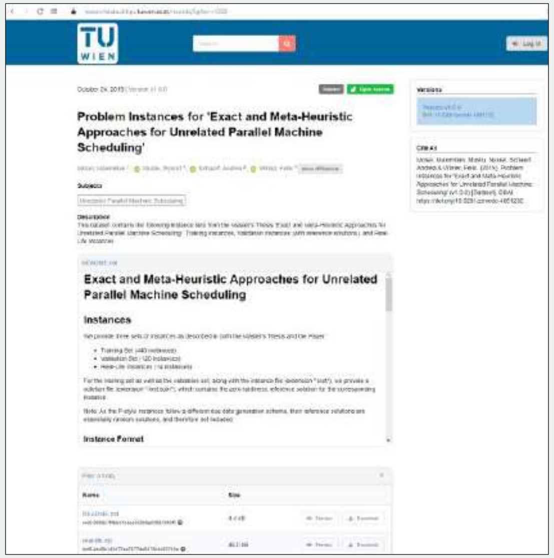
FAIR Principles



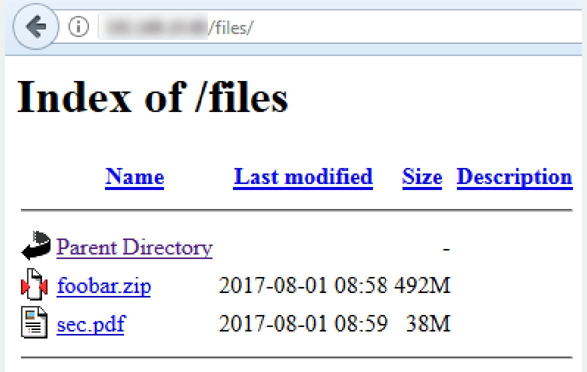
https://www.go-fair.org/fair-principles/

Findable – simplified examples

GitHub

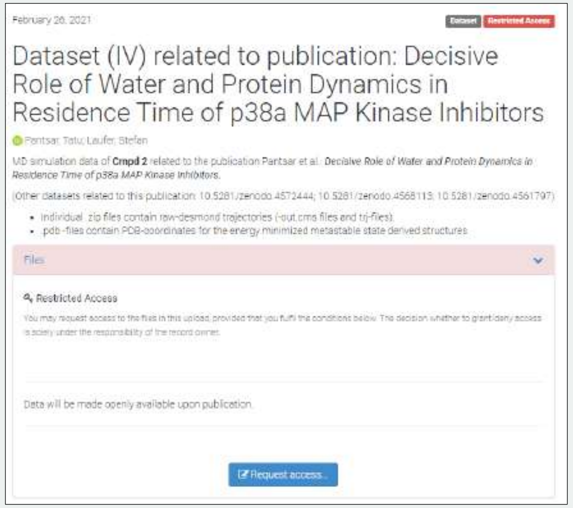


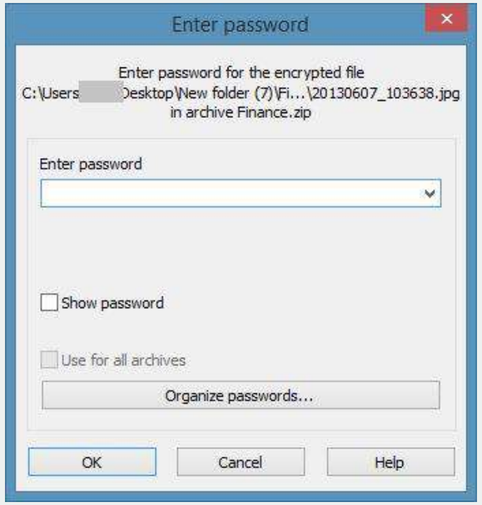
Data repository



Personal website

Accessible – simplified examples





Restricted access, but a clear way to request access

Interoperable – simplified examples

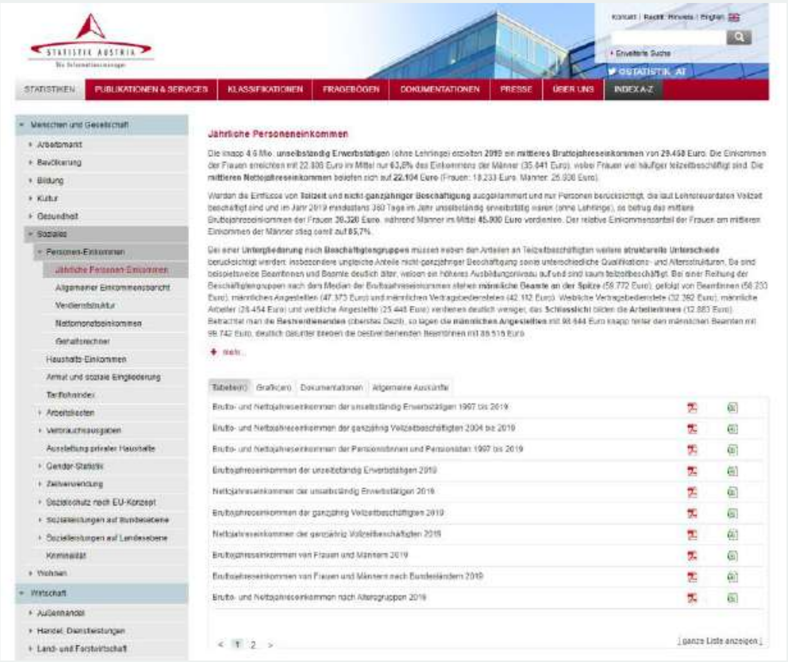
Yes

- XML following known XSD Schema
- MP3 for audio recordings
- Data model using common vocabularies

No

- Custom XML without any documentation
- M4P (Apple) for audio recordings
- Custom fields in data model with poor documentation

Reusable – simplified examples

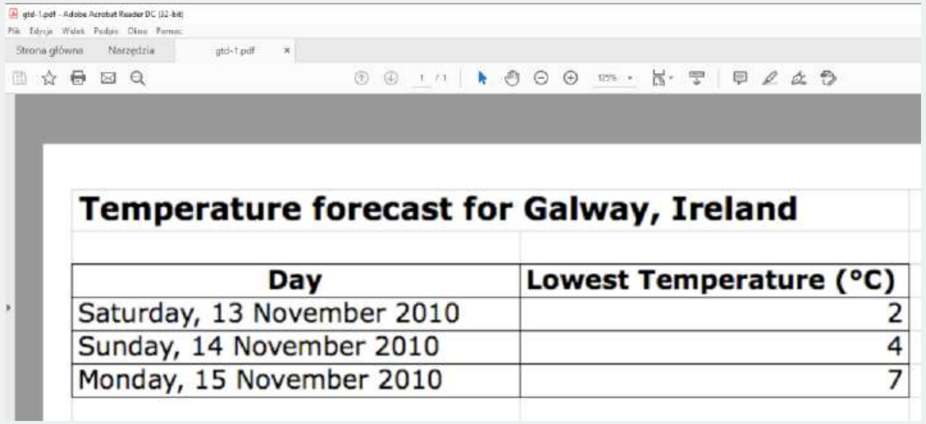


Trusted source, permission to reuse, well defined meaning of terms used

Provenance and permissions not clear

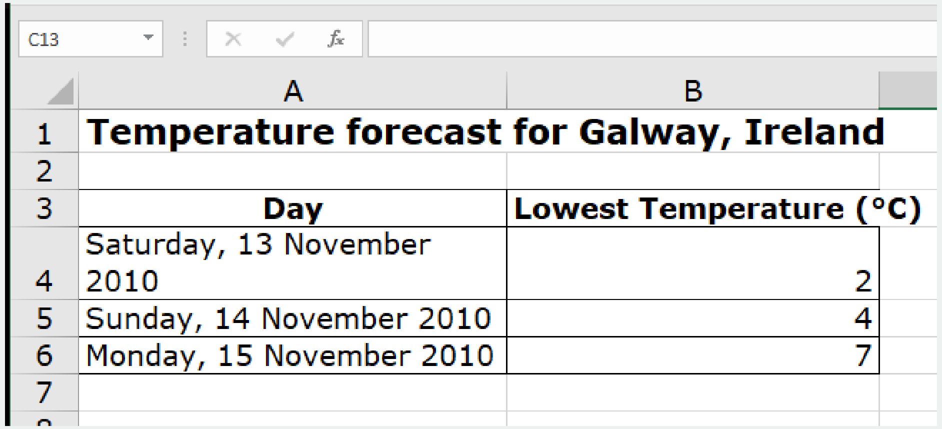
Machine-actionability lies at the heart of FAIR!

Not machine-actionable



Temperature forecast for Galway, Ireland	
Day	**Lowest Temperature (°C)**
Saturday, 13 November 2010	2
Sunday, 14 November 2010	4
Monday, 15 November 2010	7

Machine-actionable



	A	B
1	**Temperature forecast for Galway, Ireland**	
2		
3	**Day**	**Lowest Temperature (°C)**
4	Saturday, 13 November 2010	2
5	Sunday, 14 November 2010	4
6	Monday, 15 November 2010	7
7		

https://5stardata.info/en/

FAIR

More in a dedicated lecture!

In the meantime you can watch:

- Let's Make Our Data FAIR! Webinar for GO-FAIR

https://www.youtube.com/watch?v=dEV2HnraqaI

Next lecture

14 March 2022

Digital Preservation

- How to keep data available for the long term?

Digital Preservation Introduction

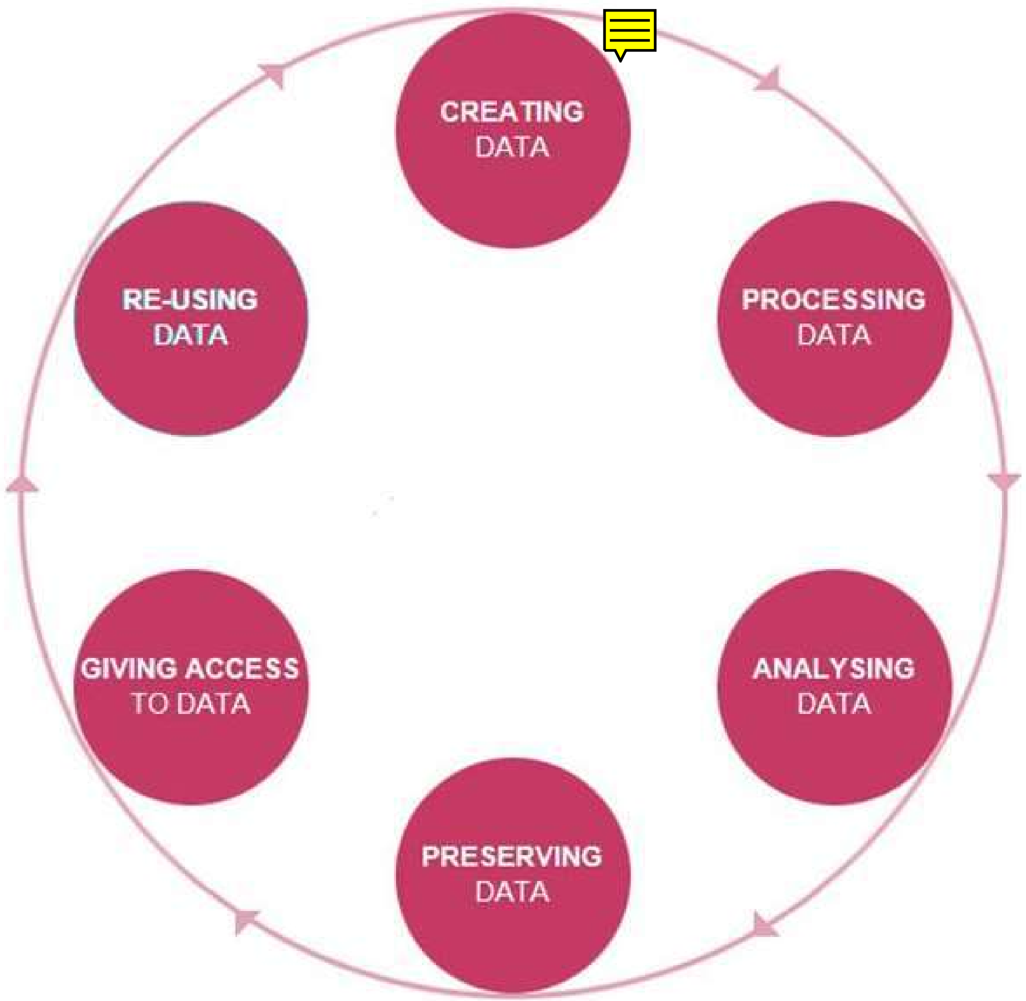
Andreas Rauber

Department of Software Technology and Interactive Systems

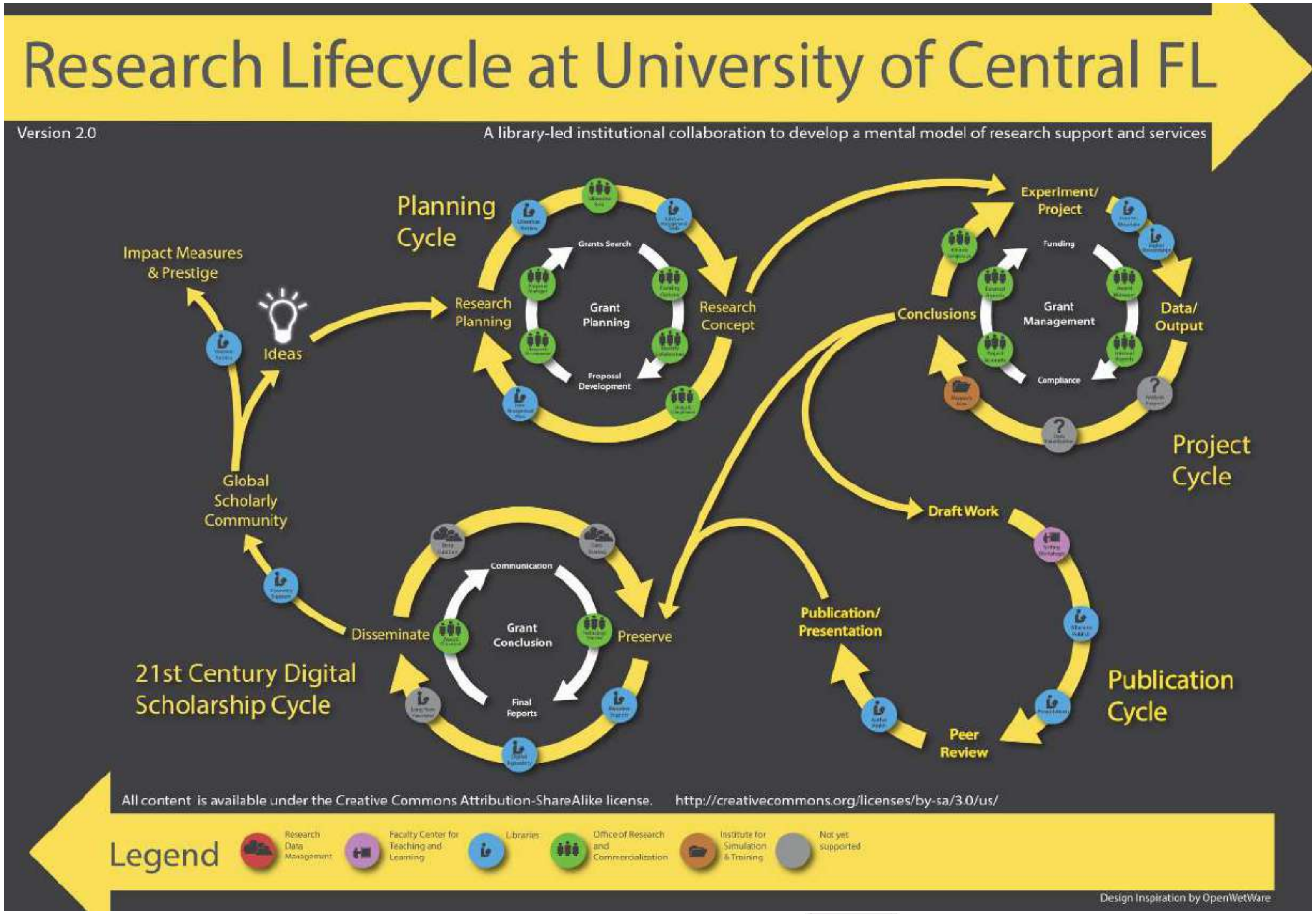
Vienna University of Technology

http://www.ifs.tuwien.ac.at/~andi

UK Data Archive Lifecycle model



University of Central Florida Lifecycle Model



http://guides.ucf.edu/ScholarlyCommunication/ResearchLifecycle

Overview

- What are the challenges in Digital Preservation?
- What methods can we use to counter them?
- What's the issue with file formats?
- What do we have in our repository?

Why do we need Digital Preservation?

Questions / discussion:

- What is *Digital Preservation?*

Why do we need Digital Preservation?

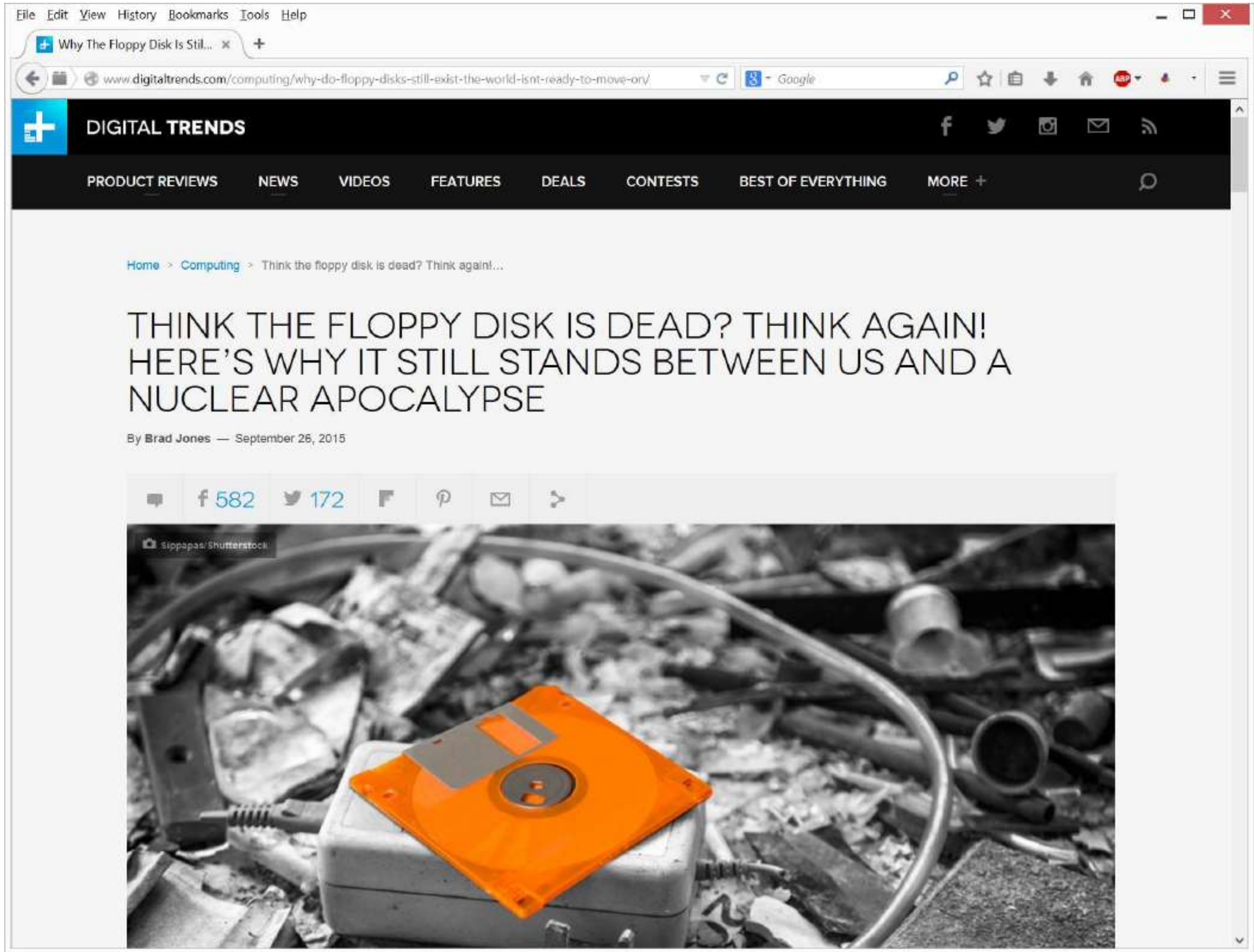
Why do we need Digital Preservation?

1. Physical Preservation (Bit-stream preservation)

- Transfering to current storage systems
  - note: transfer may not be trivial (file systems, encodings, relative references, copy protection,...)
- Ensure redundancy
  - technologically
  - geographic spread
- Access, security
- Error detection, recovery, disaster planning

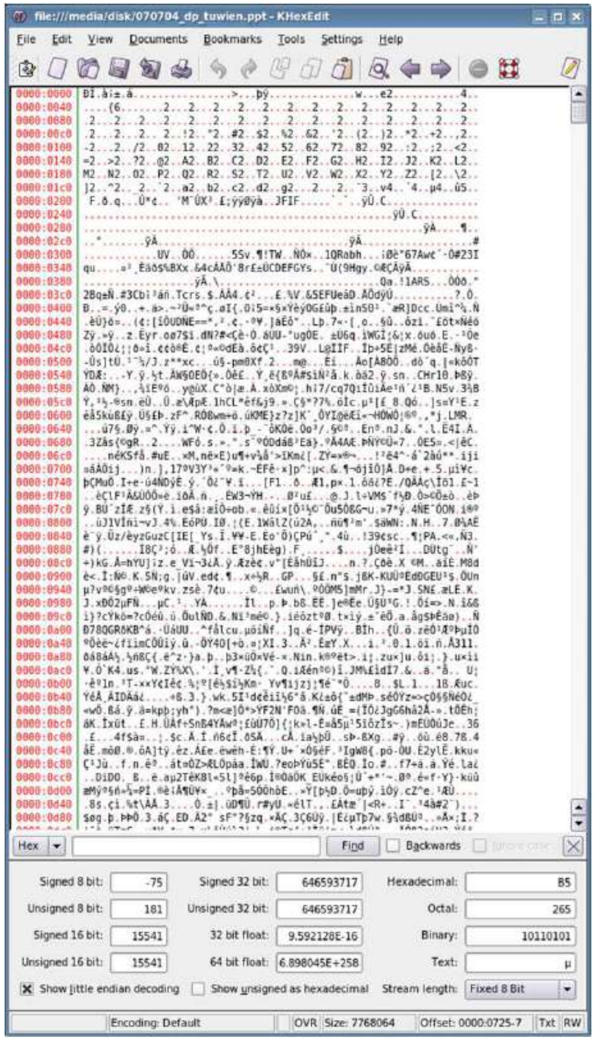
Why do we need Digital Preservation?

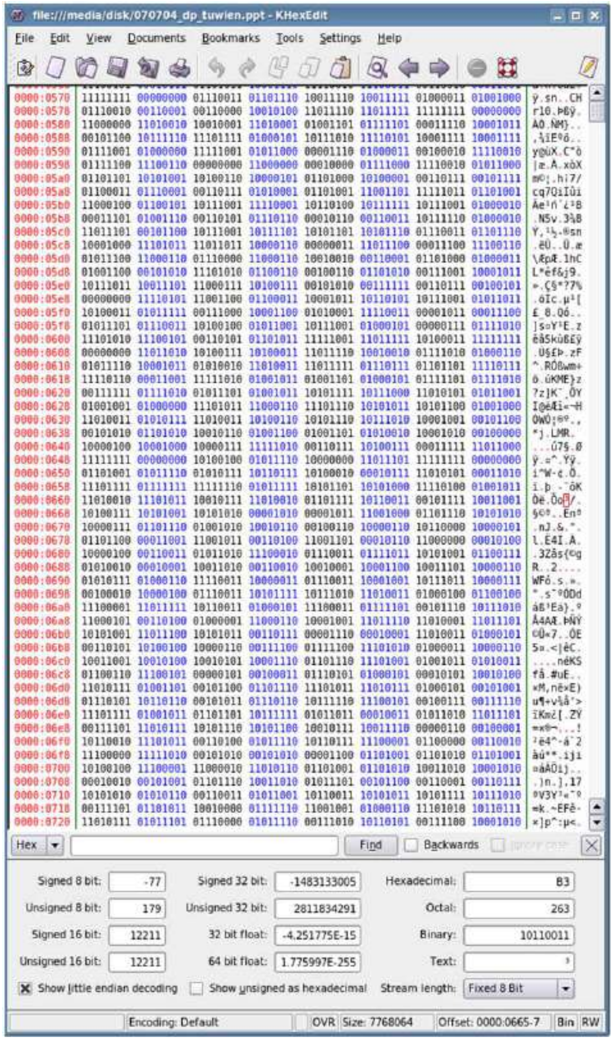
Just as a curiosity:

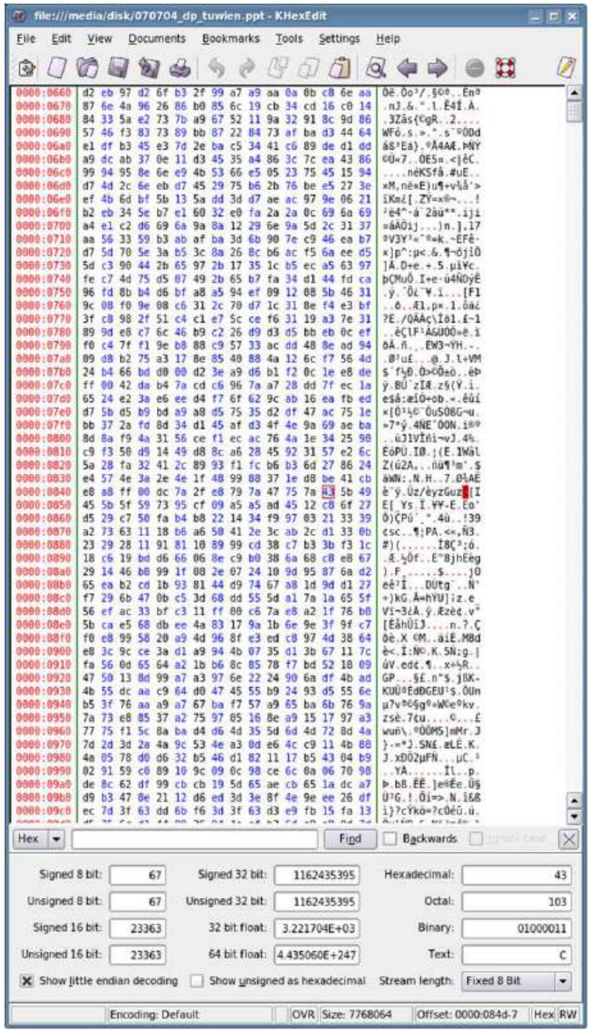


http://www.digitaltrends.com/computing/why-do-floppy-disks-still-exist-the-world-isnt-ready-to-move-on/

Why do we need Digital Preservation?







Why do we need Digital Preservation?

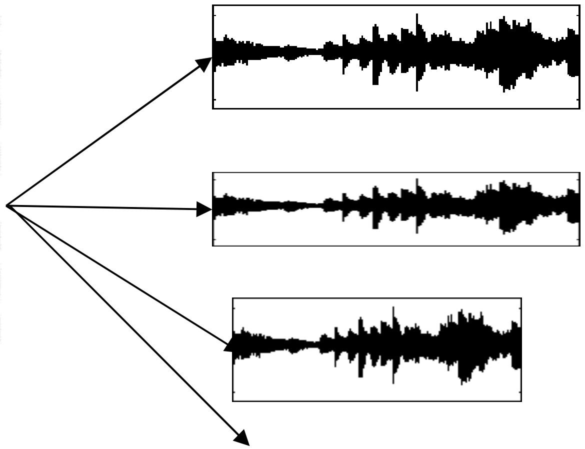
2. Logical Preservation

- Digital Objects require specific environment to be accessible :
  - Files need specific programs
  - Programs need specific operating systems (-versions)
  - Operating systems need specific hardware components
- SW/HW environment is not stable:
  - Files cannot be opened anymore
  - Embedded objects are no longer accessible/linked
  - Programs won‘t run
  - Information in digital form is lost
    (usually total loss, no degradation)
- Digital Preservation aims at maintaining digital objects authentically usable and accessible for long time periods.

Strategies for Logical Preservation

Another way of viewing this…

http://en.wikipedia.org/wiki/GNU_LilyPond



…………

https://en.wikipedia.org/wiki/Odeon_of_Herodes_Atticus

https://commons.wikimedia.org/wiki/File:Stereoanlage_Vision_2000.jpg

https://en.wikipedia.org/wiki/Konzerthaus,_Vienna

………………………………………………

Why do we need Digital Preservation?

Homann Heirs Map
Wikimedia 1747

Cary Map of Au
Wikimedia 1801

Austria 1999 CIA map
Wikimedia, 1999

Mitchell Map of Austria, Hungary and Transylvania
Wikimedia 1850

ifs FACULTY OF !NFORMATICS

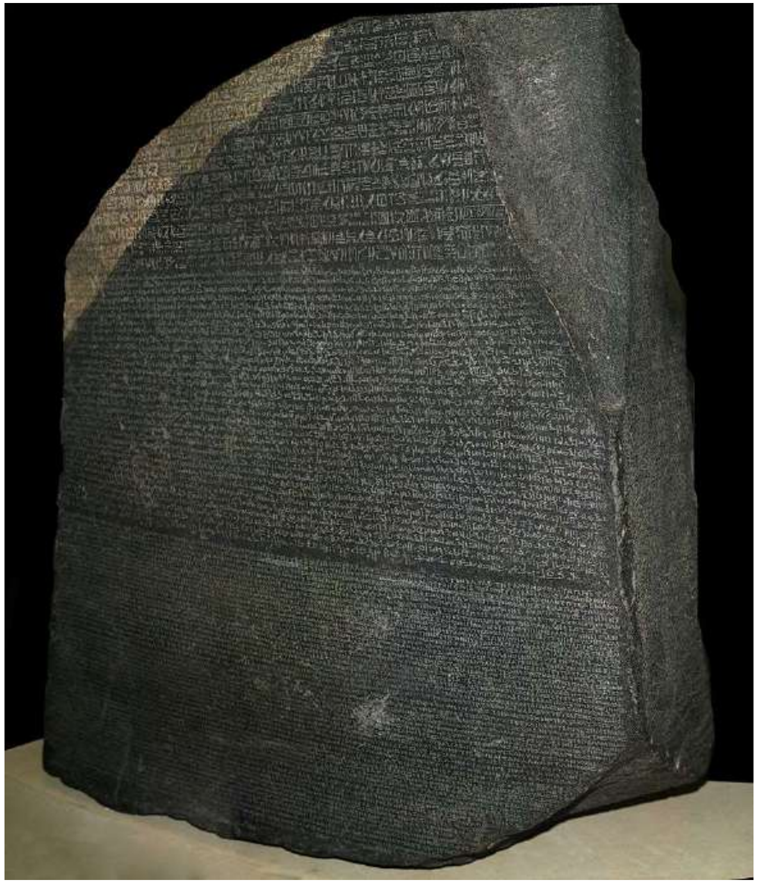
Why do we need Digital Preservation?

3. Semantic Layer: information object

- How to interpret the data (information?) in the objects?
  - terminology changes:
    changes in country names, borders, connotation of words,...
  - concept changes:
    drunk driving: before 1998: 0.8‰ , afterwards 0.5‰
  - transformations: currencies/exchange rates, sensor resolutions,
  - provenance: actions applied to objects
    sources: who? / which sensor?, transformations, post-processing
  - context of objects:
    understanding the context of decisions, side-effects, quotations, calibration timestamps
- For preserving digital information, all 3 layers need to be addressed

Why do we need Digital Preservation?

One of the most famous examples…



https://commons.wikimedia.org/wiki/File:Rosetta_Stone.JPG

Why do we need Digital Preservation

- The goal of Digital Preservation is to **maintain digital objects accessible and usable in an authentic manner for a long term** into the future.

Why do we need Digital Preservation?

Questions / discussion:

- What is *digital data*?
- What is *digital storage*?
- What do we mean by
  - *accessible*?
  - *authentic*?
  - *long-term*?

Why do we need Digital Preservation?

- Essential for all digital objects
  - Office documents, accounting, emails, …
  - Scientific datasets, sensor data, metadata, …
  - Applications, simulations, business processes, …

- All application domains
  - Cultural heritage data
  - eGovernment, public administration
  - Science / Research
  - Industry
  - Health, pharmaceutical industry
  - Aviation, control systems, construction, …
  - Private data
  - …

Why do we need Digital Preservation

3 levels of threat / preservation

1. Bit rot – physical preservation / bit preservation
   Physical Layer: how to keep the 0's and 1's
2. Object formats – logical preservation
   Logical Layer: how to remain able to open a file, run a program
3. Authenticity, interpretability – semantic preservation
   Semantic Layer: how to ensure we can understand/interpret data correctly

- What can we do?

Overview

- What are the challenges in Digital Preservation? (recap)
- What methods can we use to counter them?
- What’s the issue with file formats?
- What do we have in our repository?

Bit-level preservation

- Maintain bit-sequence
- Redundant storage:
  - Lockss: lots of copies keeps stuff safe
  - Cloud
- Distributed storage – physically separated
- Different technologies / platforms / production batches
- Controlled storage conditions
- Regular maintenance: tape rewinding, disc spinning, …
- Maintain devices for accessing storage!
- Trade-off capacity, energy, effort

Bit-level preservation

Questions / discussion:

- How long do tapes / CDs / DVDs / HDDs / SSD last?
- What are the costs of bit-level preservation?
- What are the logistic challenges?
- Is a DVD that lasts for 200 years a solution?
- What would be the most durable storage technologies?
- What is "digital storage"?
- Distribution and Trust?
- Are we allowed to store redundantly? in the cloud?
  - Copyright
  - Copy protection
  - Distributed objects, referenced via URL? DOI?

Why do we need Digital Preservation

3 levels of threat / preservation

1. Bit rot – physical preservation / bit preservation
   Physical Layer: how to keep the 0's and 1's
2. Object formats – logical preservation
   Logical Layer: how to remain able to open a file, run a program
3. Authenticity, interpretability – semantic preservation
   Semantic Layer: how to ensure we can understand/interpret data correctly

- How? What can we do?

Strategies for Logical Preservation

Technology Museum

- Keep the hardware (drives, computer,…)
- \+ Maintains full functionality
- \+ Creates time buffer to develop more permanent strategies
- \+ Requires detailed documentation of HW and SW, but this also helps
- \+ Only strategy for some types of objects? (which?)
- \- Economically and technically infeasible to maintain spare parts forever
- \- Requires huge "museum"
- \- Requires highly specialized know-how for all platforms and software

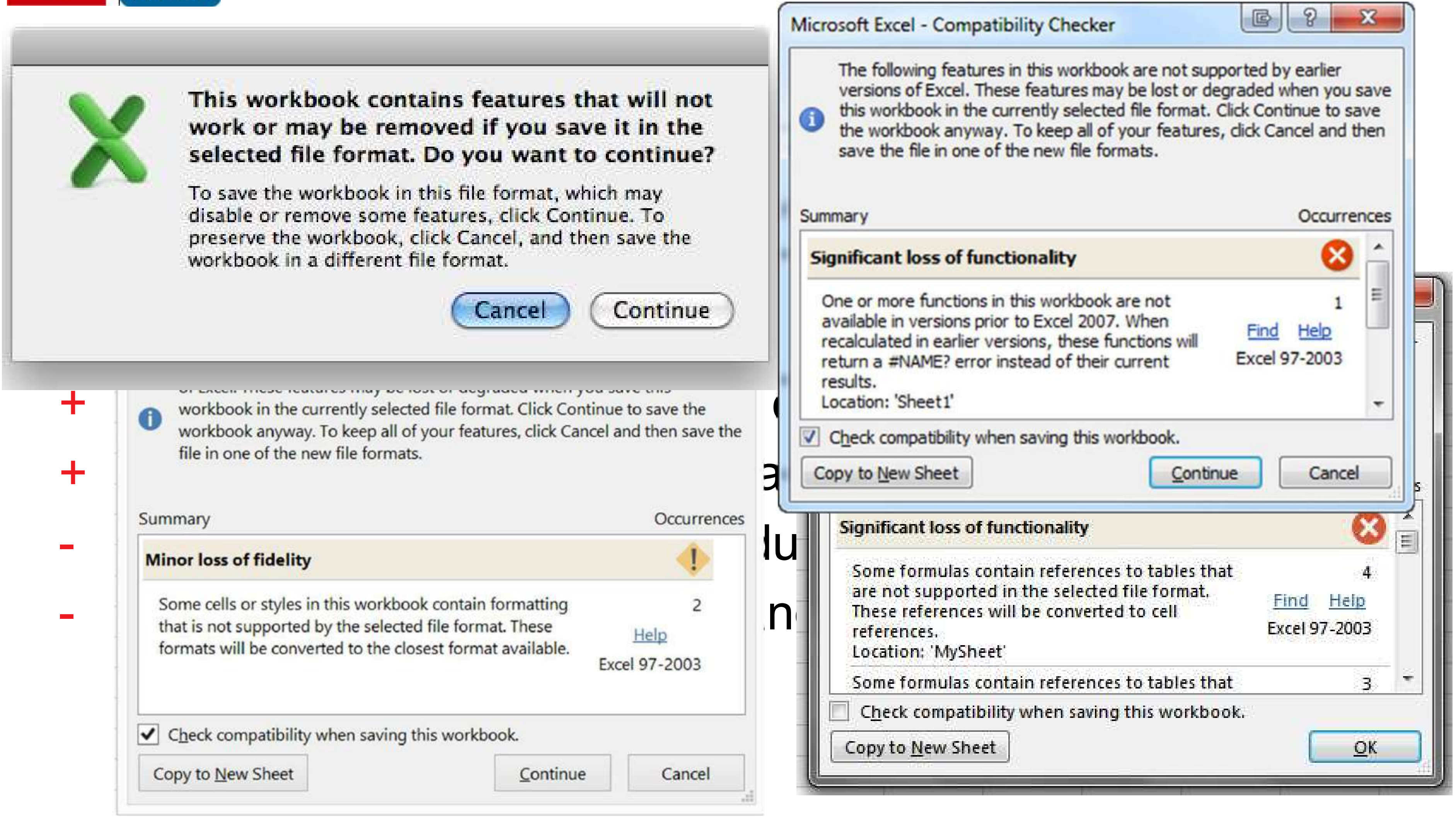
Strategies for Logical Preservation

Migration

- Transform into different format
- Continually or on demand (Viewer)

+ Widely used
+ Possibility to compare at time of migration
+ Resulting objects are always accessible
- Possibly undesired changes during migration
- Needs to be repeated again and again

Strategies for Logical Preservation



https://support.office.com/en-us/article/Use-Office-Excel-2010-with-earlier-versions-of-Excel-2fd9ffcb-6fce-485b-85af-fecfd651a5ac

Strategies for Logical Preservation

Emulation

- Emulation of Hardware or Software (OS, application)

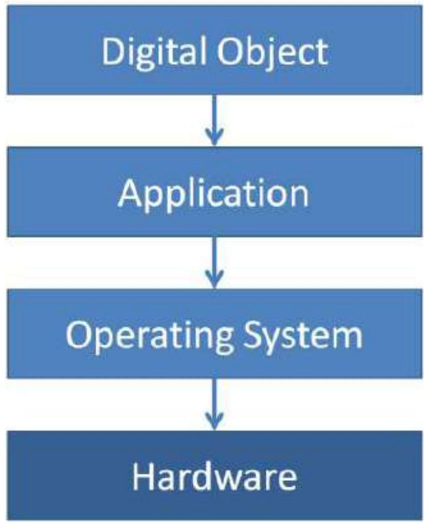
+ Widely used principle
+ Many emulators available
+ Potentially preserving complete functionality
+ *Document is unchanged*

- *Document is unchanged*
- Complex technology, lot of research required
- Requires detailed documentation of the system
- Requires experience how to interact with emulated historic system in the future
- Emulators must be migrated as well
- Emulators potentially erroneous (Complexity)

Strategies for Logical Preservation

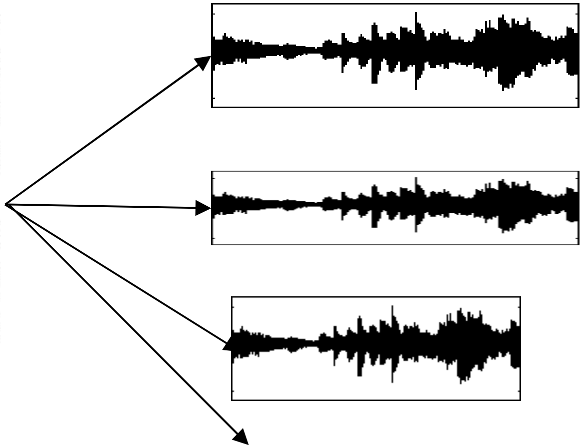
Excursion: Emulation vs. Migration

- Different on the pragmatic level, but conceptually identical
- Change occurs somewhere in the viewpath
- Have basically the same advantages/disadvantages and characteristics
- None of them guarantees identical rendering/performance of digital objects
- Many variants (e.g. viewer, virtualization)
- Need to be evaluated the same way



Emulation vs. Migration – remember this?

http://en.wikipedia.org/wiki/GNU_LilyPond



....

https://en.wikipedia.org/wiki/Odeon_of_Herodes_Atticus

https://commons.wikimedia.org/wiki/File:Stereoanlage_Vision_2000.jpg

https://en.wikipedia.org/wiki/Konzerthaus,_Vienna

....

Strategies for Logical Preservation

Standardization

- Using open or de-facto standards
- \+ Simplifies DP process
- \+ Many tools available
- \+ Tools for standards are easier to build also in the future
- \- Significant effort required for standardization
- \- Loss at converting into standard
  (who is responsible?)
- \- Some object types cannot be standardized

Strategies for Logical Preservation

Standardization - Excursion into file formats
Proprietary vs. Open

- Proprietary
  - Documentation mostly not available
  - License and patent rules
  - License agreements subject to change
  - Restrictions for use and modifications may apply
- Open
  - Documentation available!
  - Unlimited use
  - No license fee
  - Open for modifications
  - No patent owners
- But: sometimes proprietary may better than open - **why**?
- Is the concept of "file formats" still useful?

Strategies for Logical Preservation

Limiting Accepted Formats

- Similar to standardization
- \+ Reduces challenge to smaller number of formats
- \- Does not solve the problem
- \- Limits the type of objects that can be accepted
- \- Potential loss at conversion
- \- Requires strict control of formats (and what's in them!)

Strategies for Logical Preservation

Data/Information Extraction

- Create abstract representation of information (e.g. databases or documents -> XML)
+ Independent of specific infrastructure
+ Many tools available
+ Easier to develop tools in the future
- High effort to develop tools for specific abstraction scenario
- Limited functionality of tools designed to interpret information, many aspects not preservable
- Cannot be applied to all types of objects

Strategies for Logical Preservation

Encapsulation

- Add metadata, software,... (representation information) to object („onion“)

\+ Simplifies search for preservation solution on demand, offering several potential layers

\+ Always allows for the application of several other strategies at different levels

\- Does not solve the problem

\- Even with all information encapsulated we may not be able to find a solution

Strategies for Logical Preservation

Universal Computing Platform

- Example: UVC: Universal Virtual Computer (IBM)
- Abstract virtual machine, intermediate platform that can be implemented on many other platforms

\+ Works for documents and software

\+ A kind of standardization for platform, reduces development effort

\+ Can test solution at time when being developed

\- Pretty complex (cf. Java, but that's still simple)

\- High effort at time of preservation

\- Requires cooperation of the producers of information

\- High risk of loosing aspects of information

Strategies for Logical Preservation

Backwards Compatibility and Version Migration

- current SW reads old versions and performs migration

+ Usually available
+ Creates time buffer for more permanent solutions
+ sometimes equal or better functionality

- Doubtful whether this will work for a long time (why?)
- Each change might lead to unwanted changes
- No guarantee from part of the producer of the SW

Strategies for Logical Preservation

Viewer

- Migration on demand, interpretation by Viewer software
+ Original datastream unchanged, interpreted directly
+ No continuous migration
+ No cummulative errors
- Viewer sometimes cannot process all (parts of) objects
- Increasing time delay when developing viewers
- Viewer SW must be carried along with technology changes
- Hard to evaluate whether viewer is correct

Strategies for Logical Preservation

Non-digital Strategies

- Printing to paper, microfilm, ...
- \+ Reqires transformation to readable form -> stable
- \+ Coding of digital data is possible
- \+ Lots of experience in handling analog data carriers
- \+ High stability -> Bit-stream Preservation
- \- Loosing functionality, loosing advantage of digital technology
- \- Not applicable for all objects
- \- High costs for preserving some of the analog data carrier material, low storage density, ...
- \- Even this can be “buggy” (Xerox bug, manipulation)

Strategies for Logical Preservation

Data Recovery, Data Archeology

- Analysis of bit-stream to interpret data, digital forensics

\+ Probably only approach to recover "lost" information

\- No guarantee that it works

\- Without sufficient documentation close to "guessing"

\- Extremely high costs per object

\- Hard to guess whether it may be successful for a given object

Logical Preservation

Summary

- Changing object, environment
- Loss upon migration / emulation
- Decision of what to preserve → **Significant Properties**!
- How to detect/document what you lost?
- Range of strategies available, none is perfect
- Combination of strategies
- No solution forever -> DP is a process!

Logical Preservation

- Preservation Planning
- Identify objects at risk
- Standardization reduces risk (why?)
- Apply preservation actions such as migration / emulation / HW-museum
- Identify what you need to preserve (significant properties)
- Identify suitability of tools
- Find out what you can preserve / what you loose
- Do it, document it, verify it, monitor it

Logical Preservation

Questions / Discussion:

- What are the problems of logical preservation?
- What is the optimal strategy?
- What is the optimal strategy for a specific object?
- What is a good format / platform (e.g. to migrate to)?
- What are characteristics of good formats/platforms/… ?
- How can we identify objects at risk?
- When is a format "more/less risky"?
- What is a file format?
- How can we find out what we loose with a strategy?

Logical Preservation

Questions / Discussion (2):

- What is the difference between emulation and migration? Are they different? Are they not different?
- What are the significant properties of an object / process?
- “I want to preserve everything” – (how) can we do this?
- What is the “original object”?
- What is the complexity of each strategy? Costs? Effort?
- What know-how do we need to decide on a strategy?
- What would be potential risks/difficulties e.g. for construction plans? Medical imaging (DICOM)?

Logical Preservation

Questions / discussion (3):

- Which objects are most at risk?
- Which objects are most difficult to preserve?
- How do we preserve entire business processes?
- If we loose significant properties with a strategy, what is the impact on authenticity? Can we use a “changed” object?
- What is the difference to systems engineering?

Why do we need Digital Preservation

3 levels of threat / preservation

1. Bit rot – physical preservation / bit preservation
   Physical Layer: how to keep the 0's and 1's
2. Object formats – logical preservation
   Logical Layer: how to remain able to open a file, run a program
3. Authenticity, interpretability – semantic preservation
   Semantic Layer: how to ensure we can understand/interpret data correctly

- How? What can we do?

Semantic preservation

- Threats at semantic level
  - meaning of terms change: city names, …
  - measurement scales, sensor sensitivity, …change
  - interpretation of facts change: alcohol levels, …
- Rather long-term, but subtle to notice
- Consider context of objects
  - purpose, setting, limitations, cultural context, related objects, …

Semantic preservation

- Approaches / solutions:
  - Semantic enrichment
  - Metadata
  - Migration at semantic level
  - Documentation of context
  - Tracing of metadata
  - Document intended meaning / interpretation

Semantic preservation

Questions / discussion:

- How do we identify need for action?
- What is the risk of missing timely action?
- How do we solidly identify and document context?
- How can we implement semantic enrichment / semantic migration, …?
- What about security issues?
- Is PDF save? PDF/A?
- Who is allowed to have access to which documents? Who had access to them?
- Are differences in the communication protocol at an API level a problem of logical or semantic preservation?

From Data to Processes

- Assume we know how to preserve data - **Is this sufficient?**
- Preserving data: Data Management Plans
  - describing data and context: provenance, authenticity, representation information,...
  - range of (ambiguous) definitions of context
  - But: mostly not actionable, not enforcable,...
  - BUT: data are (just) results of processes!
- Processes may be needed to
  - verify data
  - understand provenance
  - re-use process on new data
  - integrate data over time
- **Process curation instead of data curation!**

What can go wrong?

- A lot……
- Many times: trivial mistakes!
- But also more serious / conceptual issues
- From mistakes to actual fraud
- Overlap with security research, digital forensics, …
- Roles and Responsibilities, Policies, …

What can go wrong?

- Ingest / Standardization:
  - Who is performing the initial migration?
  - Who is liable?
  - Who will need to manage any problems subsequently?
- Migration
  - Something added? E.g. Word -> TXT, Excel -> TXT
  - Something lost?
- What is a PDF file?
  - A malicious invoice…
  - A multi-purpose paper: https://bigdata.uni-saarland.de/publications/p1972-dittrich.html

What can go wrong?

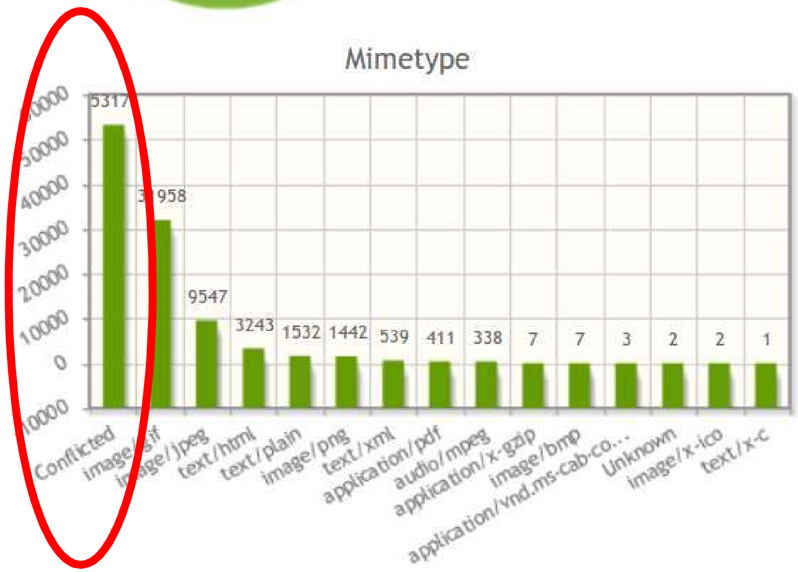
- Dissemination
  - (c.f. OAIS Model: AIP to DIP transformation)
  - Decide which information to release in which format
  - Consider metadata!
    c.f. Supriya Adhatarao, Cédric Lauradoux: Exploitation and Sanitization of Hidden Data in PDF Files. arXiv:2103.02707, March 2021
    Analyzing PDFs published by security agencies: metadata revealing weak links, less than 10% of agencies sanitized part of their documents, 65% of these still contained sensitive information)

What can go wrong?

- Collection Profiling: what is in the repository?

ffident

Droid



Overview

- What are the challenges in Digital Preservation? (recap)
- What methods can we use to counter them?
- What’s the issue with file formats?
- What do we have in our repository?

What Kind of File is This?

- By external characteristics (file extensions)
- By internal characteristics („magic number“, „signature“).

E.g. A TIFF file begins with …

- Bytes 0-1
  - The byte order used within the file. Legal values are: “II” (4949.H) / “MM” (4D4D.H)
- Bytes 2-3
  - An arbitrary but carefully chosen number (**42**) that further identifies the file as a TIFF file.

What Kind of File is This?

- What's wrong with file extensions?
  - Not necessarily unique (e.g. wks)
  - Granularity not sufficient
  - Can be altered by users
- Formats vs. Format profiles
  - PDF is not **one** format
  - DOC is not **one** format
  - TIFF is not **one** format
- A lot of things can go wrong: by coincidence or maliciously!
  - Word -> TXT, Excel -> TXT
  - Standardization: "We only use PDFs, we've no problem"
    - Not all PDFs are created equal…
    - A PDF file? https://bigdata.uni-saarland.de/publications/p1972-dittrich.html

What's Wrong with MIME Types?

- Insufficient depth of detail
  - No requirements regarding syntax and semantic description
  - No requirement for complete disclosure, especially of proprietary formats
- Insufficient granularity
  - Both tiled RGB GeoTIFF with LZW and striped bi-tonal TIFF-FX with Group 4 are typed as "image/tiff"
  - All of PDF 1.0 – 1.4, PDF/X-1, X-2, X-3, and PDF/A are typed as "application/pdf"
  - These variants might require radically different workflows

Why Do We Need a Format Registry?

- Knowledge base of file format representation information
  - properties,
    - what do they mean?
    - how to read them?
  - supporting software
- Unification of vocabulary (entity names and mappings)
- A (single?) access point to various information about formats though a common API

File Format Registries

- PRONOM
  - http://www.nationalarchives.gov.uk/pronom/
- Global Digital Format Registry (defunct)
  - http://library.harvard.edu/preservation/digital-preservation_gdfr.html
- Unified Digital Format Registry (UDFR) (defunct)
  - http://www.udfr.org/
- Sustainability of Digital Formats Planning for Library of Congress Collections
  - https://www.loc.gov/preservation/digital/formats/index.shtml
- FileExt
  - http://filext.com

PRONOM

You are here: Home > Services for professionals > Preservation > PRONOM > Search by format > Details: Summary

The technical registry
PRONOM

Welcome : About Add an entry
Search ? Help Information resources

Details: File format summary

? Help : detailed report on file format

Simple search | File format | PRONOM Unique Identifier | Software | Vendor | Lifecycles

Details for: Microsoft Word for Windows Document 97-2003

Save as... XML | CSV Print

Go to: Summary | Documentation › | Signatures › | Compression › | Character encoding › | Rights › | Reference files › | Properties ›

Summary

Name	Microsoft Word for Windows Document
Version	97-2003
Other names	Microsoft Word for Windows Document (97-XP)
Identifiers	MIME: application/msword Apple Uniform Type Identifier: com.microsoft.word.doc PUID: fmt/40
Family	
Classification	Text (Wordprocessed)
Disclosure	None
Description	With the release of Word 97, Microsoft revised the native binary word processing format, which is based on its generic OLE2 Compound Document Format. The format is proprietary and Microsoft does not make details of its structure public. The information here is derived primarily from OpenOffice.org's reverse-engineered documentation of the format and should not therefore be regarded as definitive. A Word document is stored as a 'WordDocument' stream within a Compound Document Format file. The format remained unchanged with the releases of Word 2000, 2002 and 2003.
Orientation	Binary
Byte order	Little-endian (Intel)
Related file formats	Has priority over OLE2 Compound Document Format Is subsequent version of Microsoft Word for Windows Document (6.0/95) Is subtype of OLE2 Compound Document Format

Registry Content

- Descriptive information
- Identifiers
  - MIME
  - Pronom Unique Identifier (PUID)
- Relationships to formats
- Technical environment
- References and links
- Risk factors

Registry Use Cases

- Identification
  - "I have a digital object; what format is it?"
- Validation
  - "I have an object purportedly of format *F*; is it?"
- Transformation
  - "I have an object of format *F*, but need *G*; how can I produce it?"
- Characterization
  - "I have an object of format *F*; what are its significant properties?"
- Risk assessment
  - "I have an object of format *F*; is it at risk of obsolescence?"
- Delivery
  - "I have an object of format *F*; how can I render it?"

Identification Tools

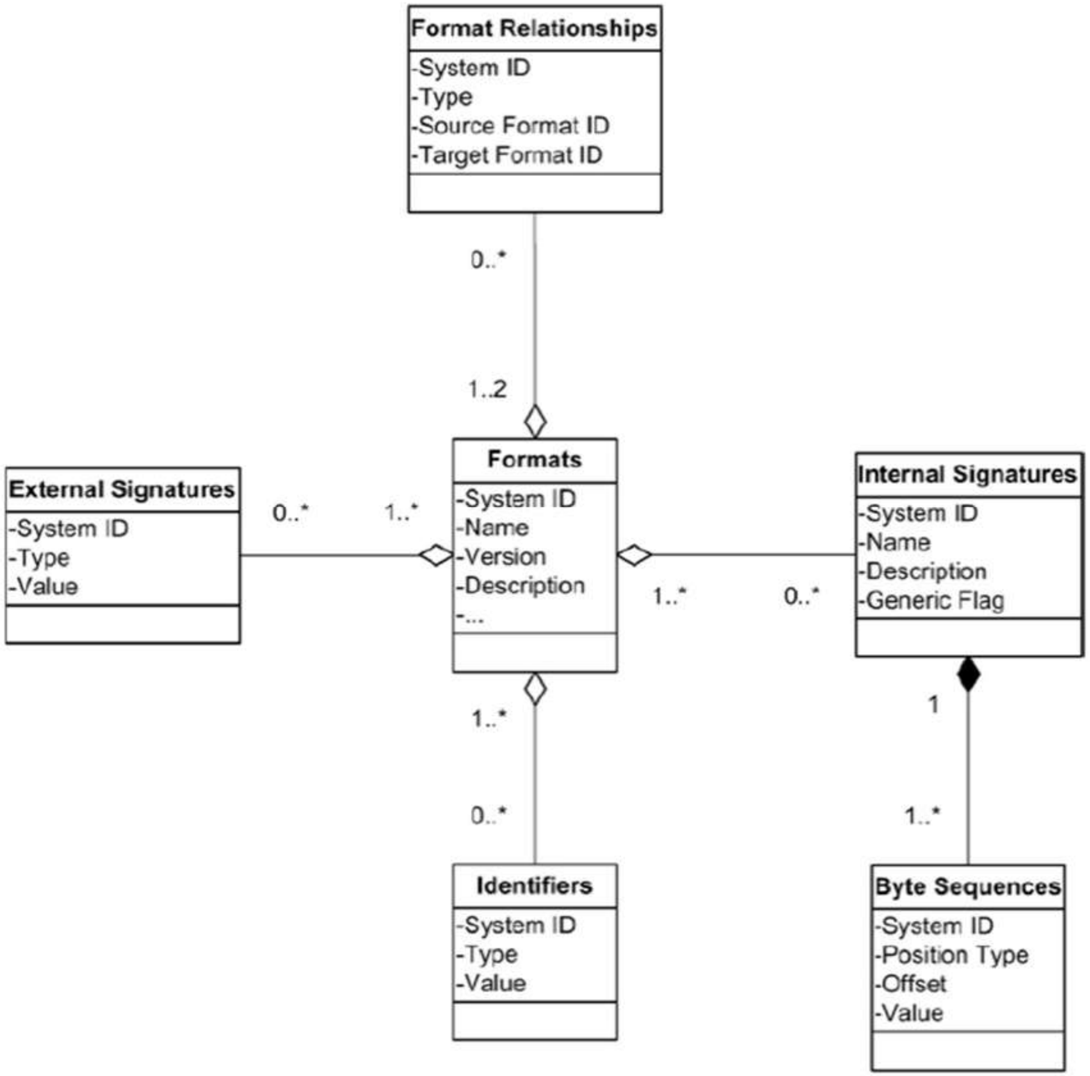
- DROID (Digital Record Object Identification)
  - relies on PRONOM
  - The National Archives, UK
- JHOVE
  - JSTOR/Harvard Object Validation Environment
  - Validation and characterisation
- FITS (File Information Tool Set)
- Apache Tika: file content analysis
- veraPDF: PDF/A validation
  http://verapdf.org/home/

ffident

Droid

Signatures in DROID

- External signatures
  - -File extensions
- Internal signatures
  - -Format indicators in the bitstream
  - -Byte sequences



JHove

- JSTOR/Harvard Object Validation Environment
- Modular and extensible Java-based architecture
  - Image modules: GIF, JPEG, JPEG2000, TIFF
  - Document modules: ASCII,HTML,PDF, UTF-8, XML
  - ...
- Three functions
  - Identification
  - Validation
  - Characterisation
- JHove2
  - Identification and validation
  - Feature extraction
  - Policy based assessment
  - Able to handle complex objects

The TIFF Module…

- Tagged Image File Format (TIFF) raster images TIFF 4.0, 5.0, and 6.0 [TIFF 4.0, TIFF 5.0, TIFF 6.0]
- Baseline 6.0 Class B, G, P, and R [TIFF 6.0]
- Extension Class Y [TIFF 6.0]
- TIFF/IT (ISO 12639:2003) [TIFF/IT] File types CT, LW, HC, MP, BP, BL, and FP, and conformance levels P1 and P2
- TIFF/EP (ISO 12234-2:2001) [TIFF/EP]
- Exif 2.0, 2.1 (JEIDA-49-1998), and 2.2 (JEITA CP-3451) [Exif 2.1, Exif 2.2]
- GeoTIFF 1.0 [GeoTIFF]
- TIFF-FX (RFC 2301) [TIFF-FX]
- Profiles C, F, J, L, M, and S
- Class F (RFC 2306) [Class F, RFC 2306]
- RFC 1314 [RFC 1314]
- DNG (Adobe Digital Negative) [DNG]

FITS - File Information Tool Set

- Main features:
  - Consolidates output
  - Can include raw output
  - Configurable/Extendable

- FITS includes:
  - Droid
  - Metadata Extra
  - Jhove
  - Exiftool
  - FFident
  - File Utility

Conflicts

3 types of conflicts:

1. Inconsistent property naming, e.g: *image_width* and *imagewidth*
2. Competing characterisation results, e.g: tool1 identifies a file as *plain text*, but tool2 identifies the file as *PDF*
3. Close, but not the same property values, e.g: *application/xhtml+xml* vs. *application/xml*.

Validation

- A digital object is **well-formed** if it meets the purely syntactic requirements for its format.
- An object is **valid** if it is well-formed and it meets additional semantic-level requirements.

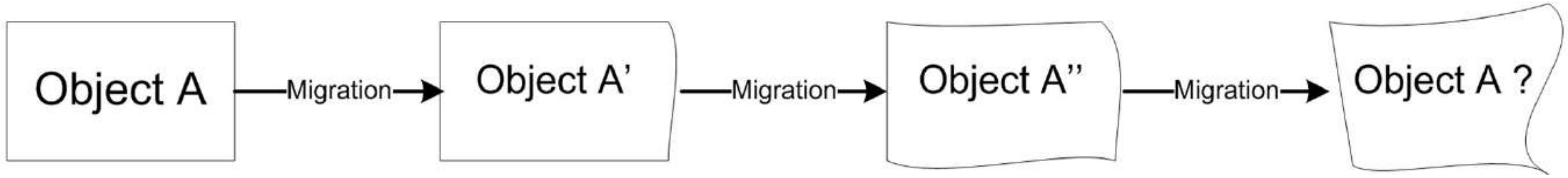
- Validation use cases:
  - "I have an object that purports to be of format *F*; is it?"
  - "I have an object of format *F*; does it meet profile *P* of *F*?"
  - "I have an object of format *F* and external metadata about *F* in schema *S*; are they consistent?"

Validating a migrated image

- Yes, it's in JPEG 2000 format

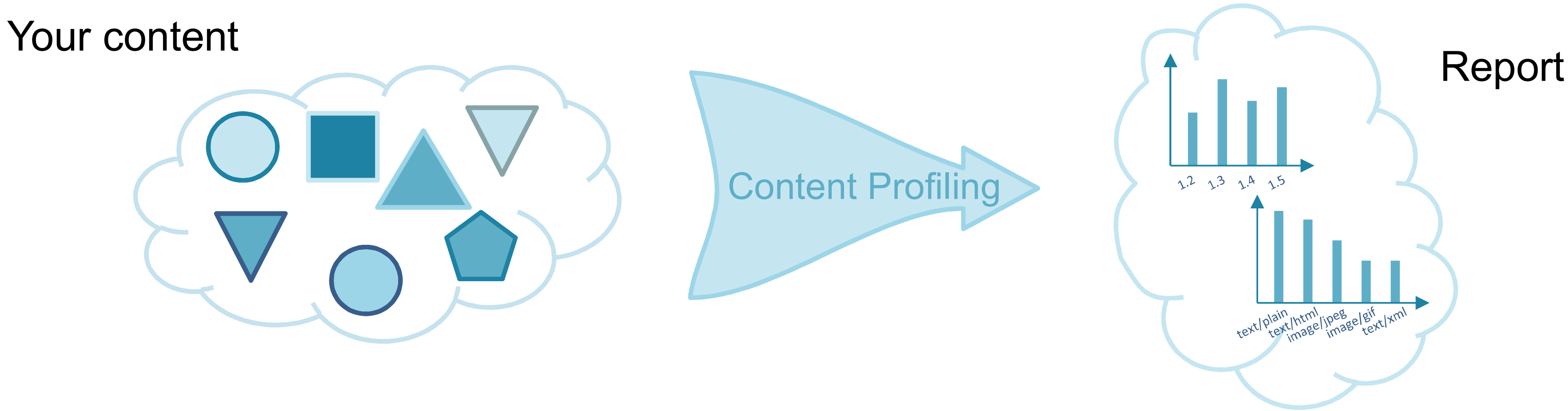
- Yes, it's well-formed
- Yes, it's valid
- Yes, it still has the same dimensions
- …. But is it still the same image?

- We need more characterisation.



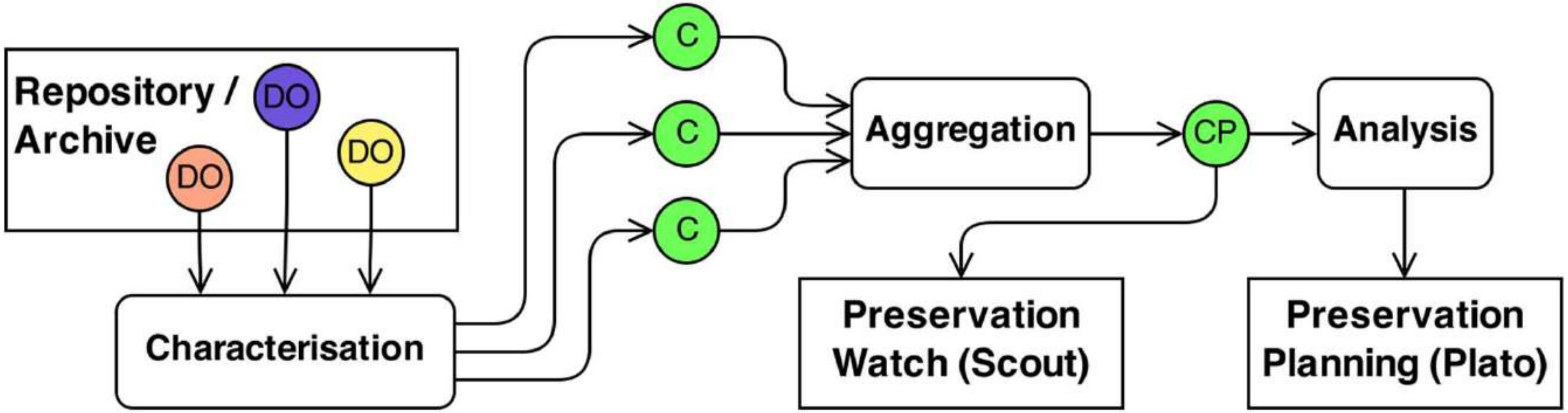
What is Content Profiling

- Better understand your content
- Reveal risks and opportunities
- Part of Preservation Planning
  - Planning and Watch, http://bit.ly/scape-suite



Content Profiling in Details

- A way of getting control over data
  - Decision support
- Consists of:
  - Characterization
  - Aggregation
  - Analysis
  - Reporting / Use for decision making



Aggregation

- Provides an overview of the content
  - Distributions of characteristics
  - Statistics (size, min, max, avg...)
- Data sizes grow dramatically
- Heterogeneity of data
- No universal characterisation tool
  - Combination of such tools

Analysis

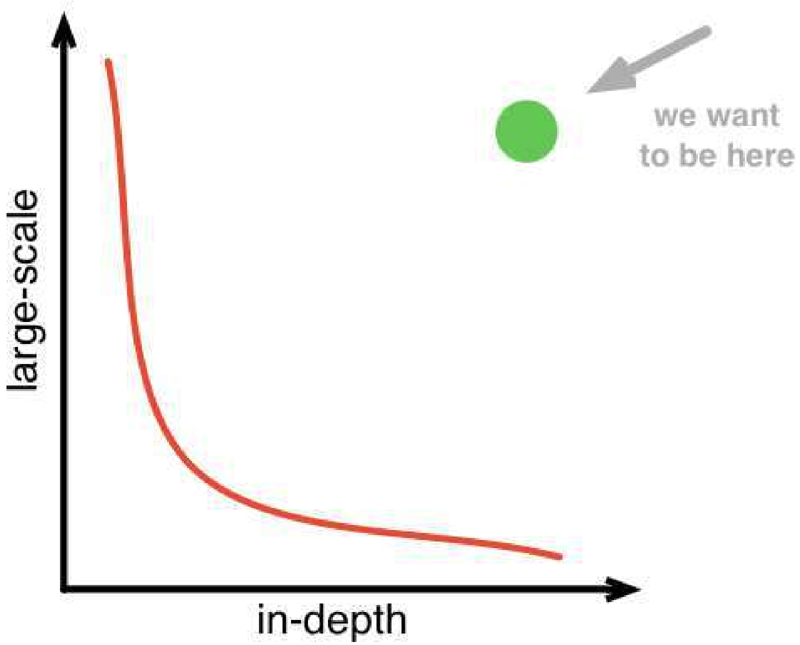
- In-depth research into your content
  - Drilling down
  - Filtering
  - SQL-like queries
- Representative samples generation
  - Based on metadata
  - Outlier detection
  - Stratification across
    - File type,
    - Size,
    - Time, or
    - Any other relevant property

Challenges

- Lack of:
  - Trustworthy tools for characterisation
- Depth
  - Address heterogeneity of data
  - Rise awareness of content properties
  - Combine several characterisation tools
- Quality
  - Conflicts due to combination of characterisation results
  - Resolve conflicting metadata

Challenges

- Scale
  - Effectively analyze substantial amount of metadata
  - Large-scale approaches for content profiling

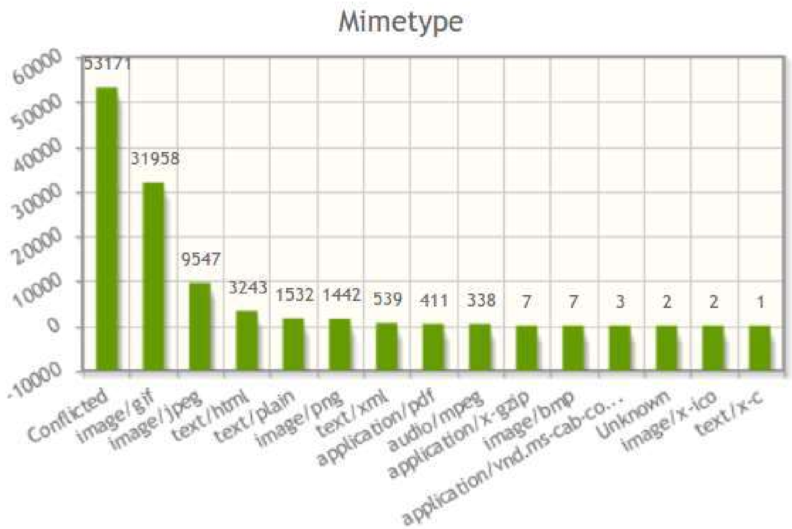


C3PO

- C3PO – Clever, Crafty, Content Profiling of Objects
- Reads and analyzes information from FITS
- Support large-scale database solutions for aggregation and analysis of characterisation metadata
  - MongoDB, http://www.mongodb.org
  - HBase, http://hbase.apache.org
- Aggregation-only mode
  - Useful to fast and explorative generation of a content profile
  - Statistics calculation using predefined filters
  - Single read of data, without computationally expensive ingest and further analysis

C3PO

- Uses characterisation results
- Interface to support other characterisation tools
- Deeper content analysis with interactive visuals through a web-app
- Representative sampling
- Open-source
  - http://ifs.tuwien.ac.at/imp/c3po



C3PO

- Stored metadata property mapping to the existing vocabulary, Planning and Watch Ontology
  - http://purl.org/DP/quality/measures

compression algorithm (Individual)

Definition

Definition

The **URI** of this individual is `http://purl.org/DP/quality/measures#118`

compression algorithm	http://www.w3.org/2004/02/skos/core#prefLabel	compression algorithm
compression algorithm	http://purl.org/DP/quality#scale	http://purl.org/DP/quality/scales#FREETEXT
compression algorithm	http://purl.org/DP/quality#attribute	http://purl.org/DP/quality/attributes#39

C3PO

- Rule-based engine to resolve conflicts in characterization metadata
  - Drools, http://www.jboss.org/drools
- Preservation-specific rules

Target tool	Execution condition	Action
Droid, Exiftool, all	Droid and Exiftool identify a file as “Microsoft Powerpoint Presentation”	Ignore format identifications by other tools
Jhove, all	Jhove reports “text/html” mimetype, other tools report “application/xhtml+xml"	Ignore the “text/html“ mimetype provided by Jhove

Overview

- What are the challenges in Digital Preservation? (recap)
- What methods can we use to counter them?
- What's the issue with file formats?
- What do we have in our repository?

Digital Preservation - Summary

- Is a complex task
- Requires a concise understanding of the objects, their intellectual characteristics, the way they were created and used and how they will most likely be used in the future
- Requires a continuous commitment to preserve objects to avoid the „digital dark hole“
- Requires a solid, trusted infrastructure and workflows to ensure digital objects are not lost
- Is essential to maintain electronic publications & data accessible
- Will become more complex as digital objects become more complex
- Needs to be defined in a preservation plan

Digital Preservation

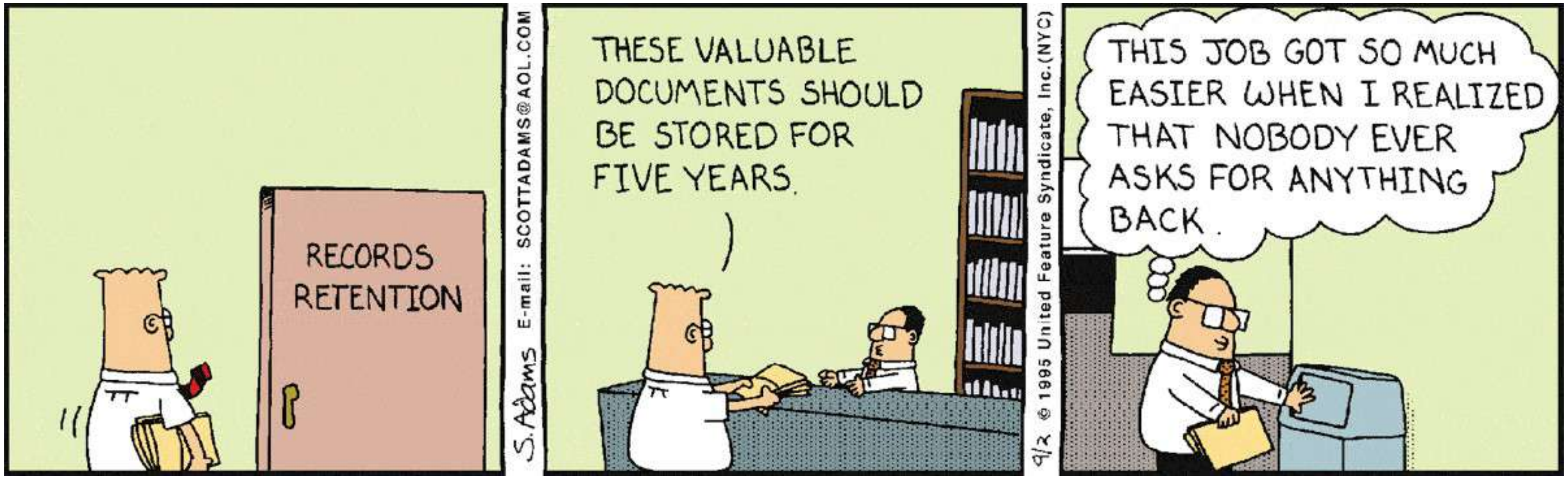
Questions / Discussion:

- At what levels are digital objects threatened?
- What are the time intervals at each level?
- How can we identify objects at risk?
- What can we do to mitigate the risk?
- How can we recover if mitigation fails / is missed?
- What competences do we need?
- How would a training/education program look like?
- How do we know if somebody is doing a good job at DP?

Current Issues

- Atomic file formats, stability of file formats
  - What are the atomic building blocks of information?
  - Can we split information objects?
  - Can we synthesize them? - Help for benchmarking?
- Scalability, Semantics
- Digital forgetting
  - how to decide what to keep and what to forget?
  - keep all? just storage? how to find? utilize? unterstand?
- Sustainable Systems Engineering
  - How can we build preservation-ready systems?
  - How to integrate DP-considerations into software engineering?
- Costs: what does DP cost?
  - cost factors?
  - How to model? evaluate?

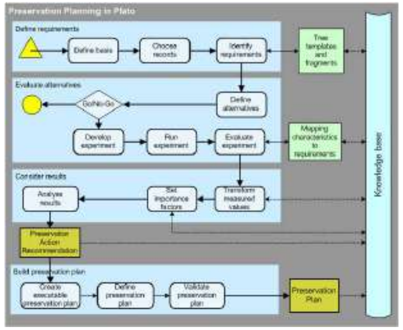
Thank you!

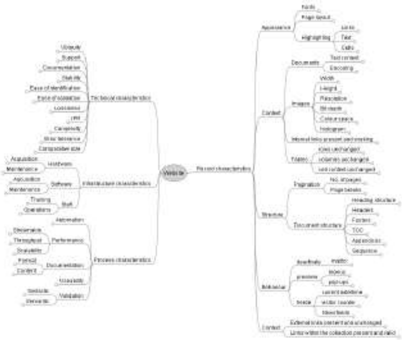


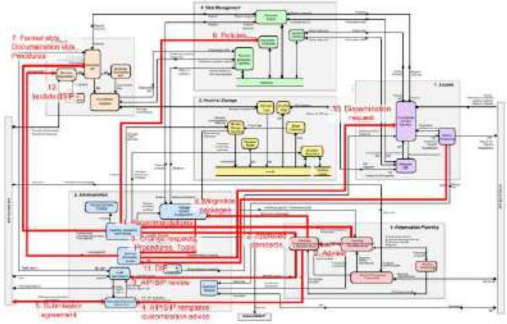
Source: http://dilbert.com/strip/1995-09-02

http://www.ifs.tuwien.ac.at/dp

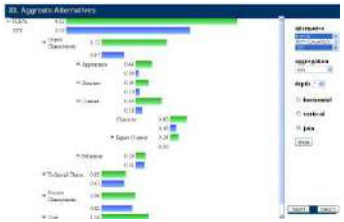
Thank you!







http://www.ifs.tuwien.ac.at/dp



Digital Preservation

OAIS Reference Model

Andreas Rauber
Institute of Software Technology and Interactive Systems
Vienna University of Technology
http://www.ifs.tuwien.ac.at/dp

Outline

- Principles of the OAIS Model
- Technical Overview
- Functional Overview
- Information Modell
- Summary

OAIS and the role of NASA

- National Space Science Data Center
  - NASA's first digital archive
  - has gone through many technology changes since 1966
- Consultative Committee for Space Data Systems
  - International group of Space Agencies
  - developed a set of standards across disciplines
  - evolved into working group ISO TC 20/ SC 13 around 1990

    TC20: Aircraft and Space Vehicles

    SC13: Space Data and Information Transfer Systems

What's a reference model

- A Framework
  - to understand the relationship between significant entities in an environment
  - for the development of consistent standards or spezifikations to support this environment.
- A reference model
  - is based on a small number of unified concepts
  - is an abstraction of the core concepts, their relationships and interfaces within as well as external to the framework
  - can be used as a basis for training and to explain standards.

OAIS

- OAIS is a reference model
- No design specification, no data model, no set of functional requirements!
- Describes elements and concepts that are relevant for a project
- Goal: determine, which parts of the reference model map to which subsystems, functions and responsibilities in a desired solution.

OAIS Sources of Information

- Reference Model for an Open Archival Information System (OAIS), ISO 14721:2012
- Blue Book, CCSDS 650.0-B-1, January 2002
- Pink Book, CCSDS 650.0-P-1.1, August 2009
- http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206500P11/Attachments/650x0p11.pdf
- Slides based on Blue Book, Pink Book and:
  - Don Sawyer, Lou Reich: ISO Reference Model for an Open Archival Information System (OAIS) Tutorial Presentation, LOC, June 13 2003
- http://ssdoo.gsfc.nasa.gov/nost/isoas/overview.html

Outline

- Principles of the OAIS Model
- Technical Overview
- Functional Overview
- Information Modell
- Summary

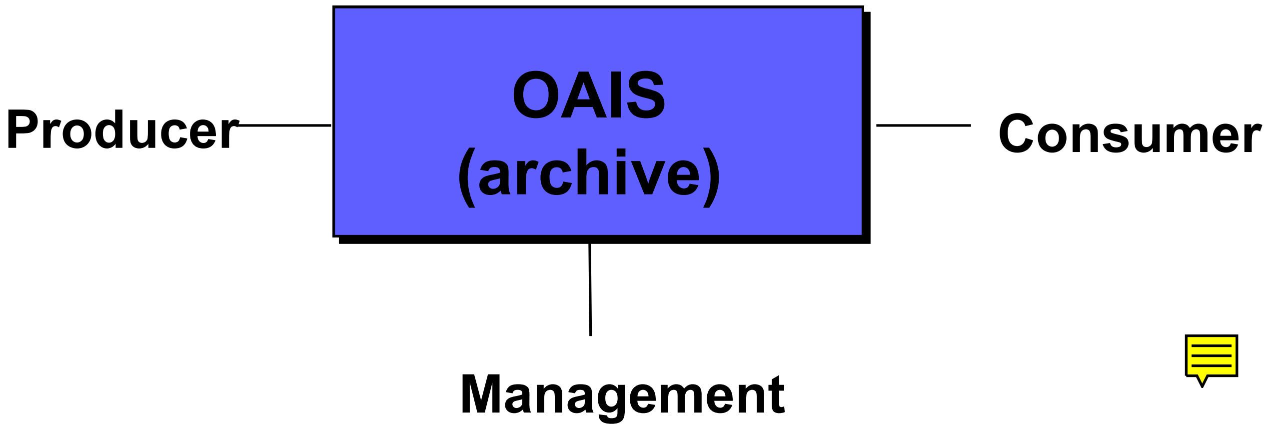
Open Archival Information System (OAIS)

- Open
  - Reference Model standard(s) are developed using a public process and are freely available
- Information
  - Any type of knowledge that can be exchanged
  - Independent of the forms (i.e., physical or digital) used to represent the information
  - Data are the representation forms of information
- Archival Information System
  - Hardware, software, and people who are responsible for the acquisition, preservation and dissemination of the information

Purpose, Scope, and Applicability

- Framework for understanding and applying concepts needed for long-term digital information preservation
  - Long-term is long enough to be concerned about changing technologies
  - Starting point for model addressing non-digital information
- Provides set of minimal responsibilities to distinguish an OAIS from other uses of ‘archive’
- Framework for comparing architectures and operations of existing and future archives
- Basis for development of additional related standards
- Addresses a full range of archival functions
- Applicable to all long-term archives and those organizations and individuals dealing with information that may need long-term preservation
- Does NOT specify an implementation

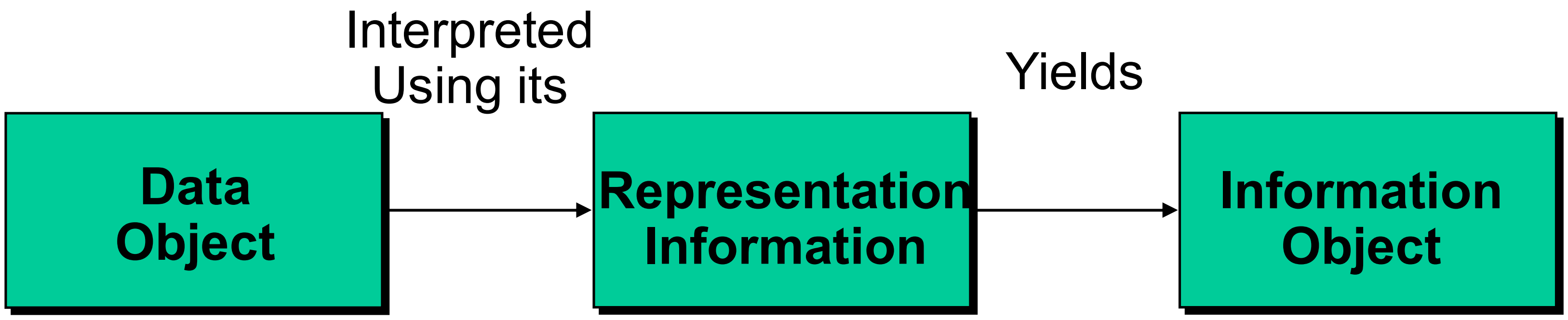
Model View of an OAIS Environment



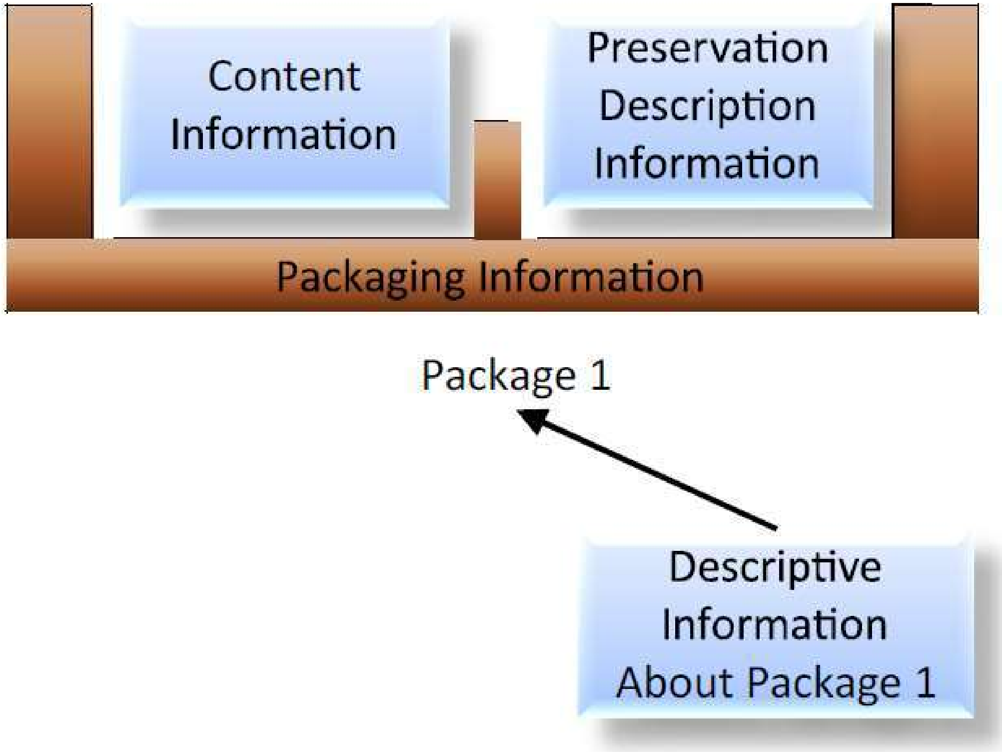
- Producer is the role played by those persons, or client systems, who provide the information to be preserved
- Management is the role played by those who set overall OAIS policy as one component in a broader policy domain
- Consumer is the role played by those persons, or client systems, who interact with OAIS services to find and acquire preserved information of interest

OAIS Information Definition

- Information is always expressed (i.e., represented) by some type of data
- Data interpreted using its Representation Information yields Information
- Information Object preservation requires clear identification and understanding of the Data Object and its associated Representation Information



Information Package Definition

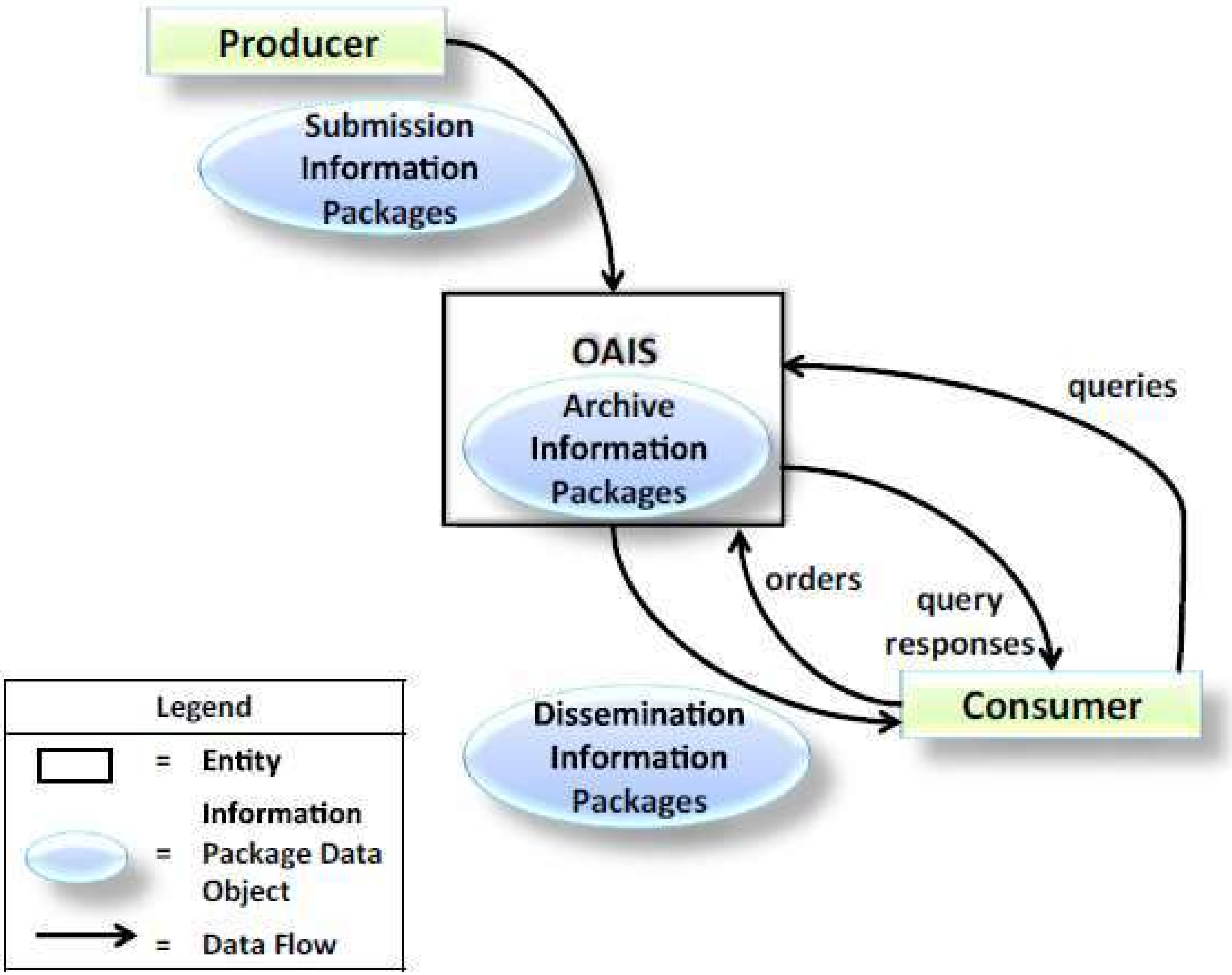


- An Information Package is a conceptual container holding two types of information
  - Content Information
  - Preservation Description Information (PDI)
- Plus descriptive information

Information Package Variants

- **SIP**: Submission Information Package
  - Negotiated between Producer and OAIS
  - Sent to OAIS by a Producer
- **AIP**: Archival Information Package
  - Information Package used for preservation
  - Includes complete set of Preservation Description Information (PDI) for the Content Information
- **DIP**: Dissemination Information Package
  - Includes part or all of one or more Archival Information Packages
  - Sent to a Consumer by the OAIS

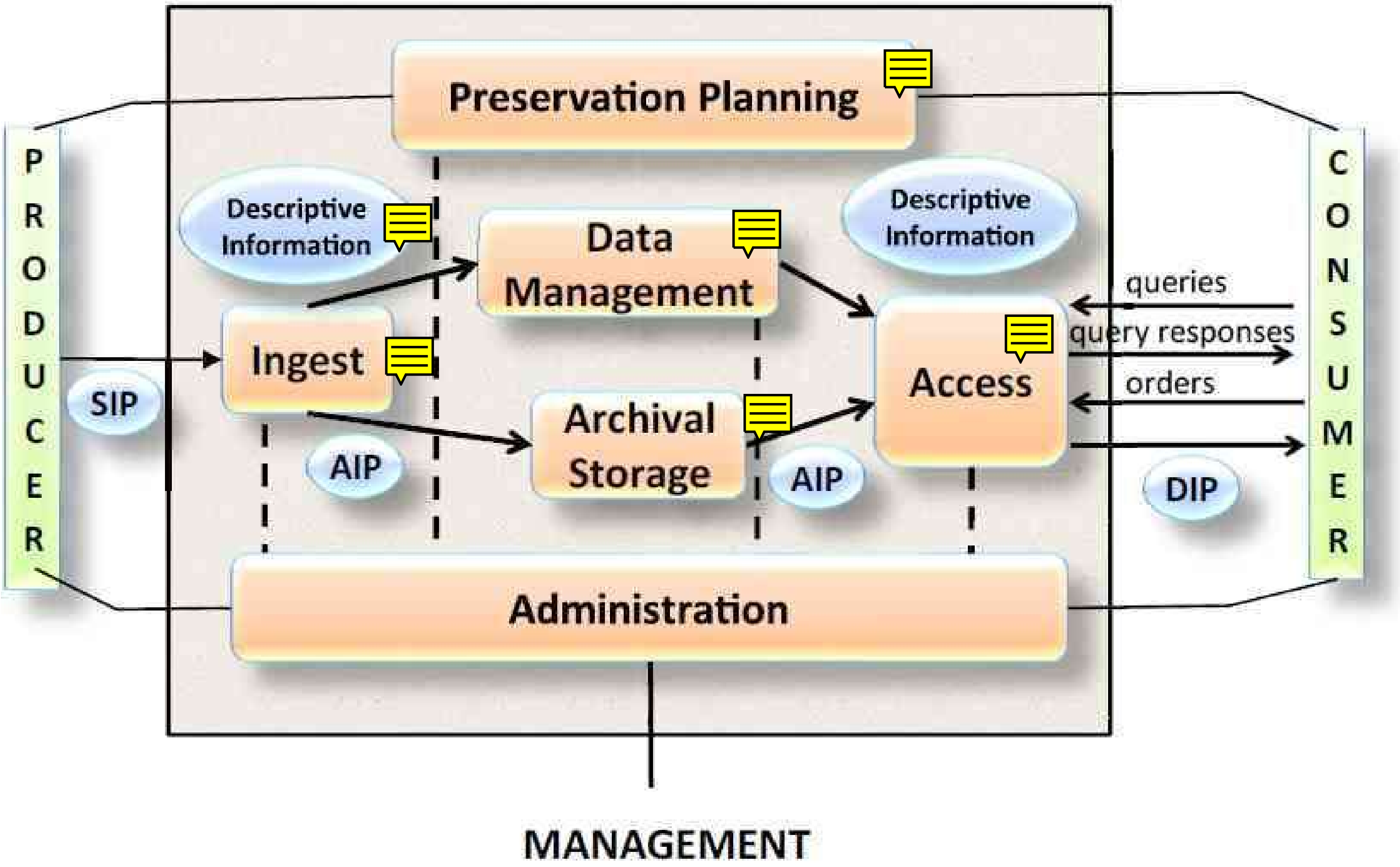
External Data Flow View



Outline

- Principles of the OAIS Model
- Technical Overview
- Functional Overview
- Information Modell
- Summary

Open Archival Information System: Six Functional Entities



SIP = Submission Information Package

AIP = Archival Information Package

DIP = Dissemination Information Package

Functional Entities in an OAIS (1/2)

- Ingest: This entity provides the services and functions to accept Submission Information Packages (SIPs) from Producers and prepare the contents for storage and management within the archive

- Archival Storage: This entity provides the services and functions for the storage, maintenance and retrieval of Archival Information Packages

- Data Management: This entity provides the services and functions for populating, maintaining, and accessing both descriptive information which identifies and documents archive holdings and internal archive administrative data.

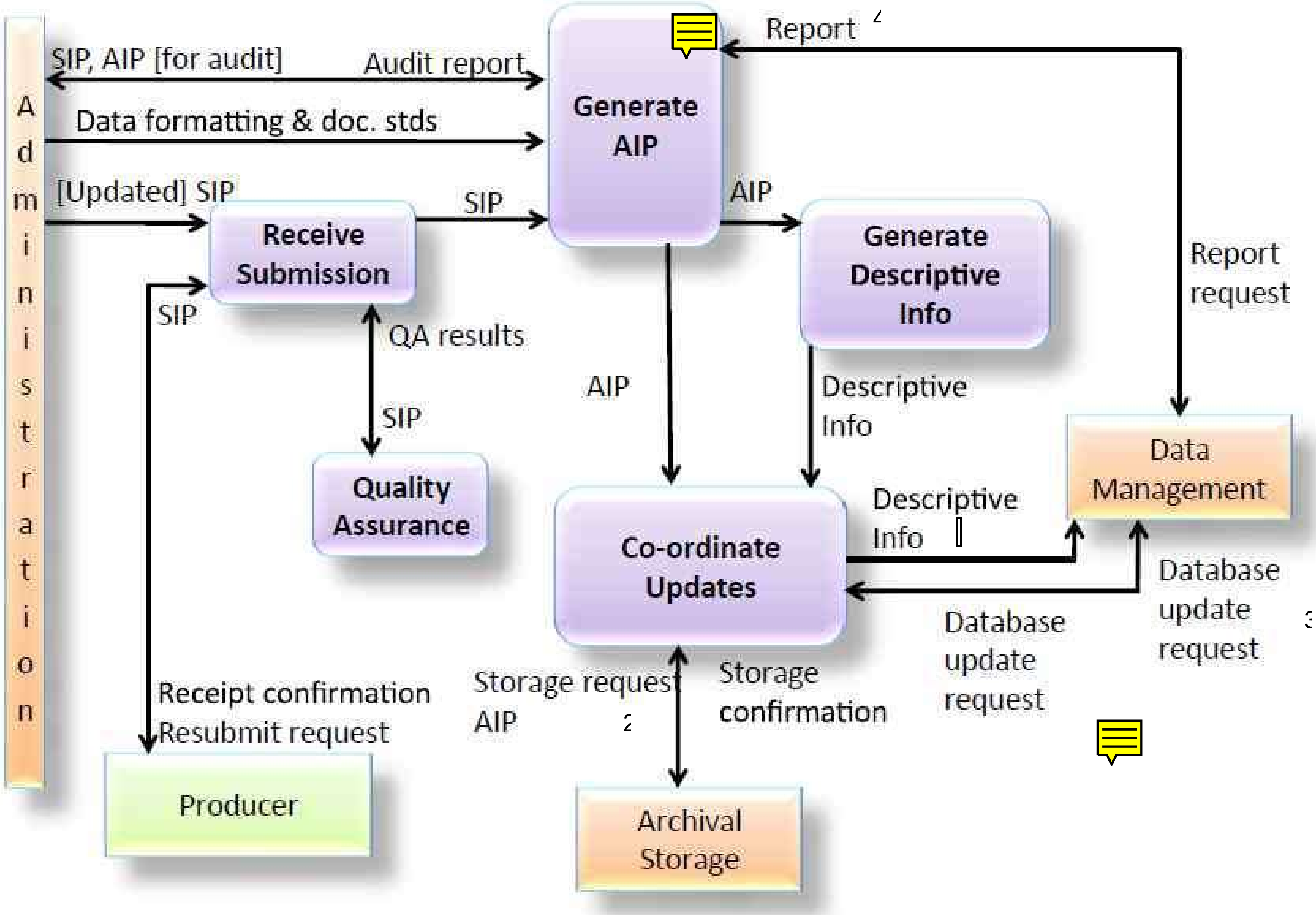
Functional Entities in an OAIS (2/2)

- Administration: This entity manages the overall operation of the archive system

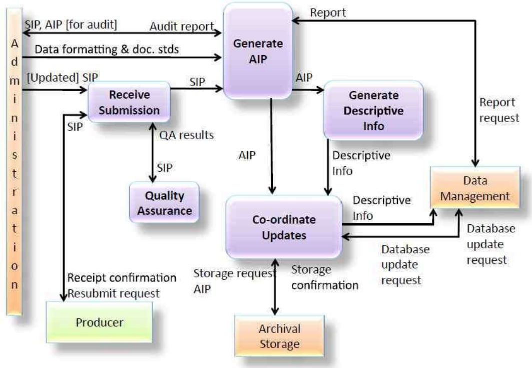
- Preservation Planning: This entity monitors the environment of the OAIS and provides recommendations to ensure that the information stored in the OAIS remain accessible to the Designated User Community over the long term even if the original computing environment becomes obsolete.

- Access: This entity supports consumers in determining the existence, description, location and availability of information stored in the OAIS and allowing consumers to request and receive information products

Ingest Data Functions



Ingest Data Functions

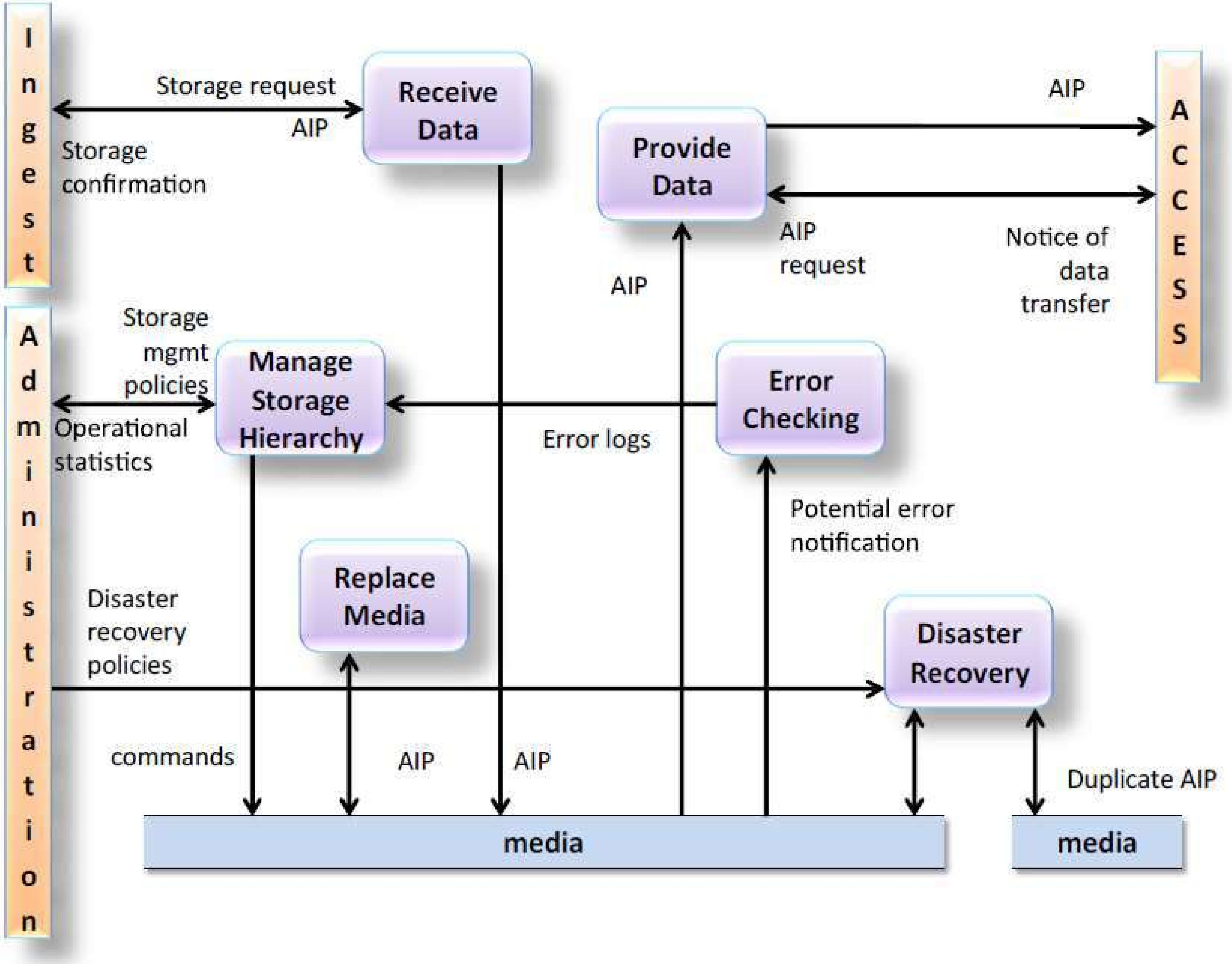


- Receive Submissions:
  - Staging Area for submissions
  - Confirmation on acceptance into staging area
- Quality Assurance
  - Validation of submission (CRC, logs, identity checks,media)
- Generate AIP
  - Transformation of SIPs into AIPs according to standard/policies (Transformation, Migration, Transcoding)
  - Forwarding of AIPs to Audit (Administration)
- Generate Descriptive Information
  - Collection / extraction of descriptive information on AIP for Data Management and Access Aids
- Coordinate Updates
  - Transfer of AIPs to Archival Storage
  - Confirmation -> Descriptive Information -> Data Management

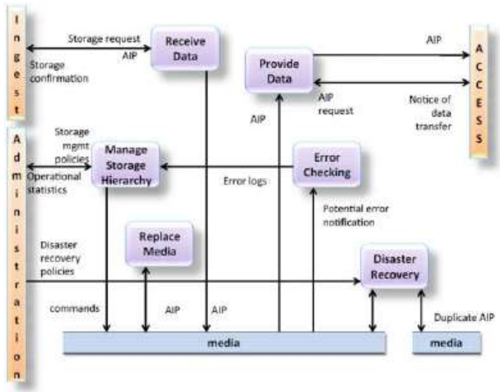
Ingest Data Functions

- Remember: what can go wrong?!
- Ingest / Standardization:
  - Who is performing the initial migration?
  - Who is liable?
  - Who will need to manage any problems subsequently?
- Migration
  - Something added? E.g. Word -> TXT, Excel -> TXT
  - Something lost?
- What is a PDF file?
  - A malicious invoice…
  - A multi-purpose paper: https://bigdata.uni-saarland.de/publications/p1972-dittrich.html

Archival Storage Functions

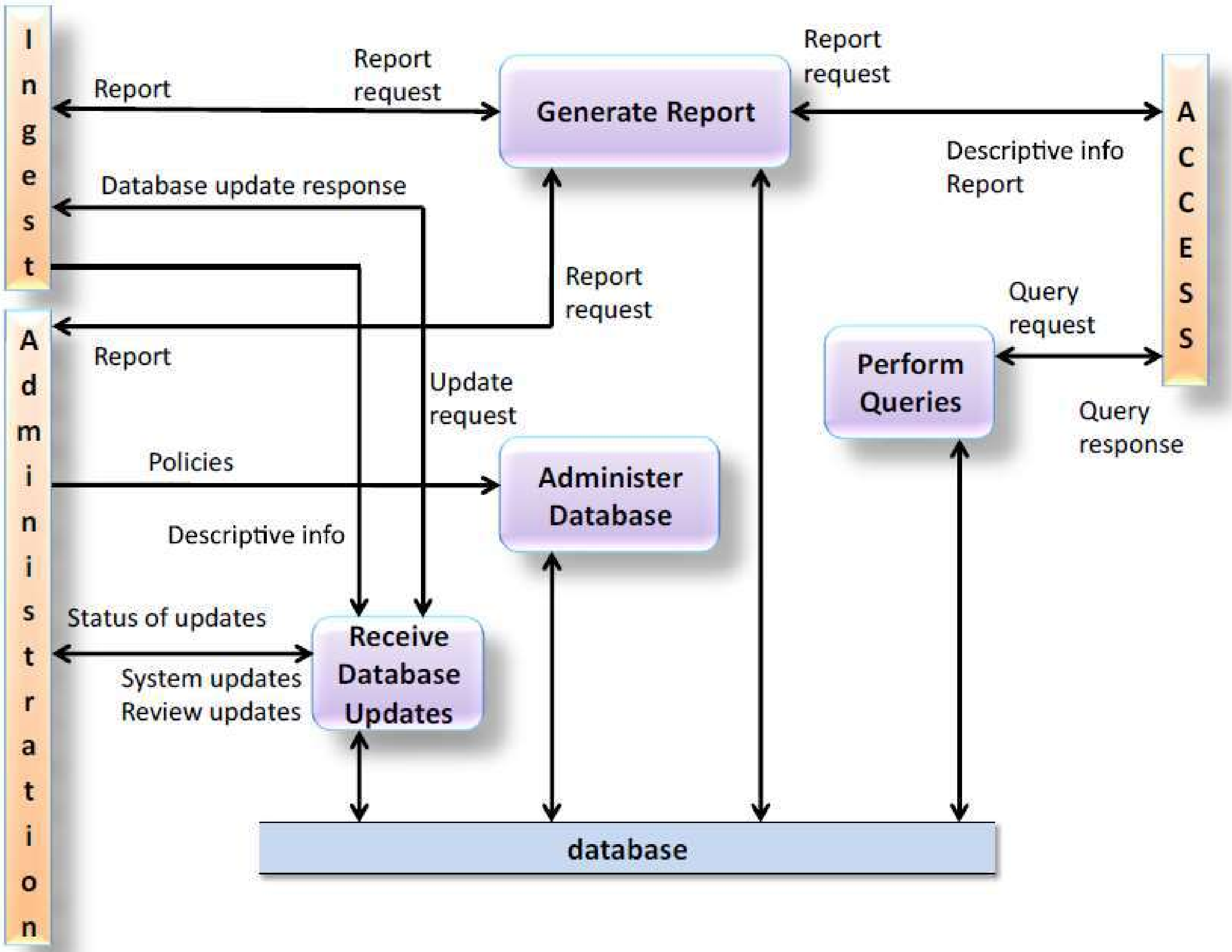


Archival Storage Functions

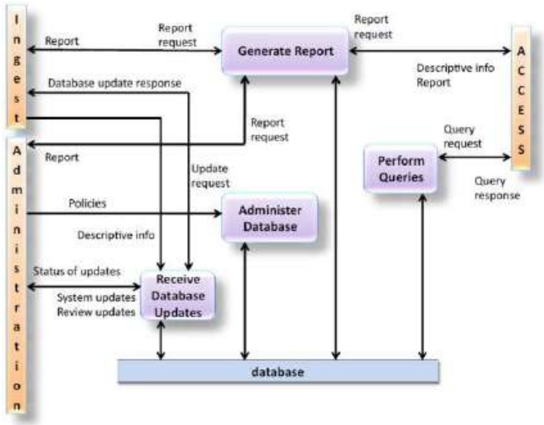


- Receive Data:
  - accept storage request for AIP
  - Decides on storage location, media
  - Retourns confirmation messages
- Manage Storgae Hierarchy
  - Management of storage according to policies
  - Monitoring of error messages, operational statistics
- Replace Media
  - Reproduction of AIPs over time (no changes of content or Preservation Description Information, only Packaging Information – other changes need to go via Administration)
- Error Checking
  - PDI Fixity Information (CRCs, error-correcting codes, ...)
- Disaster Recovery
  - Duplicating of storage media content (back-up)
  - Transport to physically seperated location
- Provide Data
  - Generate copies of AIPs for Access

Data Management Functions

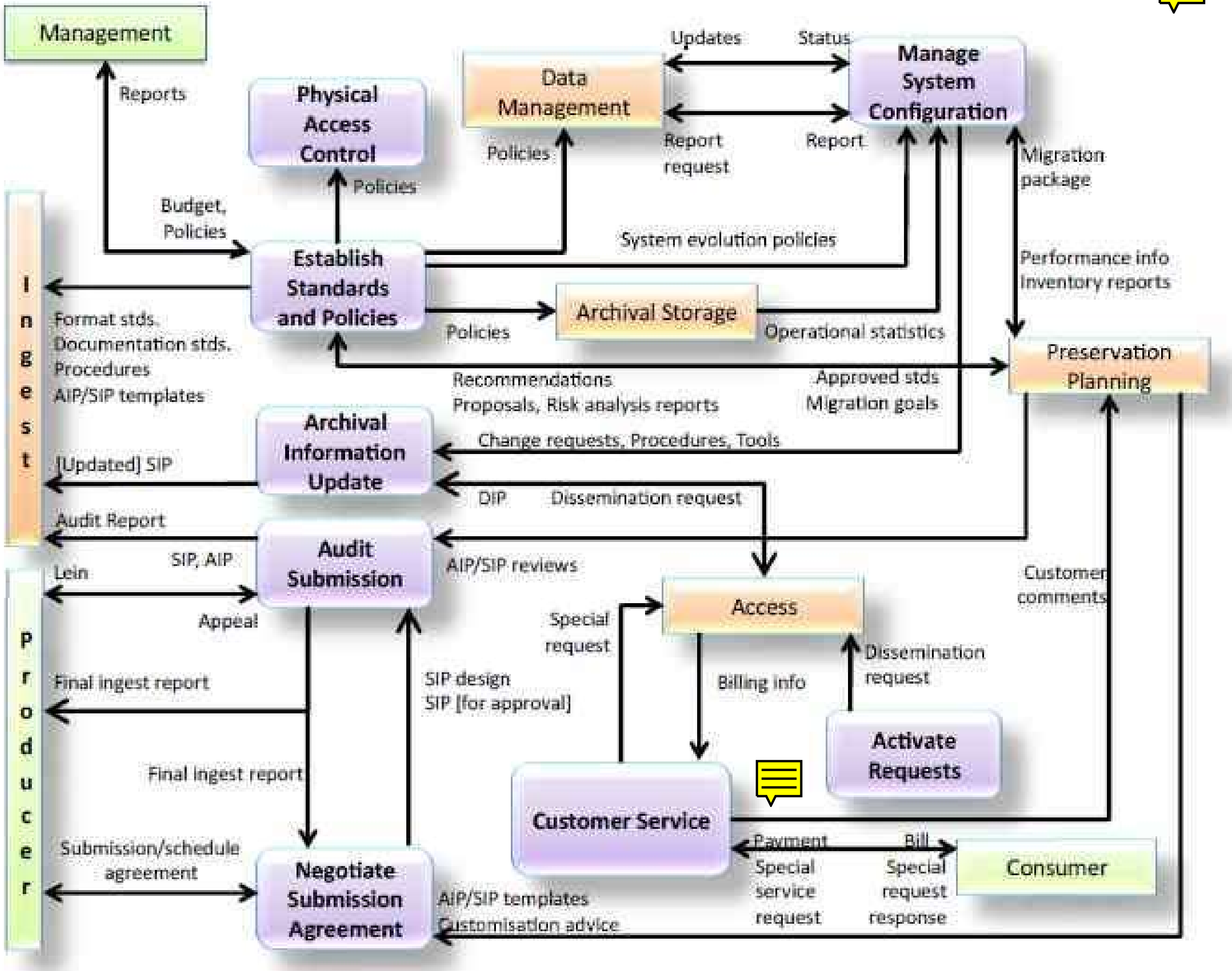


Data Management Functions

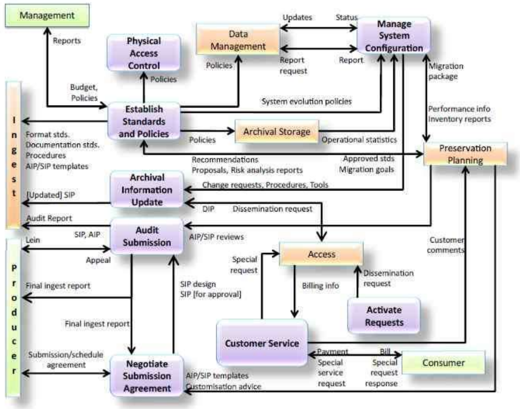


- Administer Database
  - Integrity of DB for Descriptive Information and system information
- Perform Queries
  - Processing of queries from Access
- Generate Reports
  - Reports for Ingest, Access, Administration
- Receive DB Updates
  - Add/delete/modify information in Management DB
  - Ingest: new AIPs, Administration: updates

Administration Functions

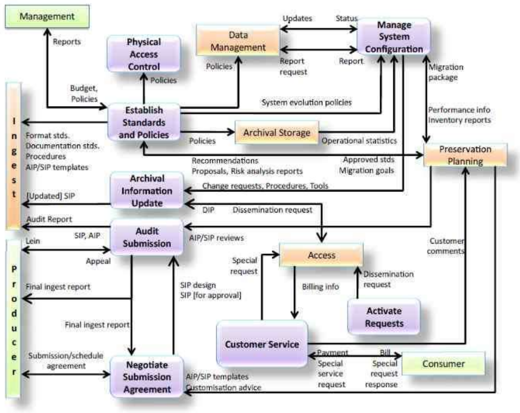


Administration Functions



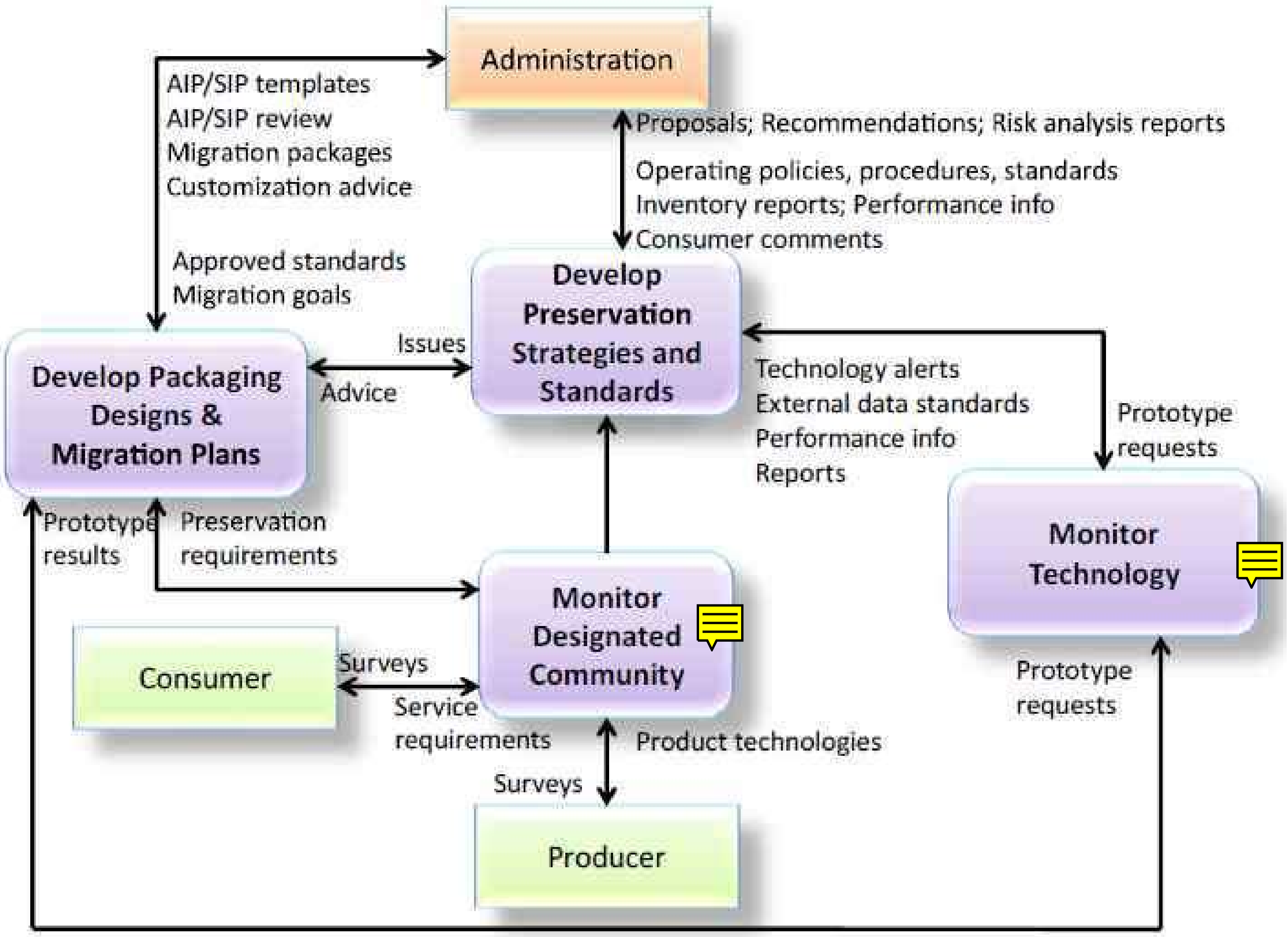
- Negotiate Submission Agreement
  - Contracts with Producers, submission prozedures
- Manage System Configuration
  - System evolution, monitoring
  - Provide information for Policies
- Archival Information Update
  - Update content of the archive: Modifying DIPs and Re-Submission -> Migration
- Establish Standards and Policies
  - Budget, Standards, Policies
- Audit Submission
  - Analyse whether SIPs and AIPs conform to policies and regulations
  - Verifia Representation and Package Information

Administration Functions

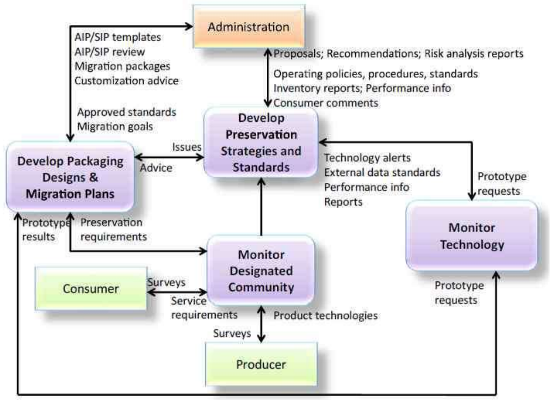


- Activate Requests
  - Protocol of triggered queries/evaluations
  - Periodic checks/queries to archive to verify cc
  - Ordering data/reports periodically
- Customer Service
  - Manage customer accounts
  - Collect costs from Access, create invoices for customers

Preservation Planning Functions

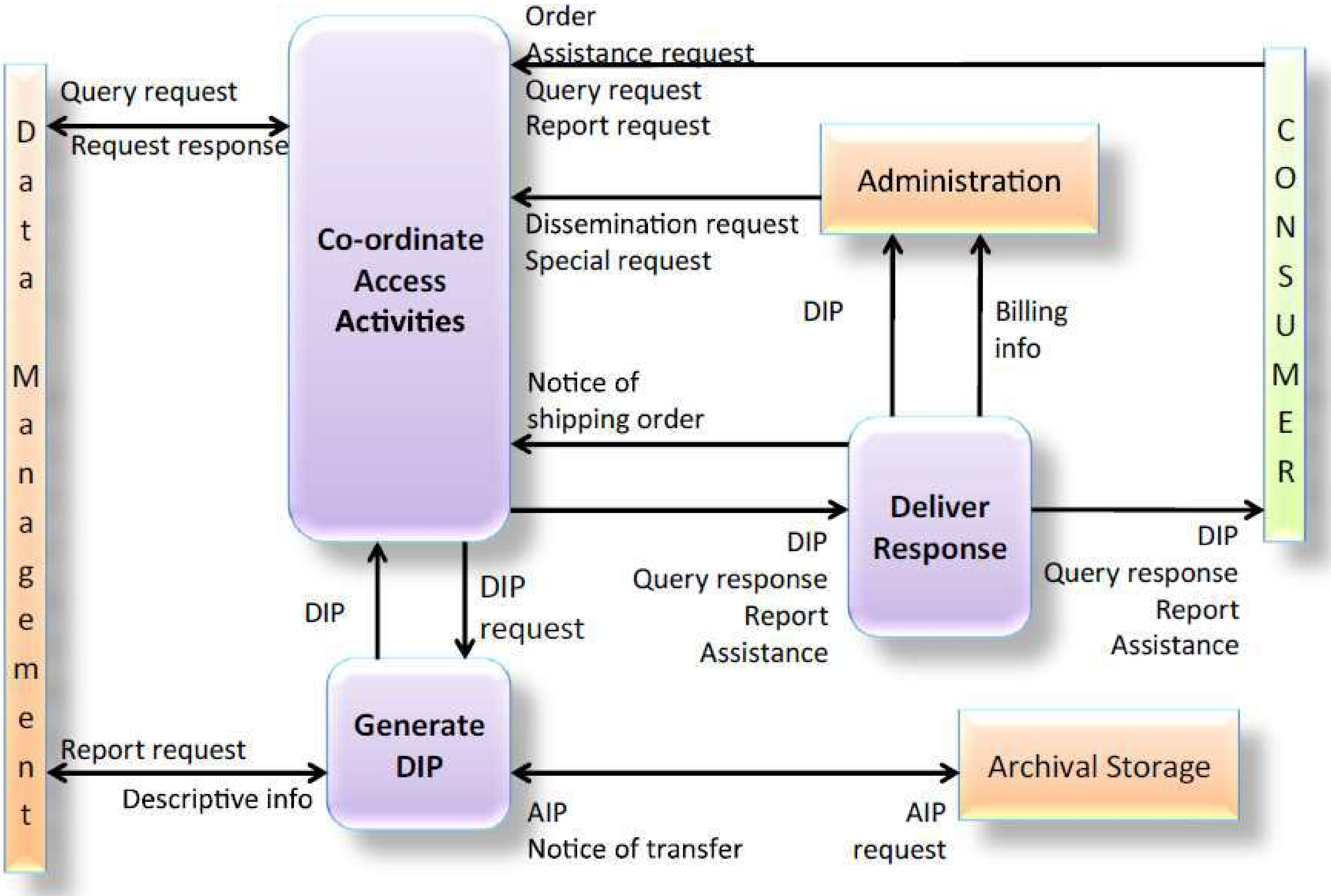


Preservation Planning Functions

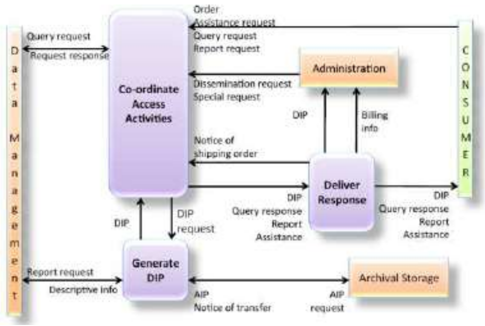


- Monitor Designated Community
  - Interaction with Producer and Consumer
- Monitor Technology
  - Technology evolution: HW, SW, Formats
- Develop Preservation Strategies and Standards
  - Strategies, trend analysis
- Develop Packaging Designs and Migration Plans
  - Migration paths, tools
  - Create Preservation Description Information

Access Functions



Access Functions

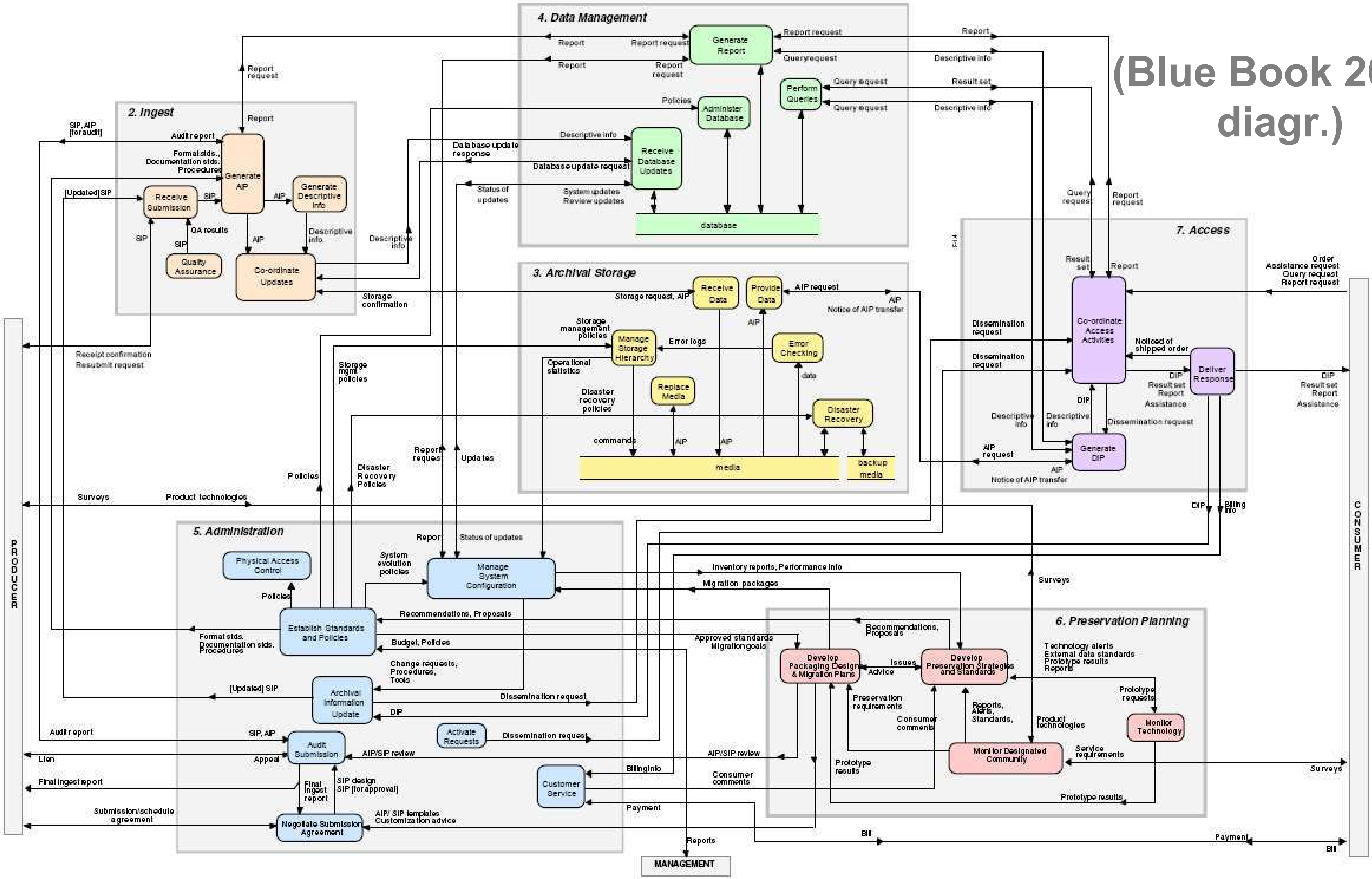


- Coordinate Access Activities
  - User interface, authorization
  - 3 types of Requests:
    - Queries to Data Management for Result Set
    - Order for Data Management and Archival Storage
    - Dissemination Requests by Administration for Archival Information Update
- Generate DIP
  - Get data from Archival Storage into Staging Area
  - Get Descriptive Information from Data Management
  - Apply processes to transform AIPs into a suitable DIP depending on query / consumer
- Deliver Response
  - On-line and off-line responses
  - Forward results

Access Functions

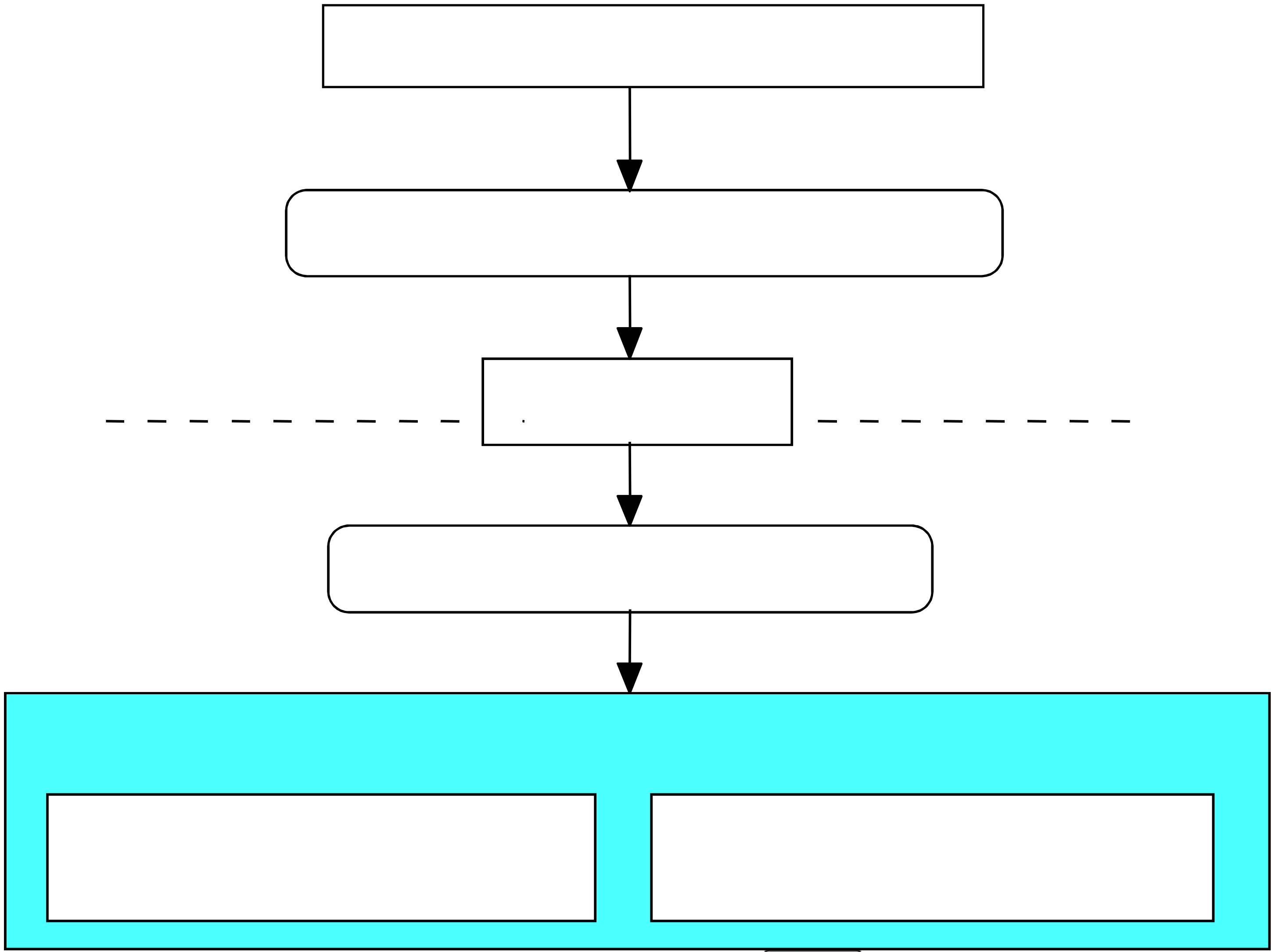
- AIP – DIP conversion
  - Decide which information to release in which format
  - Remember: Consider metadata!
    c.f. Supriya Adhatarao, Cédric Lauradoux: Exploitation and Sanitization of Hidden Data in PDF Files. arXiv:2103.02707, March 2021 (Analyzing PDFs published by security agencies: metadata revealing weak links, less than 10% of agencies sanitized part of their documents, 65% of these still contained sensitive information)

OAIS Composite Functional Entities



(Blue Book 2002 diagr.)

Migration Context



Digital Migration

Digital Migration is defined to be the transfer of digital information, while intending to preserve it, within the OAIS.

- Focus on preservation of the full information content
- New information implementation replaces the old
- OAIS has full control and responsibility over all aspects of the transfer

Migration Motivators

- Motivators driving digital migrations
  - Media Decay
    - Often this is superceded by escalating media drive maintenance costs
  - Increased Cost Effectiveness
    - More cost-effective media types with higher volumes and lower drive maintenance costs
  - New User/Consumer Service Requirements
    - New formats more compatible with user’s technology and applications
  - Proprietary software evolution
    - New software versions used to ‘upgrade’ formats of the information objects being preserved

Digital Migration Approaches

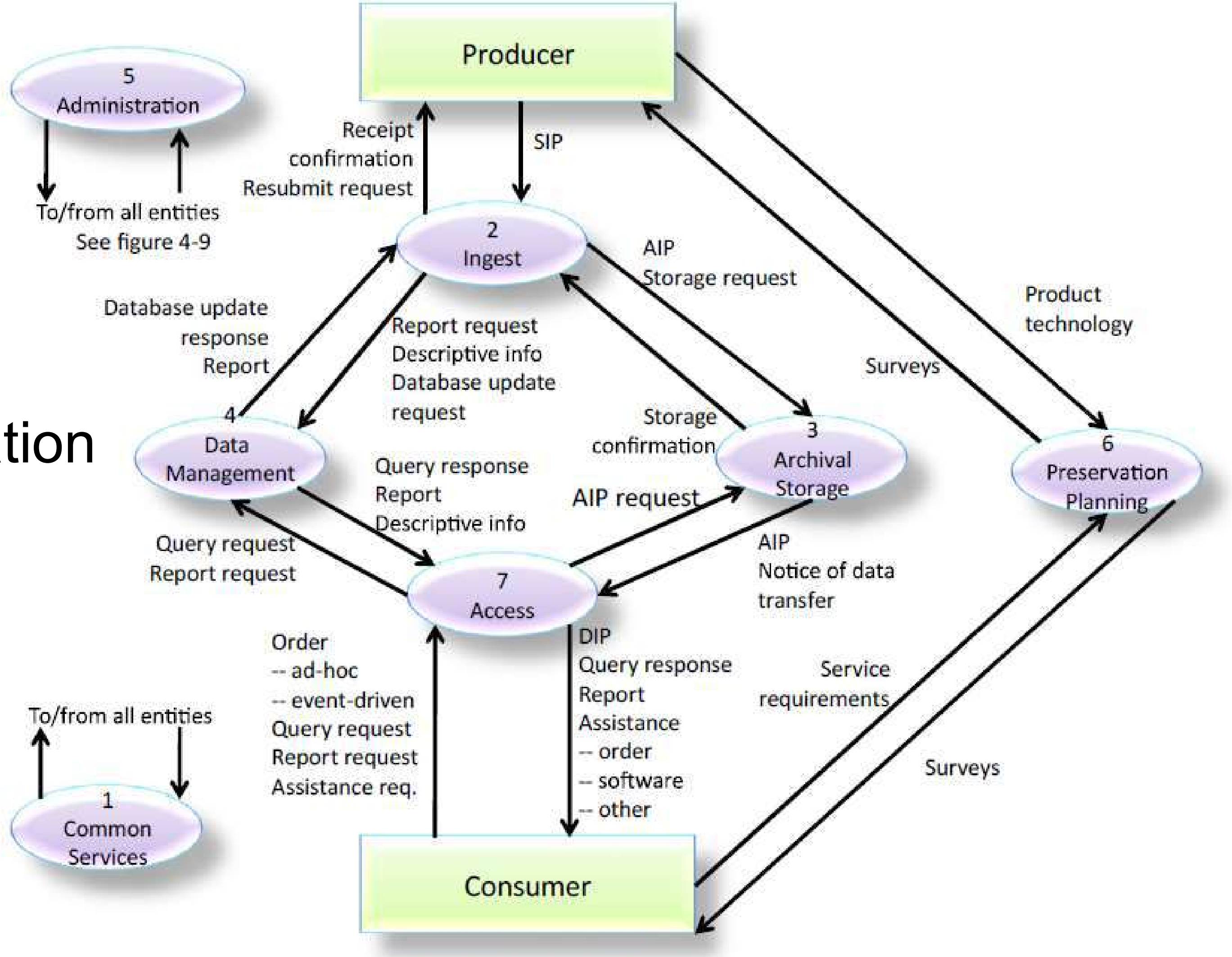
- Four primary types of digital migration in response to motivators, ordered by increasing risk of information loss:
  - Refreshment
    - Media replacement with no bit changes
  - Replication
    - No change to Packaging Information or Content Information bits (e.g. copying to new file / new location)
  - Repackaging
    - Some bit changes in Packaging Information (e.g multiple files packaged in directory structure get copied to other carrier)
  - Transformation
    - Reversible: Bit changes in Content Information are reversible by an algorithm
    - Non-reversible: Bit changes in Content Information are not reversible by an algorithm

Digital Migration and AIPs

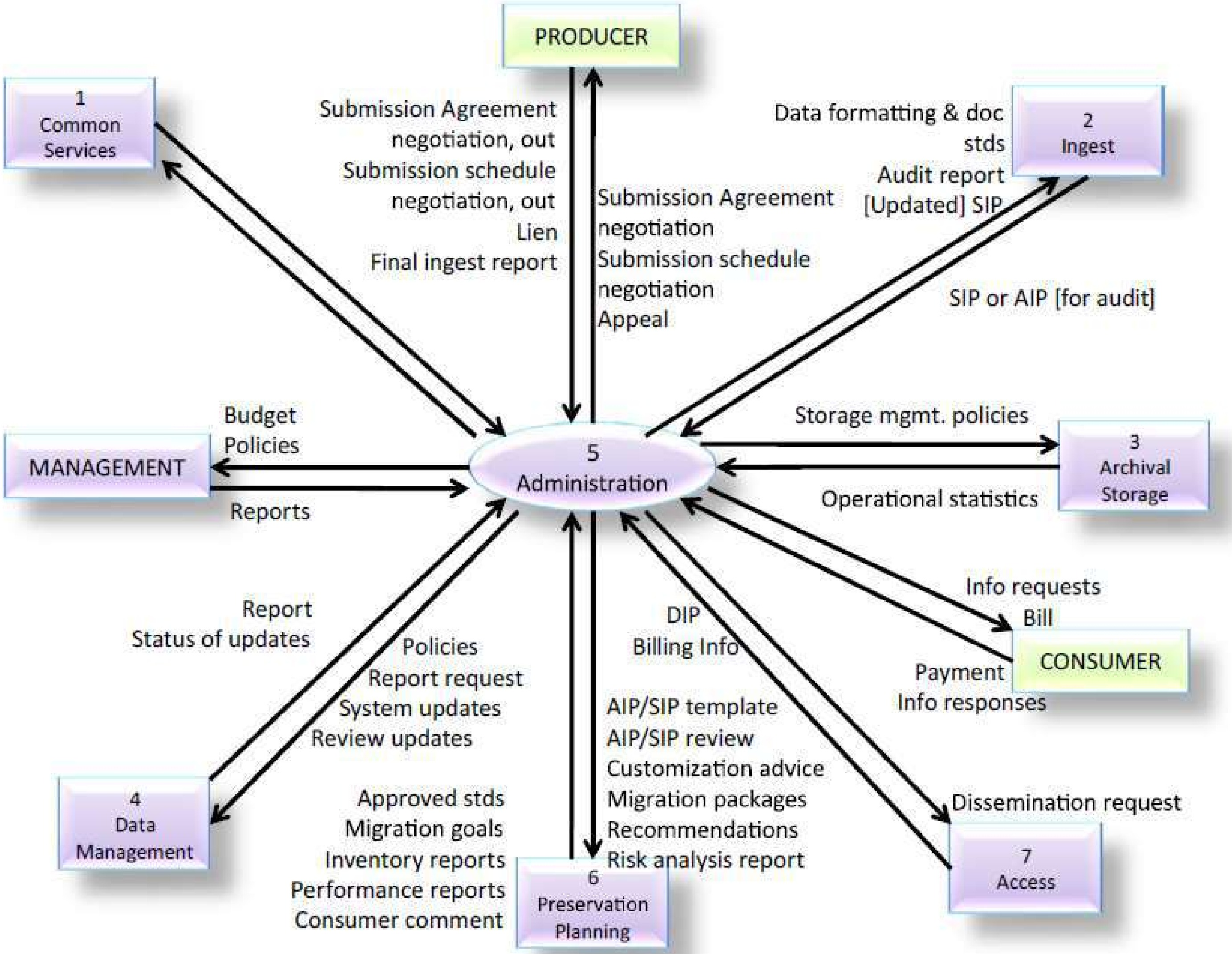
- Unless migration involves transformation:
  no new AIP version
- Transformation:
  new **AIP Version**
- Upgrading or improvement of AIPs:
  new **AIP Edition**
- Extracting or aggregating from multiple AIPs:
  Derived AIP

OAIS Data Flow

- w/o administration



OAIS Data Flow



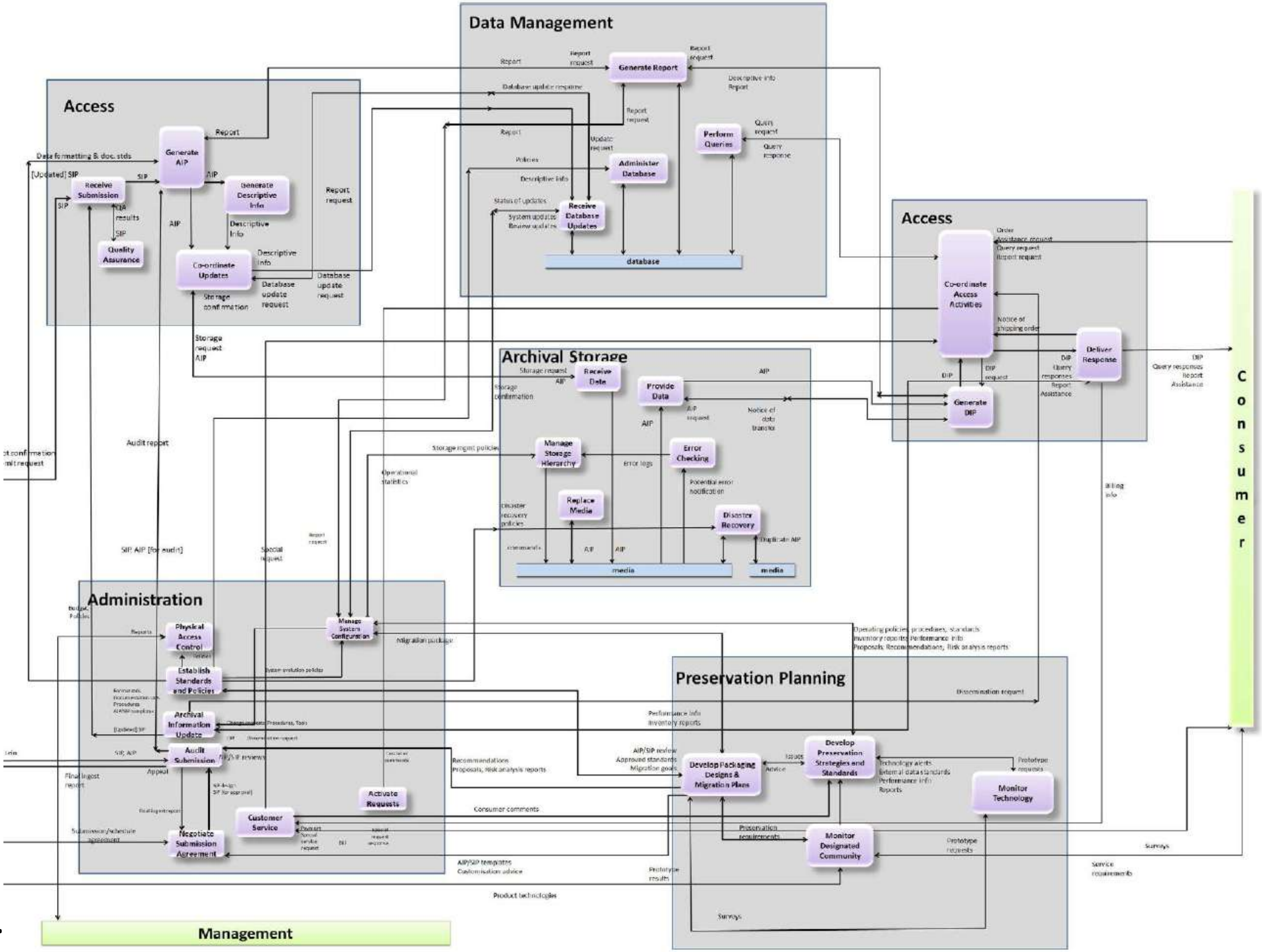
Common Services

- Modern, distributed computing applications assume a number of supporting services
- Examples of Common Services include:
  - inter-process communication
  - name services
  - temporary storage allocation
  - exception handling
  - security
  - file and directory services

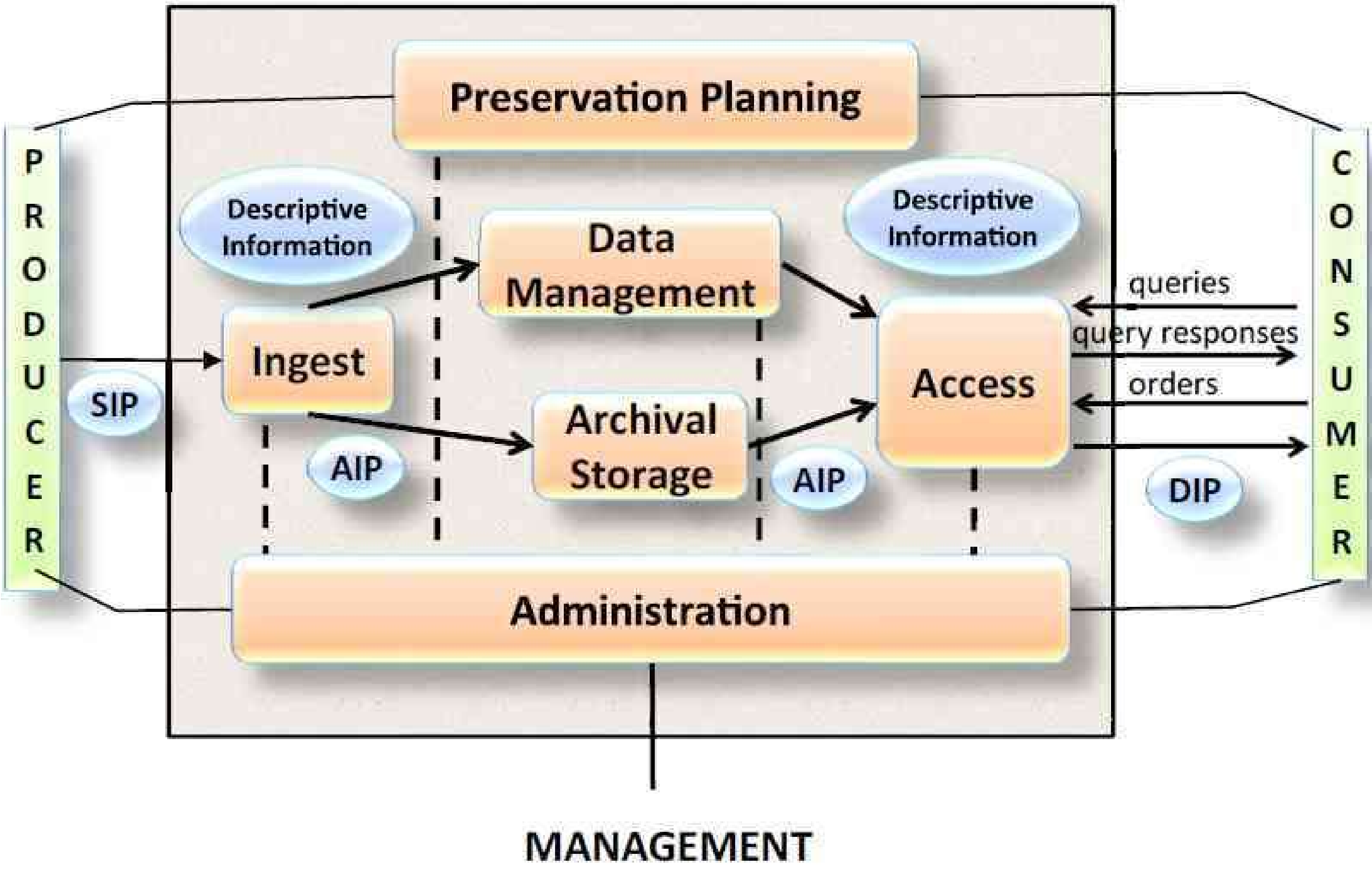
Common Services

- Important:
  - All processing steps taken need to be documented (logs, protocols)
  - Reporting
  - Confirmations
  - This documentation is part of the archive as well

OAIS Composite Functional Entities



Open Archival Information System: Summary



SIP = Submission Information Package

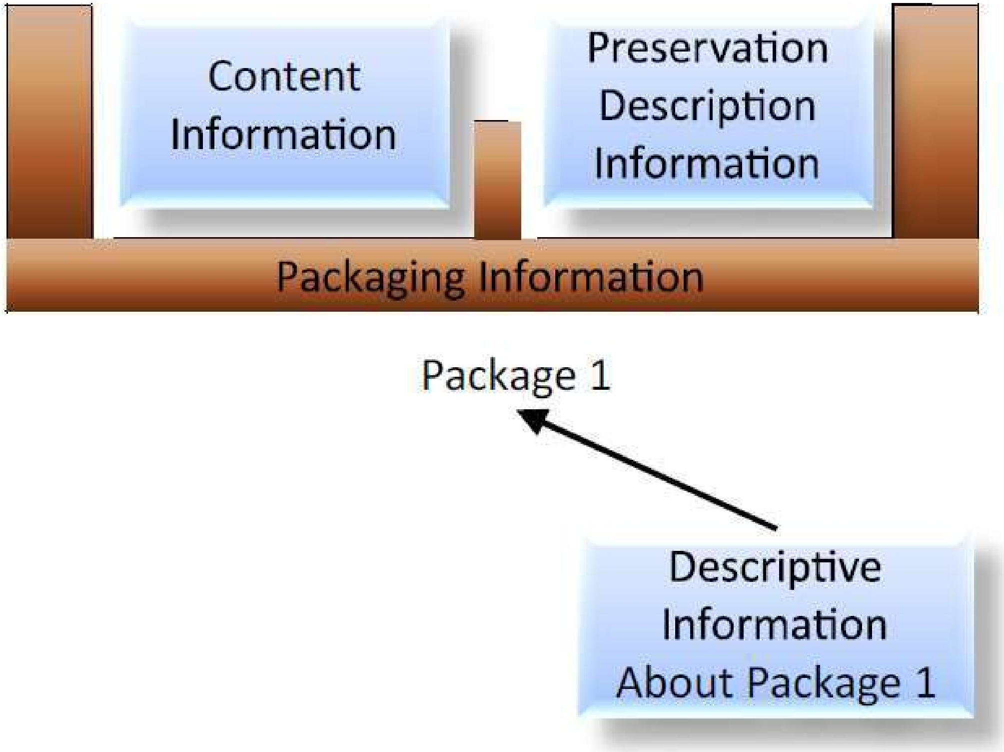
AIP = Archival Information Package

DIP = Dissemination Information Package

Outline

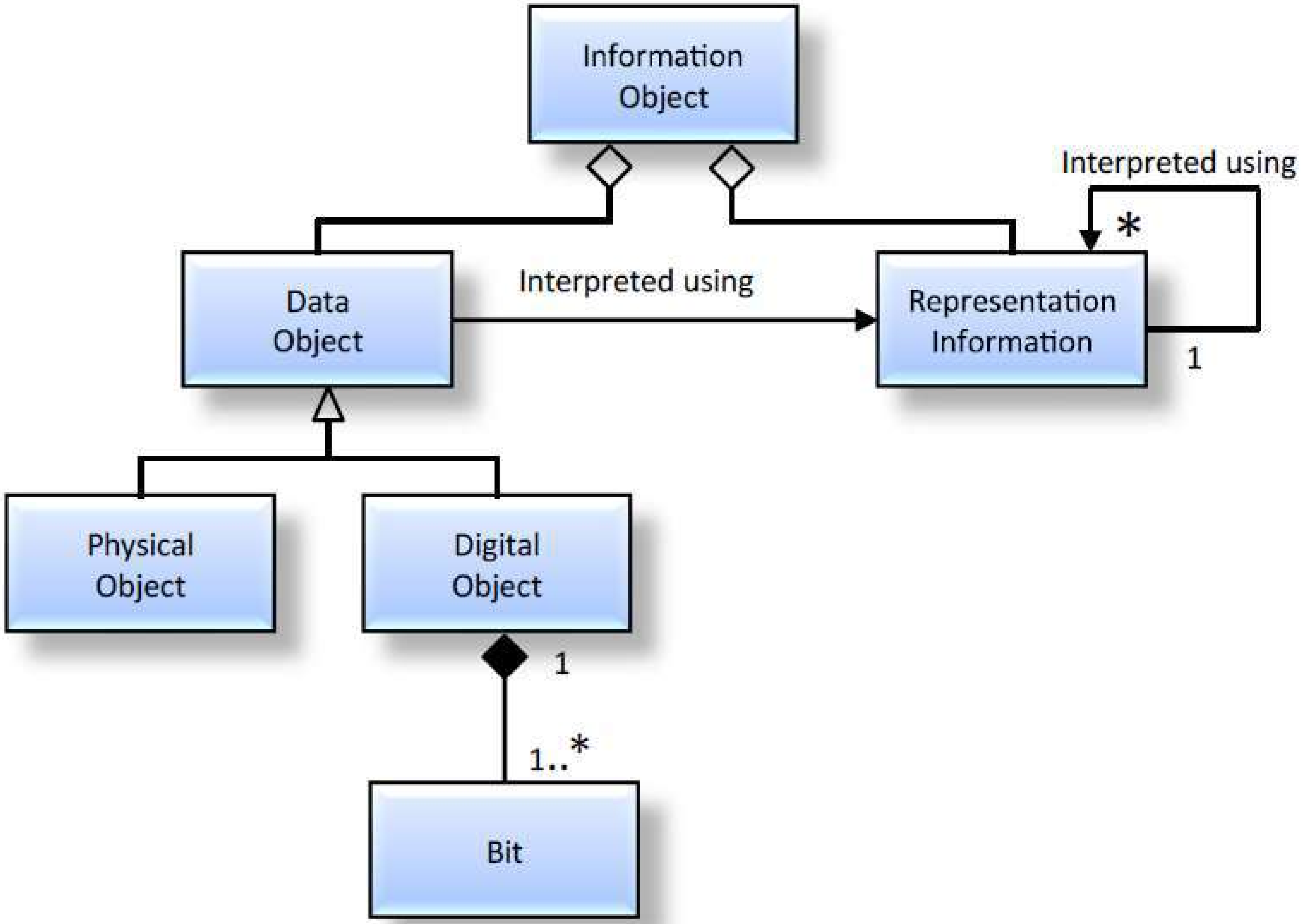
- Principles of the OAIS Model
- Technical Overview
- Functional Overview
- Information Modell
- Summary

Information Package Definition



- An Information Package is a conceptual container holding two types of information
  - Content Information
  - Preservation Description Information (PDI)
- Plus descriptive information

Information Object

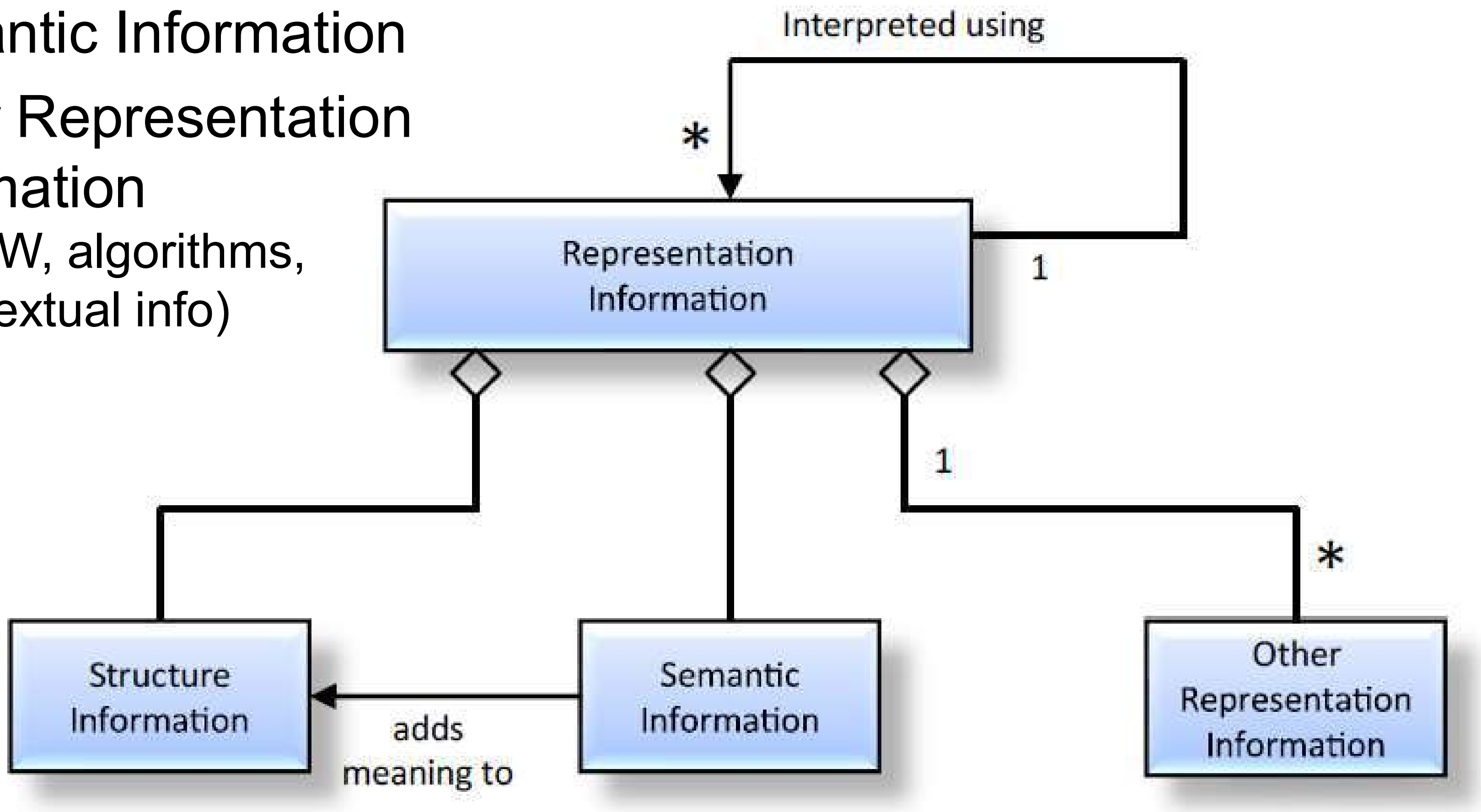


Representation Information

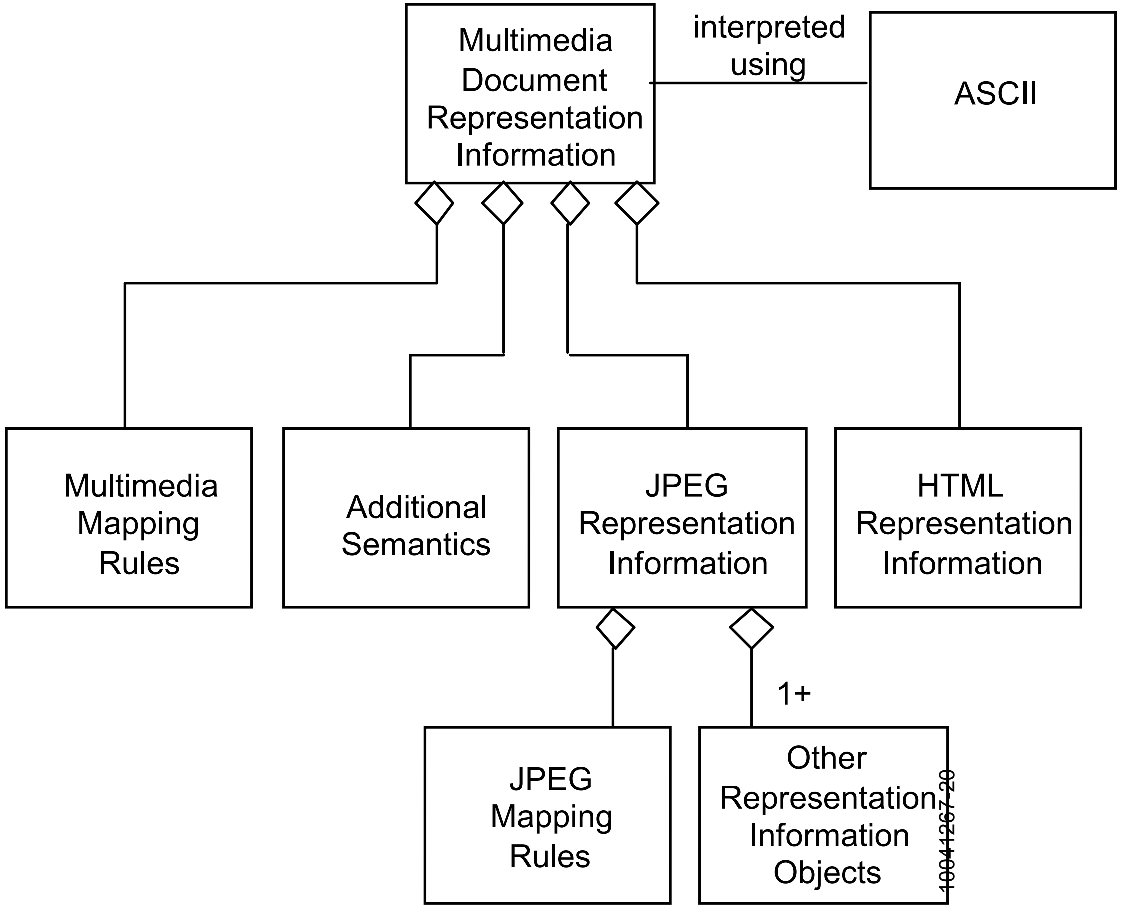
- The Representation Information accompanying a physical object, like a moon rock, may give additional meaning
  - It typically is a result of some analysis of the physically observable attributes of the rock
- The Representation Information accompanying a digital object, or sequence of bits, is used to provide additional meaning.
  - It typically maps the bits into commonly recognized data types such as character, integer, and real and into groups of these data types.
  - It associates these with higher level meanings which can have complex inter-relationships that are also described

Recursive Nature of Representation Information

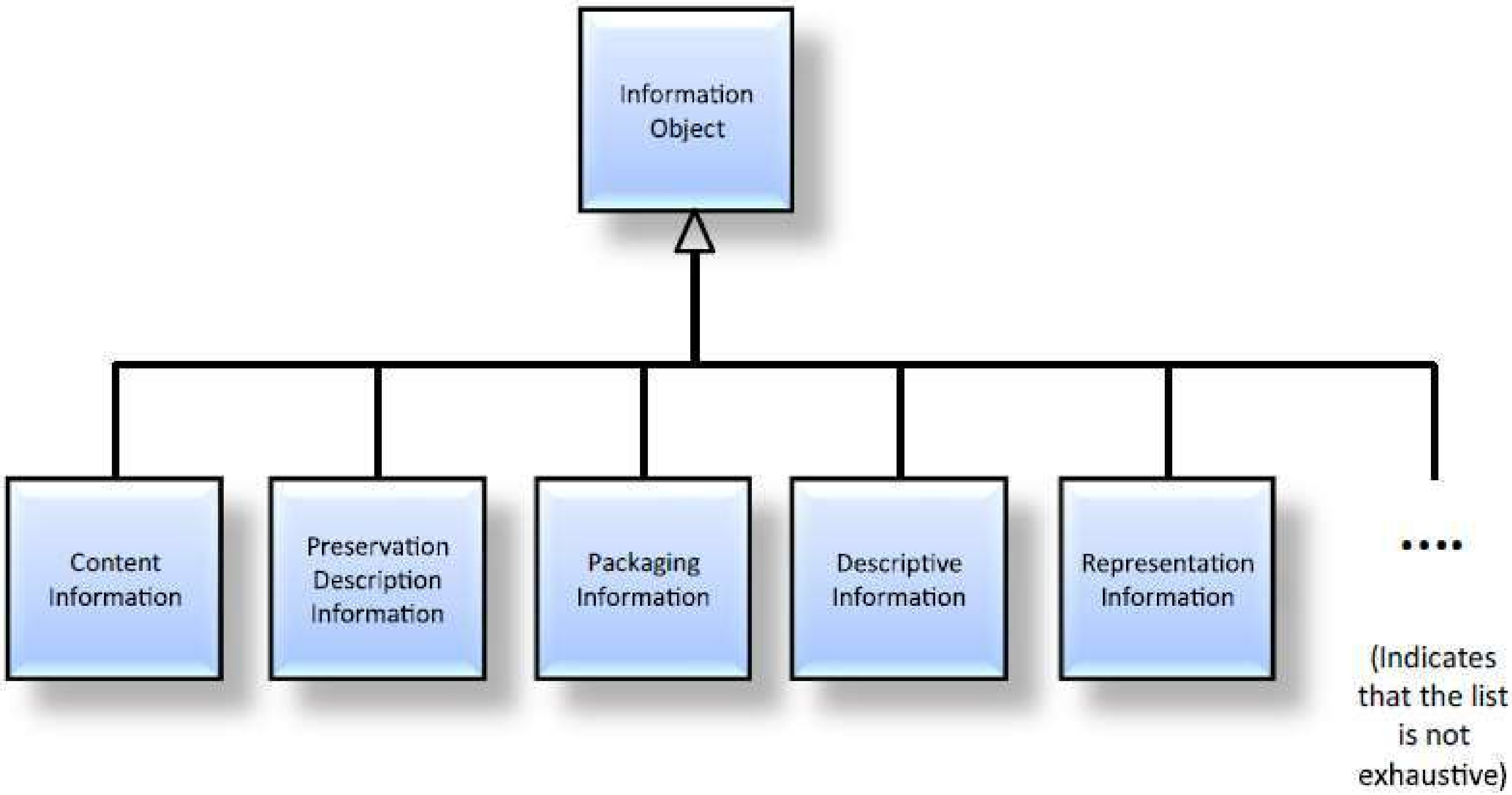
- Structure Information
- Semantic Information
- Other Representation Information
  (e.g. SW, algorithms, other textual info)



Sample Representation Net



Types of Information Used in OAIS



Content Information

- The information which is the primary object of preservation
- An instance of Content Information is the information that an archive is tasked to preserve.
- Deciding what is the Content Information may not be obvious and may need to be negotiated with the Producer
- The Data Object in the Content Information may be either a Digital Object or a Physical Object (e.g., a physical sample, microfilm)

Preservation Description Information

- Provenance Information
  - Describes the source of Content Information, who has had custody of it, what is its history
- Context Information
  - Describes how the Content Information relates to other information outside the Information Package
- Reference Information
  - Provides one or more identifiers, or systems of identifiers, by which the Content Information may be uniquely identified
- Fixity Information
  - Protects the Content Information from undocumented alteration

PDI Examples

Content Information Type	Reference	Provenance	Context	Fixity	Access Rights
Space Science Data	• Object identifier • Journal reference • Mission, instrument, title, attribute set	• Instrument description • Principal Investigator • Processing history • Storage and handling history • Sensor description • Instrument • Instrument mode • Decommutation map • Software interface specification • Information Property Description	• Calibration history • Related data sets • Mission • Funding history	• CRC • Checksum • Reed-Solomon coding	• Identification of the properly authorized Designated Community (Access Control) • Permission grants for preservation and for distribution • Pointers to Fixity and Provenance Information (e.g., digital signatures, and rights holders)

PDI Examples

Digital Library Collections	• Bibliographic description
	• Persistent identifier
		• For scanned collections:
		  • metadata about the digitization process
		  • pointer to master version
		• For born-digital publications:
		  • pointer to the digital original
		• Metadata about the preservation process:
		  • pointers to earlier versions of the collection item
		  • change history
		  • Information Property Description
			• Pointers to related documents in original environment at the time of publication	• Digital signature
				• Checksum
				• Authenticity indicator
					• Legal framework(s)
					• Licensing offers
					• Specifications for rights enforcement measures applied at dissemination time
					• Permission grants for preservation and for distribution
					• Information about watermarking applied at submission and preservation time
					• Pointers to Fixity and Provenance Information (e.g., digital signatures, and rights holders)

PDI Examples

Content Information Type	Reference	Provenance	Context	Fixity	Access Rights
Software Package	• Name • Author/ Originator • Version number • Serial number	• Revision history • Registration • Copyright • Information Property Description	• Help file • User guide • Related software • Language	• Certificate • Checksum • Encryption • CRC	• Designated Community • Legal framework(s) • Licensing offers • Specifications for rights enforcement measures applied at dissemination time • Pointers to Fixity and Provenance Information (e.g., digital signatures, and rights holders)

Packaging Information

- Information which, either actually or logically, binds and relates the components of the package into an identifiable entity on specific media

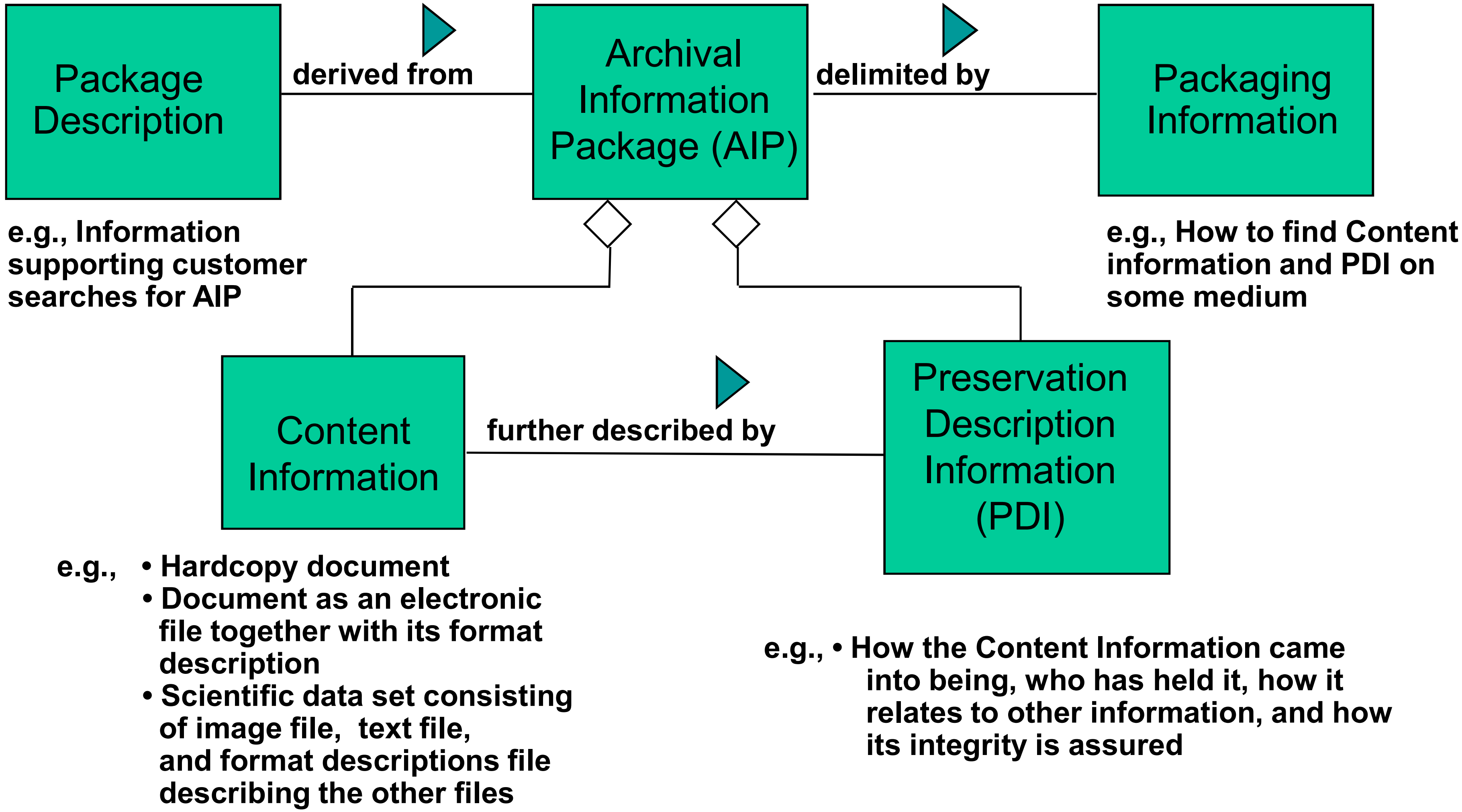
- Examples of Packaging Information include tape marks, directory structures and filenames

Descriptive Information

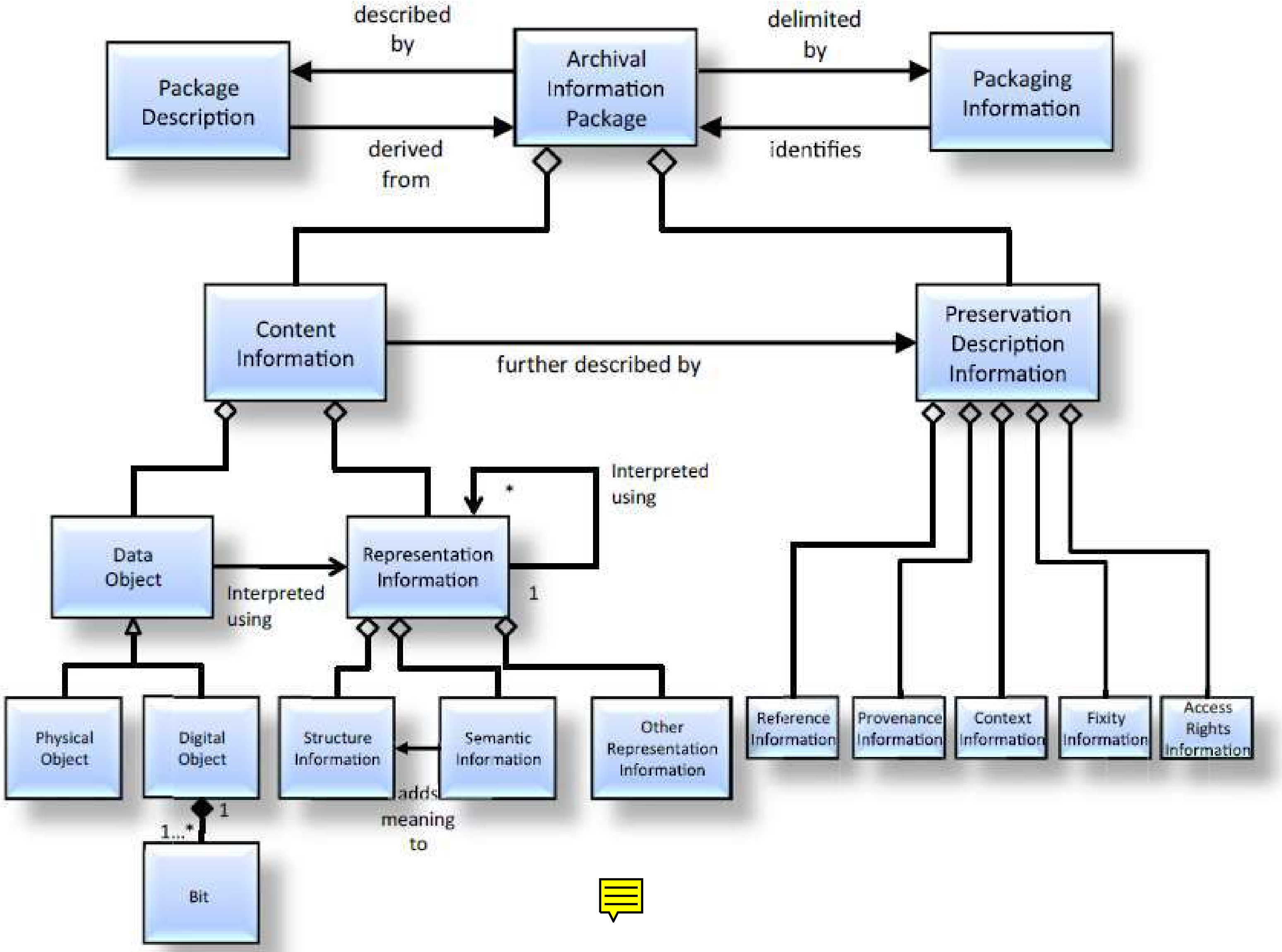
- Contain the data that serves as the input to documents or applications called Access Aids.

- Access Aids can be used by a consumer to locate, analyze, retrieve, or order information from the OAIS.

OAIS Archival Information Package

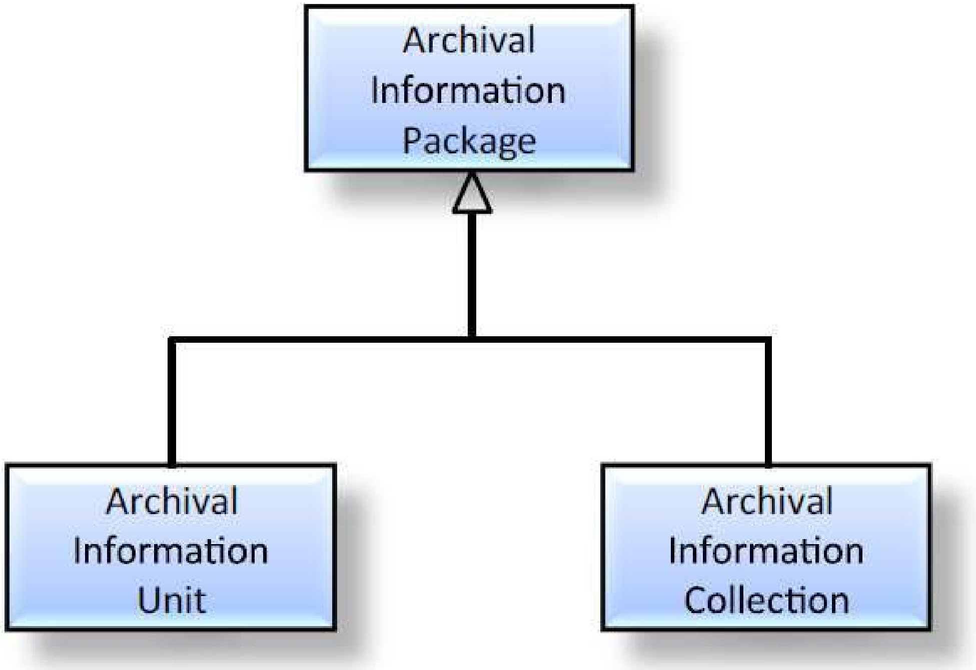


AIP detailed view



AIP Types

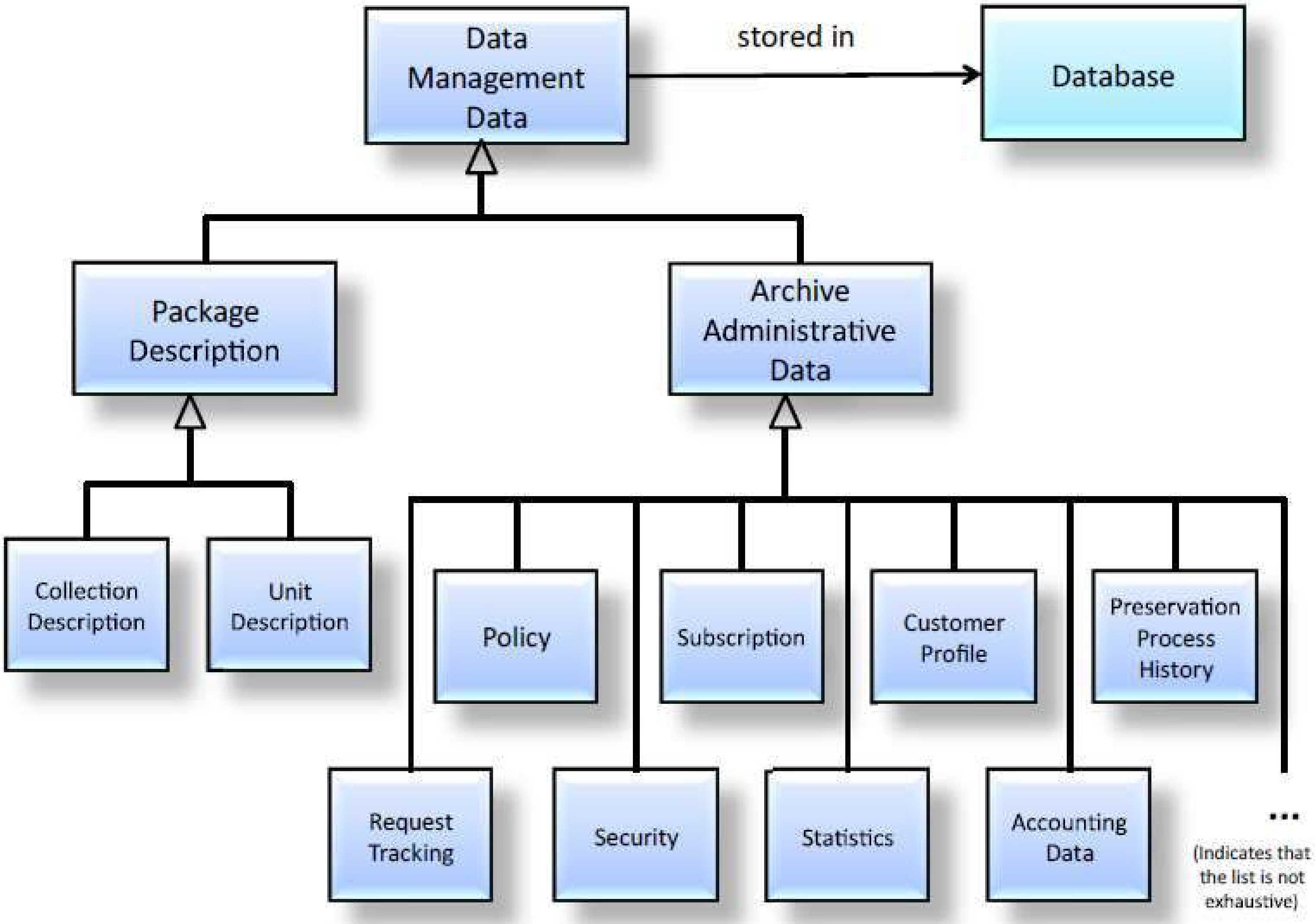
- Archival Information Unit (AIU) contains a single Data Object as the Content Object
- Archival Information Collection (AIC) contains multiple AIPs in its Content Object
  - Each member of an AIC is an AIP containing Content Information and PDI
  - The AIC contains unique PDI on the collection process



Package Descriptions and Access Aids

- Package Descriptions are needed by an OAIS to provide visibility and access to the OAIS holdings
- Package Descriptions contain 1 or more Associated Descriptions which describe the AIP Content Information from the point of view of a single Access Aid
- Some example of Access Aids Include:
  - Finding Aids - assist the consumer in locating information of interest
  - Ordering Aids - allow the consumer to discover the cost of and order AIUs of interest
  - Retrieval Aids - enable authorized users to retrieve the AIU described by the Unit Descriptor from Archival Storage

Data Management Information



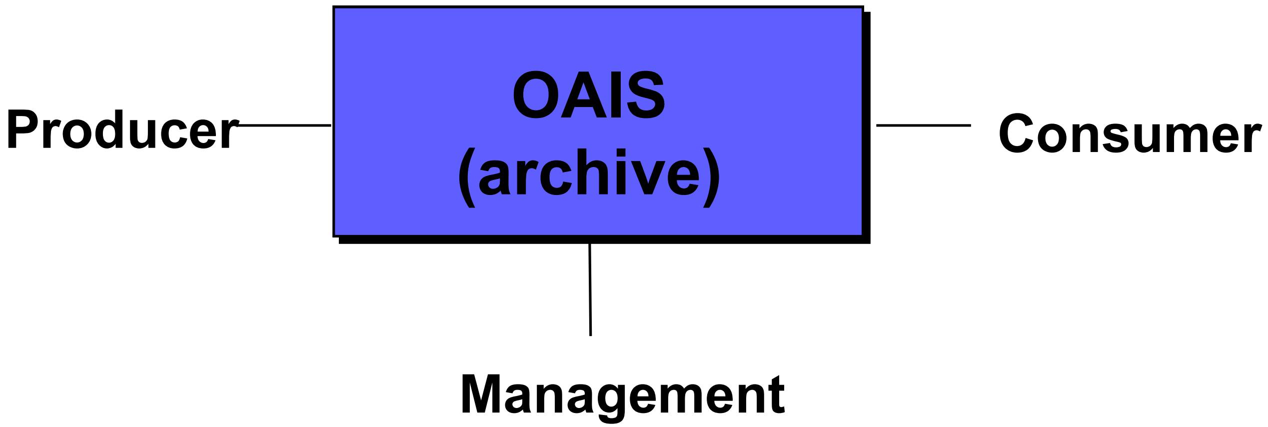
Information Model Summary

- Presented a model of information objects as containing data objects and representation objects
- Classified information required for Long-term archiving into 4 classes: Content Information, PDI, Packaging Information and Descriptive Information
- Described how these classes would be aggregated and related in an AIP to fully describe an instance of Content Information
- Presented information needed for Access, in addition to that needed for Long-term Preservation
- Put the Access oriented structures in the context of the other data needed to operate an OAIS

Outline

- Principles of the OAIS Model
- Technical Overview
- Functional Overview
- Information Modell
- Summary

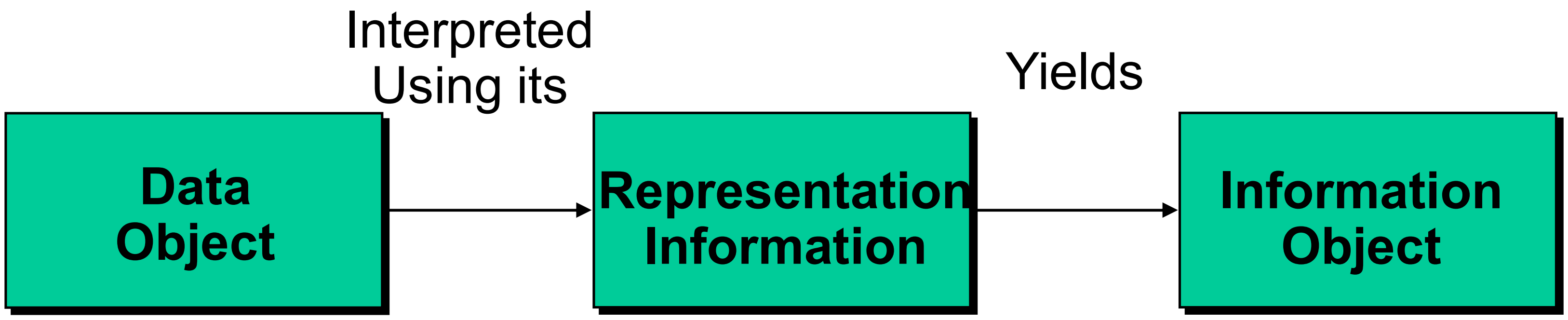
Model View of an OAIS Environment



- Producer is the role played by those persons, or client systems, who provide the information to be preserved
- Management is the role played by those who set overall OAIS policy as one component in a broader policy domain
- Consumer is the role played by those persons, or client systems, who interact with OAIS services to find and acquire preserved information of interest

OAIS Information Definition

- Information is always expressed (i.e., represented) by some type of data
- Data interpreted using its Representation Information yields Information
- Information Object preservation requires clear identification and understanding of the Data Object and its associated Representation Information



Summary

- OAIS is a reference model
- OAIS no implementation specification
- Defines language, responsibilities, functionalities,...
- Can be used for all kind of archives, institutions, organizations, systems
- Can be used for all kinds of objects, physical or digital

Data Management Plans

Tomasz Miksa

tomasz.miksa@tuwien.ac.at

Agenda

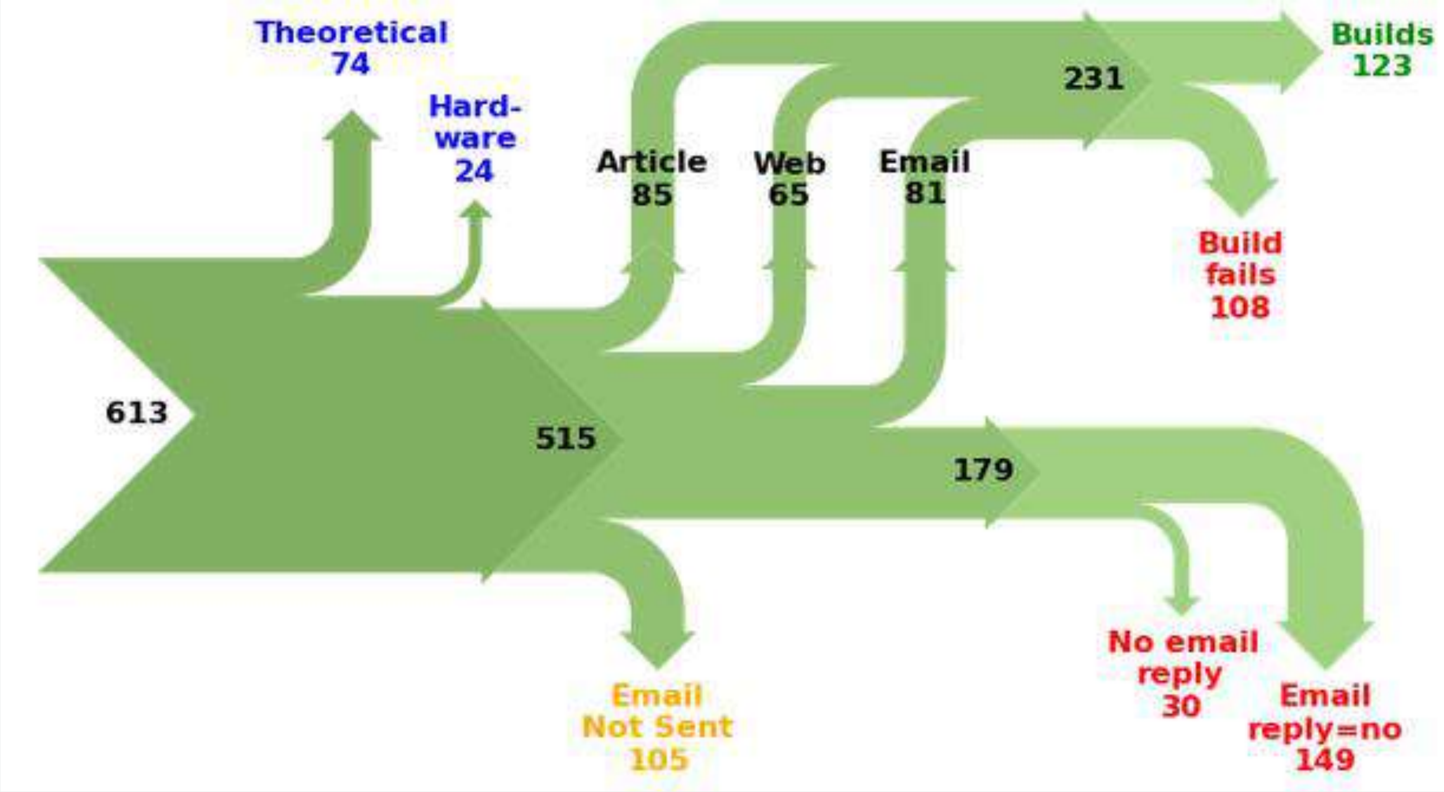
Why do we need to manage data properly?

What are Data Management Plans (DMPs)?

How to create a DMP?

Reproducibility - Computer Science

613 papers in 8 ACM conferences



C. Collberg and T. Proebsting, “Measuring reproducibility in computer systems research,” 2014. [Online]. Available: http://reproducibility.cs.arizona.edu/tr.pdf

Reproducibility - Computer Science

E-mail responses from authors

- Wrong version
- Code will be available soon
- Programmer left
- Bad backup practices
- Commercial code
- Proprietary academic code
- Intellectual property
- No intention to release
- …

Variety of solutions

In response to these needs many solutions were proposed and are being implemented

- **FAIR principles**
- **open access** to scientific publications and data
- research **data repositories** to host the data
- **persistent identifiers** to locate the data
- **data management plans**
- …

WHAT IS A DATA MANAGEMENT PLAN (DMP)?

Data Management Plan

DMP is a formal document

It outlines what you will do with your data **during** and **after** you complete your research

It ensures your data is safe for the **present** and the **future**

[from University of Virginia Library]

DMP is an awareness tool!

DMP makes you think

- what data you will use and where you get it from
- what infrastructure, software, licenses are needed
- what will be the output of your research
- how you will share your research outputs

DMP helps you organise yourself better

DMP can reveal how solid your methodology is

- is it a ‘fishing expedition’?

DMPs are used worldwide

- Required by
  - research funders
  - institutions, e.g. universities

Research Data Management

General Information

Research data management is an integral part of good research practice (see » Research Integrity & Research Ethics). The FWF therefore requires a data management plan (DMP) for all projects approved as of 1 January 2019. A DMP describes how data and their metadata are collected, organised, stored, published, shared, and archived for a specific project. Furthermore, the DMP outlines how the data will be made FAIR, which means Findable, Accessible, Interoperable and Reusable. The » FWF's Open Access Policy to Research Data must be taken into account when drafting the DMP.

https://www.fwf.ac.at/en/research-funding/open-access-policy/research-data-management

Example: Projects funded by European Commission

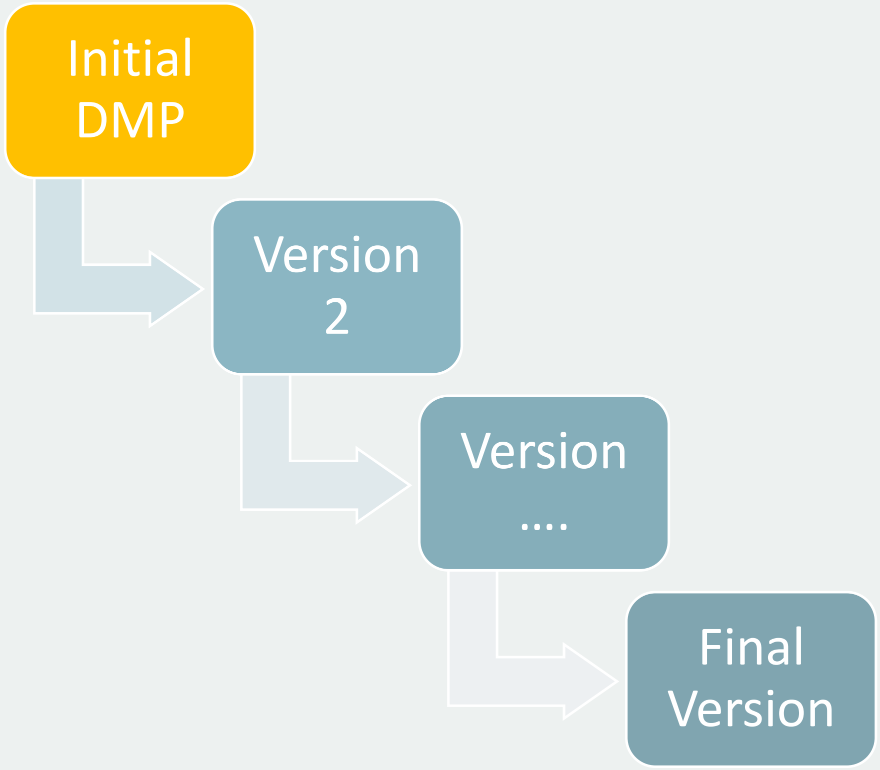
DMP is a living document

First version

- within the first 6 months

Updated versions

- when significant changes occur
  - new datasets
  - changes in policies
- periodic reporting
  - project reviews
- end of project



DMPs are not for research only

- DMPs are the requirement of research funders
  - But it does not mean they are not useful elsewhere!
- Fire extinguisher is an obligatory car equipment
  - But it does not mean it is not useful elsewhere!
- Goal of this lecture
  - Not to make you experts on funder requirements
    - Examples provided are meant for illustration only
  - To help you create DMPs whenever you work with data
  - To improve your (Research) Data Management (R)DM!
- Overlaps with FAIR

HOW TO CREATE A DMP?

How to create a DMP?

Most cases by

- filling out a template
- answering questions from a checklist

Using software tools

- users choose appropriate funders template
- only relevant questions and guidance is presented

Science Europe Guidelines

Basis for many funder templates

Table of Contents

4 For procedural elements of implementing DMPs, see the RDA DMP Common Standards Working Group: https://www.rd-alliance.org/groups/dmp-common-standards-wg

FWF Example

Based on the SE requirements

I General Information		
I.1 Administrative information	Provide information such as name of principal investigator, FWF project number, and version of DMP	- Provide the relevant grant information. - Consider regular updates of the DMP.
I.2 Data management responsibilities and resources	Who (for example, role, position, and institution) will be responsible for data management? What resources will be dedicated to data management and ensuring that data will be FAIR (Findable, Accessible, Interoperable, Re-usable)?	- Indicate who is responsible for implementing the DMP, and for ensuring it is reviewed and, if necessary, revised. - For collaborative projects, explain the co-ordination of data management responsibilities across partners. - Explain how the necessary resources (for example, time) to prepare the data for sharing/preservation have been costed in. Carefully consider and justify any resources needed to deliver the data. These may include storage costs, hardware, staff time, and repository charges.
II Data Characteristics		
II.1 Data description and collection or re-use of existing data	How will new data be collected or produced and/or how will existing data be re-used? What data (types, formats, and volumes) will be collected or produced?	- Explain which methodologies or software will be used if new data are collected or produced. - State any constraints on re-use of existing data if there are any. - Explain how data provenance will be documented. - Give details on the kind of data: for example, numeric (databases), textual (documents), image, audio, or video. - Give details on the data format: the way in which the data is encoded for storage, often reflected by the filename extension (for example, pdf, xls, doc, txt, or rdf).

DMP tools

DMP Online

- https://dmponline.dcc.ac.uk/

Data Stewardship Wizard

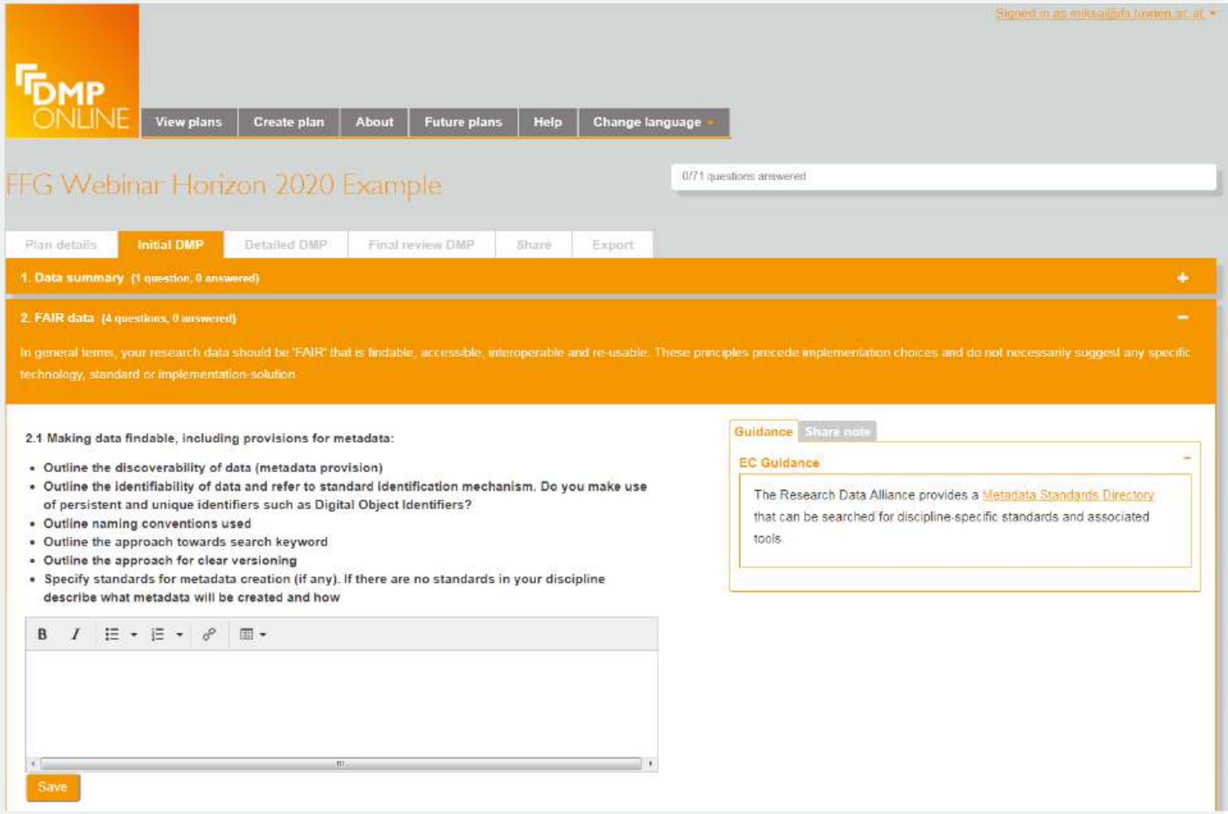
- https://ds-wizard.org

Argos

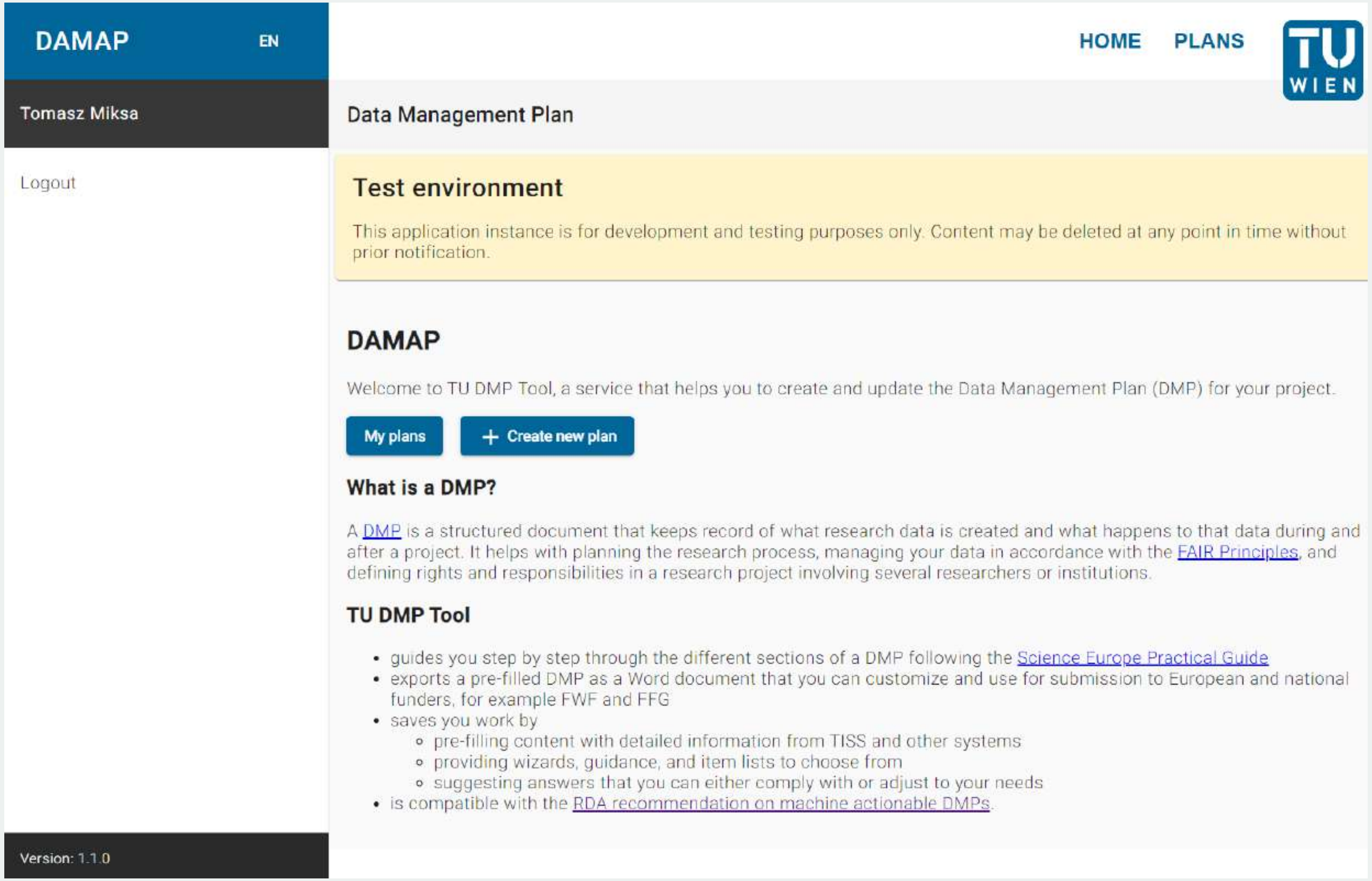
- https://argos.openaire.eu/splash/

RDMO

- https://rdmorganiser.github.io/en/



DMP tool at TUW

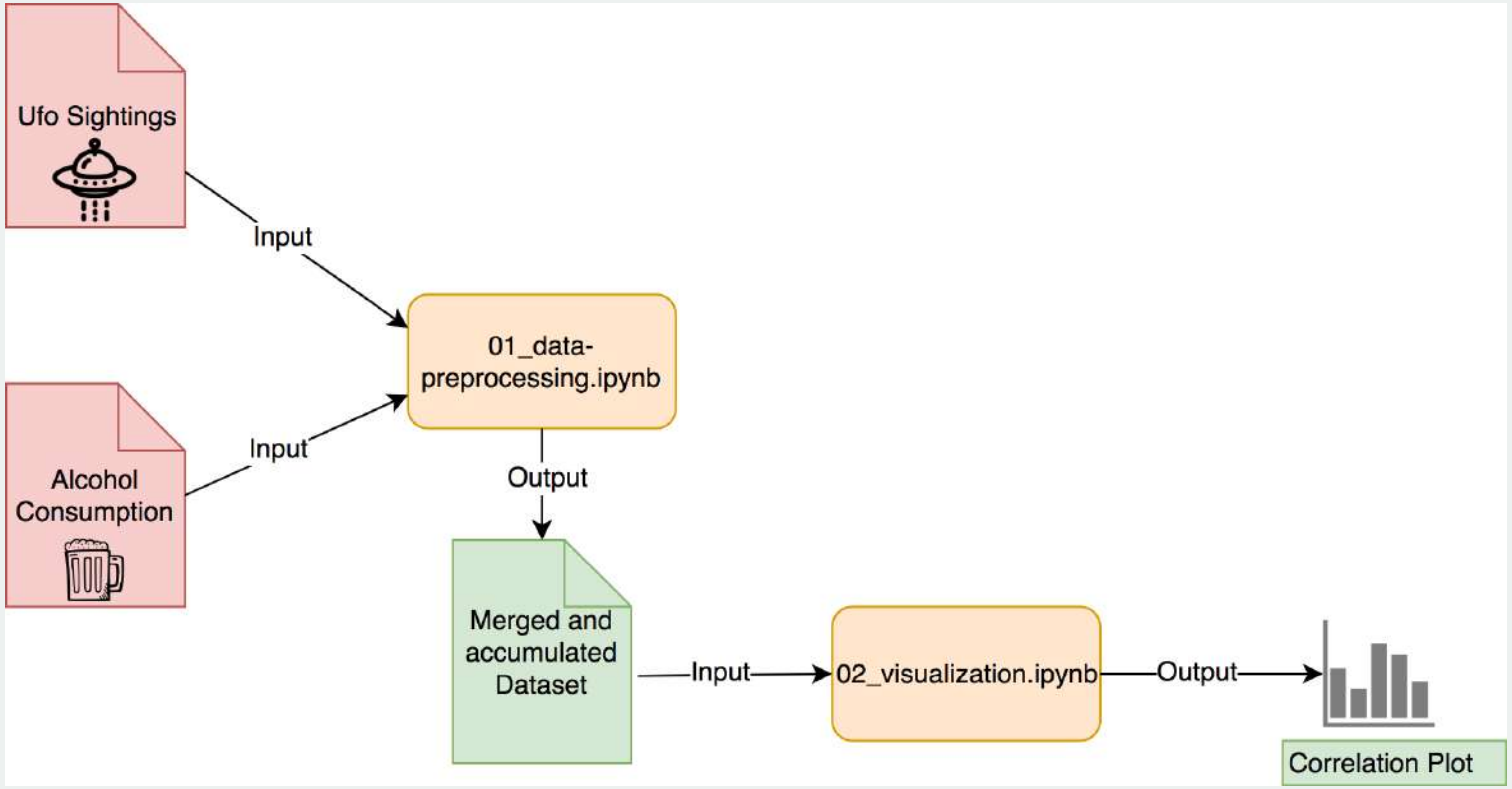


https://damap-qa.apps.dev.csd.tuwien.ac.at

(and what I should also do!)

WHAT SHOULD I WRITE IN FACT?

Correlating Alcohol Consumption and UFO Sightings in the USA (running example)



https://github.com/mdietrichstein/digitalpreservation-dmp

Running example

This example is ***very simplified***

- Few inputs, few outputs
  - Compared to thousands of file processed
- ‘Digital objects’ are identical with files (which is not the case usually)
  - Compared to collections of files with a complex structure
- Experiment is identical with a single notebook
  - Compared to several workflows, software tools, etc.

DP Exercise

- Do not use the example as an easy copy-paste!

FWF Template – running example

I. General information

II. Data Characteristics

III. Documentation and Data Quality

IV. Data Storage, Sharing and Long-Term Preservation

V. Legal and Ethical Aspects

This lecture: not an exhaustive walk-through! Only interesting/relevant aspects!

I General Information	
I.1 Administrative information	
I.2 Data management responsibilities and resources	
II Data Characteristics	
II.1 Data description and collection or re-use of existing data	
III Documentation and Data Quality	
III.1 Metadata and documentation	
III.2 Data quality control	
IV Data Storage, Sharing, and Long-Term Preservation	
IV.1 Data storage and backup during the research process	
IV.2 Data sharing and long-term preservation	
V Legal and Ethical Aspects	
V.1 Legal aspects	
V.2 Ethical aspects	

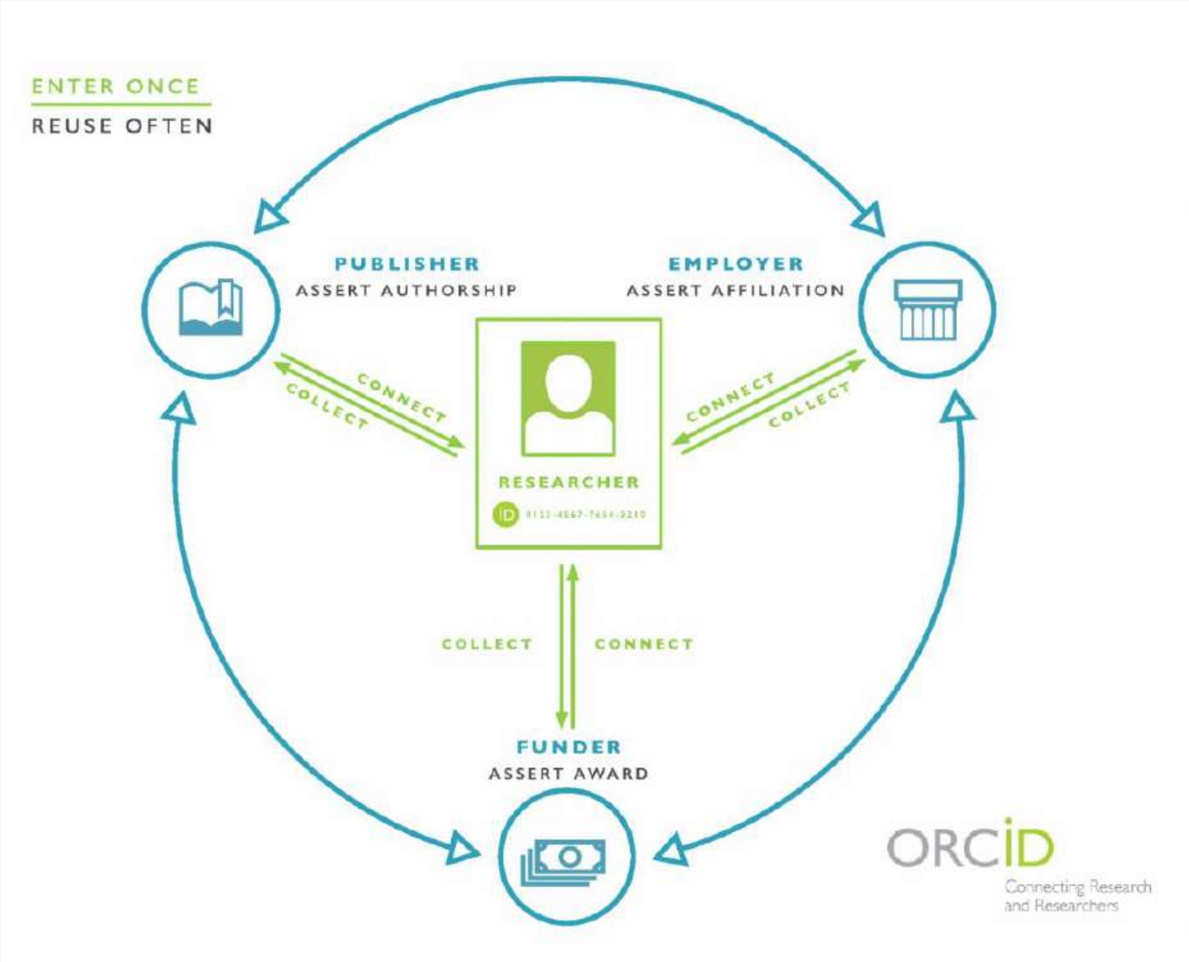
I. GENERAL INFORMATION

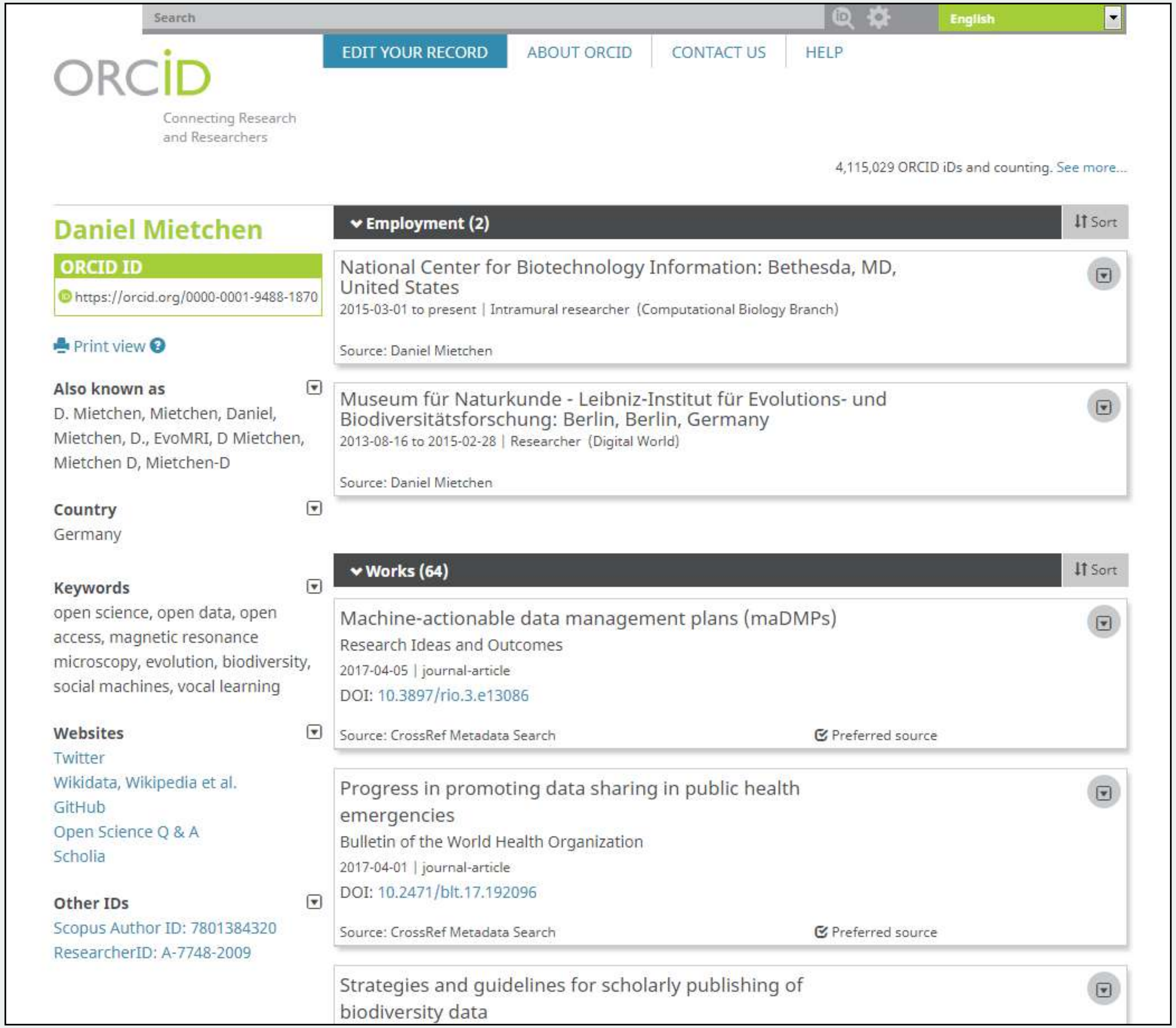
ORCID – persistent identifier for people

ORCID ID

- Unique person ID
- ORCID assigned once
- Person can change affiliations (jobs)
- Example: 0000-0002-4929-7875

ORCID
Connecting Research
and Researchers




Search
English
EDIT YOUR RECORD
ABOUT ORCID
CONTACT US
HELP
ORCID
Connecting Research and Researchers
4,115,029 ORCID iDs and counting. See more...
Daniel Mietchen
ORCID ID
https://orcid.org/0000-0001-9488-1870
Print view
Also known as
D. Mietchen, Mietchen, Daniel, Mietchen, D., EvoMRI, D Mietchen, Mietchen D, Mietchen-D
Country
Germany
Keywords
open science, open data, open access, magnetic resonance microscopy, evolution, biodiversity, social machines, vocal learning
Websites
Twitter
Wikidata, Wikipedia et al.
GitHub
Open Science Q & A
Scholia
Other IDs
Scopus Author ID: 7801384320
ResearcherID: A-7748-2009
Employment (2)
Sort
National Center for Biotechnology Information: Bethesda, MD, United States
2015-03-01 to present | Intramural researcher (Computational Biology Branch)
Source: Daniel Mietchen
Museum für Naturkunde - Leibniz-Institut für Evolutions- und Biodiversitätsforschung: Berlin, Berlin, Germany
2013-08-16 to 2015-02-28 | Researcher (Digital World)
Source: Daniel Mietchen
Works (64)
Sort
Machine-actionable data management plans (maDMPs)
Research Ideas and Outcomes
2017-04-05 | journal-article
DOI: 10.3897/rio.3.e13086
Source: CrossRef Metadata Search
Preferred source
Progress in promoting data sharing in public health emergencies
Bulletin of the World Health Organization
2017-04-01 | journal-article
DOI: 10.2471/blt.17.192096
Source: CrossRef Metadata Search
Preferred source
Strategies and guidelines for scholarly publishing of biodiversity data

II. DATA CHARACTERISTICS

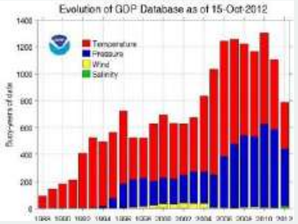
slido

What is data?

ⓘ Start presenting to display the poll results on this slide.

What is data?

•Instrument measurements

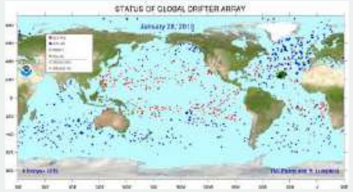


•Experimental observations

•Still images, video and audio

•Text documents, spreadsheets, databases

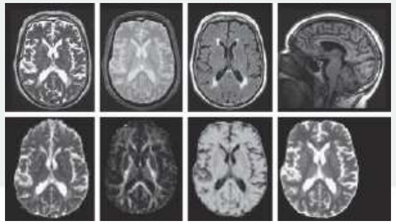
•Quantitative data (e.g. survey data)



•Survey results & interview transcripts

•Simulation data, models & software

•Slides, artefacts, specimens, samples



•Questionnaires

•Sketches, diaries, lab notebooks …

Data Summary

Type

- text, spreadsheets, software, models, images, movies, audio, patient records, etc.

Source

- human observation, laboratory, field instruments, experiments, simulations, compilations, etc.

Volume

- total volume of data, number of files, etc.

Data and file formats

- non-proprietary formats
- used within community

Data Summary - example

Produced Data

This project produces aggregated dataset in CSV format (Filesize ~800K) that contains data points that combine alcohol consumption data with the UFO sighting data and a correlation plot of these in PNG format (Filesize (~100K)).

Input Data

Project accesses two external CSV datasets. Both datasets have been downloaded and saved along with the source code in the folder *data/raw*.

1. Alcohol Consumption: OECD (2018), Alcohol consumption (indicator).

    DOI: 10.1787/e6895909-en (Accessed on 22 March 2018)

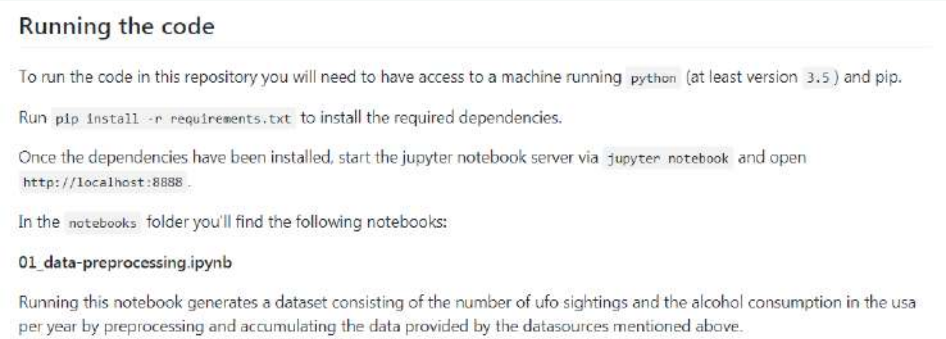
    File Location: data/raw/DP_LIVE_22032018202902423.csv

    File Size: 112K

Data Summary - example

The experiment has been conducted with Jupyter notebooks. The notebooks contain the experiment's code, accompanying documentation, tables and plots.

We have included instructions (README.md) on how to run the experiment either directly or via Docker.



Running the code

To run the code in this repository you will need to have access to a machine running `python` (at least version `3.5`) and pip.

Run `pip install -r requirements.txt` to install the required dependencies.

Once the dependencies have been installed, start the jupyter notebook server via `jupyter notebook` and open `http://localhost:8888`.

In the `notebooks` folder you'll find the following notebooks:

01_data-preprocessing.ipynb

Running this notebook generates a dataset consisting of the number of ufo sightings and the alcohol consumption in the usa per year by preprocessing and accumulating the data provided by the datasources mentioned above.

III. DOCUMENTATION AND DATA QUALITY

What is in the picture?

slido

How is this bird called?

ⓘ Start presenting to display the poll results on this slide.

slido

What is the sex of this bird?

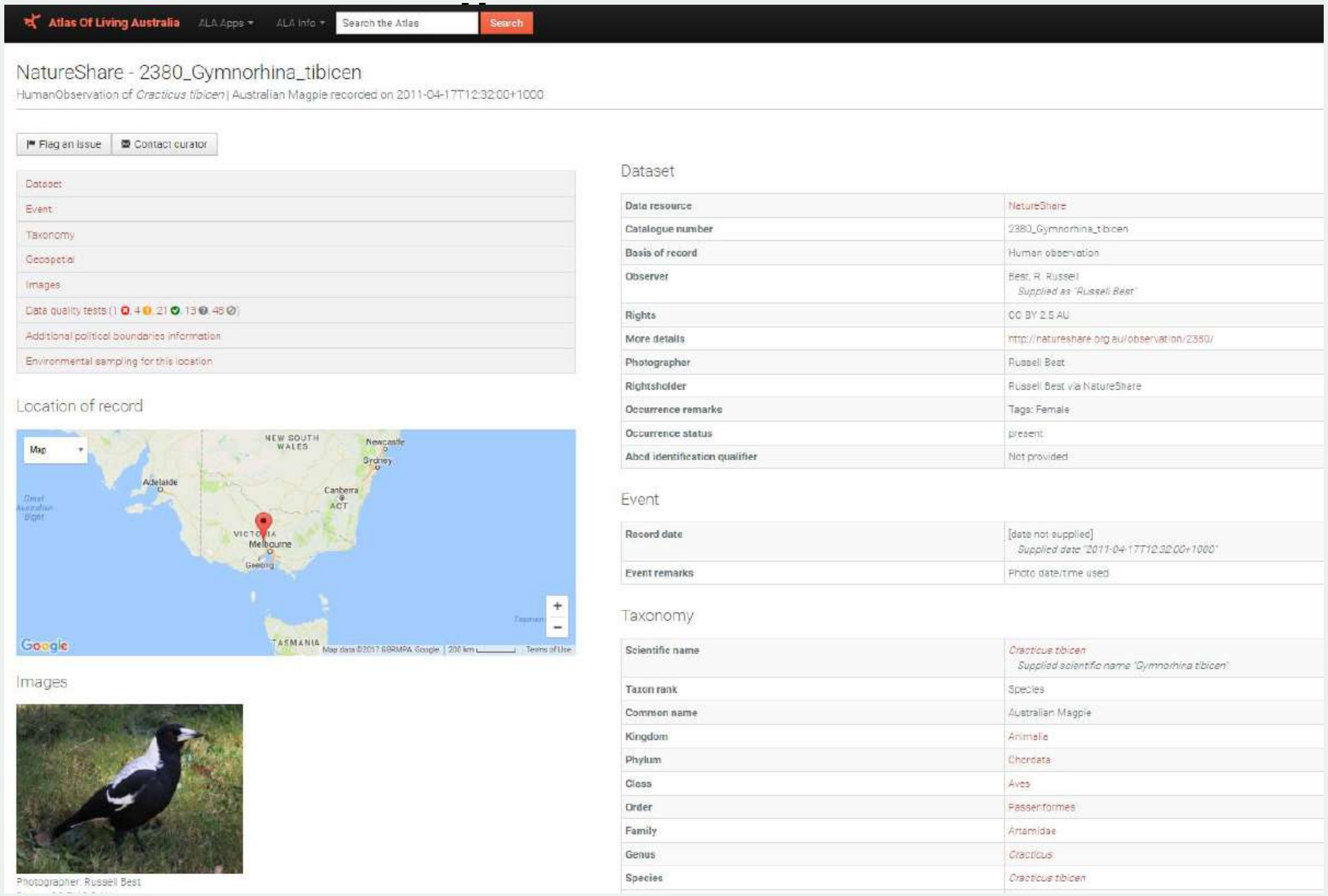
ⓘ Start presenting to display the poll results on this slide.

slido

On which continent was the picture taken?

ⓘ Start presenting to display the poll results on this slide.

Metadata – Atlas Of Living



http://biocache.ala.org.au/occurrences/544b0271-5f04-47ab-9d8b-0dbe3b5f59d7

Metadata – Atlas Of Living Australia

Dataset

Data resource	NatureShare
Catalogue number	2380_Gymnorhina_tibicen
Basis of record	Human observation
Observer	Best, R. Russell *Supplied as "Russell Best"*
Rights	CC BY 2.5 AU
More details	http://natureshare.org.au/observation/2380/
Photographer	Russell Best
Rightsholder	Russell Best via NatureShare
Occurrence remarks	Tags: Female
Occurrence status	present
Abcd identification qualifier	Not provided

Metadata – Atlas Of Living Australia

Event

Record date	[date not supplied] *Supplied date "2011-04-17T12:32:00+1000"*
Event remarks	Photo date/time used.

Taxonomy

Scientific name	*Cracticus tibicen* *Supplied scientific name "Gymnorhina tibicen"*
Taxon rank	Species
Common name	Australian Magpie
Kingdom	Animalia
Phylum	Chordata
Class	Aves
Order	Passeriformes
Family	Artamidae
Genus	*Cracticus*
Species	*Cracticus tibicen*
Taxonomic issues	No issues
Name match metric	Exact match The supplied name matched the name exactly.

Metadata – Atlas Of Living Australia

Geospatial

Country	Australia
State or territory	Victoria
Local government area	Macedon Ranges (S)
Latitude	-37.421078
Longitude	144.61954
Geodetic datum	EPSG:4326
Biome	Terrestrial
Verbatim longitude	144.619541
Verbatim latitude	-37.421077

Location of record

Standards and metadata

Metadata

- helps to understand and interpret data
- provides details about experiment setup
  - who, when, in which conditions, tools, versions, etc.
- helps identify and discover new data

Use community standards to enable interoperability

Metadata is also covered in the lecture on FAIR

http://www.dcc.ac.uk/resources/metadata-standards

Metadata - example

The metadata file can be found inside the project folder (/documentation/metadata.xml).

- experiment title, authors, date, tools, coverage, rights, etc.

Additionally a descriptive file is added, which <u>explains the axes and units</u> used in the output files, this file can be found inside the project folder as well (/documentation/description.txt).

- The alcohol consumption is the average consumption rate in liters/capita of USA inhabitants, which are older than 17.

Metadata-example

```xml
<?xml version="1.0" encoding="UTF-8"?>

<metadata xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:dc="http://pur
<dc:title>UFOs & Alcohol</dc:title>
<dc:creator>Marc Dietrichstein (https://orcid.org/0000-0003-4890-3498)</dc:creator>
<dc:creator>Markus Neumeyer (https://orcid.org/0000-0002-4081-0716)</dc:creator>
<dc:subject>Correlation of alcohol consumtion and UFO sightings</dc:subject>
<dc:description>Automated tool that investigates and computes the correlation betwee
<dc:date>23.03.2018</dc:date>
<dc:type>DataGeneration</dc:type>
<dc:format>Jupyternotebook</dc:format>
<dc:source>Ufo Sightings: Sigmond Axel. (2014)</dc:source>
<dc:source>Alcohol Consumption: OECD (2018)</dc:source>
<dc:language>English</dc:language>
<dc:coverage>1960 - 2014 </dc:coverage>
<dc:rights>Free access</dc:rights>
```

Some comments

- Sometimes it is hard to tell the difference between data and metadata
  - NetCDF (network Common Data Form)
    - is a file format for storing multidimensional scientific data (variables) such as temperature, humidity, pressure, wind speed, and direction.
    - Is self-describing, meaning that a netCDF file includes information about the data it contains, such as when data elements were captured and what units of measurement were used.
  - https://youtu.be/K1_8EqCJlwo
- Very often metadata is reduced to domain independent metadata
  - e.g. only Dublin Core or Data Cite metadata in a repository
    - Author, title, description, license, etc.
  - What about domain specific metadata?

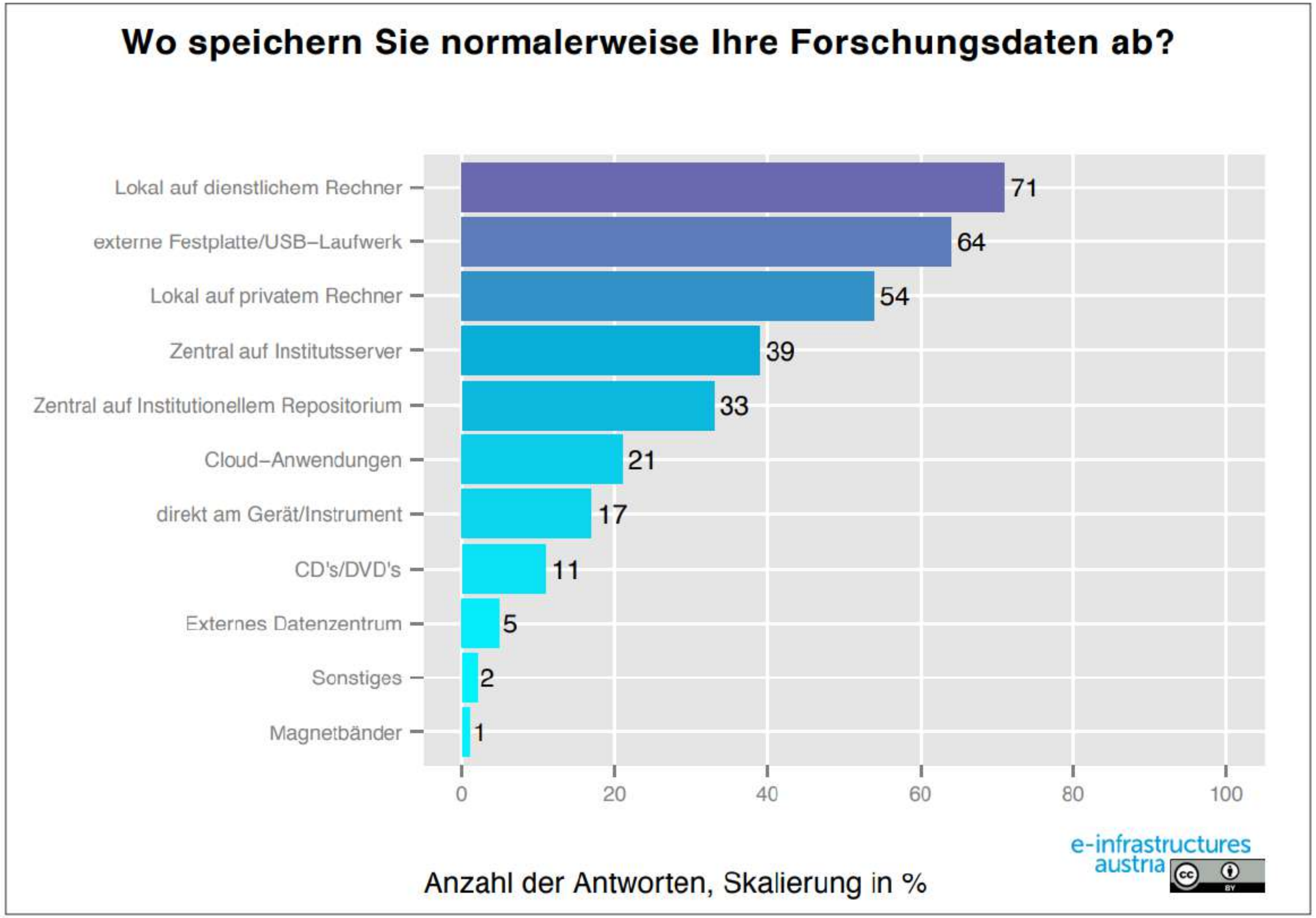
IV. DATA STORAGE, SHARING AND LONG-TERM PRESERVATION

slido

Where do you keep your data?

ⓘ Start presenting to display the poll results on this slide.

Managing data during research



http://phaidra.univie.ac.at/o:407513

Managing data during research

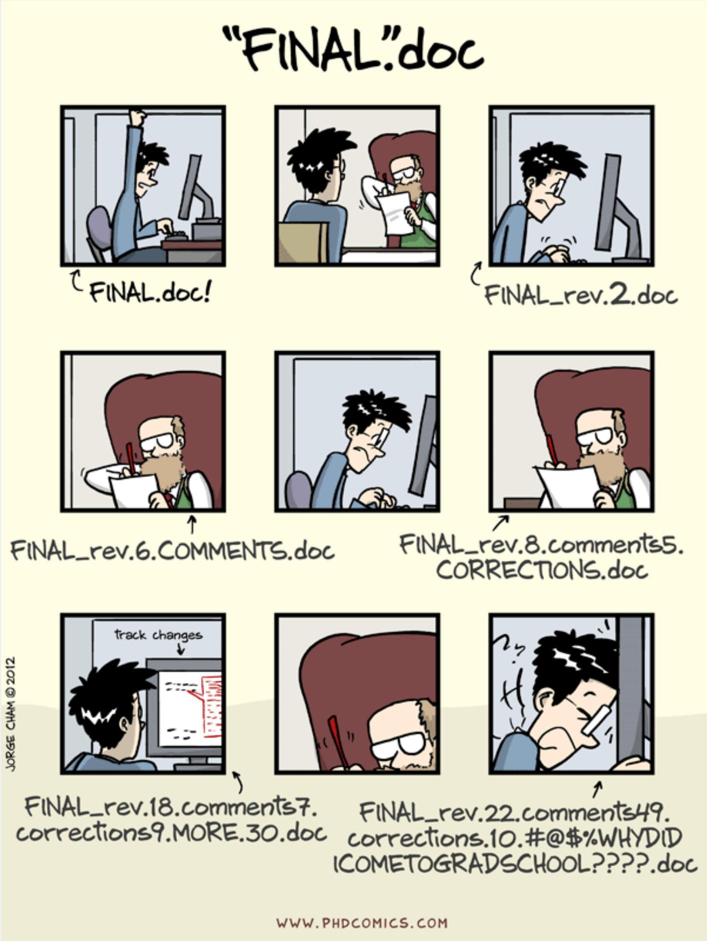
If you loose your data there will be nothing to share!

Recreating or recollecting data can be

- impossible
  - e.g. observational data
- too expensive
  - e.g. cost of computational power

How do you manage data during the project?

- file naming convention
- versioning
- backups
- should the access be restricted?
- who is responsible?



Data sharing - example

Code and data are hosted in a public git repository on GitHub.

Read access is open to everyone. Write-access is limited to the researchers working on the project.

Permissions are managed via Github's account system using SSH keys.

Backup vs archiving and preservation (traditional view)

Data managed during the project
- Changed/deleted
- Backup

Stable snapshot of data
- Moved into a repository
- Enriched with metadata and licensing details
- Not only backups

What makes a system a repository?

Backup vs archiving and preservation (new approaches)

- No need to differentiate between project and post-project phases
- One system can be used for managing and preserving data

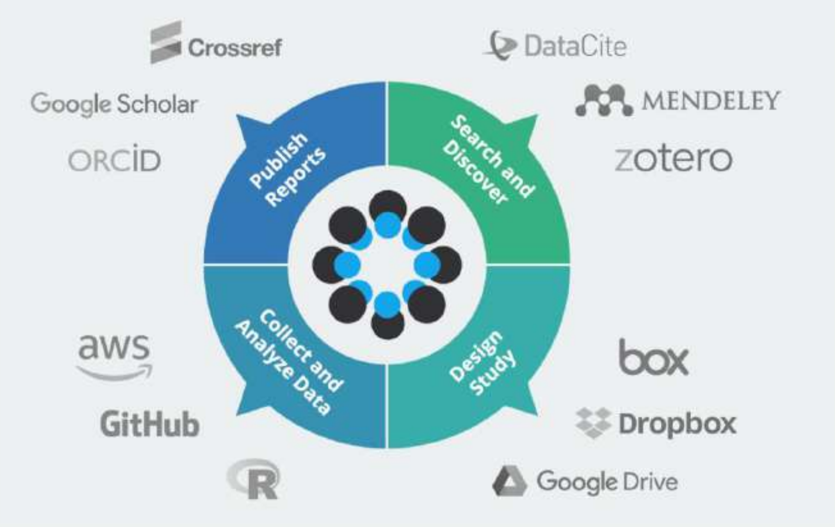

Designing Research Repositories Using Automated Workflows and Machine Actionable Data Management Plans

MASTER'S THESIS

Master of Science

Software Engineering and Internet Computing

Asztrik Bakos



https://www.cos.io/products/osf

Archiving and preservation

Which data will be shared?

- What has to be kept?
- What can’t be recreated?
- What is potentially useful to others?
- What legally must be destroyed?

Where will the data be deposited?

- not all of the data must be shared in the same way

Are there any embargo periods?

For how long?

What is the cost and who will pay for it?

Which license to use?

Preservation, repositories, costs are discussed in separate lectures!

Archiving and preservation - example

The following files are relevant to reproduce the experiment and should be preserved

- *README.md* – Text file containing instructions on how to run the experiment
- Both *Jupyer notebooks* - The experiment's code and documentation
- *Dockerfile* - To build a docker container for running the experiment
- *requirements.txt* - List of python dependencies required by the experiment
- *documentation/architecture.pn*g - Architectural diagram of the experiment
- *documentation/description.txt* – Text file describing the correlation plot's content
- *documentation/metadata.xml* - Metadata relevant to the experiment

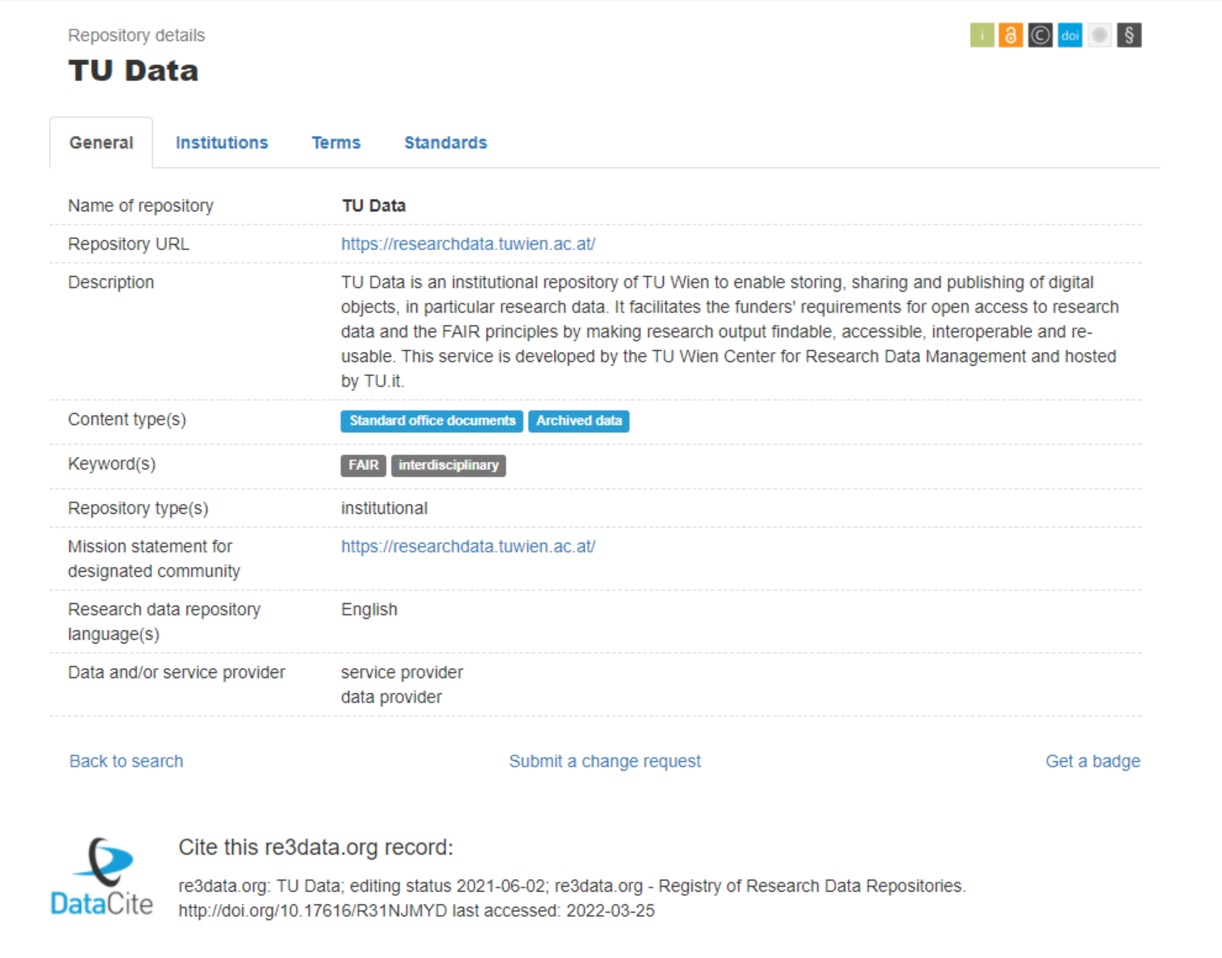
Note: input data was not selected for preservation

- it is maintained by existing repository (*easy* to get)

Where to find a repository?

1. Use Domain specific repository
   - e.g. chEMBL (if you work with molecules)
2. Use Institutional repository
   - e.g. phaidra.univie.ac.at (if you work at Uni Wien)
3. Search registry to find a relevant one
   - e.g. re3data.org
4. Use *catch-all* repository
   - e.g. zenodo.org

re3data.org



Repository details

TU Data

General | Institutions | Terms | Standards

Name of repository	**TU Data**
Repository URL	https://researchdata.tuwien.ac.at/
Description	TU Data is an institutional repository of TU Wien to enable storing, sharing and publishing of digital objects, in particular research data. It facilitates the funders' requirements for open access to research data and the FAIR principles by making research output findable, accessible, interoperable and re-usable. This service is developed by the TU Wien Center for Research Data Management and hosted by TU.it.
Content type(s)	Standard office documents, Archived data
Keyword(s)	FAIR, interdisciplinary
Repository type(s)	institutional
Mission statement for designated community	https://researchdata.tuwien.ac.at/
Research data repository language(s)	English
Data and/or service provider	service provider data provider

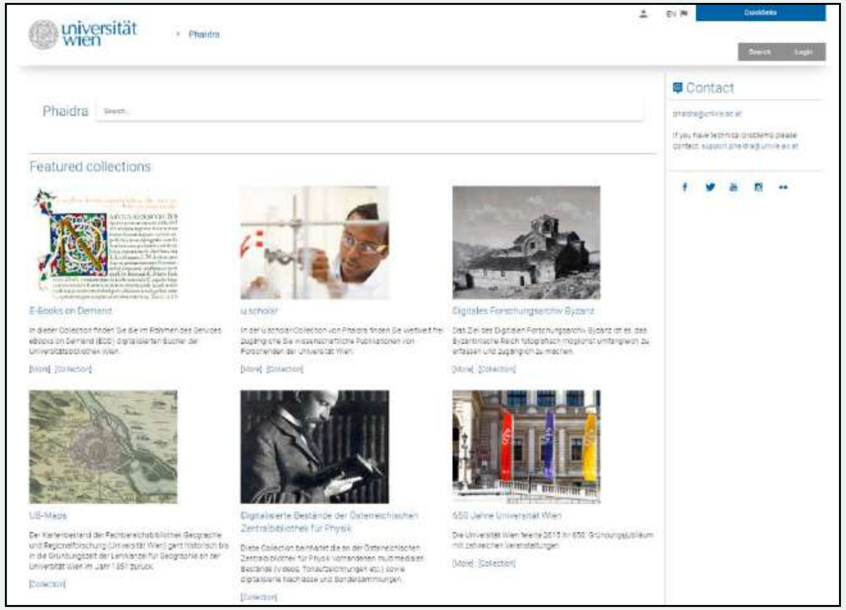
Back to search   Submit a change request   Get a badge

DataCite

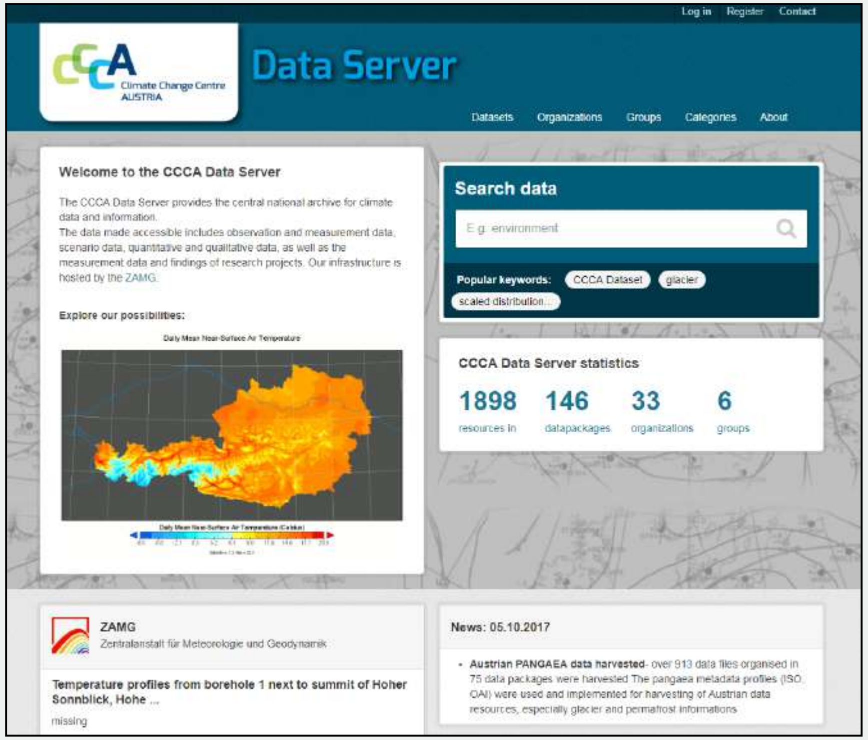
Cite this re3data.org record:

re3data.org: TU Data; editing status 2021-06-02; re3data.org - Registry of Research Data Repositories. http://doi.org/10.17616/R31NJMYD last accessed: 2022-03-25

Repositories in Austria - examples



https://phaidra.univie.ac.at



https://data.ccca.ac.at

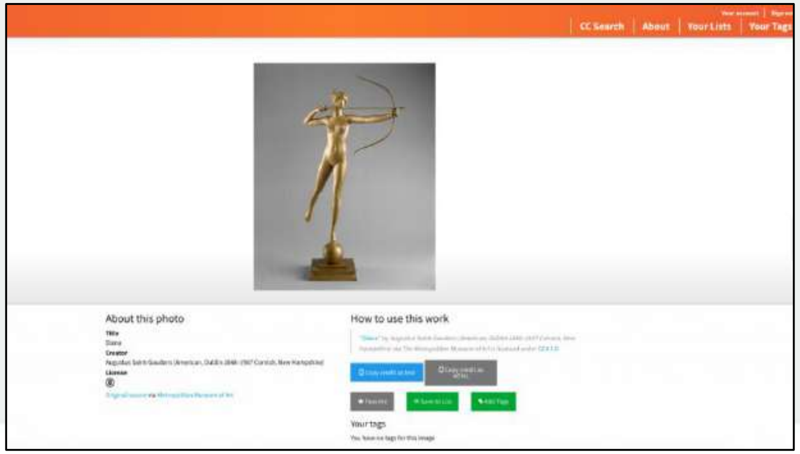
V. LEGAL AND ETHICAL ASPECTS

Licenses

CC-0

- waives creator rights -> public domain
- allows anyone to use, re-use, and remix a work without restriction
- Example: all images from Metropolitan Museum of Art in New York

  https://creativecommons.org/2017/02/07/met-announcement/



Creative Commons

CC-BY (Attribution)

- allows anyone to use, re-use, and remix a work without restriction, also commercially
- You must give appropriate credit, provide a link to the license, and indicate if changes were made.

CC BY-SA (Attribution – ShareAlike)

- all new works must care the same license

CC BY-ND (Attribution- NoDerivs)

- reuse, but no changes

CC BY-NC

- no commercial use

CC BY-NC-SA

CC BY-NC-ND

Software Licenses

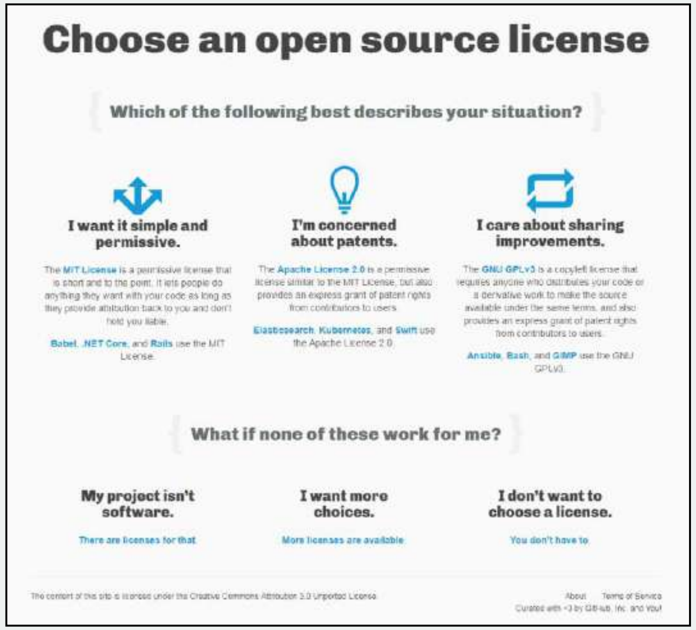
Choose correct license for your software

- Apache, MIT, GNU, BSD, …

Check licenses of libraries you reuse in your software

- Example: GNU GPL vs GNU **L**GPL
    - GPL enforces the reusing software to be GPL (also public)
    - LGPL code must be clearly marked, rest of the software can have different license (can be private)

Software licenses can also be used for data



https://choosealicense.com

License - example

The external datasets are using permissible licenses which allows us the usage and redistribution of the following data:

- * Ufo Sightings - Creative Commons Attribution 4.0
- * Alcohol Consumption - Free to use and distribute according to http://www.oecd.org/termsandconditions/ -Section C - Permitted use

All code, data and documentation is available on Github and is licensed under the MIT license.

More on legal aspects of data management

Verena Dolovai will give a lecture in May

SUMMARY

Tips for writing DMPs

DMP can reveal how solid your work is

Seek advice - consult and collaborate

When answering questions from checklists write coherent text

Be specific when referring to tools and standards

Assign responsibilities and name responsible personnel

Tips for writing DMPs

Think about things early…

- Negotiation on licenses and consent agreement may preclude later sharing if not careful
- Manage your data correctly from the very beginning
  - backups, file naming conventions, access restrictions, metadata collection
- Plan your budget

Decisions made early on affect what you can do later

DMPs are not that perfect

Data Management Plans

- are manually created
- depend on scientific honesty
- focus mainly on input and output data
- provide very general overview of the experiment
- have scarce information about the process
- cannot be automatically validated
- do not support sufficiently the reproducibility of research

Lecture on maDMPs will focus on how to fix some of these problems

You should know and be able to explain

- Why and what for we need DMPs
- how to improve your own data management
- what a DMP is and what kind of information it contains
  - Data
  - Metadata
  - Repositories
  - Licensing
  - …

Useful resources

Center for RDM at TUW (login first!)

- **https://www.tuwien.at/en/research/rti-support/research-data/center-for-rdm**

Managing and sharing data by UK Data Archive

- http://www.data-archive.ac.uk/media/2894/managingsharing.pdf

EUDAT webinars on data management

- https://www.eudat.eu/events/webinar/research-data-management-an-introductory-webinar-from-openaire-and-eudat

FFG-Akadmie: Der Data Management Plan (DMP) in Horizon 2020 (Webinar)

- https://www.ffg.at/europa/veranstaltungen/ffg-akademie_2017-10-18

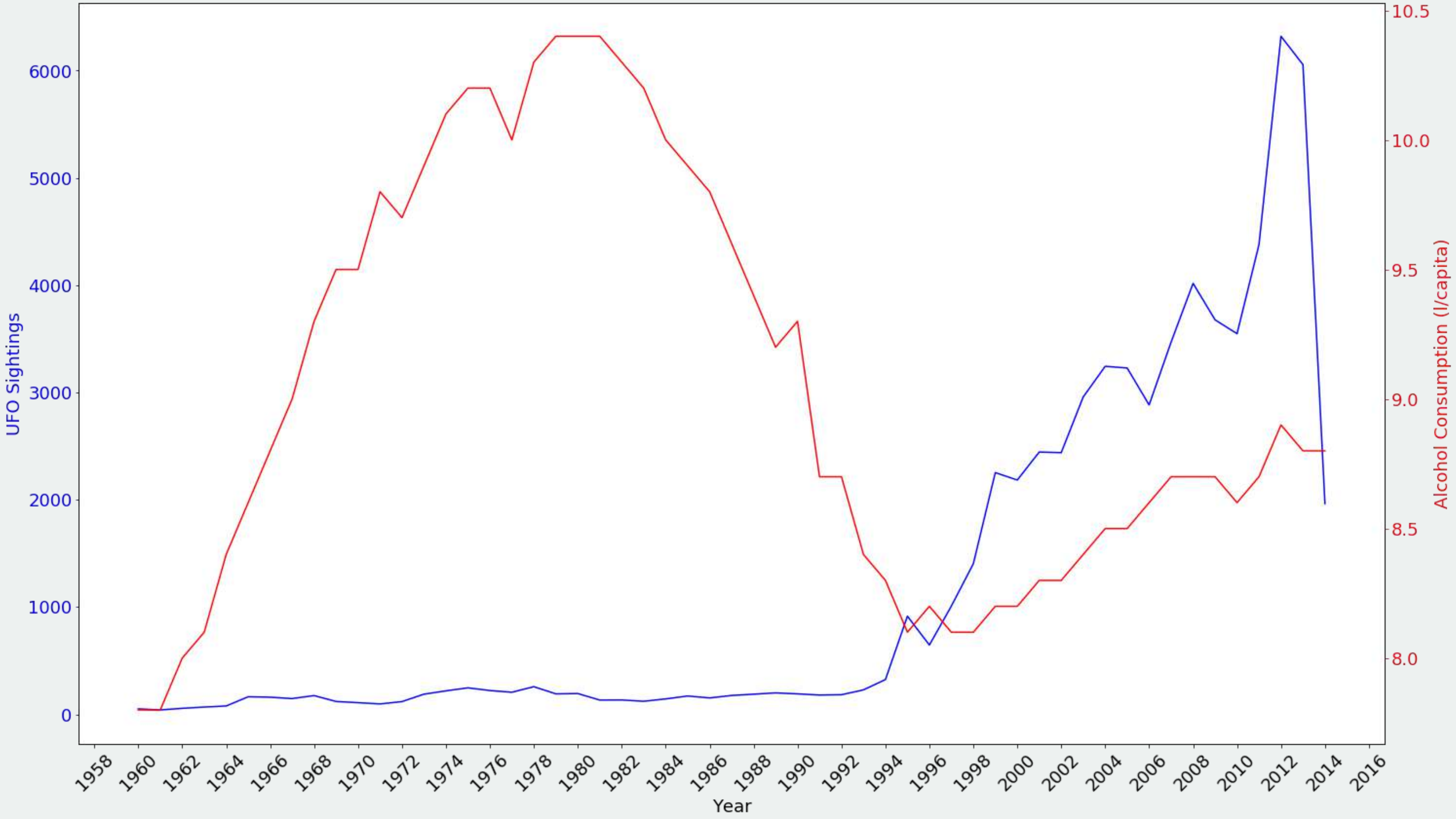
Ten Simple Rules

- http://dx.doi.org/10.1371/journal.pcbi.1004525

DMP Checklist

- http://www.dcc.ac.uk/sites/default/files/documents/resource/DMP/DMP_Checklist_2013.pdf

Correlating Alcohol Consumption and UFO Sightings in the USA



https://github.com/mdietrichstein/digitalpreservation-dmp

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

ifs FACULTY OF !NFORMATICS

Outline

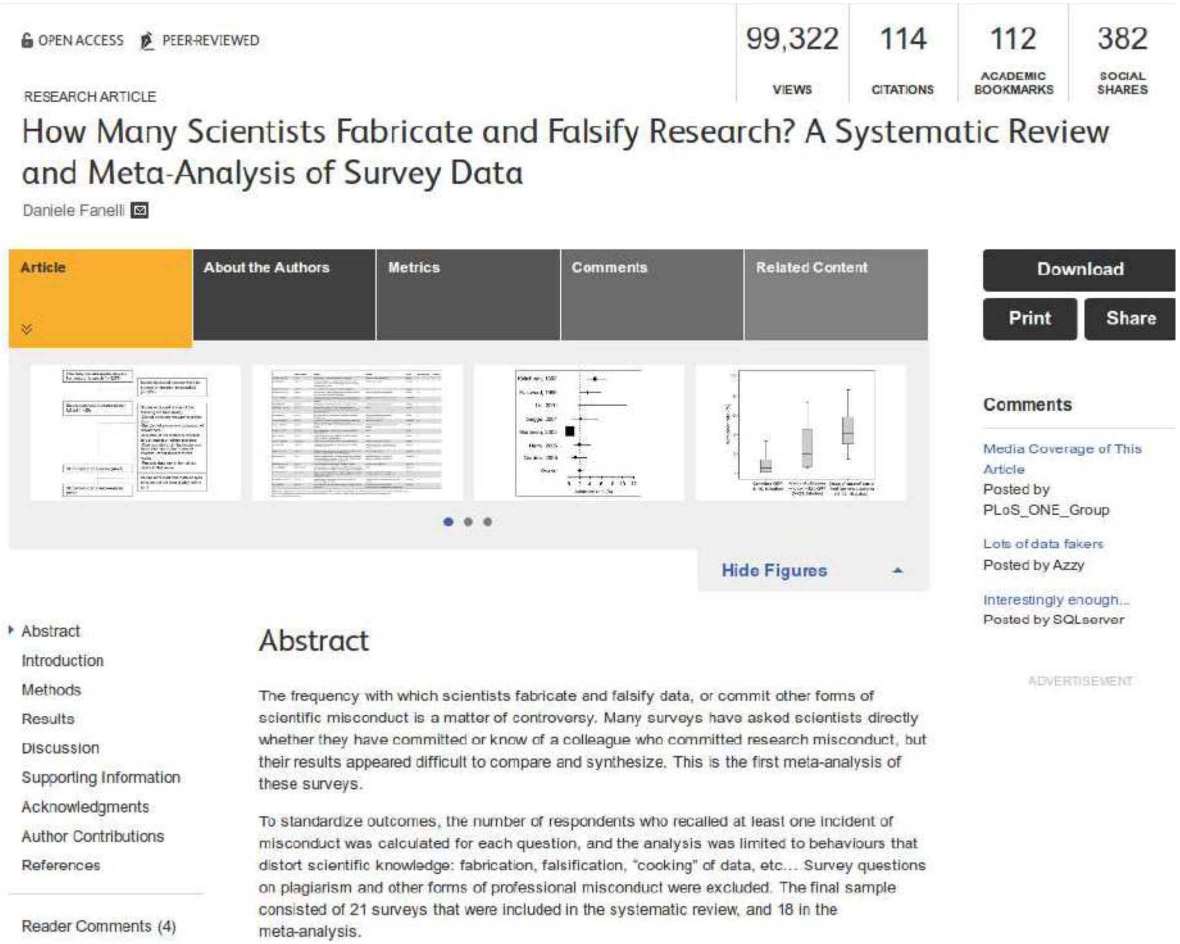
- Why should we want to cite data?
- What identifier system should I use?
- What are the challenges in data identification and citation?
- How should we do it, according to the RDA WG?
- Who is doing it so far, and how?
- Summary

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

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

RESEARCH ARTICLE

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

Daniele Fanelli

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

Download | Print | Share

Hide Figures

Comments

Media Coverage of This Article
Posted by PLoS_ONE_Group

Lots of data fakers
Posted by Azzy

Interestingly enough...
Posted by SQLserver

ADVERTISEMENT

Abstract
Introduction
Methods
Results
Discussion
Supporting Information
Acknowledgments
Author Contributions
References

Reader Comments (4)

Abstract

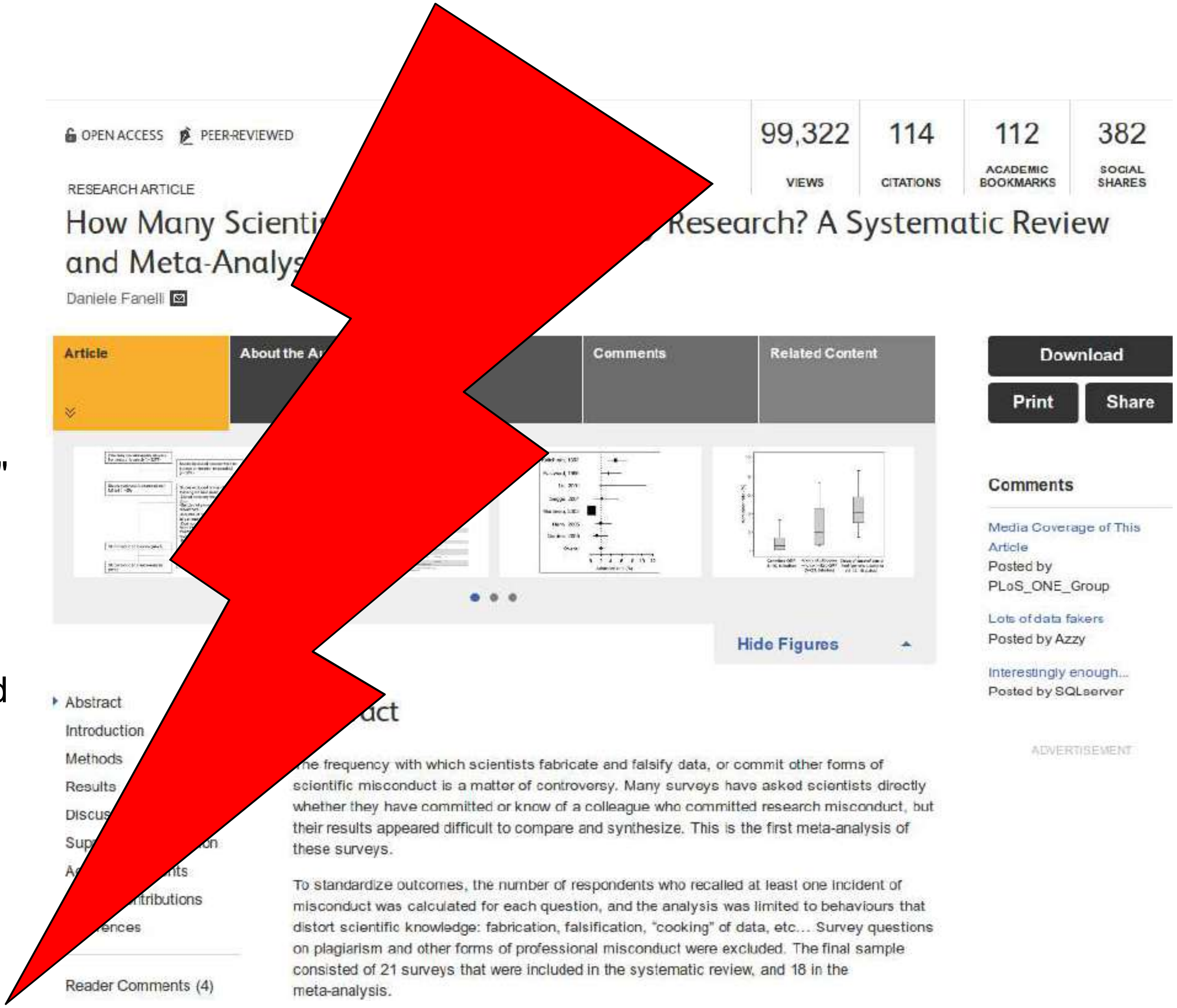
The frequency with which scientists fabricate and falsify data, or commit other forms of scientific misconduct is a matter of controversy. Many surveys have asked scientists directly whether they have committed or know of a colleague who committed research misconduct, but their results appeared difficult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

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

Source: http://www.plosone.org

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.



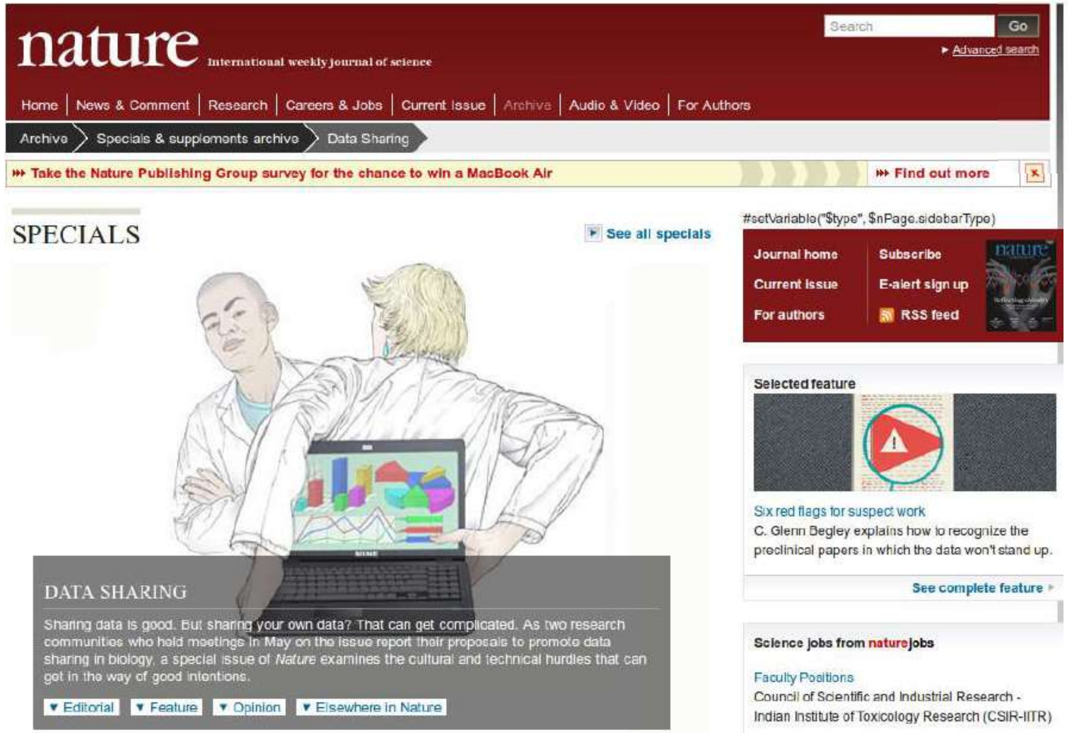
Source: http://www.plosone.org

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

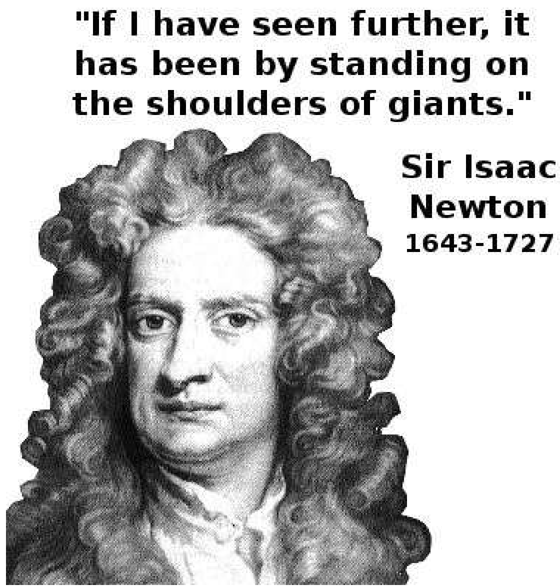
Why to cite data?

- 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, …



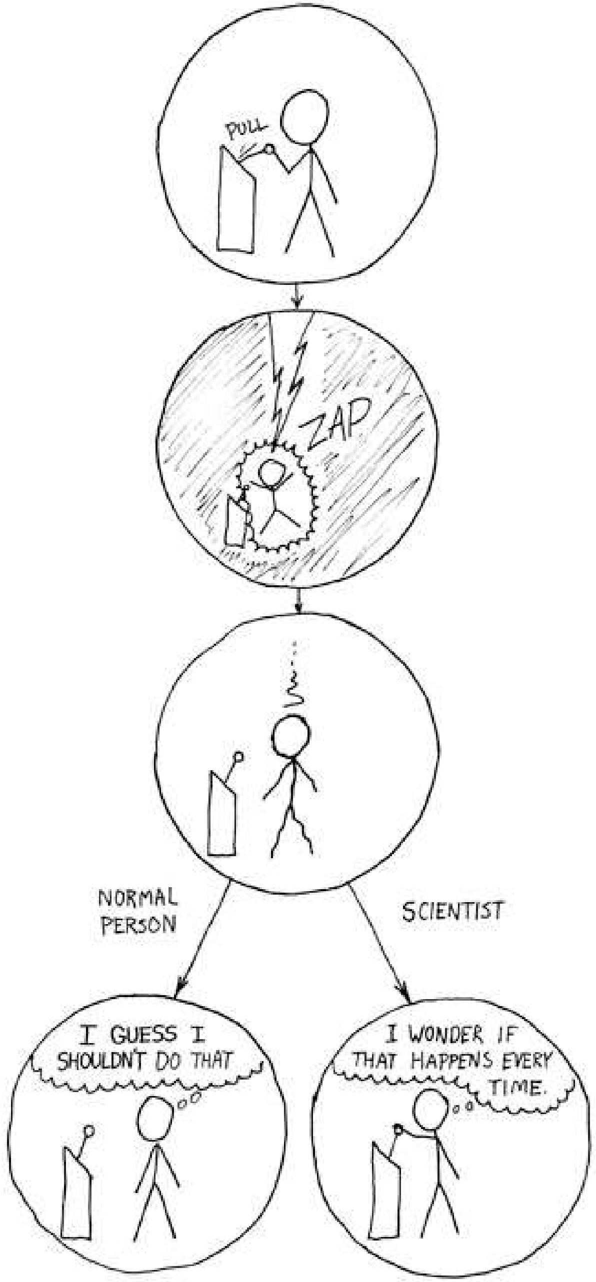
Why to cite data?

- 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/

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) ✔

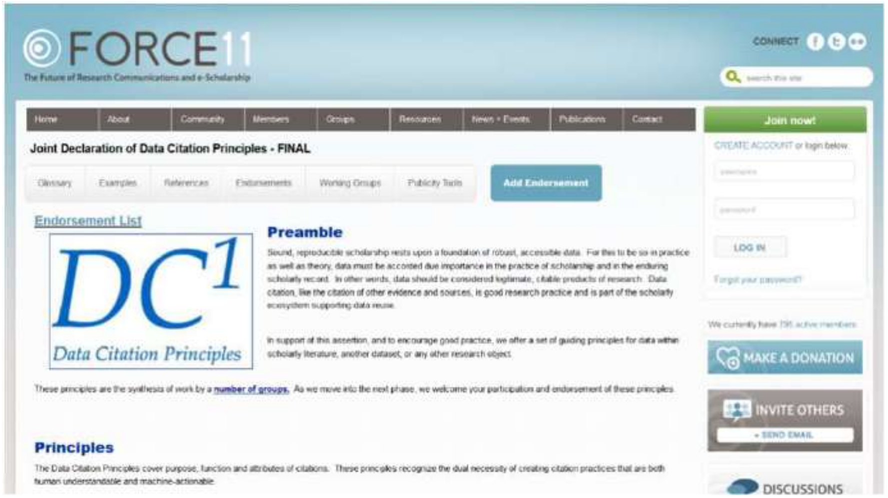
Why to cite data?

- 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”)**

Joint Declaration of Data Citation Principles

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



Joint Declaration of Data Citation Principles (cont'd)

DC1 Data Citation Principles

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)

DC1 Data Citation Principles

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)

DC1 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)

DC[1]
Data Citation Principles

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)

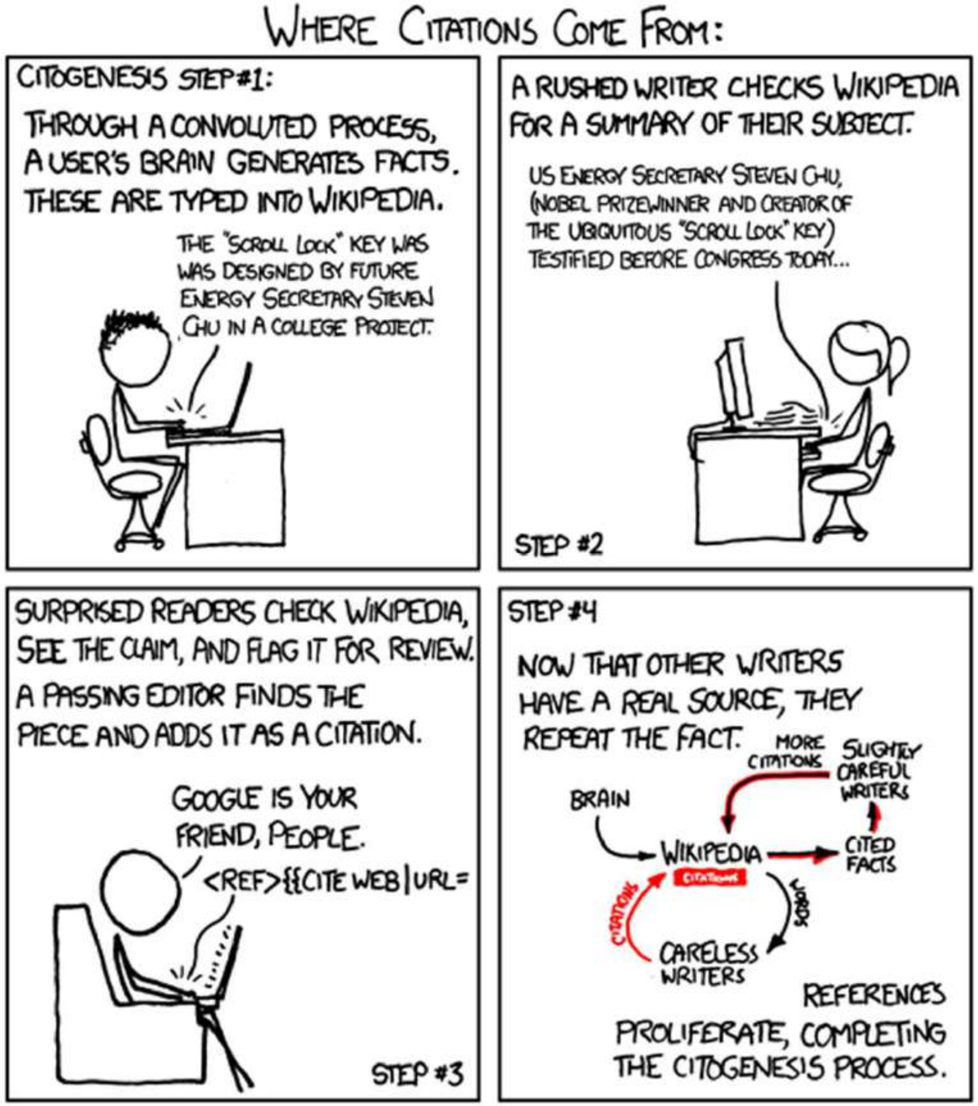
DC[1]
Data Citation Principles

8) Interoperability and flexibility

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

Benefits of Citation

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



https://xkcd.com/978/

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"

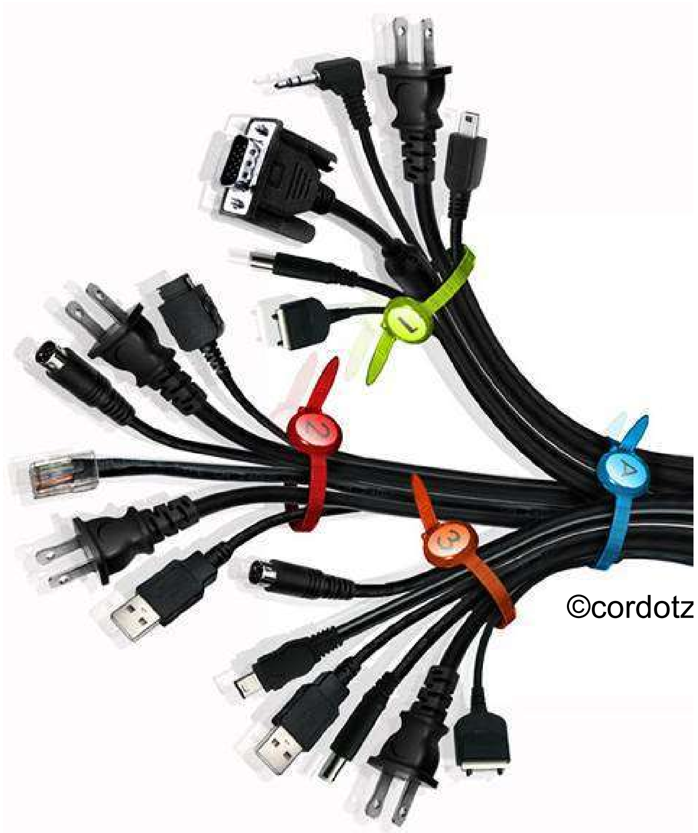
Source: http://www.dcc.ac.uk/resources/how-guides/cite-datasets

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



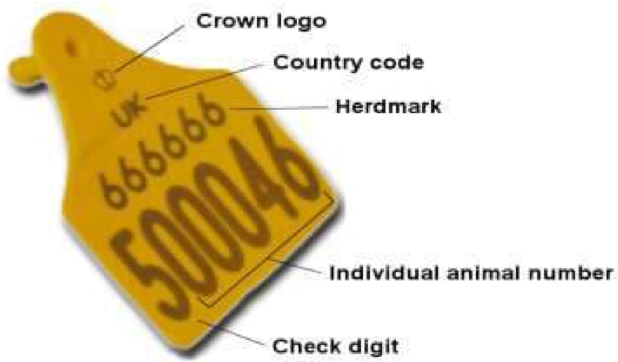
©cordotz

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 (basis for all others)
  - 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):

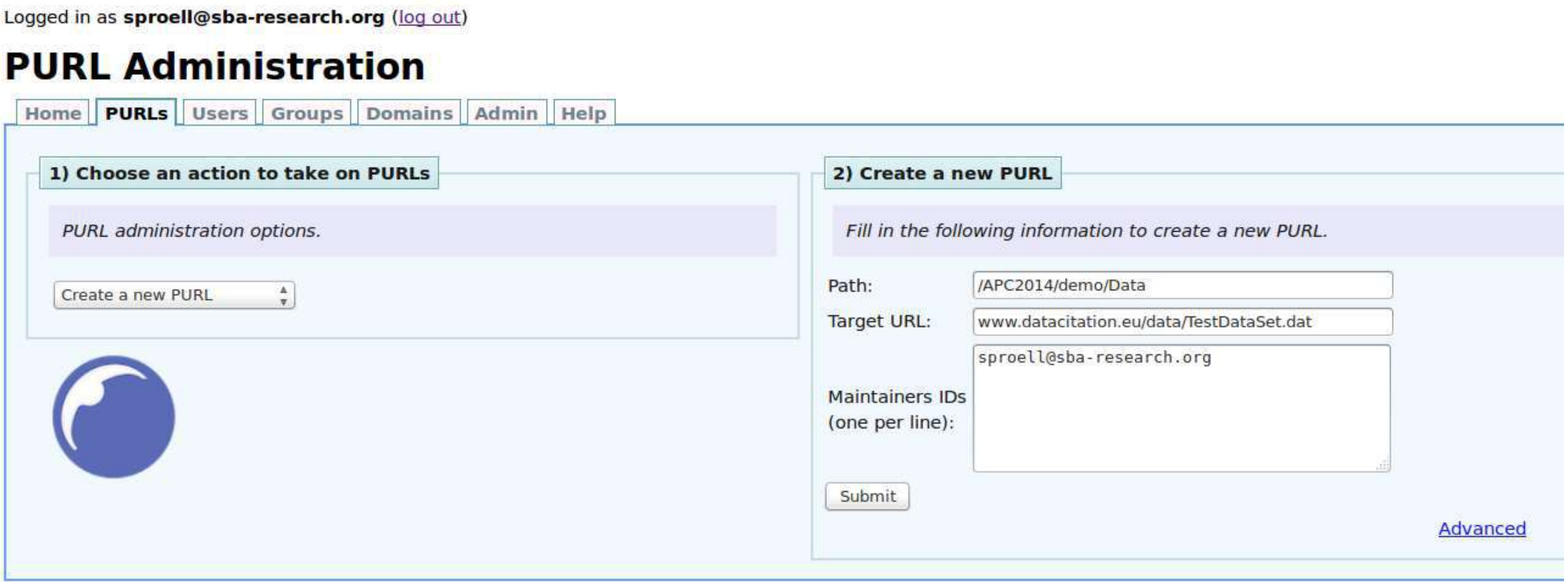
Submit

Create Successful

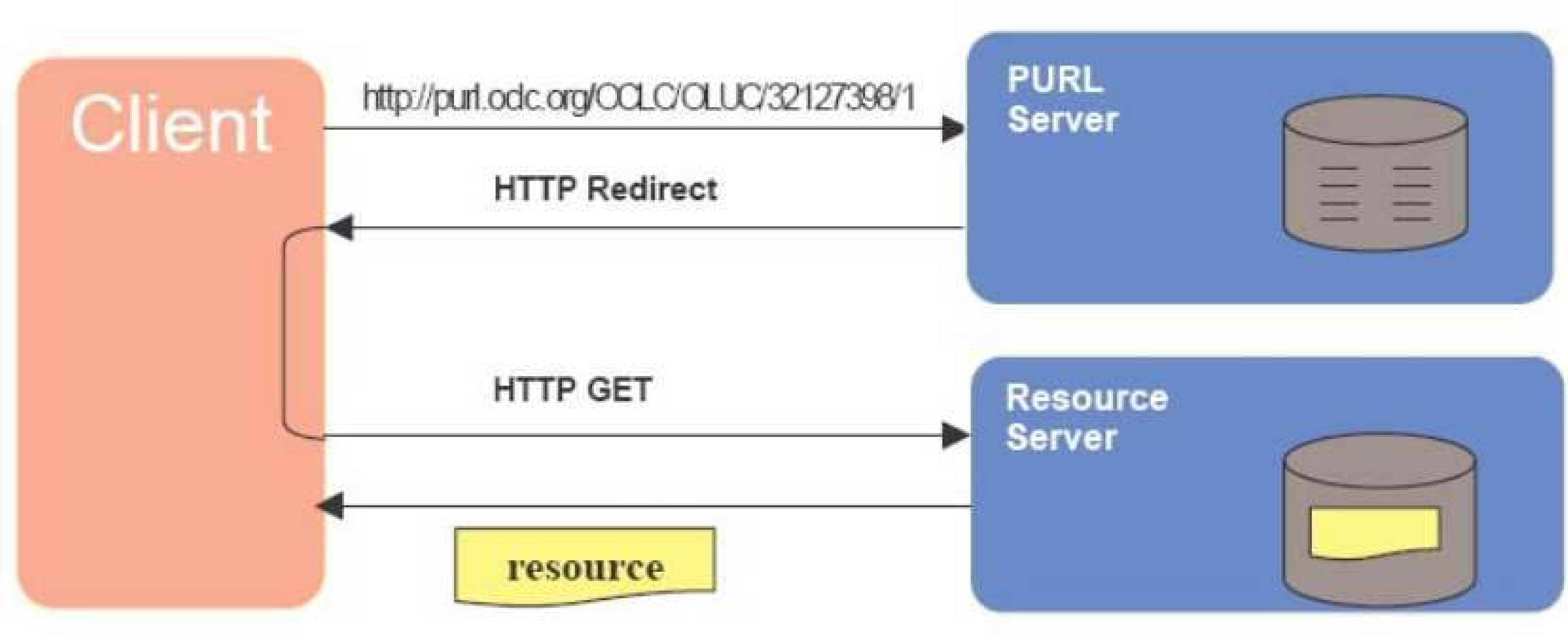
status: Pending approval

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

PURL Registration



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>
    - urn: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 system)

Handle

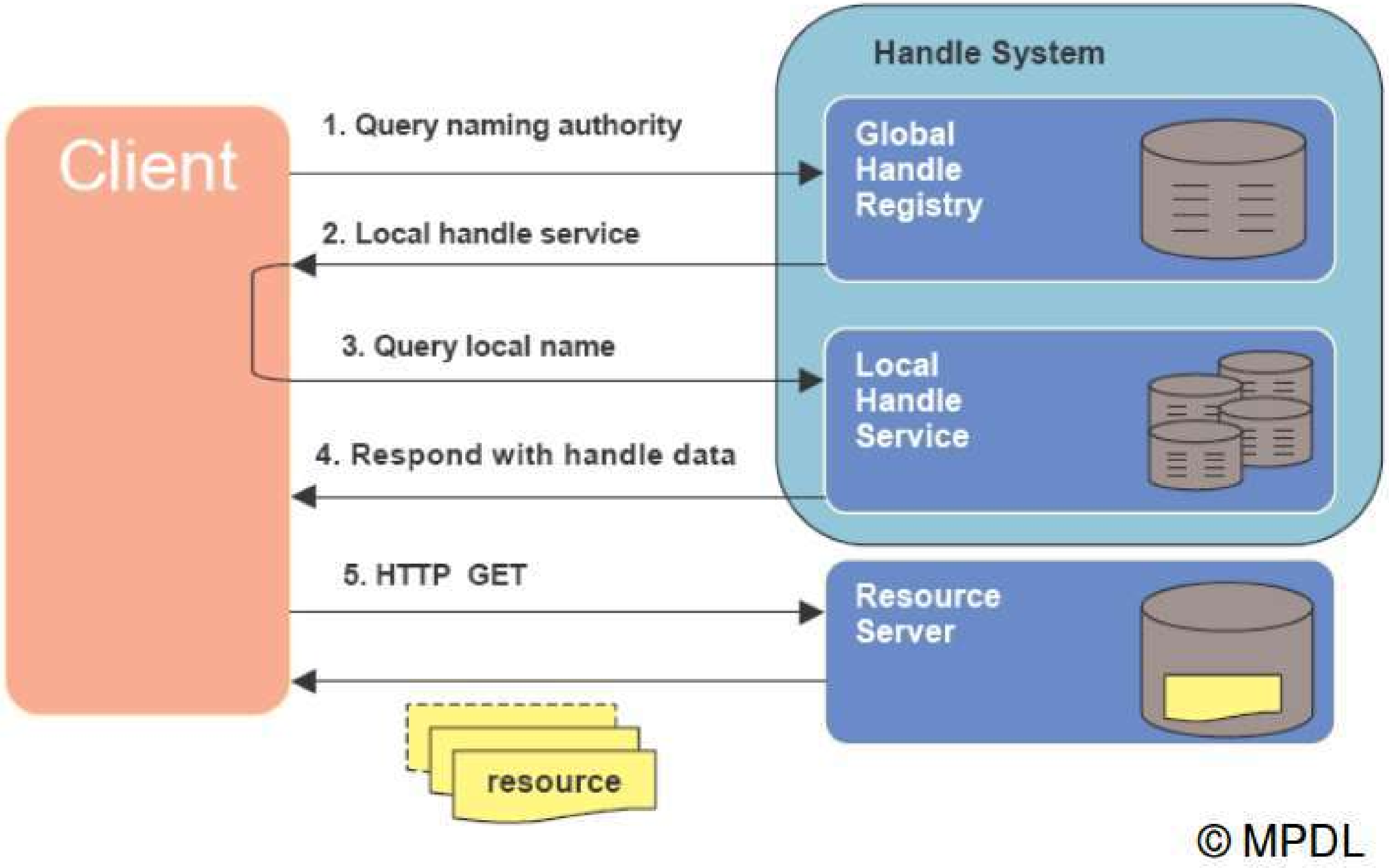
- Distributed persistent naming system
- Conforms to URN framework
- Used by most identifier systems
- Persistent identifier consists of two parts:
  - Naming authority
  - Name (must be unique string to the authority)
- Digital objects on the Internet can be assigned, managed and resolved by handles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

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
- International Standard: ISO 26324 (May 2012)
- TU Wien Library provides DOI Services for Austria
  https://www.ub.tuwien.ac.at/pid/pid_result.html

Digital Object Identifier (DOI)

- Consists of three parts:

  http://dx.doi.org/10.1016/S0921-4526(00)00753-5

Resolver Service	Prefix (Assigning Body)	Suffix (Resource)
http://dx.doi.org/	10.1016/	S0921-4526(00)00753-5

- 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

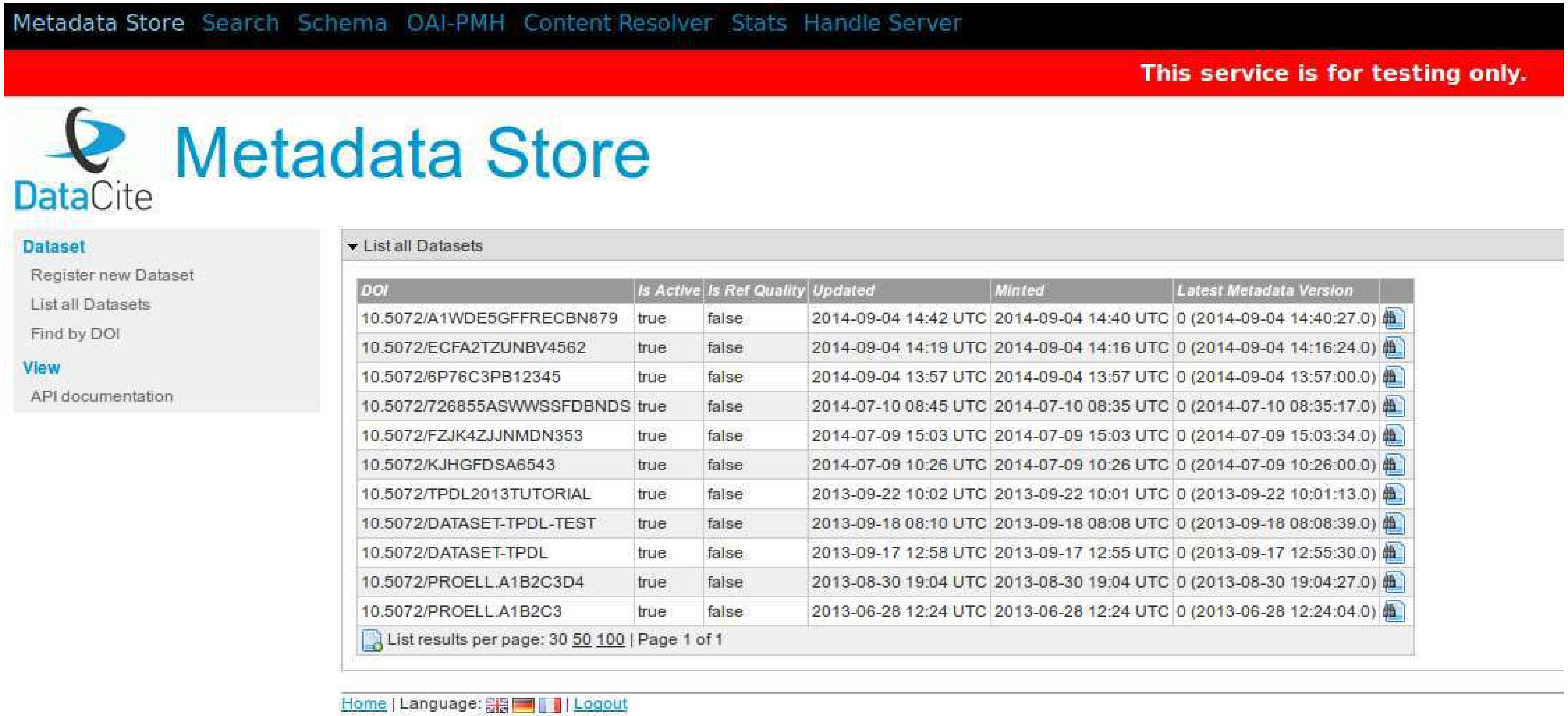
DOI

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



Metadata Store Search Schema OAI-PMH Content Resolver Stats Handle Server

This service is for testing only.

DataCite Metadata Store

Dataset
- Register new Dataset
- List all Datasets
- Find by DOI

View
- API documentation

▾ List all Datasets

DOI	Is Active	Is Ref Quality	Updated	Minted	Latest Metadata Version	
10.5072/A1WDE5GFFRECBN879	true	false	2014-09-04 14:42 UTC	2014-09-04 14:40 UTC	0 (2014-09-04 14:40:27.0)	
10.5072/ECFA2TZUNBV4562	true	false	2014-09-04 14:19 UTC	2014-09-04 14:16 UTC	0 (2014-09-04 14:16:24.0)	
10.5072/6P76C3PB12345	true	false	2014-09-04 13:57 UTC	2014-09-04 13:57 UTC	0 (2014-09-04 13:57:00.0)	
10.5072/726855ASWWSSFDBNDS	true	false	2014-07-10 08:45 UTC	2014-07-10 08:35 UTC	0 (2014-07-10 08:35:17.0)	
10.5072/FZJK4ZJJNMDN353	true	false	2014-07-09 15:03 UTC	2014-07-09 15:03 UTC	0 (2014-07-09 15:03:34.0)	
10.5072/KJHGFDSA6543	true	false	2014-07-09 10:26 UTC	2014-07-09 10:26 UTC	0 (2014-07-09 10:26:00.0)	
10.5072/TPDL2013TUTORIAL	true	false	2013-09-22 10:02 UTC	2013-09-22 10:01 UTC	0 (2013-09-22 10:01:13.0)	
10.5072/DATASET-TPDL-TEST	true	false	2013-09-18 08:10 UTC	2013-09-18 08:08 UTC	0 (2013-09-18 08:08:39.0)	
10.5072/DATASET-TPDL	true	false	2013-09-17 12:58 UTC	2013-09-17 12:55 UTC	0 (2013-09-17 12:55:30.0)	
10.5072/PROELL.A1B2C3D4	true	false	2013-08-30 19:04 UTC	2013-08-30 19:04 UTC	0 (2013-08-30 19:04:27.0)	
10.5072/PROELL.A1B2C3	true	false	2013-06-28 12:24 UTC	2013-06-28 12:24 UTC	0 (2013-06-28 12:24:04.0)	

List results per page: 30 50 100 | Page 1 of 1

Home | Language: | Logout

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

DOI

DataCite

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** (remains even if object is no longer available)
  - 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

Name Mapping Authority

Label

Name

Qualifier

Name Assigning Authority Number (NAAN)

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

ARK - Scheme

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

```
          _____     slips on    _____
       .-'    ,_,'-..  ---->  .-'      '-.
     /     (o,o)  \\          /           \
    :      {`"'}  ||        :              `____
   /  .-.  -"-"-  ||        /  .-.               '--^.   .^--^.          .^.
  {  (   )        ||       {  (   )                   `-'      `-^--^-'   '--^.
   \  `-'    _o   ||        \  '-'         ===================================}
    :      _|<,_  ||         :             ___________________________________/
     \   (*)/(*) /            \           /
      `-.______.-'             `-.______.-'
|....................|...............|....|.........................|..|
         ^                     ^        ^                 ^            ^
         :                     :        :                 :            :
       Cover=                Bow=    Shoulder  .------ Blade          Tip
        NMA              Scheme+NAAN    :      :  .-------------------'
         :                  :    :      :      :  :
         v                  v    v      v      v  v
|.........................|....+.....|...|......|.|
 http://OwlBike.example.org/ark:/13030/tqb3kh97gh8w   <----  Example Key
                            doi:10.30/tqb3kh97gh8w           with parallel
                            hdl:13030/tqb3kh97gh8w          parts in other
                            urn:13030:tqb3kh97gh8w            id schemes.
|.........................|.......................|....
   Name Mapping Authority        Base identifier      ...
```

UC3

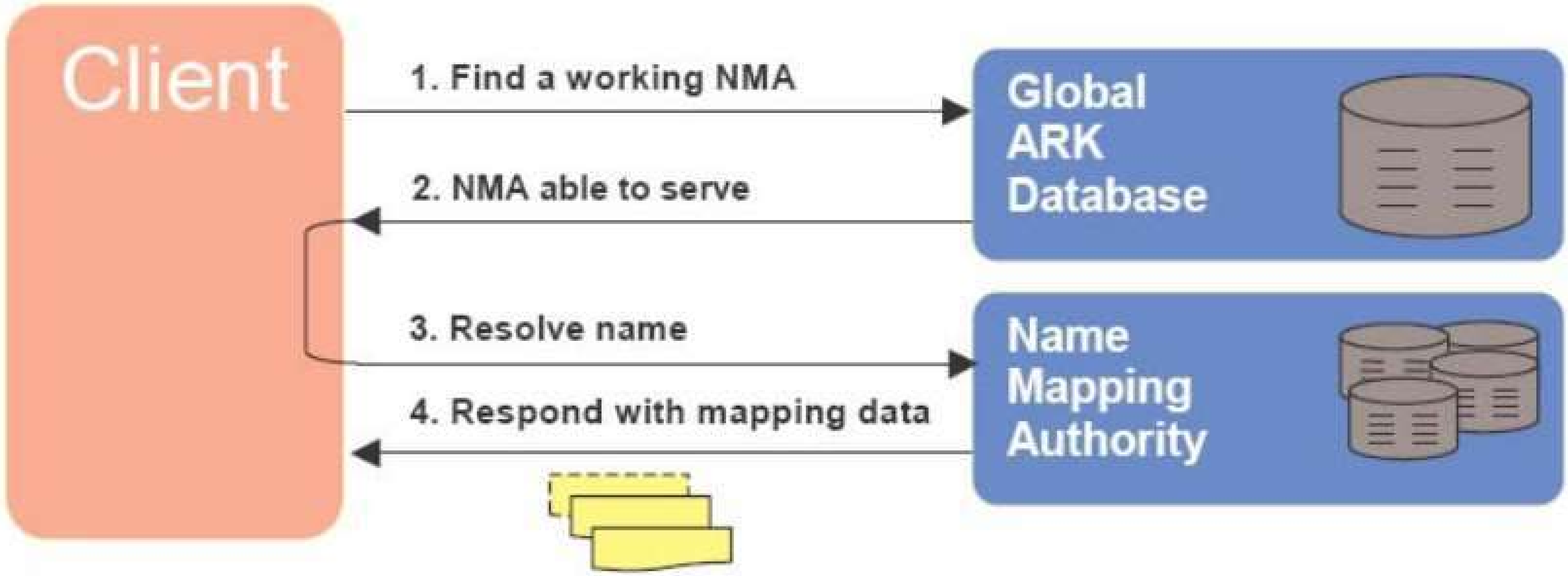
University of California
CDL
California Digital Library

ifs FACULTY OF !NFORMATICS

ARK

- Currently there are more than 600 NAANs
  - Universities
  - Libraries (e.g. Austrian National Library)
  - Companies (e.g. Google)
  - Organizations (e.g. IEEE)
  - https://n2t.net/e/pub/naan_table.html
  - 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



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

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 (common names, name changes)
  - Link research activities and output
  - ORCID Austria: https://www.tuwien.at/kooperationen/orcid/en/home
  - Examples: (increasing order of detailed information provided)
    - https://orcid.org/0000-0002-9272-6225
    - https://orcid.org/0000-0002-4929-7875
- Other Person ID systems
  - Social Security Number ☺
  - Web of Science ResearcherID
  - Scopus Author ID

Persistent Identifiers @ TU Wien

- DOI Service Austria, ORCID-Austria

DOI Service Austria, ORCID Austria

To improve the visibility of Austrian researchers and their academic performance, TU Wien Bibliothek is leading two national initiatives: the DOI Service Austria and ORCID Austria. Not only should this raise awareness of the significance of persistent identifiers (PIDs) in academic communication, it should also create a community of practice.

DOI Service Austria

Since January 2020, with the DOI Service Austria, TU Wien Bibliothek has been providing all Austrian universities, research institutions and other non-profit organisations in the research and education sector domiciled in Austria with an attractive opportunity to register and use Digital Object Identifiers (DOIs) to ensure stable retrieval of academic output via the internet.

For the first time, there is a central point of contact in Austria providing advice on the subject of DOIs, organising international developments and relaying information promptly to Austrian institutions. In order to be able to provide this service as a local authority, TU Wien Bibliothek is a member of the DataCite Association. DataCite is a DOI provider that focuses specifically on the persistent identification of objects stored in repositories and relies on uniform metadata.

Why use DOIs? DOIs are recognised and used internationally. The use of DOIs for research output published on the internet ensures reliable citations and promotes the visibility and stable findability of the document on the internet. The use of DOIs together with other persistent identifiers, such as ORCID iDs for authors and ROR for institutions, also enables improved, reliable and stable attribution of research output to particular persons, research facilities and institutions.

The DOI Service Austria enables Austrian institutions to use the Fabrica registration platform and the DataCite interfaces (MDS API, REST API): this enables both manual and automatic registration of DOIs. Customers of the DOI Service Austria receive the prefixes from us for the independent DOI assignment in the respective institutional repositories. TU Wien Bibliothek provides technical support as well as support for the quality assurance of metadata in Austrian information systems and the interoperability between IT applications. Fees for the DOI Service Austria are based on the DataCite cost model. Contact us for more details.

Search in CatalogPlus

Any field

Search term *

SEARCH

DOI Service Austria

doi

Contact

pid-services@tuwien.ac.at

ORCID Austria

Home | Consortium | ORCID Resources | Contact | Deutsch

WHAT IS ORCID?

ORCID (Open Researcher and Contributor ID) is an open and non-profit project dedicated to creating and maintaining a registry of persistent digital identifiers for researchers (ORCID iD). The name ORCID applies to both the identification system and the organisation behind it. ORCID iD stands for the unique alpha-numeric identifier itself.

Researchers and all individuals with research-related tasks may receive a free ORCID iD. This identifier uniquely differentiates individual researchers from others and allows them to manage their own contributions and to search for other researchers in the registry.

With over 10 million entries, ORCID has established itself as a recommended (sometimes even required) standard (ISO 27729) for person identification.

For further information on ORCID or to register you own ORCID iD, please visit the ORCID website.

https://www.tuwien.at/en/library/doi-service-austria-orcid-austria/

https://www.tuwien.at/kooperationen/orcid/
13 Institutions in Austria

Persistent Identifiers for Organizations

- GRID: Global Research Identifier Database
  - https://www.grid.ac/institutes
  - grid.5329.d
- ROR: Research Organization Registry
  - https://ror.org
  - 04d836q62
- Ringgold:
  - https://www.ringgold.com
  - RIN 508192

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

1978 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

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

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

Tools and Resources

Buy this Article

Request Permissions

TOC Service:
Email RSS

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

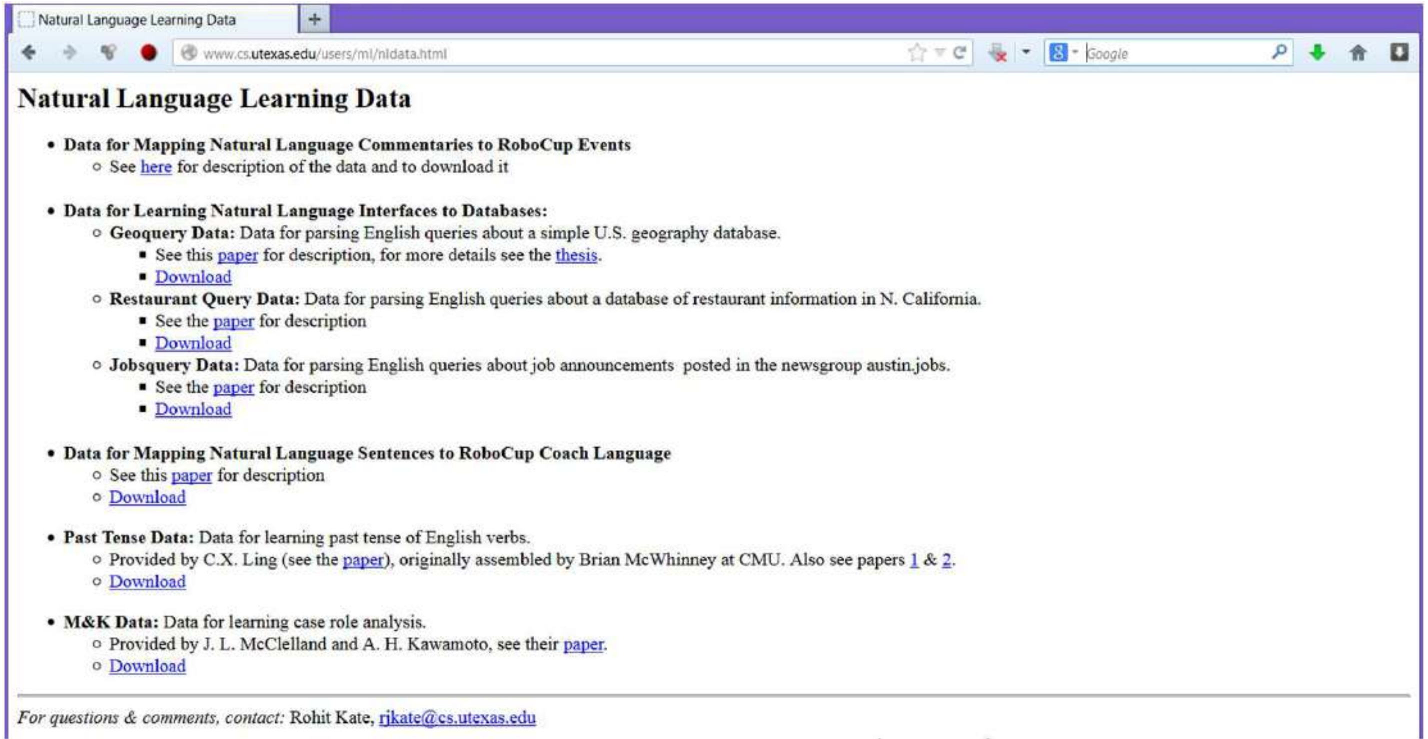
To provide a quantitative analysis of the problem of constructing an NLI we proceeded as follows: we downloaded a dataset which has been frequently used for the evaluation of natural language interfaces, i.e. the Geobase dataset collected by Mooney and his students[1]. The Geobase dataset describes states, 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/



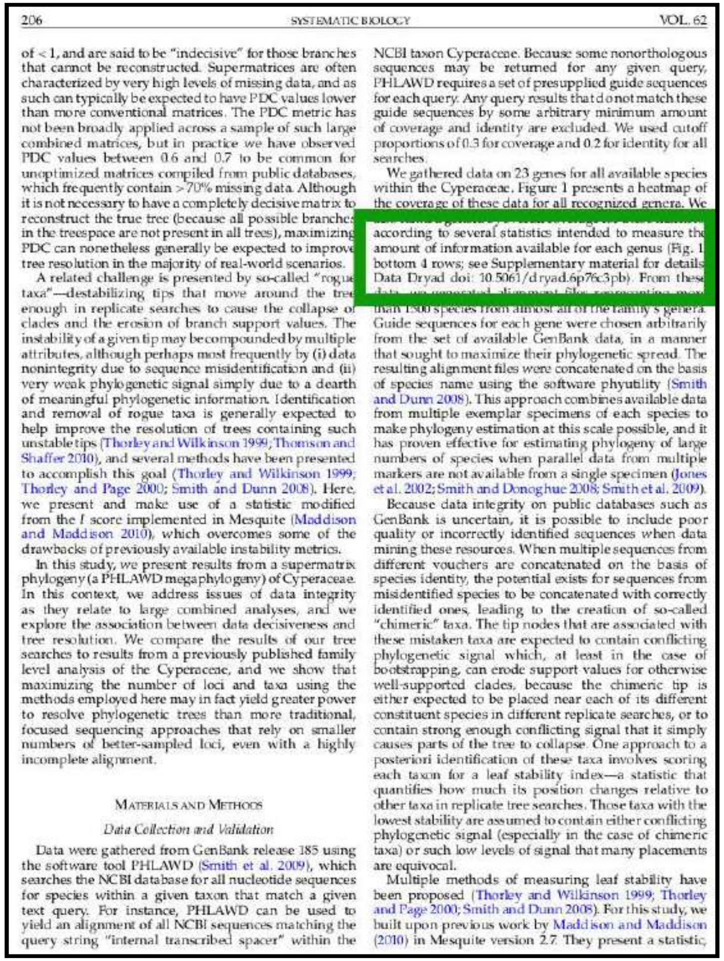
Natural Language Learning Data

- **Data for Mapping Natural Language Commentaries to RoboCup Events**
  - See here for description of the data and to download it
- **Data for Learning Natural Language Interfaces to Databases:**
  - **Geoquery Data:** Data for parsing English queries about a simple U.S. geography database.
    - See this paper for description, for more details see the thesis.
    - Download
  - **Restaurant Query Data:** Data for parsing English queries about a database of restaurant information in N. California.
    - See the paper for description
    - Download
  - **Jobsquery Data:** Data for parsing English queries about job announcements posted in the newsgroup austin.jobs.
    - See the paper for description
    - Download
- **Data for Mapping Natural Language Sentences to RoboCup Coach Language**
  - See this paper for description
  - Download
- **Past Tense Data:** Data for learning past tense of English verbs.
  - Provided by C.X. Ling (see the paper), originally assembled by Brian McWhinney at CMU. Also see papers 1 & 2.
  - Download
- **M&K Data:** Data for learning case role analysis.
  - Provided by J. L. McClelland and A. H. Kawamoto, see their paper.
  - Download

For questions & comments, contact: Rohit Kate, rjkate@cs.utexas.edu

Example: Sharing Platform

- Example: Data sharing platforms



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.

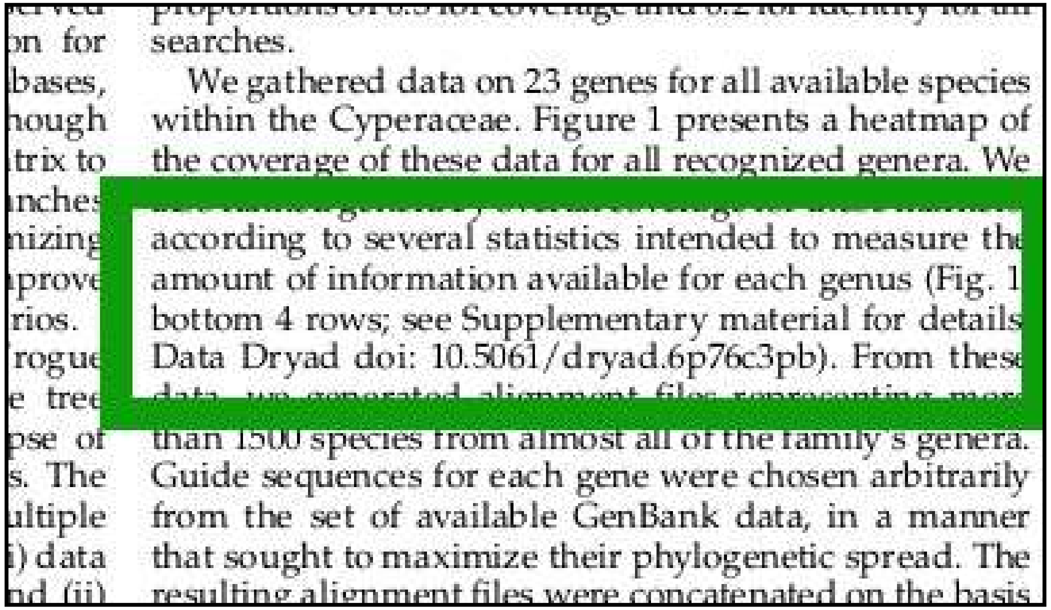
We gathered data on 23 genes for all available species within the Cyperaceae. Figure 1 presents a heatmap of the coverage of these data for all recognized genera. We

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 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.

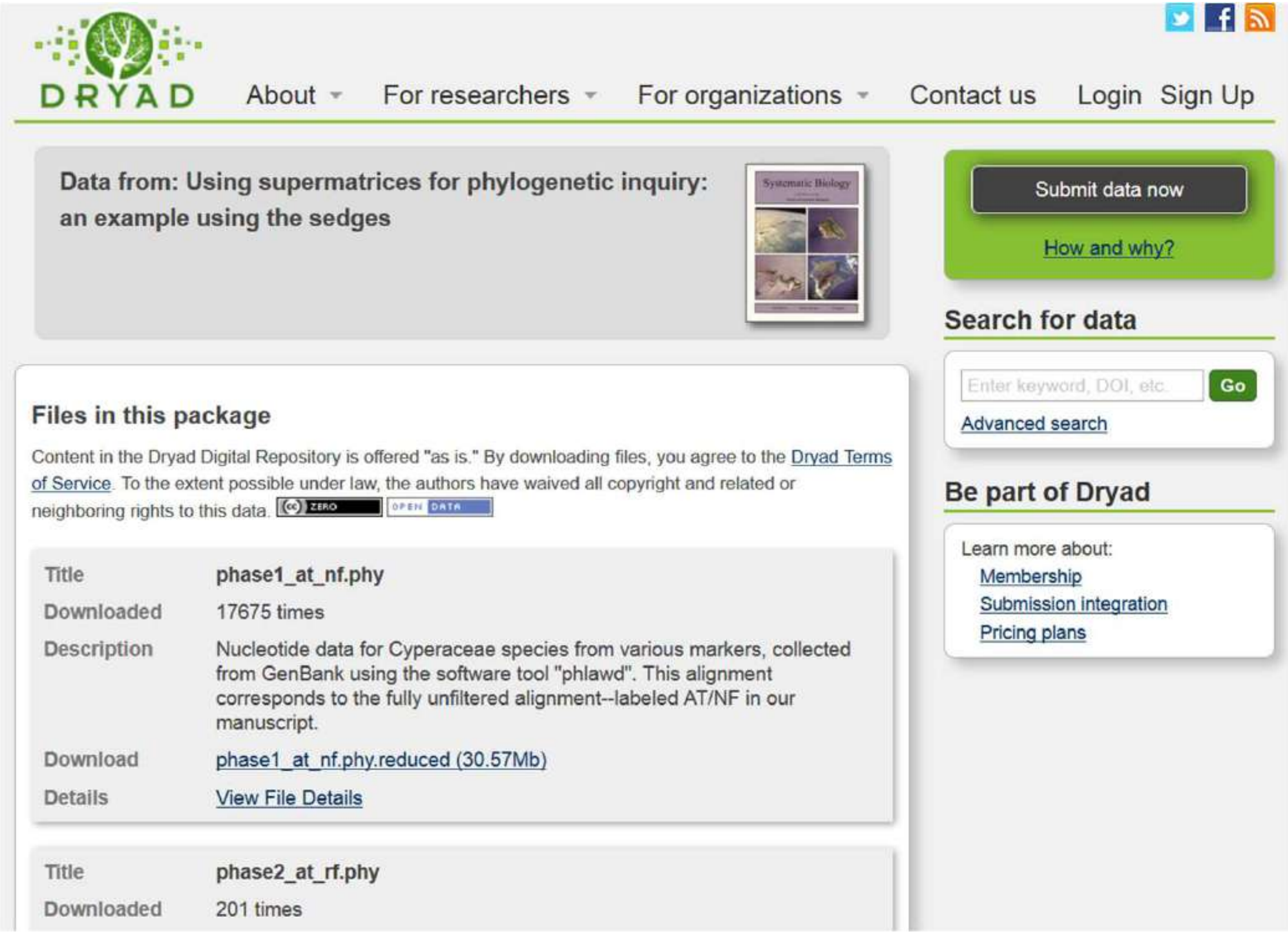
Multiple methods of measuring leaf stability have been proposed (Thorley and Wilkinson 1999; Thorley and Page 2000; Smith and Dunn 2008). For this study, we built upon previous work by Maddison and Maddison (2010) in Mesquite version 2.7. They present a statistic,



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 n 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

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

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

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

ELSEVIER

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.

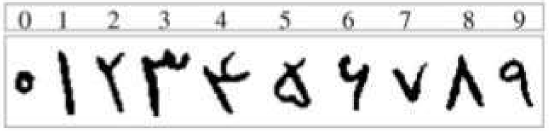


Fig. 1. Sample handwritten Farsi digits.

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

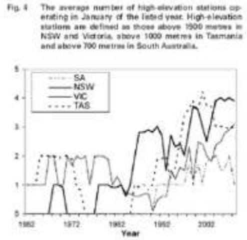
The dataset is available in four separate files, **Total.cdb, Training.cdb, Test.cdb, Remaining.cdb**. The file format is described here with a pseudo code:

```
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,

        }

}
```

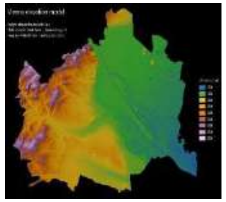
Source codes for reading the dataset files are available in Matlab, C++ and Pascal. To get the dataset please contact kabir@modares.ac.ir. or see the homepage http://www.modares.ac.ir/eng/kabir.

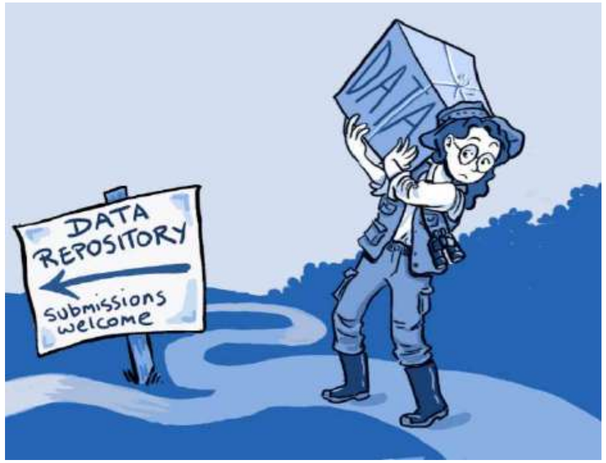
Khosravi, Hossein, and Ehsanollah Kabir. "Introducing a very large dataset of handwritten Farsi digits and a study on their varieties." Pattern Recognition Letters 28.10 (2007): 1133-1141.

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

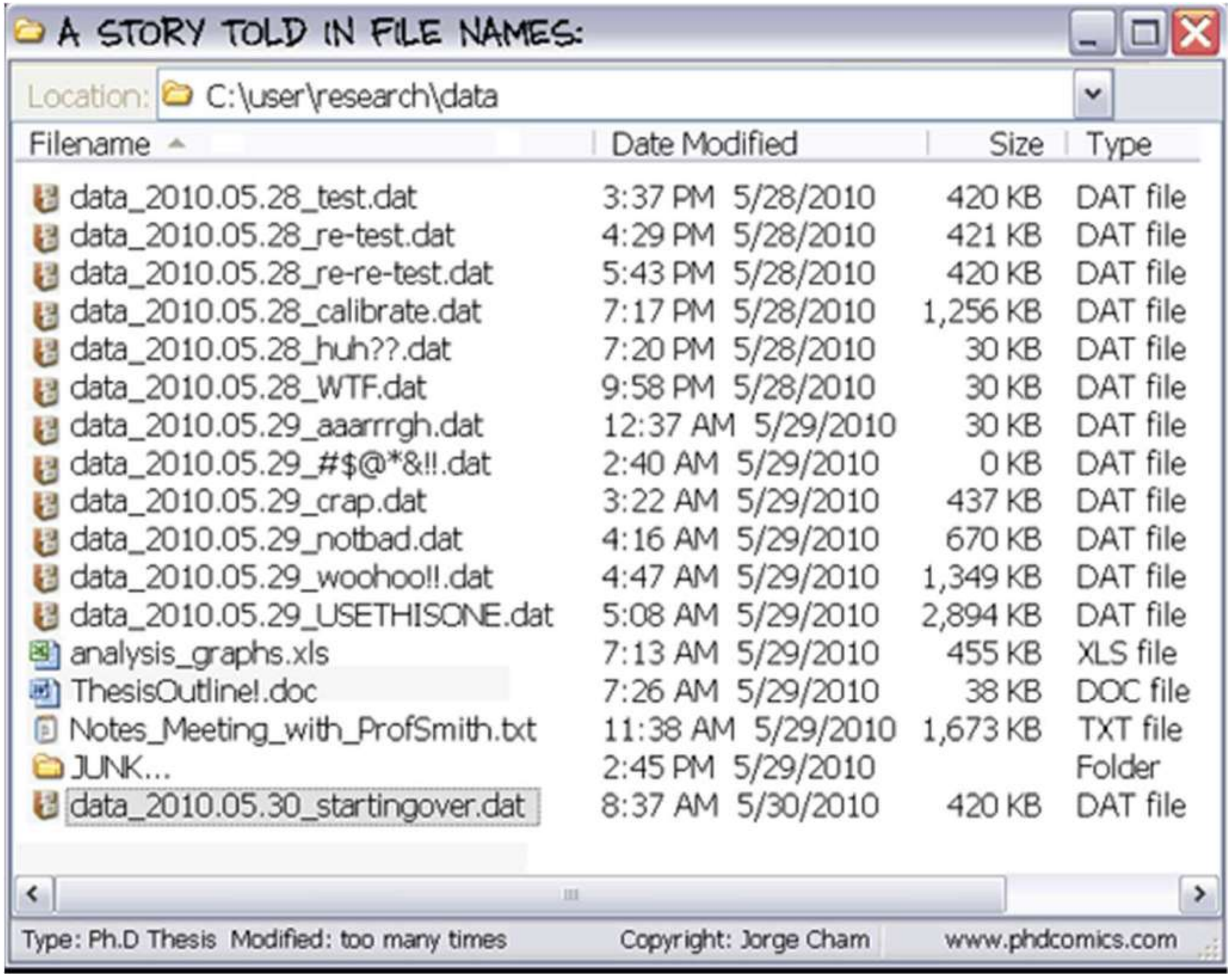
- What about the **granularity** of data to be identified?
  - Enormous amounts of CSV data
  - Researchers use specific subsets of data
  - Need to identify precisely the subset used
- Current approaches
  - Storing a copy of subset as used in study -> scalability
  - Citing entire dataset, providing 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…

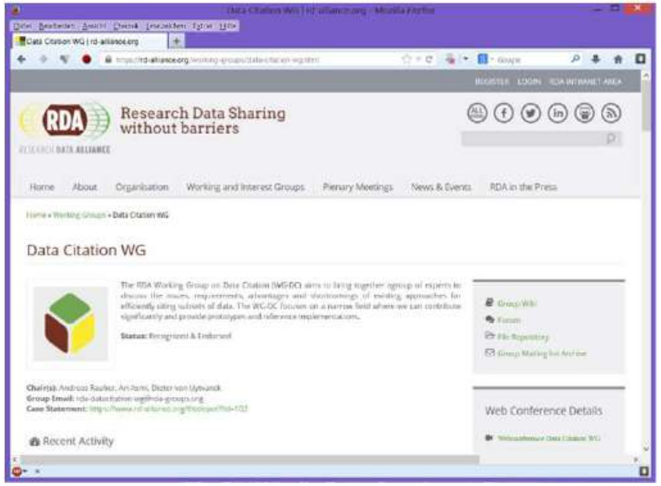
- Common approaches to data management…
  (from PhD Comics: A Story Told in File Names, 28.5.2010)
  Source: http://www.phdcomics.com/comics.php?f=1323



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
  - Recomme
- PID resolves to land g page
  - Provides detailed metadata, link o 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

This is an important advantage over traditional approaches relying on, e.g. storing a list of identifiers/DB dump!!!

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...
  - Recomme... 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...n 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:
- 2-page flyer
  https://rd-alliance.org/recommendations-working-group-data-citation-revision-oct-20-2015.html
- Detailed report: Bulletin of IEEE TCDL 2016
  http://www.ieee-tcdl.org/Bulletin/v12n1/papers/IEEE-TCDL-DC-2016_paper_1.pdf
- Adopter's reports, webinars
  https://www.rd-alliance.org/group/data-citation-wg/webconference/webconference-data-citation-wg.html
- Review / Lessons Learned
  Andreas Rauber et al., Precisely and Persistently Identifying and Citing Arbitrary Subsets of Dynamic Data
- Harvard Data Science Review, 3(4), 2021.
  DOI 10.1162/99608f92.be565013.

Data Citation – Recommendations

Preparing Data & Query Store

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

When Resolving a PID

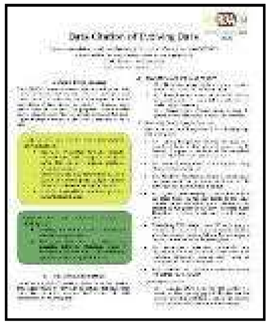
- R11 – Landing Page
- R12 – Machine Actionability

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

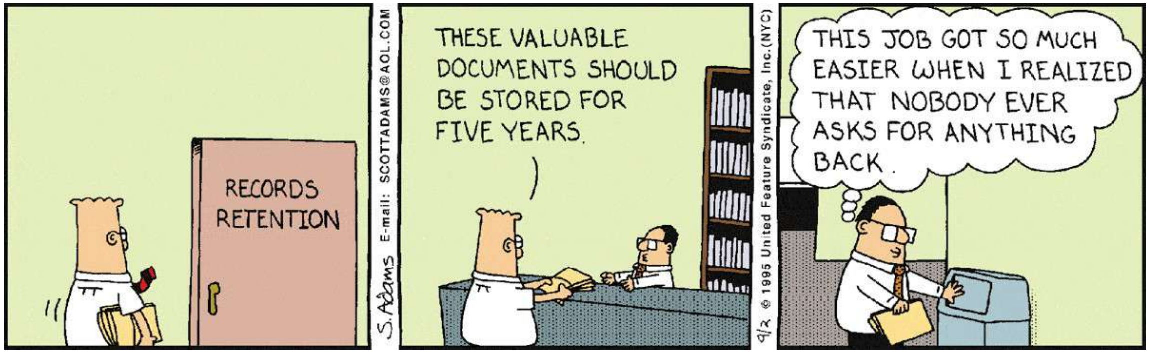
Upon Modifications to the Data Infrastructure

- 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

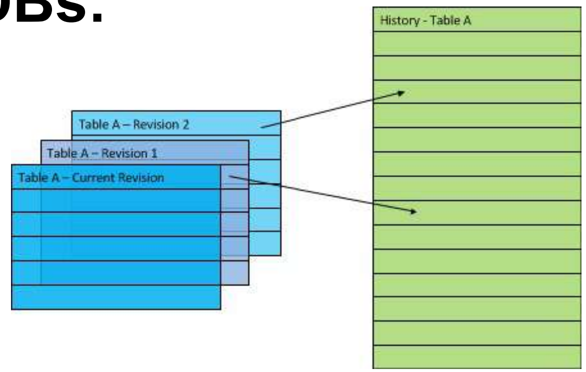
Data Versioning (cont.)

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:



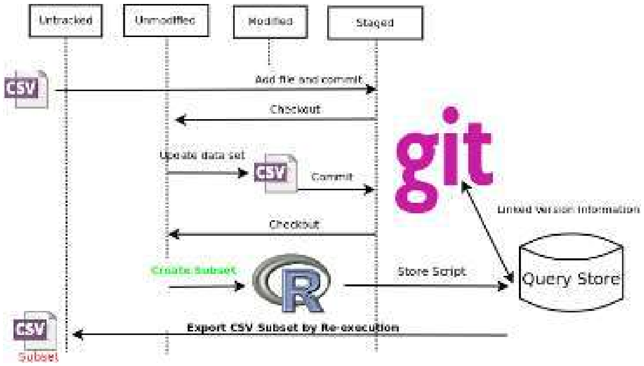
- History Table
  - Utilizes full history table
  - Also inserts reflected in history table
  - Doubles storage space, no API adaptions
- Integrated
  - Extend original tables by temporal metadata
  - Expand primary key by timestamp/version column
  - Minimal storage footprint, changes to all APIs
- 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

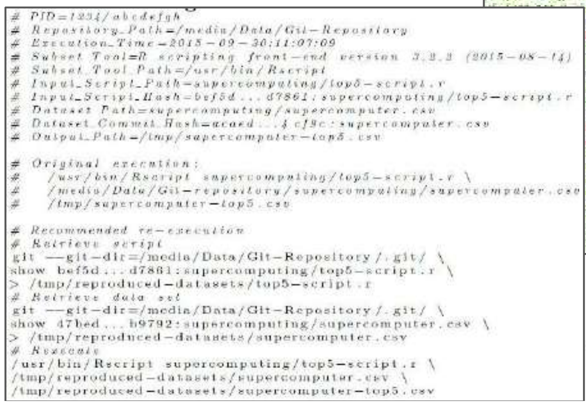
- Solution to be adopted depends on trade-off
  - Storage Demand
  - Query Complexity
  - Software/API adaption

R1 & R2: Git-Based 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



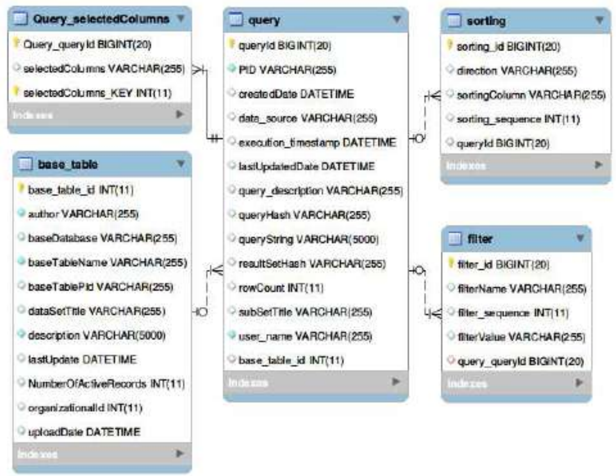


```
# PID=1234/abcdefgh
# Repository_Path=/media/Data/Git-Repository
# Execution_Time=2015-09-30:11:07:09
# Subset Tool=R scripting front-end version 3.2.2 (2015-08-14)
# Subset Tool Path=/usr/bin/Rscript
# Input_Script_Path=supercomputing/top5-script.r
# Input_Script_Hash=bef5d...d7861:supercomputing/top5-script.r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

Slide by Stefan Pröll

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

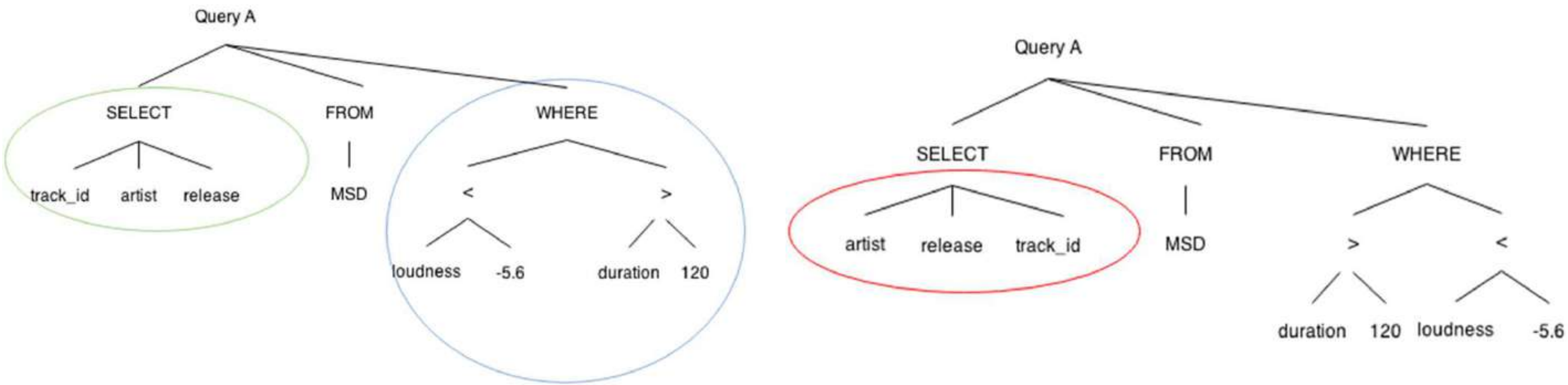
- **Re-write the query to a normalized form so that identical queries can be detected.**
  Compute checksum of the normalized query to efficiently detect identical queries
- Detecting identical queries can be challenging
  - Query semantics can be expressed in different ways
  - Different queries can 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

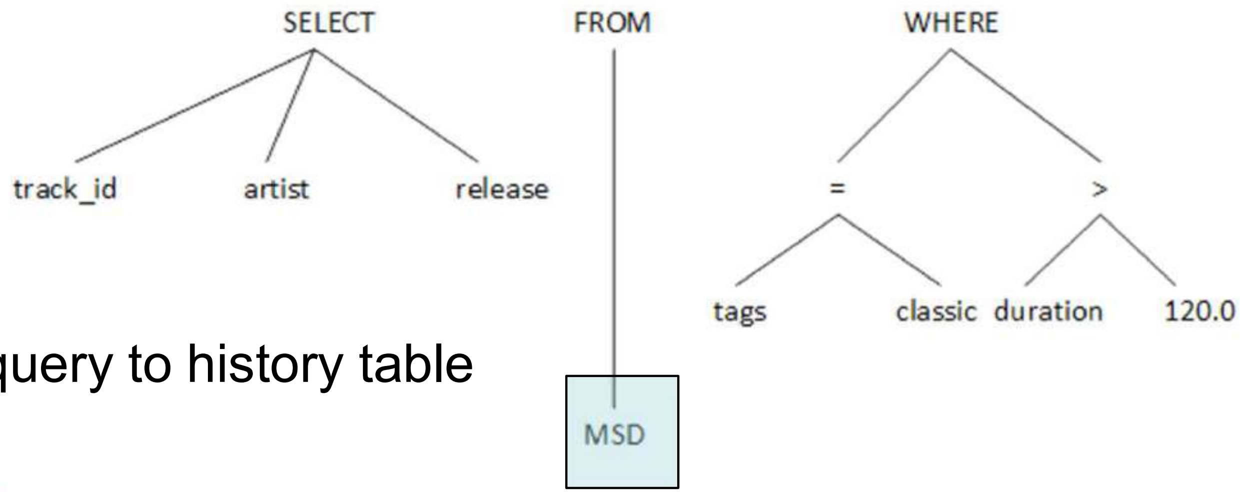
- Query re-writing needed to
  - **Standardization/Normalization** of query to help with identifying semantically identical queries
    - upper/lower case spelling, sorting of filter parameters, …
  - Re-write to **adapt to versioning approach** chosen (versioning in operational tables, separate history 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

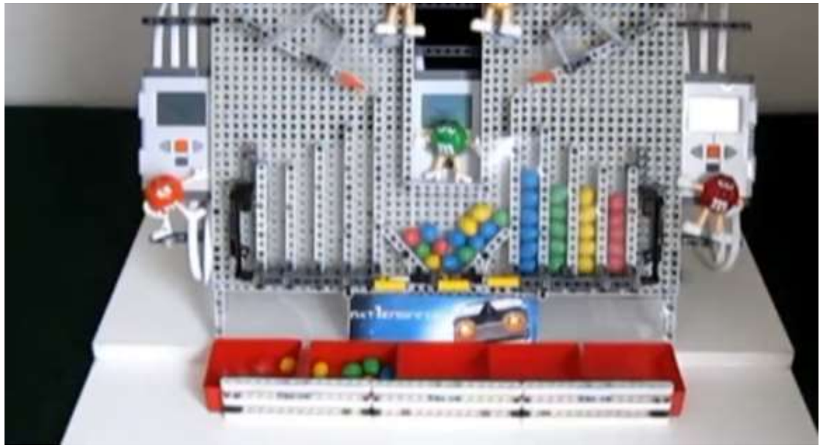


- Adapt query to history table

```
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!



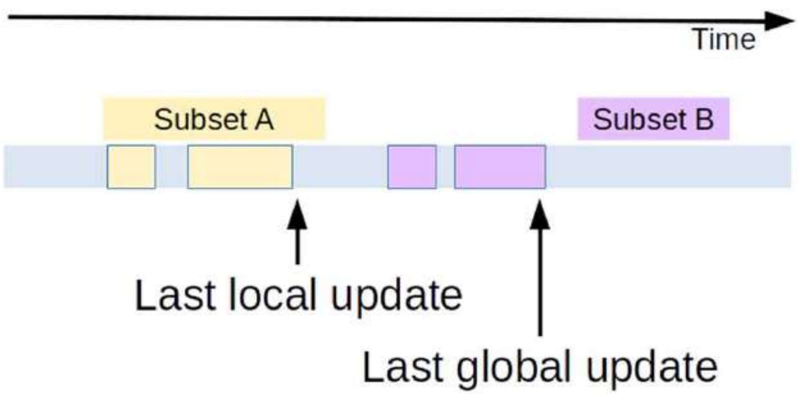
http://www.geek.com/

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

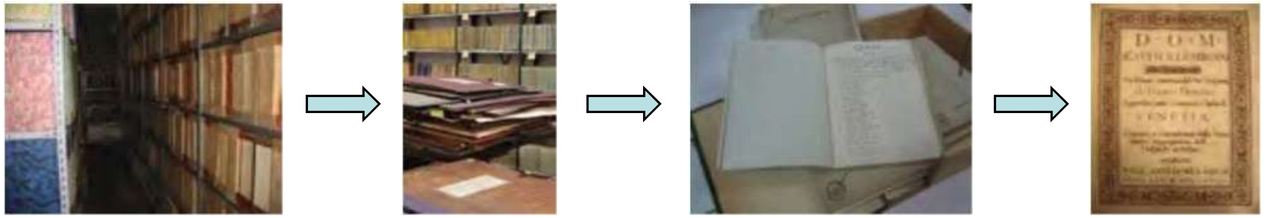
Suggested citation text:	Stefan Proell (2015) "Austria Facts" created at 2015-10-07 10:51:55.0, PID [ark:12345/qmZi2wO2vv]. Subset of CIA: "The CIA WorldFactbook", PID [ark:12345/cLfH9FjxnA]

R10: Automated Citation Texts

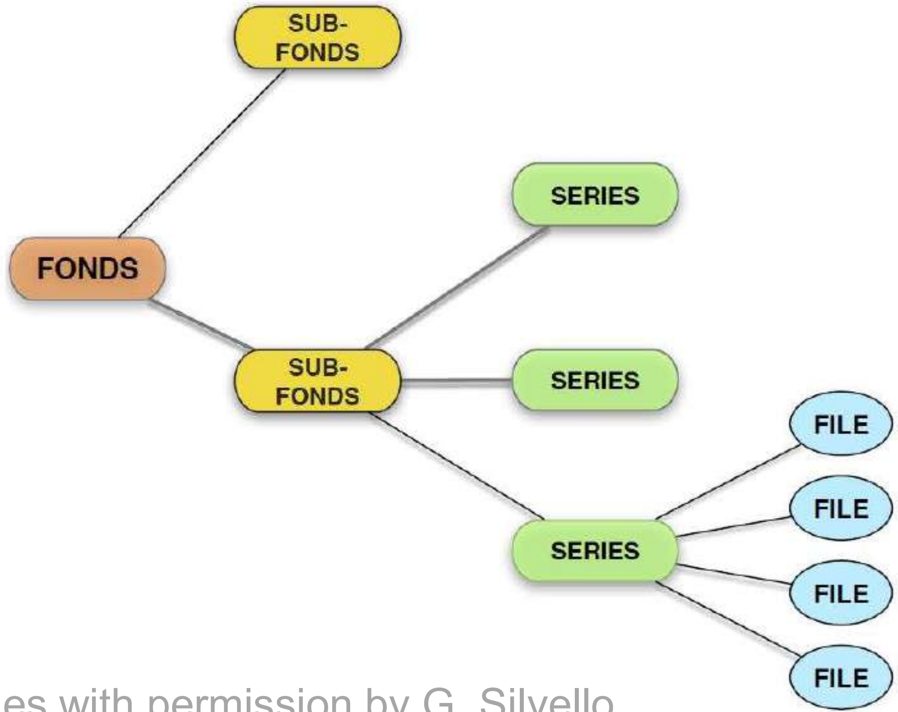
- Can be created automatically
  - relatively simple for relational
  - more complex for hierarchical/XML
- Learning to Cite:
  - Gianmaria Silvello. Learning to Cite Framework: How to Automatically Construct Citations for Hierarchical Data. Journal of the Association for Information Science and Technology (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>
```



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
```

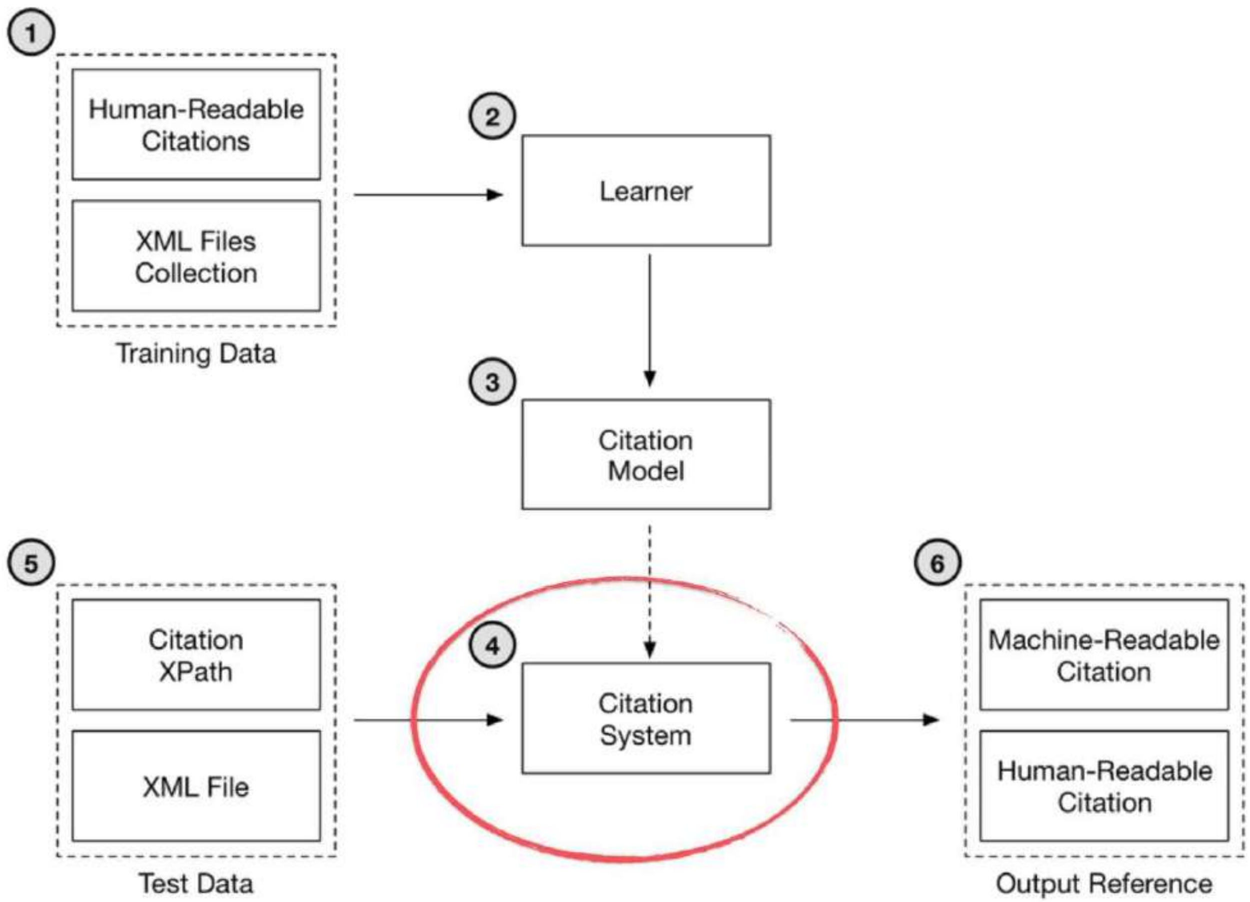
R10: Automated Citation Texts

- 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

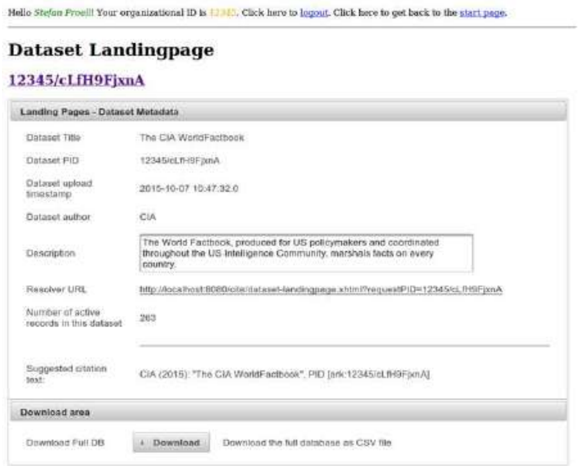
R10: Automated Citation Texts

- Learning citation models



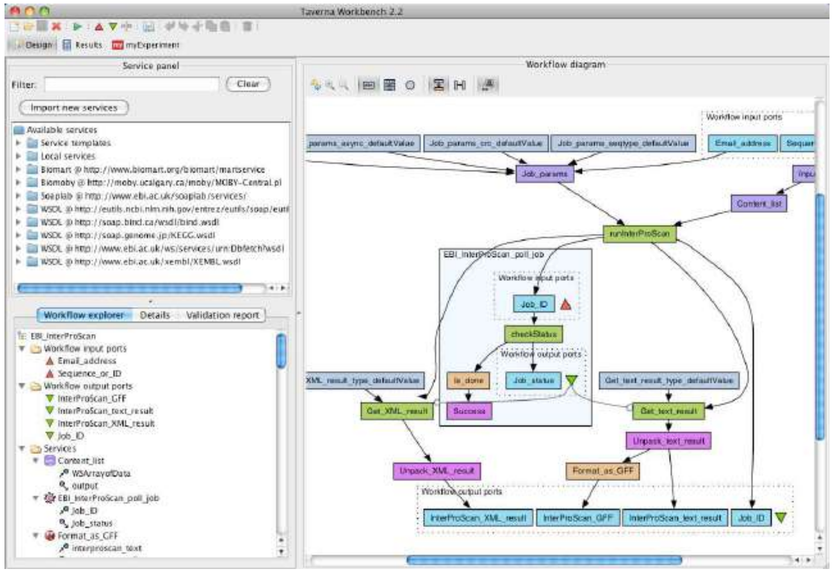
R11: Landing Page

- **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

Large Number of Adoptions

- **Standards / Reference Guidelines / Specifications:**
  - Joint Declaration of Data Citation Principles:
    Principle 7: Specificity and Verifiability (https://www.force11.org/datacitation)
  - ESIP:Data Citation Guidelines for Earth Science Data Vers. 2 (P14)
  - ISO 690, Information and documentation - Guidelines for bibliographic references and citations to information resources (P13)
  - EC ICT TS5 Technical Specification (pending) (P12)
  - DataCite Considerations (P8)
- **Reference Implementations**
  - MySQL/Postgres (P5, P6)
  - CSV files: MySQL, Git (P5, P6, P8, Webinar)
  - XML (P5)
  - CKAN Data Repository (P13)
  - RDF/SPARQL (P17)

Large Number of Adoptions

- **Pilot implementations, Use cases**
  - DEXHELPP: Social Security Records (P6)
  - NERC: ARGO Global Array (P6)
  - LNEC: River dam monitoring (P5)
  - CLARIN: Linguistic resources, XML (P5)
  - MSD: Million Song Database (P5)
  - many further individual ones discussed …

Large Number of Adoptions

- **Adoptions deployed**
  - CBMI: Center for Biomedical Informatics, WUSTL (P8, Webinar)
  - VMC: Vermont Monitoring Cooperative (P8, Webinar)
  - CCCA: Climate Change Center Austria (P10/P11/P12, Webinar)
  - EODC: Earth Observation Data Center (P14, Webinar)
  - VAMDC: Virtual Atomic and Molecular Data Center (P8/P10/P12, Webinar)
  - Ocean Networks Canada (P12, Webinar)
- **In progress**
  - NICT Smart Data Platform (P10/P14)
  - Dendro System (P13)
  - Deep Carbon Observatory (P12)

WGDC Webinar Series

- https://www.rd-alliance.org/group/data-citation-wg/webconference/webconference-data-citation-wg.html
  - Implementation of the RDA Data Citation Recommendations by **Ocean Networks Canada (ONC)**
  - Implementation of the RDA Data Citation Recommendations the **Earth Observation Data Center (EODC) for the openEO platformby**
  - **Automatically generating citation text from queries for RDBMS and XML data sources**
  - 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**

RESEARCH DATA ALLIANCE

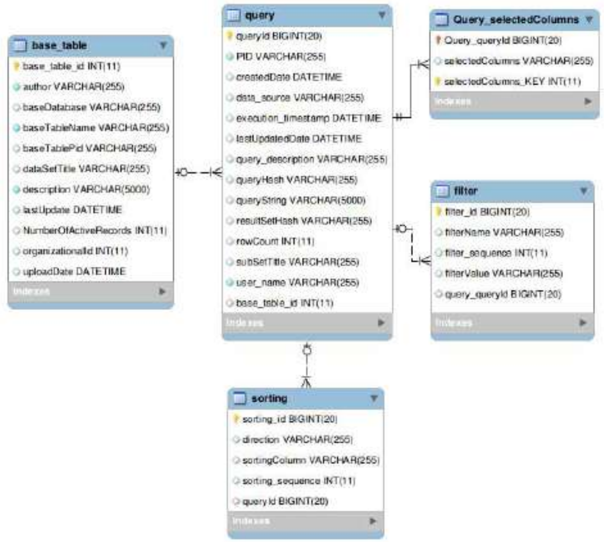
Reference Implementation for CSV Data (and SQL)

Stefan Pröll, SBA
Christoph Meixner, TU Wien

research data sharing without barriers
rd-alliance.org

Large Scale Research Settings

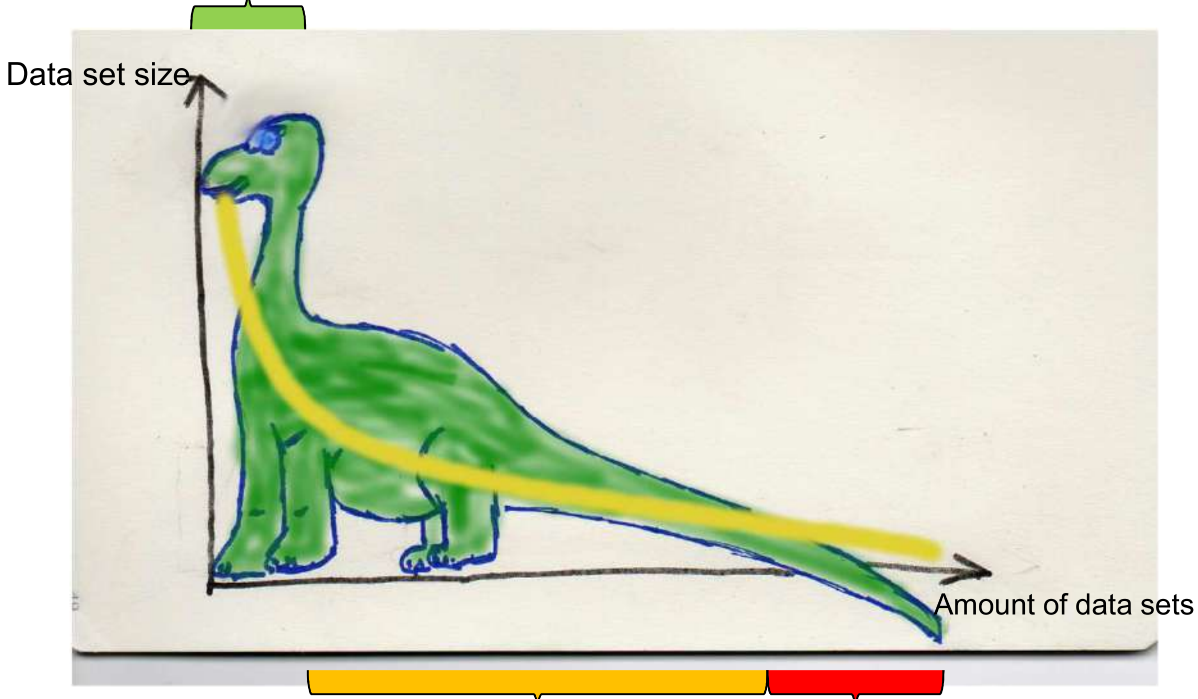
- RDA recommendations implemented in data infrastructures
- Required adaptions
  - Introduce versioning, if not already in place
  - Capture sub-setting process (queries)
  - Implement dedicated query store to store queries
  - A bit of additional functionality (query re-writing, hash functions, …)



- Done! ?
  - “Big data”, database driven
  - Well-defined interfaces
  - Trained experts available
  - “*Complex, only for professional research infrastructures*” ?

Long Tail Research Data

Big data,
well organized,
often used and cited



Less well organized,
non-standardised
no dedicated infrastructure

"Dark data"

Prototype Implementations

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

2 Reference implementations :

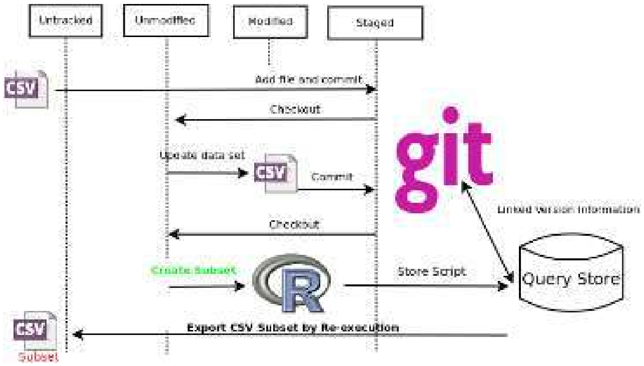
- **Git** based Prototypes: widely used versioning system
  - A) Using separate folders
  - B) Using branches
- **MySQL** based Prototype:
  - C) Migrates CSV data into relational database

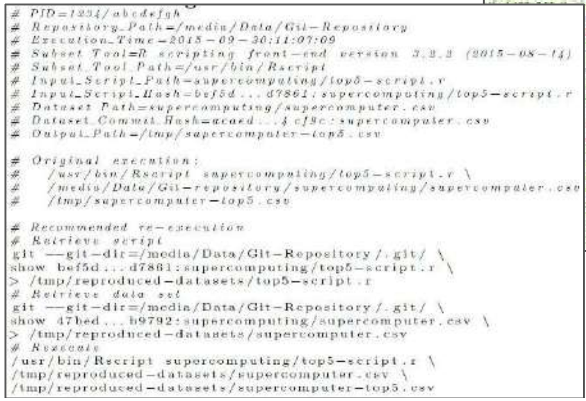
- Data backend responsible for versioning data sets
- Subsets are created with scripts or queries
  via API or Web Interface
- Transparent to user: always CSV

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



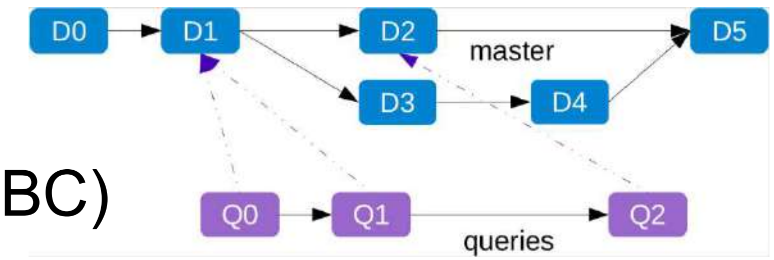


```
# PID=1234/abcdefgh
# Repository_Path=/media/Data/Git-Repository
# Execution_Time=2015-09-30:11:07:09
# Subset_Tool=R scripting front-end version 3.2.2 (2015-08-14)
# Subset_Tool_Path=/usr/bin/Rscript
# Input_Script_Path=supercomputing/top5-script.r
# Input_Script_Hash=bef5d...d7861:supercomputing/top5-script.r
# 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
# Reexecute
/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

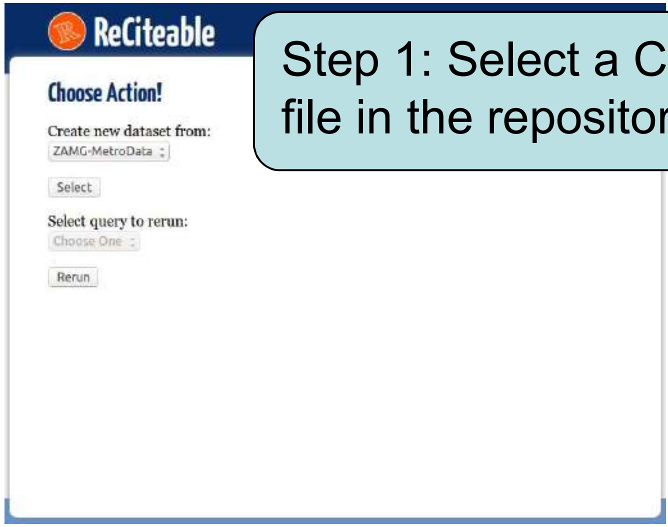
- Addresses issues
  - common commit history, branching data
- Using Git branching model:
  Orphaned branches for queries and data
  - Keeps commit history clean
  - Allows merging of data files



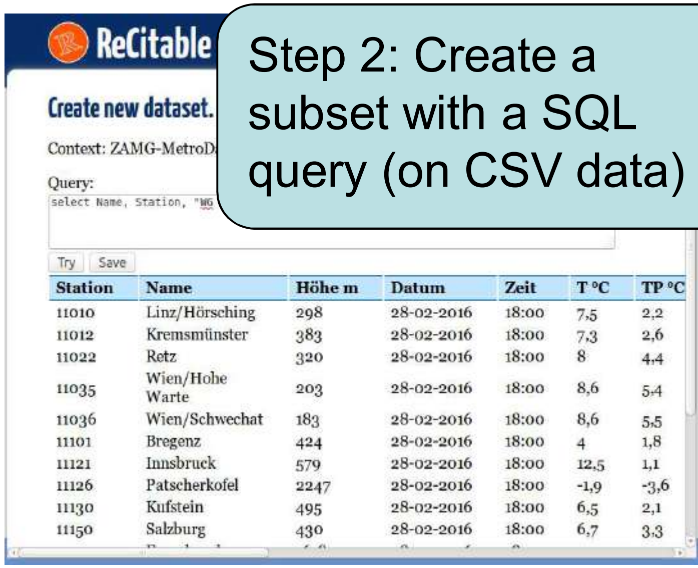
- Web interface for queries (CSV2JDBC)
- Use commit hash for identification
  - Assigned PID hashed with SHA1
  - Use hash of PID as filename (ensure permissible characters)

Git-Based Reference Implementation

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

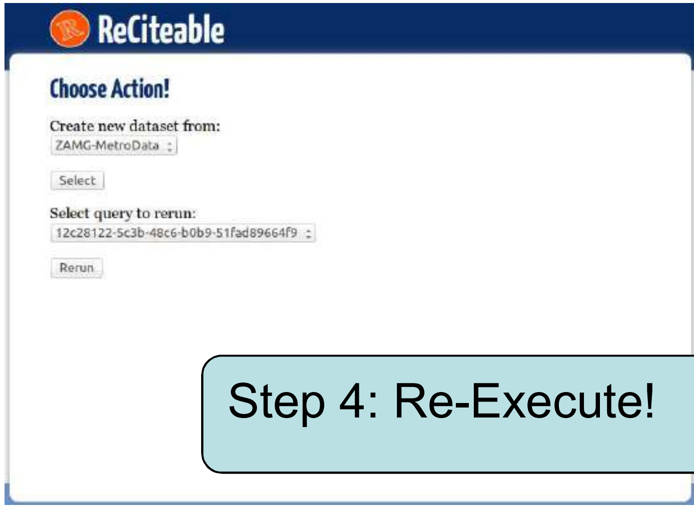


Step 1: Select a CSV file in the repository



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

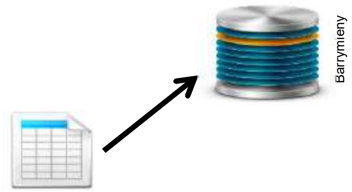
Step 3: Store the query script and metadata



Step 4: Re-Execute!

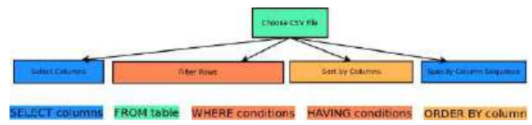
Slide by Stefan Pröll

MySQL-Based Reference Implementation



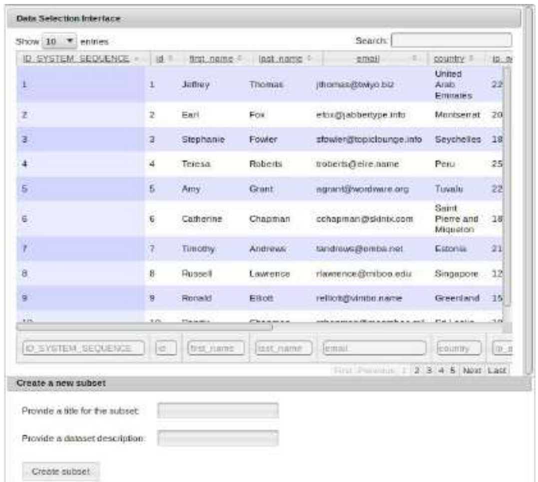
MySQL Prototype

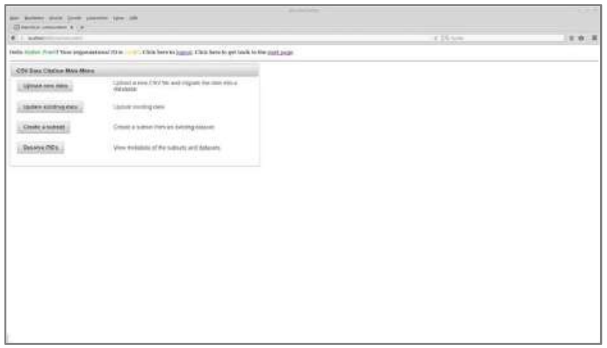
- Data upload
  - User uploads a CSV file into the system
- Data migration from CSV file into RDBMS
  - Generate table structure
  - Add metadata columns (versioning)
  - Add indices (performance)
- Dynamic data
  - Insert, update and delete records
  - Events are recorded with a timestamp
- Subset creation
  - 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

RESEARCH DATA ALLIANCE

WG Data Citation Pilot
CBMI @ WUSTL

Cynthia Hudson Vitale, Leslie McIntosh, Snehil Gupta

Washington University in St.Luis

research data sharing without barriers

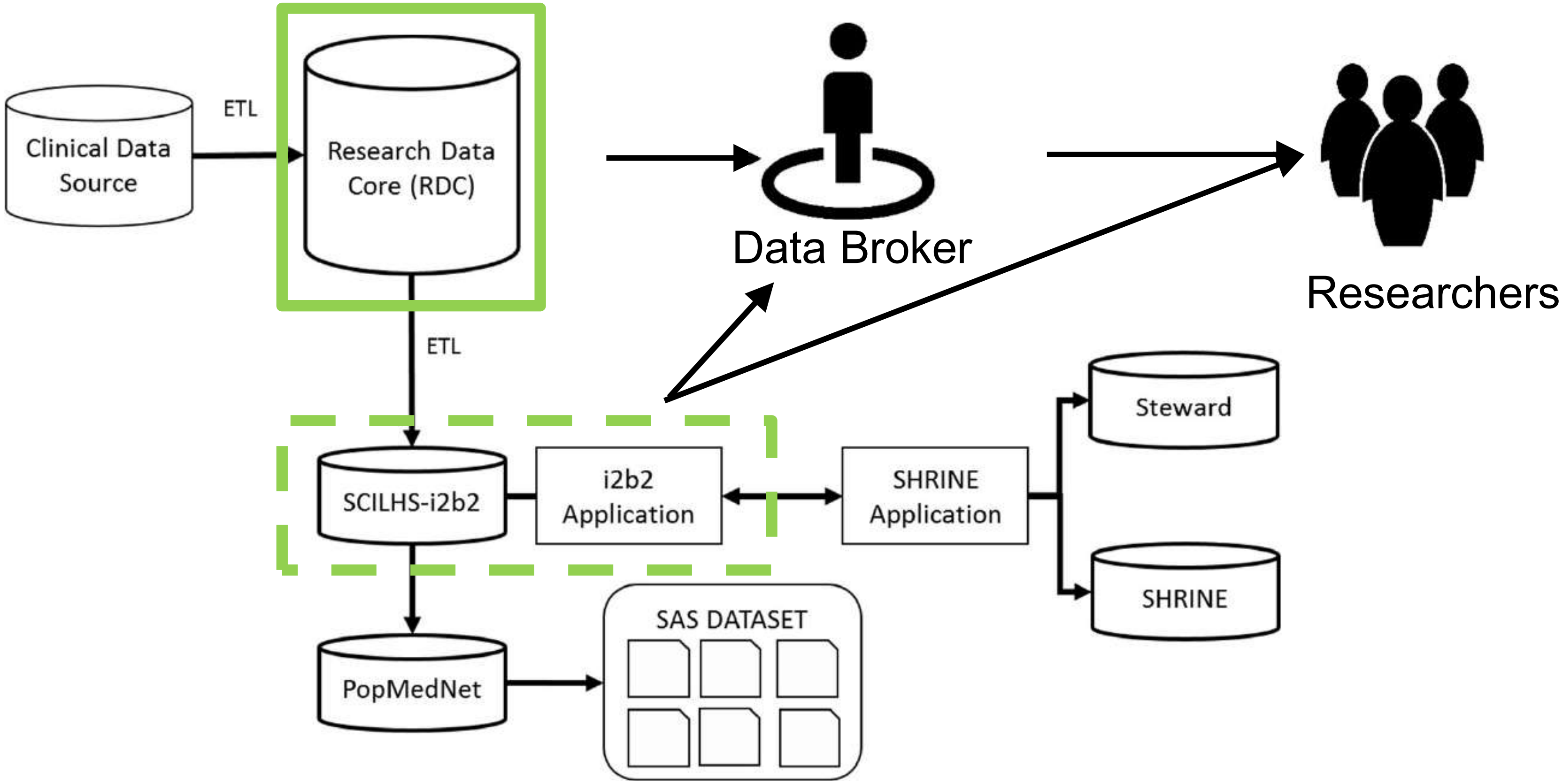
rd-alliance.org

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

RDA-MacArthur Grant Focus



Slide by Leslie McIntosh

R1 and R2 Implementation

1

PostgreSQL Extension
“temporal_tables”

2

RDC.table

c1	c2	c3

+

sys_period

triggers

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**

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

RESEARCH DATA ALLIANCE

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

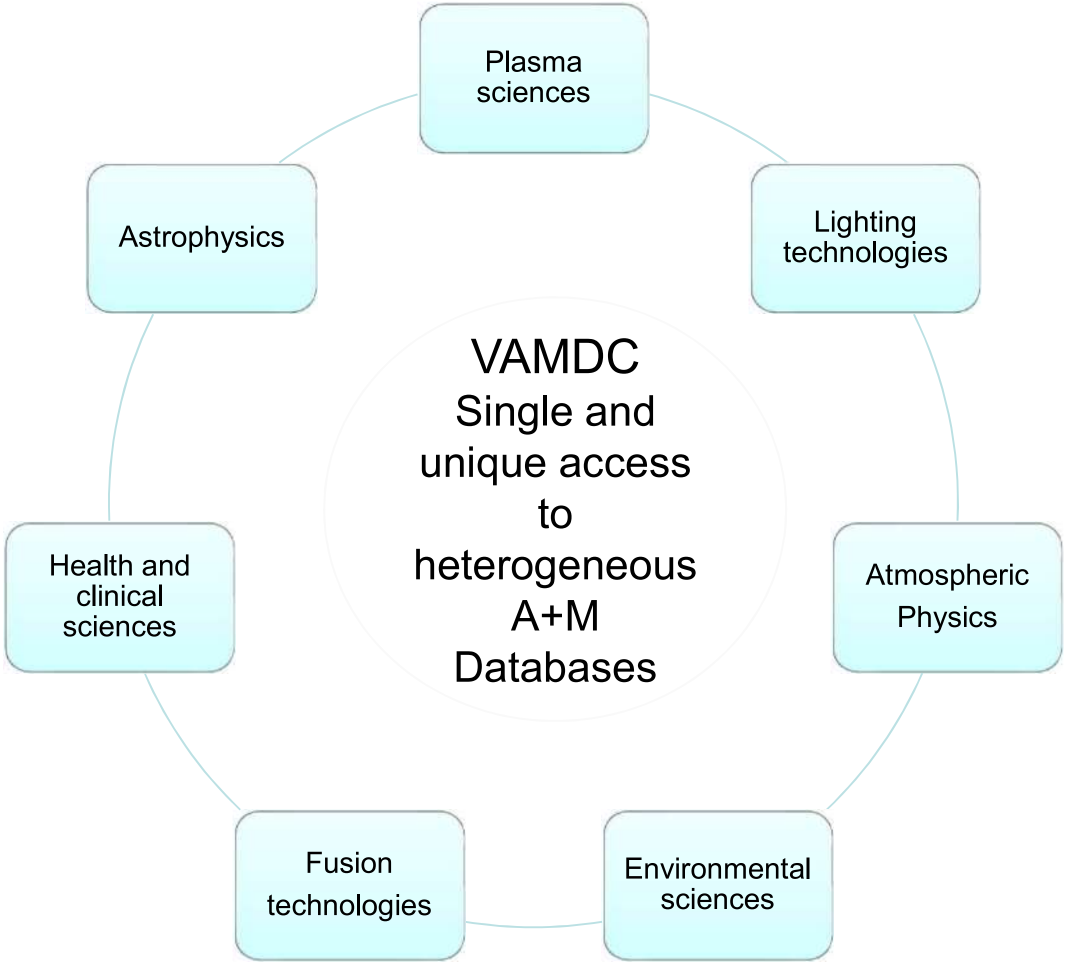
C.M. Zwölf and VAMDC Consortium

carlo-maria.zwolf@obspm.fr

research data sharing without barriers

rd-alliance.org

The Virtual Atomic and Molecular Data Centre

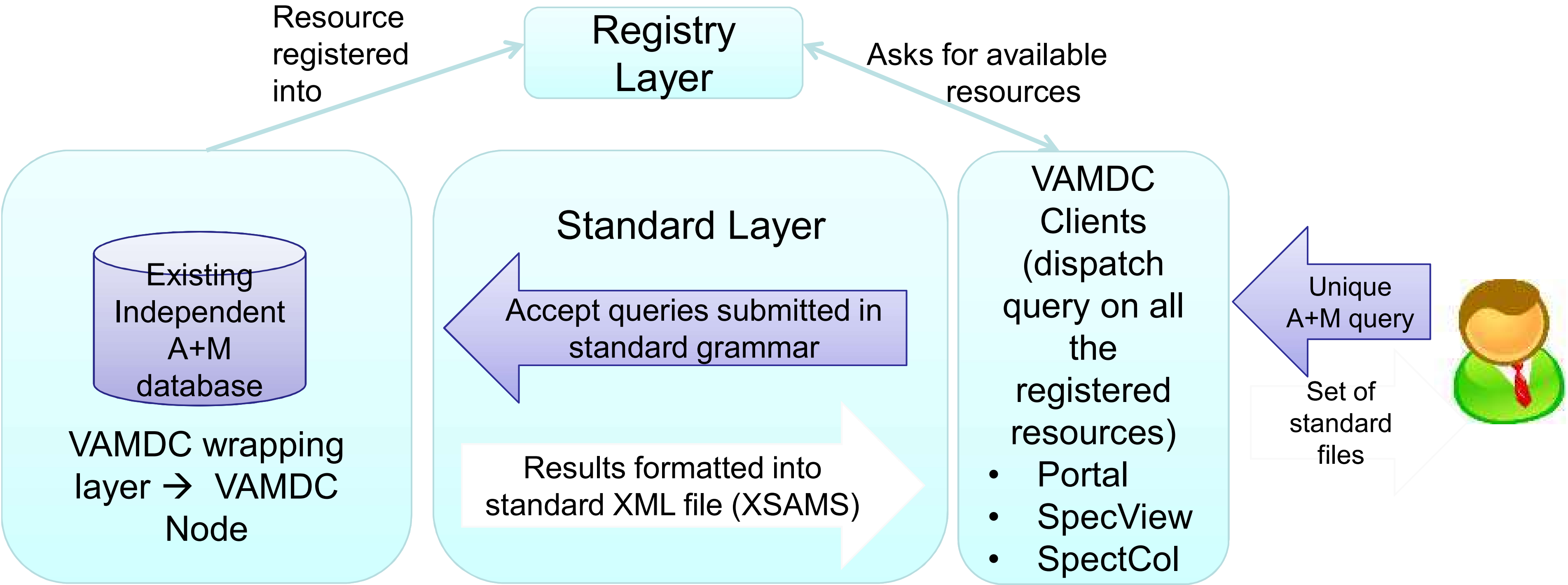


- Federates 29 heterogeneous databases http://portal.vamdc.org/
- The "V" of VAMDC stands for Virtual in the sense that the e-infrastructure does not contain data. The infrastructure is a wrapping for exposing in a unified way a set of heterogeneous databases.
- The consortium is politically organized 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

IfS FACULTY OF !NFORMATICS

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

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

Query Store

Is built over the versioning of Data

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

Slide by Carlo Maria Zwölf

Data-Versioning: Overview of the fine grained mechanisms

This approach has several advantages:

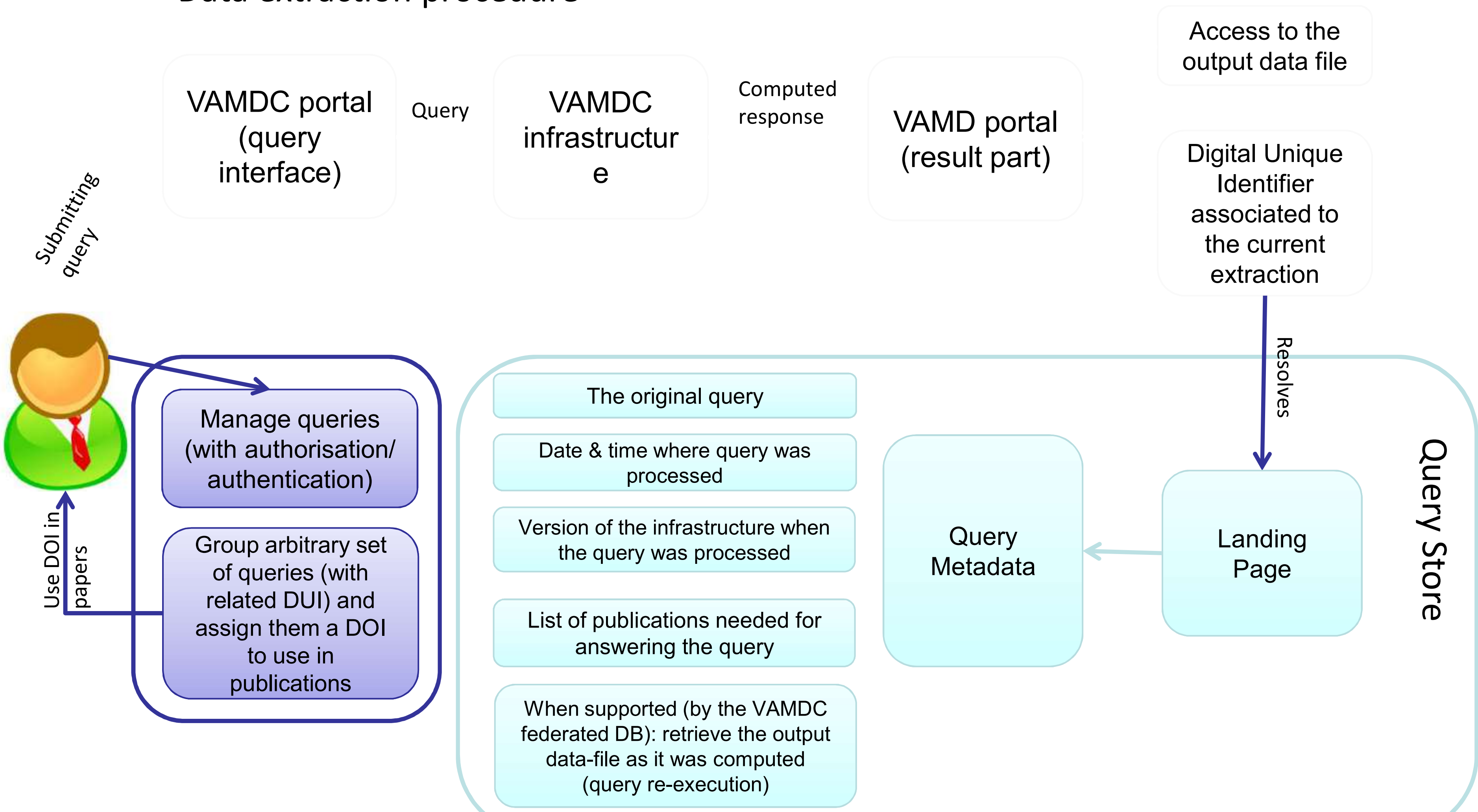
- It solves the data tagging granularity problem
- It is independent from what is considered a dataset
- The new files are compliant with old libraries & processing programs
    - We add a new feature, an overlay to the existing structure
    - We induce a structuration, without changing the structure

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:

Data extraction procedure



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)

RESEARCH DATA ALLIANCE

Climate Change Centre Austria (CCCA)

Chris Schubert

chris.Schubert@ccca.ac.at

research data sharing without barriers

rd-alliance.org

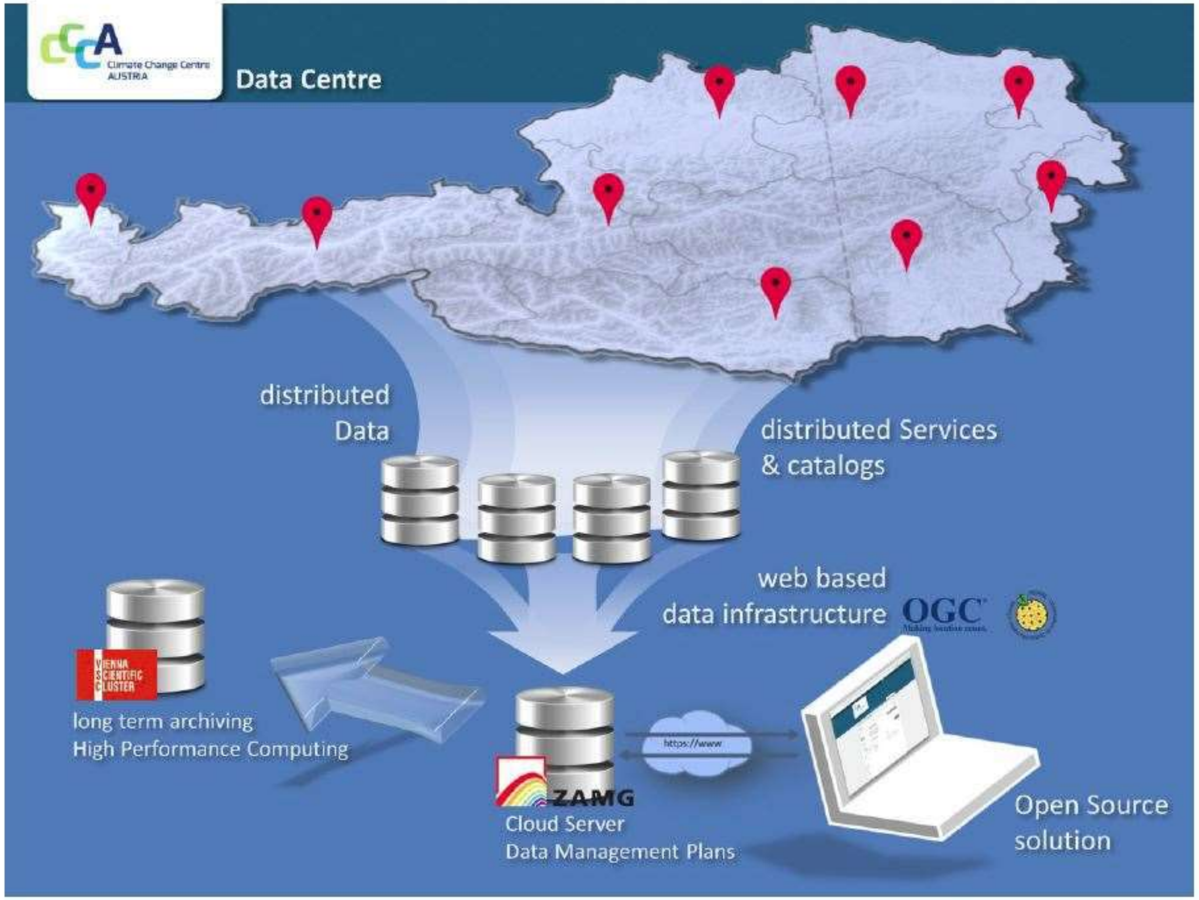
CCCA

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

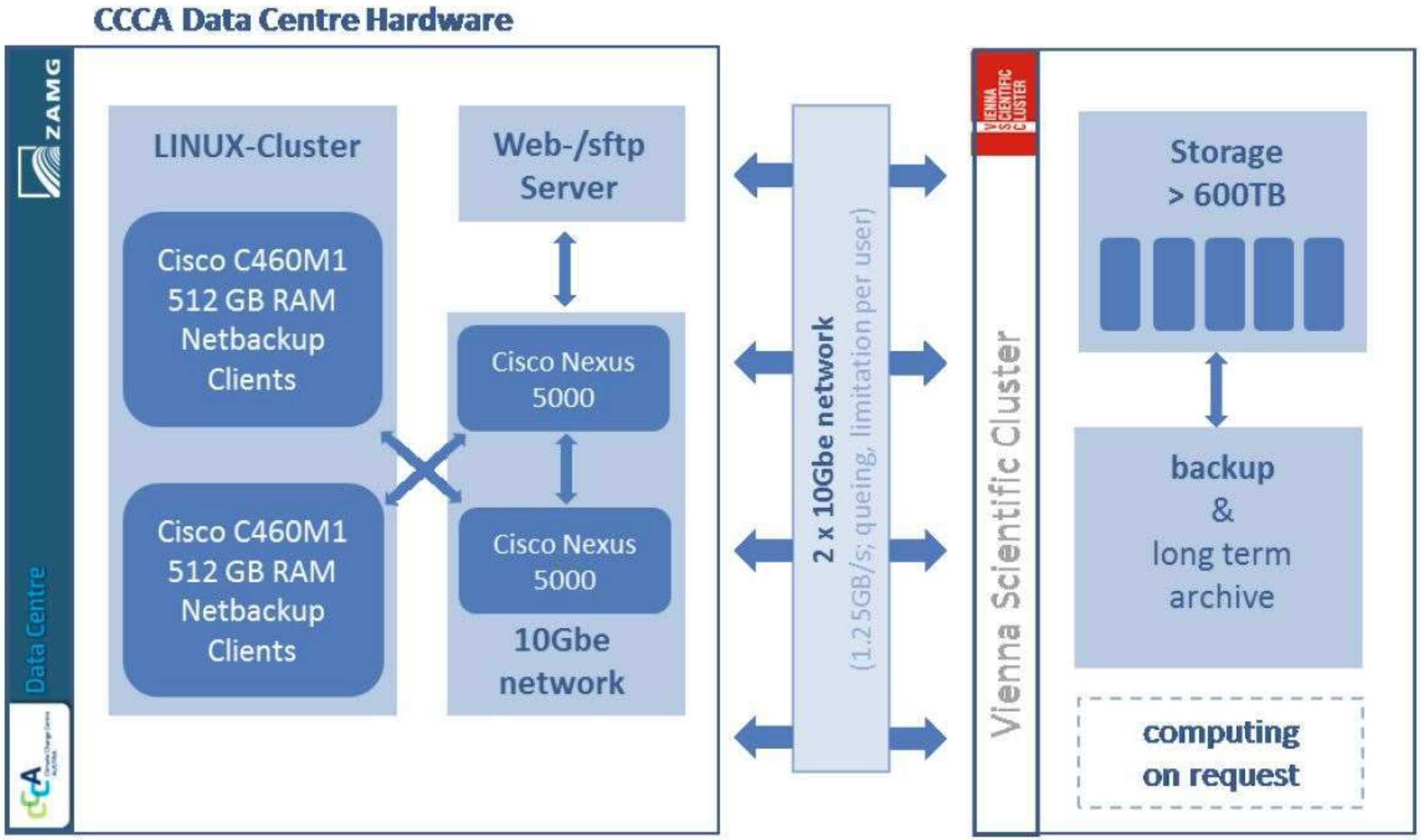
CCCA



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**

CCCA



Slide by Chris Schubert

FACULTY OF !NFORMATICS

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:

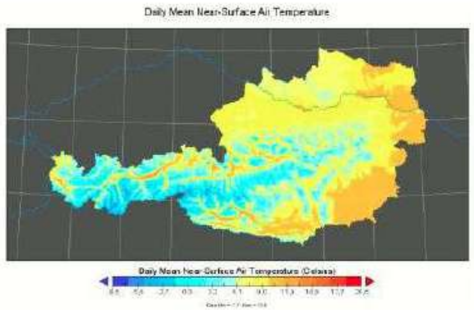
Using this data set or resource, you should cite this data set according to the given copyright conditions with following citation rules.

> 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

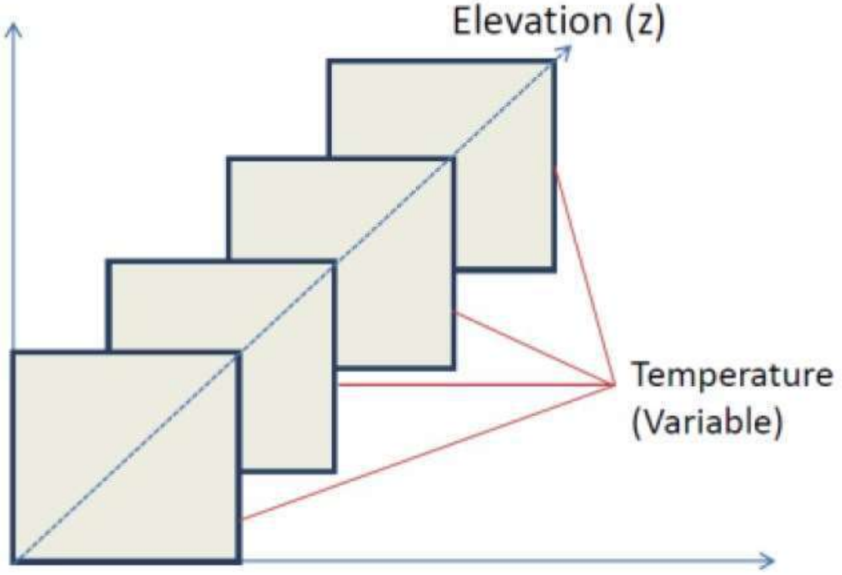
Your Publication

formal Data Citation

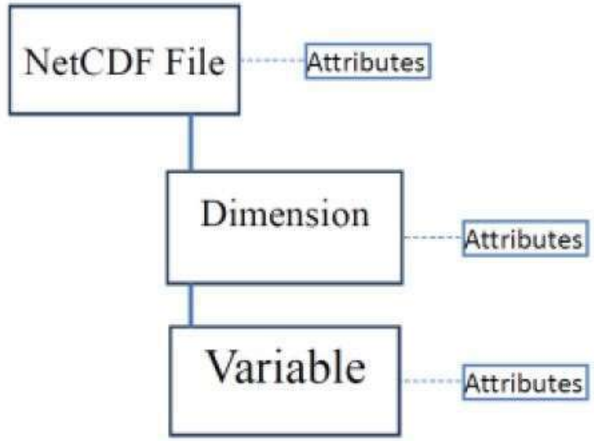
Your Data



- NetCDF Files:

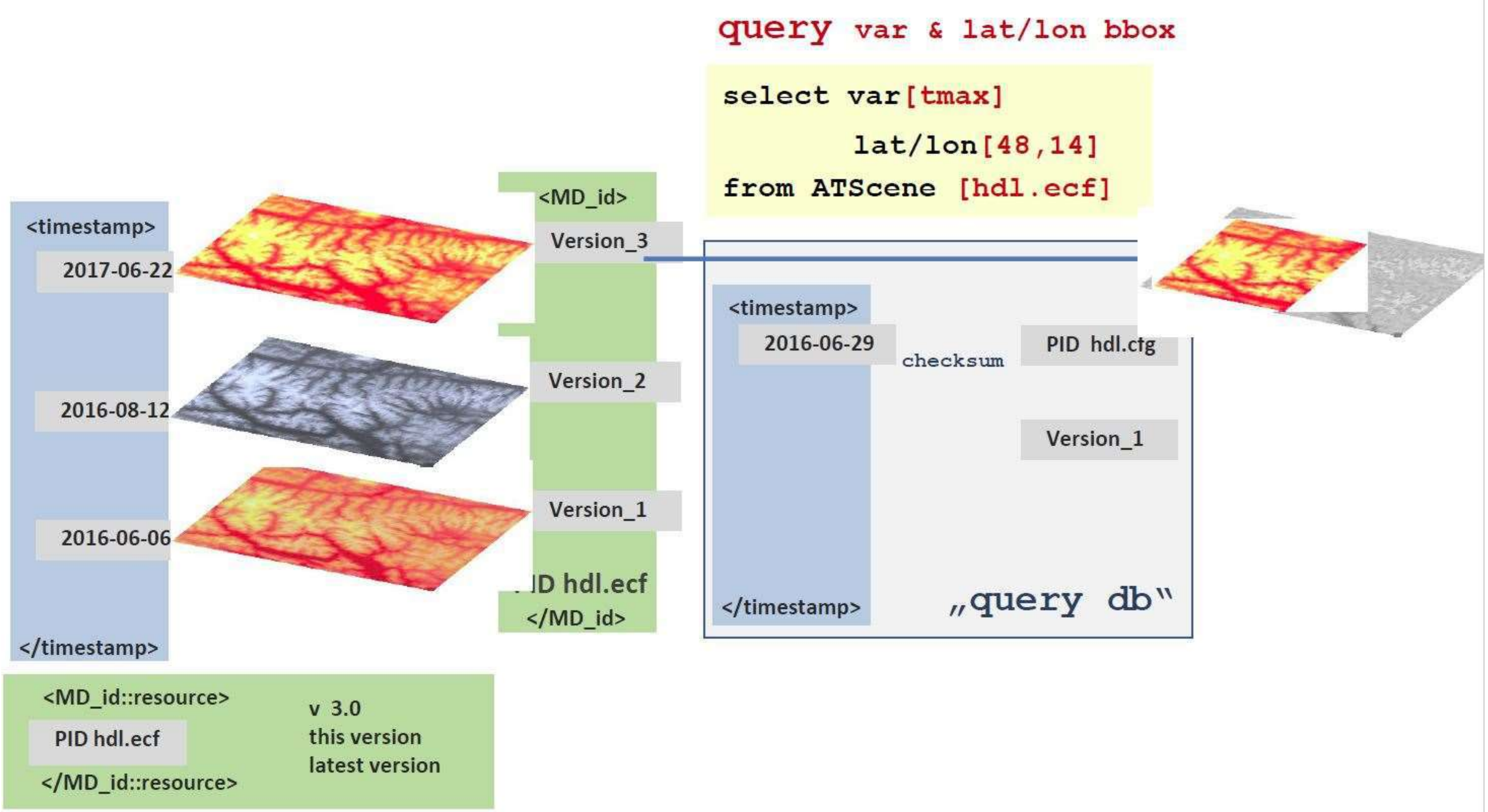


modified and based on UCAR Unidata, www.unidata.ucar.edu/



```
List of 10
 $ id          : int 4
 $ ndims       : int 3
 $ natts       : int 7
 $ unlimdimid  : num 3
 $ filename    : chr "C:\\PIK\\Data\\n4 177f9060-a351-11e1-b55e-e1515196aa7..
 $ varid2Rindex: num [1:7] 0 0 0 1 2 3 4
 $ writable    : logi FALSE
 $ dim         :List of 3
 ..$ longitude:List of 8
 .. ..$ name         : chr "longitude"
 .. ..$ len          : int 720
 .. ..$ unlim        : logi FALSE
 .. ..$ id           : int 1
 .. ..$ dimvarid     : num 1
 .. ..$ units        : chr "degrees_east"
 .. ..$ vals         : num [1:720(1d)] 0.25 0.75 1.25 1.75 2.25 2.75 3.25 3.75 4.25 4.75 ...
 .. ..$ create_dimvar: logi TRUE
 .. ..- attr(*, "class")= chr "dim.ncdf..
 ---------------------------
 $ nvars       : num 4
 $ var         :List of 4
 ..$ biomass_carbon_burning_nonCF:List of 16
 .. ..$ id         : int 4
 .. ..$ name       : chr "biomass_carbon_burning_nonCF"
 .. ..$ ndims      : int 3
 .. ..$ natts      : int 4
 .. ..$ size       : int [1:3] 720 279 4
 .. ..$ prec       : chr "float"
 .. ..$ dimids     : num [1:3] 1 2 3
 .. ..$ units      : chr "kg m-2"
 .. ..$ longname   : chr "biomass carbon burning"
 .. ..$ dims       : list()
 .. ..$ dim        :List of 3
```

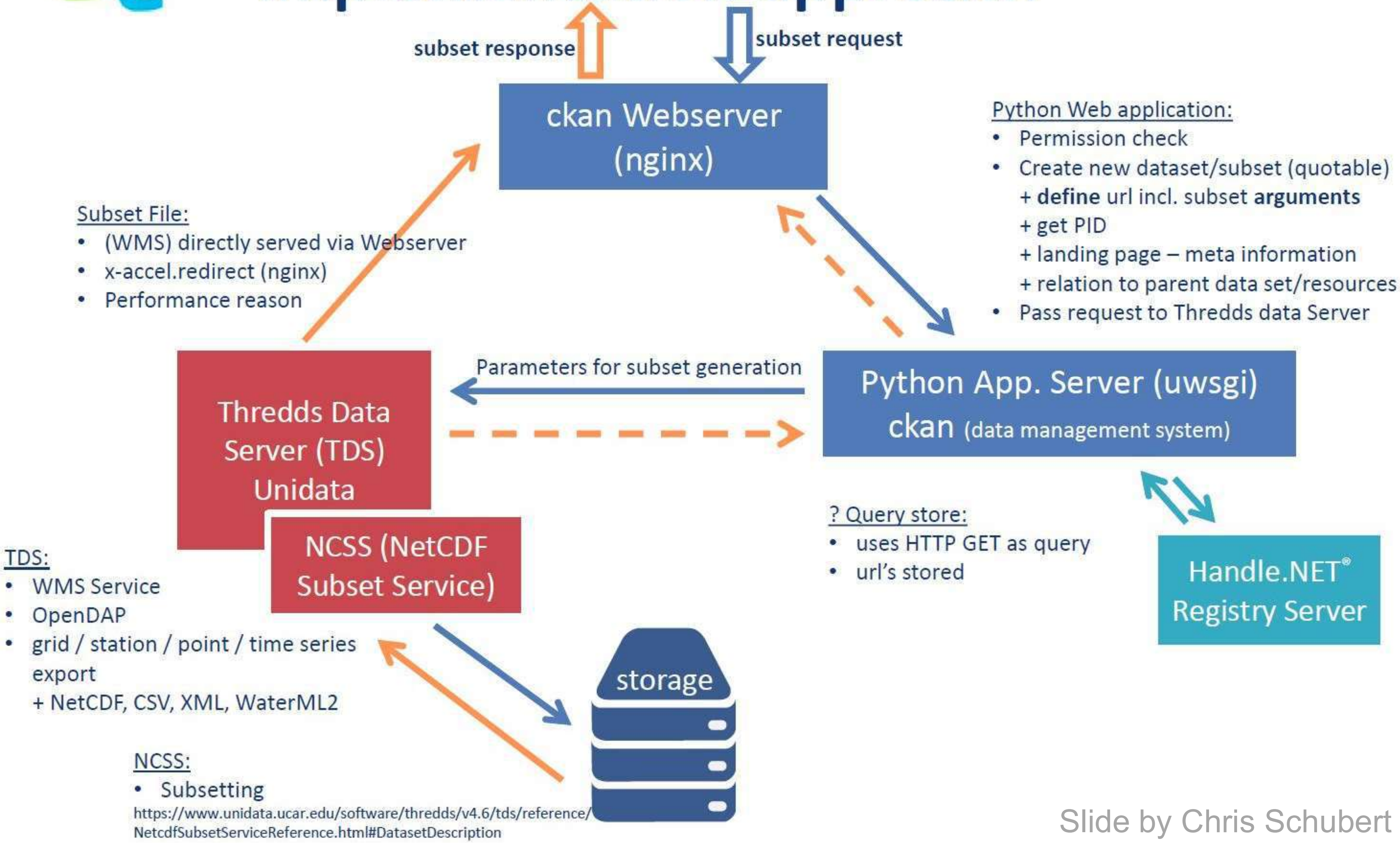
CCCA



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

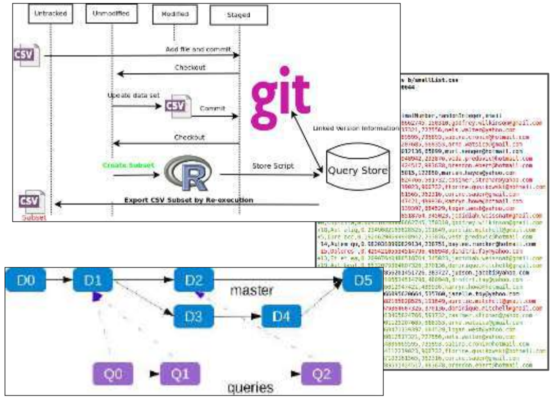
Summary

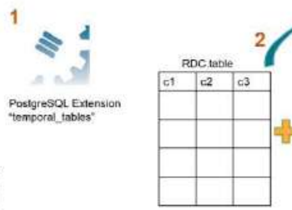
- Data citation essential for **solid** and **efficient** science (but not just for science!)
- It is more than just giving credit
- Human-readable and machine-actionable
- RDA recommendations
  - Time-stamp and version data if it is evolving
  - Provide PIDs to arbitrary subsets via selection mechanism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

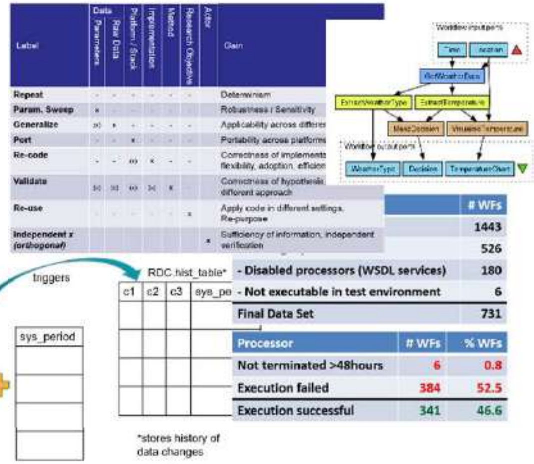
Benefits

- **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!

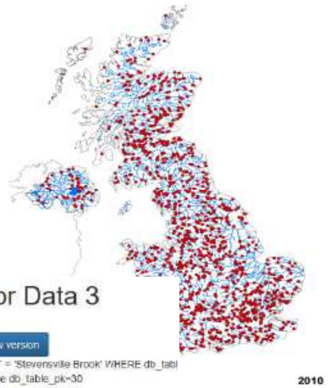


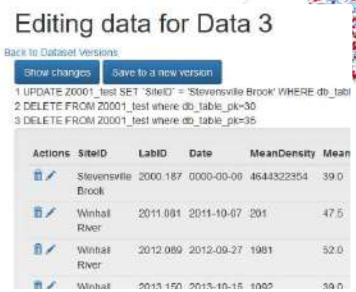


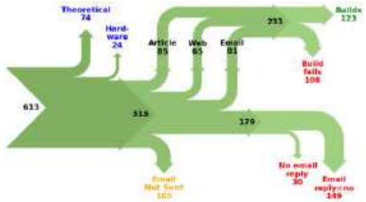


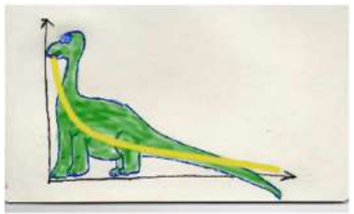
Thanks!

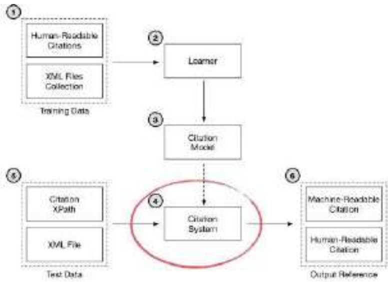
https://rd-alliance.org/working-groups/data-citation-wg.html











FAIR principles

Tomasz Miksa

TU Wien & SBA Research

tmiksa@sba-research.org

Informatics

Agenda

- Introduction
- FAIR principles in detail
  - Important concepts underpinning the principles
    - Persistent identifiers, Metadata, Vocabularies, etc.
- FAIR assessment
- FAIR Digital Object
- Summary
- Literature and useful resources

INTRODUCTION

Hans Rosling and Data Science

Talk held in 2006

16 years later the problems are not solved

- (But a lot is going on to change this)

TED

https://www.ted.com/talks/hans_rosling_the_best_stats_you_ve_ever_seen#t-1144167

Hans Rosling

“Because the **data is hidden down in the databases**. And the public is there, and the internet is there, but we have still not used it effectively.”

“There are some web pages like this,(…), but people **put prices** on them, **stupid passwords** and **boring statistics**.”

Hans Rosling

“Some countries accept that their databases can go out on the world. But what we really need is, of course, a search function, a search function where **we can copy the data up to a searchable format** and get it out in the world.”

“The **publicly funded data** is down here. (…) One of the crucial points is to **make them searchable**, and then people can use the different design tools to animate it there.”

Variety of solutions

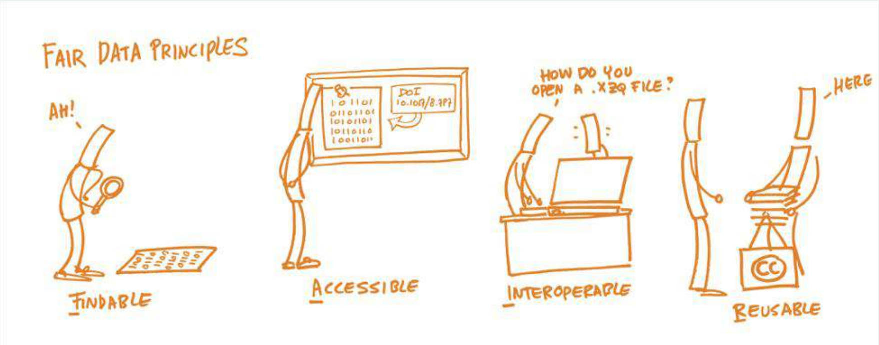
In response to these needs many solutions were proposed and are being implemented

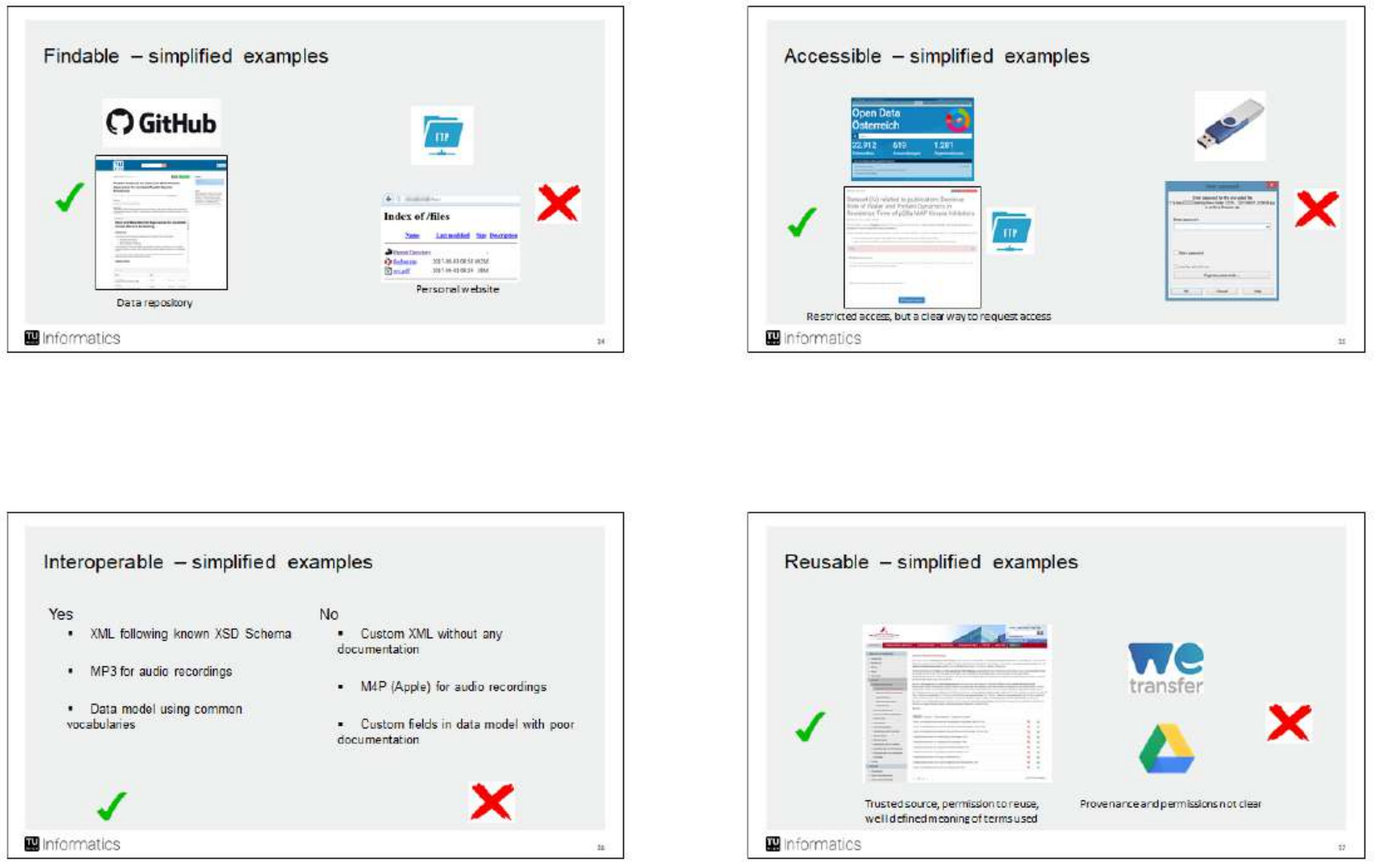
- **open access** to scientific publications and data
- research **data repositories** to host the data
- **persistent identifiers** to locate the data
- **data management plans**
- **FAIR principles**
- …

Simplified view on FAIR

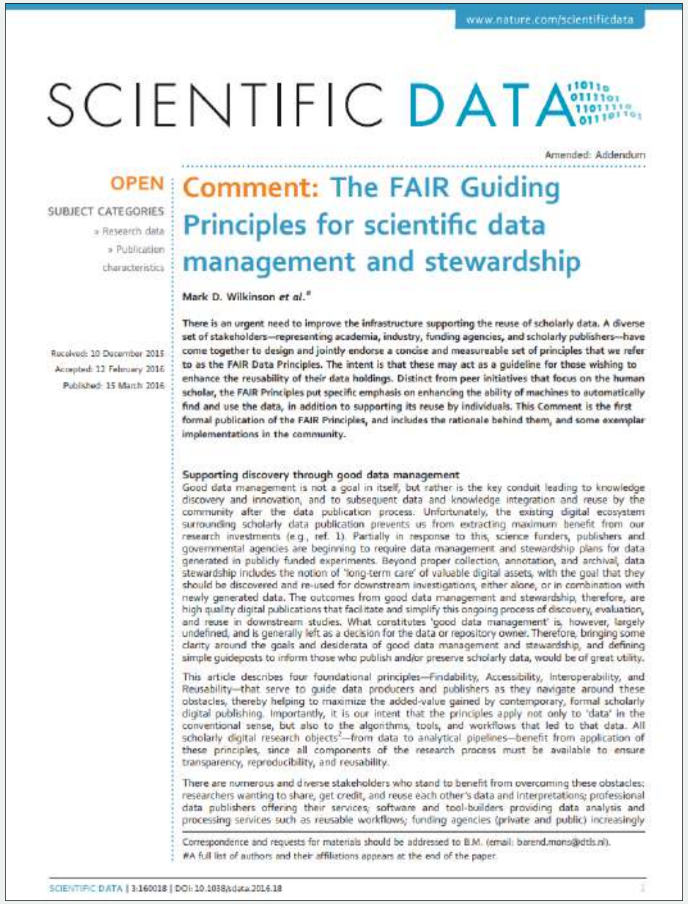
Discussed in the introductory lecture!

- Superficial
- For non-tech people





FAIR principles



www.nature.com/scientificdata

SCIENTIFIC DATA

Amended: Addendum

OPEN

SUBJECT CATEGORIES
» Research data
» Publication characteristics

Comment: The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson *et al.*[#]

Received: 10 December 2015
Accepted: 12 February 2016
Published: 15 March 2016

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measureable set of principles that we refer to as the FAIR Data Principles. The intent is that these may act as a guideline for those wishing to enhance the reusability of their data holdings. Distinct from peer initiatives that focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals. This Comment is the first formal publication of the FAIR Principles, and includes the rationale behind them, and some exemplar implementations in the community.

Supporting discovery through good data management

Good data management is not a goal in itself, but rather is the key conduit leading to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse by the community after the data publication process. Unfortunately, the existing digital ecosystem surrounding scholarly data publication prevents us from extracting maximum benefit from our research investments (e.g., ref. 1). Partially in response to this, science funders, publishers and governmental agencies are beginning to require data management and stewardship plans for data generated in publicly funded experiments. Beyond proper collection, annotation, and archival, data stewardship includes the notion of 'long-term care' of valuable digital assets, with the goal that they should be discovered and re-used for downstream investigations, either alone, or in combination with newly generated data. The outcomes from good data management and stewardship, therefore, are high quality digital publications that facilitate and simplify this ongoing process of discovery, evaluation, and reuse in downstream studies. What constitutes 'good data management' is, however, largely undefined, and is generally left as a decision for the data or repository owner. Therefore, bringing some clarity around the goals and desiderata of good data management and stewardship, and defining simple guideposts to inform those who publish and/or preserve scholarly data, would be of great utility.

This article describes four foundational principles—Findability, Accessibility, Interoperability, and Reusability—that serve to guide data producers and publishers as they navigate around these obstacles, thereby helping to maximize the added-value gained by contemporary, formal scholarly digital publishing. Importantly, it is our intent that the principles apply not only to 'data' in the conventional sense, but also to the algorithms, tools, and workflows that led to that data. All scholarly digital research objects[2]—from data to analytical pipelines—benefit from application of these principles, since all components of the research process must be available to ensure transparency, reproducibility, and reusability.

There are numerous and diverse stakeholders who stand to benefit from overcoming these obstacles: researchers wanting to share, get credit, and reuse each other's data and interpretations; professional data publishers offering their services; software and tool-builders providing data analysis and processing services such as reusable workflows; funding agencies (private and public) increasingly

Correspondence and requests for materials should be addressed to B.M. (email: barend.mons@dtls.nl).
#A full list of authors and their affiliations appears at the end of the paper.

SCIENTIFIC DATA | 3:160018 | DOI: 10.1038/sdata.2016.18

https://www.nature.com/articles/sdata201618

https://www.go-fair.org/fair-principles/

MACHINE-ACTIONABILITY

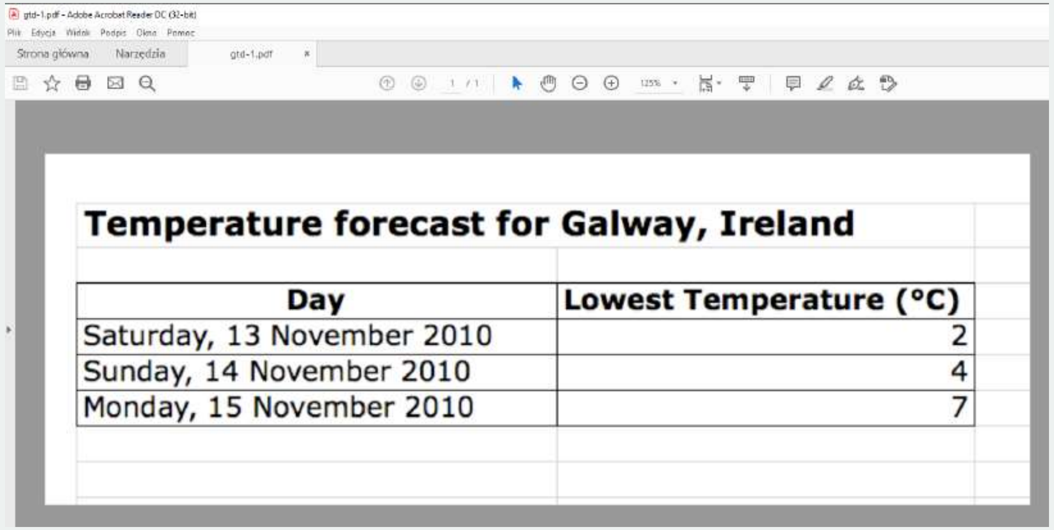
Machine actionability

“the capacity of computational systems to find, access, interoperate, and reuse data with none or minimal human intervention” https://www.go-fair.org/fair-principles

“information that is structured in a consistent way so that machines, or computers, can be programmed against the structure.” https://ddialliance.org/taxonomy/term/198

Machine-actionability is core to each of the FAIR principles

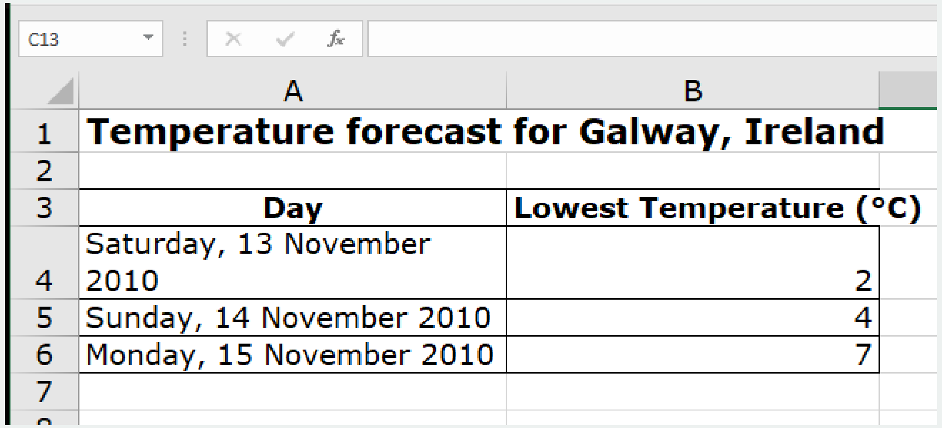
Machine-actionability - example



Temperature forecast for Galway, Ireland

Day	Lowest Temperature (°C)
Saturday, 13 November 2010	2
Sunday, 14 November 2010	4
Monday, 15 November 2010	7

Not machine-actionable



	A	B
1	**Temperature forecast for Galway, Ireland**	
2		
3	**Day**	**Lowest Temperature (°C)**
4	Saturday, 13 November 2010	2
5	Sunday, 14 November 2010	4
6	Monday, 15 November 2010	7
7		

Machine-actionable

Machine-actionability – example (Linked Open Data)

Temperature forecast for Galway, Ireland
https://5stardata.info/en/examples/gtd-5/
→ dc:title → "Temperature forecast for Galway, Ireland"
→ xhtml:stylesheet → https://5stardata.info/css/style.css
→ dcterms:title → "Temperature forecast for Galway, Ireland"
→ dcterms:created → "2012-01-22"^^xsd:date
→ dcterms:creator → http://mhausenblas.info/#i
→ dcterms:modified → "2015-08-31"^^xsd:date
→ dcterms:contributor → http://jayg.me/
→ dcterms:license → http://creativecommons.org/publicdomain/zero/1.0/

https://5stardata.info/en/examples/gtd-5/#Galway
→ rdf:type → meteo:Place
→ owl:sameAs → dbpedia:Galway
→ meteo:forecast → https://5stardata.info/en/examples/gtd-5/#forecast20101113, https://5stardata.info/en/examples/gtd-5/#forecast20101114, https://5stardata.info/en/examples/gtd-5/#forecast20101115

https://5stardata.info/en/examples/gtd-5/#forecast20101113
→ meteo:predicted → "2010-11-13T00:00:00Z"^^xsd:dateTime
→ meteo:temperature → https://5stardata.info/en/examples/gtd-5/#temp20101113
← is meteo:forecast of ← https://5stardata.info/en/examples/gtd-5/#Galway

https://5stardata.info/en/examples/gtd-5/#forecast20101114
→ meteo:predicted → "2010-11-14T00:00:00Z"^^xsd:dateTime
→ meteo:temperature → https://5stardata.info/en/examples/gtd-5/#temp20101114
← is meteo:forecast of ← https://5stardata.info/en/examples/gtd-5/#Galway

https://5stardata.info/en/examples/gtd-5/#forecast20101115
→ meteo:predicted → "2010-11-15T00:00:00Z"^^xsd:dateTime
→ meteo:temperature → https://5stardata.info/en/examples/gtd-5/#temp20101115
← is meteo:forecast of ← https://5stardata.info/en/examples/gtd-5/#Galway

https://5stardata.info/en/examples/gtd-5/#temp
→ rdfs:seeAlso → dbpedia:Temperature
→ owl:sameAs → dbpedia:Celsius

Unit definition and link to common definition

https://5stardata.info/en/examples/gtd-5/#temp20101113
→ meteo:celsius → "2"^^xsd:decimal
← is meteo:temperature of ← https://5stardata.info/en/examples/gtd-5/#forecast20101113

https://5stardata.info/en/examples/gtd-5/#temp20101114
→ meteo:celsius → "4"^^xsd:decimal
← is meteo:temperature of ← https://5stardata.info/en/examples/gtd-5/#forecast20101114

https://5stardata.info/en/examples/gtd-5/#temp20101115
→ meteo:celsius → "7"^^xsd:decimal
← is meteo:temperature of ← https://5stardata.info/en/examples/gtd-5/#forecast20101115

Values, types, link to forecast

http://graphite.ecs.soton.ac.uk/browser/?uri=http://5stardata.info/en/examples/gtd-5/

Machine actionability – different shades

5-star model shows importance and benefits of

- machine-actionability
- open data
- semantic modelling

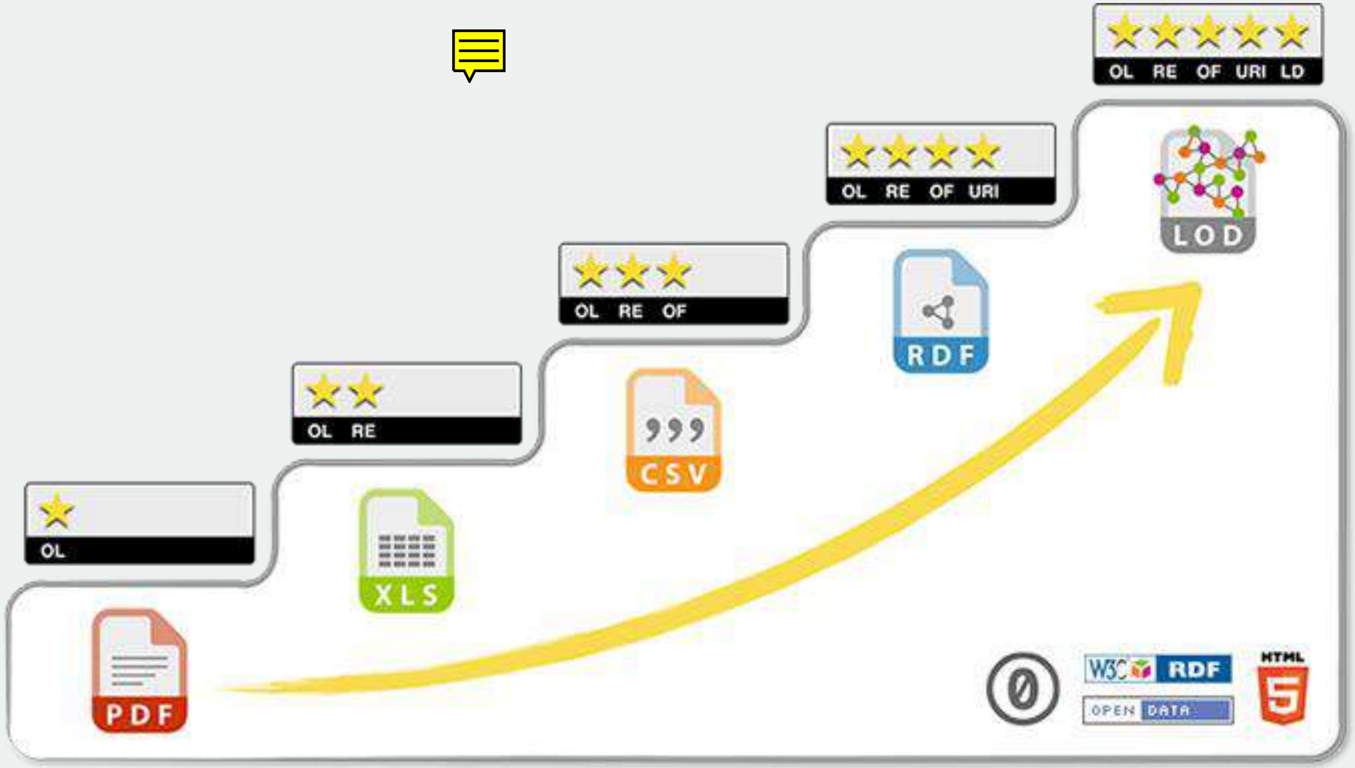
★ make your stuff available on the Web (whatever format) under an open license[1]

★★ make it available as structured data (e.g., Excel instead of image scan of a table)[2]

★★★ make it available in a non-proprietary open format (e.g., CSV instead of Excel)[3]

★★★★ use URIs to denote things, so that people can point at your stuff[4]

★★★★★ link your data to other data to provide context[5]



Explore examples on the website to learn more on differences between each level!

https://5stardata.info/en/

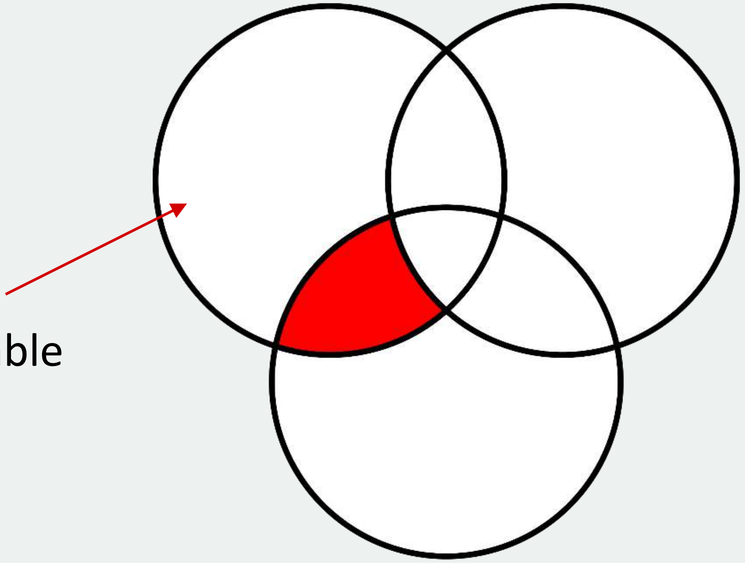
Machine actionability and FAIR

- Machine-actionability is core to each of the FAIR principles
- The more machine-actionable data is, the better it is
- FAIR does not require data to be open
  - 5-star model suggests openness – don’t confuse those two!

Machine actionability and FAIR
(how people often see it)

FAIR

Open



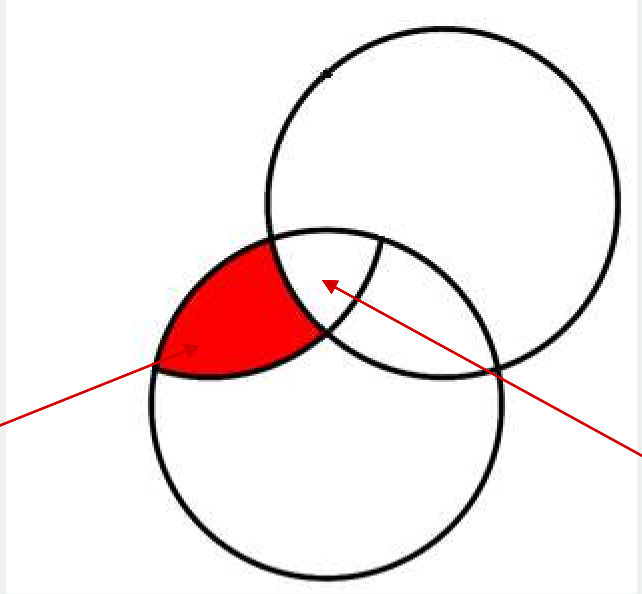
FAIR and not machine-actionable
(no such thing!)

Machine-actionable

Machine actionability and FAIR
(how it really is)

FAIR

Open



FAIR must always be machine-actionable and does not have to be open

FAIR must be machine-actionable and can be open

Machine-actionable

FAIR IN DETAIL

Findable

F1. (Meta)data are assigned a globally unique and **persistent identifier**

F2. Data are described with rich **metadata**

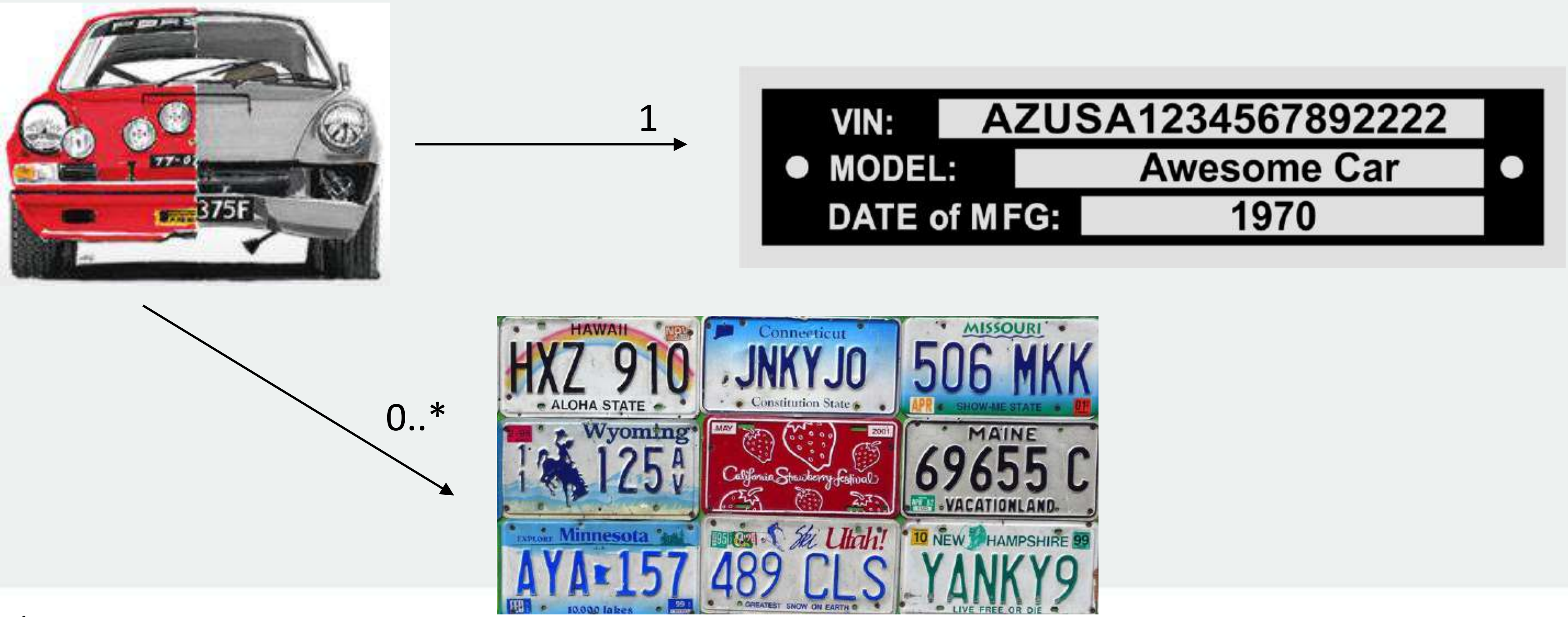
F3. Metadata clearly and explicitly include the identifier of the data they describe

F4. (Meta)data are registered or indexed in a **searchable resource**

Persistent identifiers (F1)

Example

- A car has only one VIN (PID), but can have many number plates over its lifetime (URL)



Persistent Identifiers (F1)

Digital Object Identifier (DOI)

- Uniquely identify objects
- DOI assigned once
- Physical location of data can change

10.5281/zenodo.1068223

ORCID ID

- Unique person ID
- ORCID assigned once
- Person can change affiliations (jobs)

Connecting Research and Researchers

0000-0002-4929-7875

Persistent Identifiers - DOI

Unlike the URLs, DOIs are associated to objects and not to locations

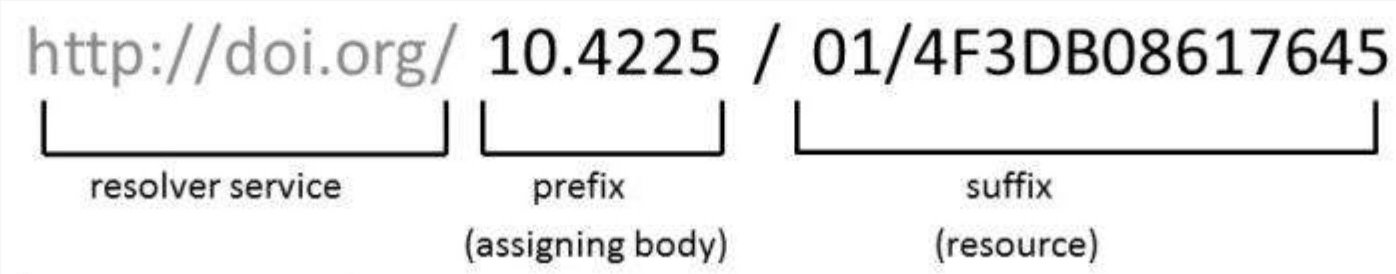
URLs are unique, but not persistent

DOIs are never deleted

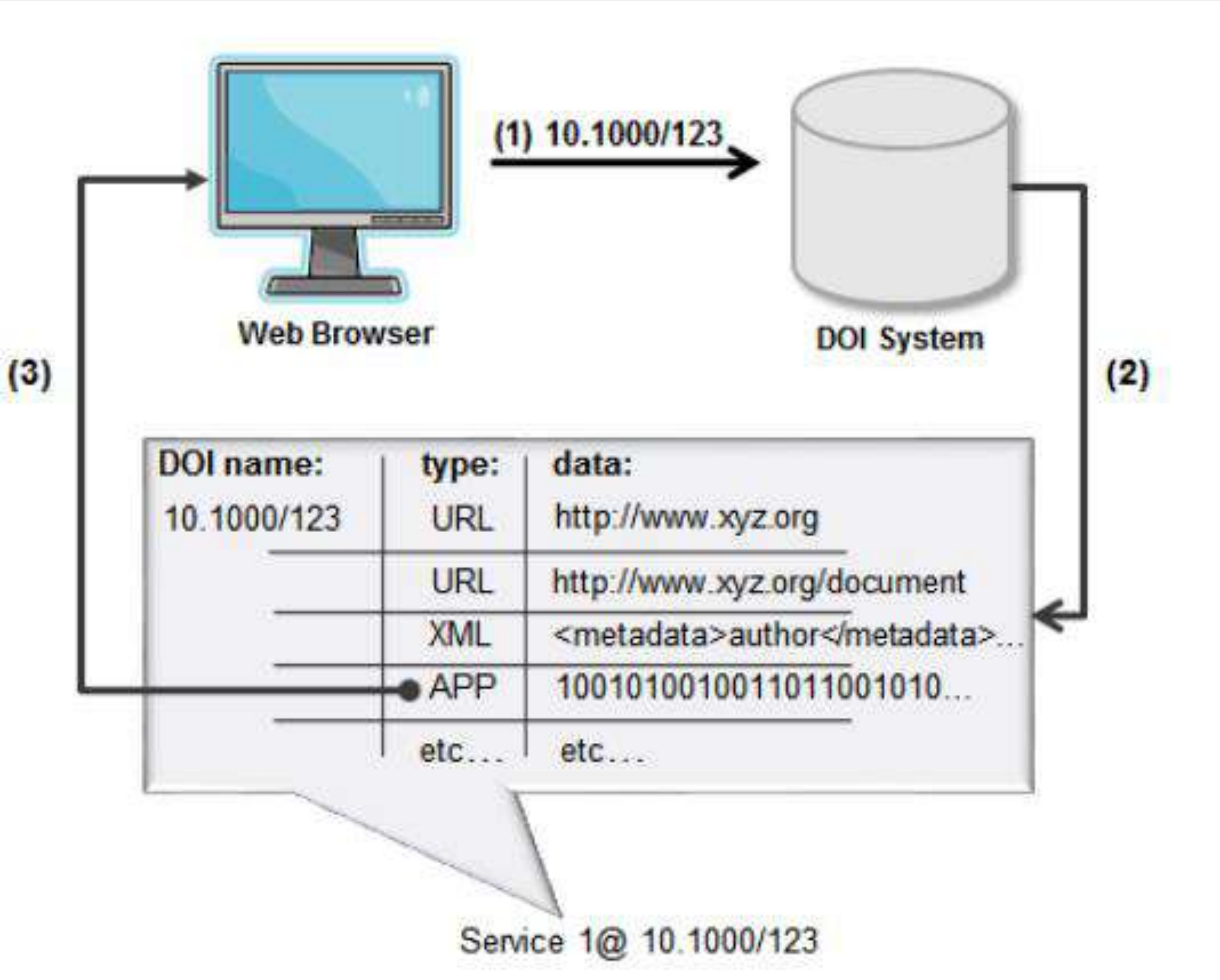
- if resource does not exist then a message is provided

Resolver service

Metadata

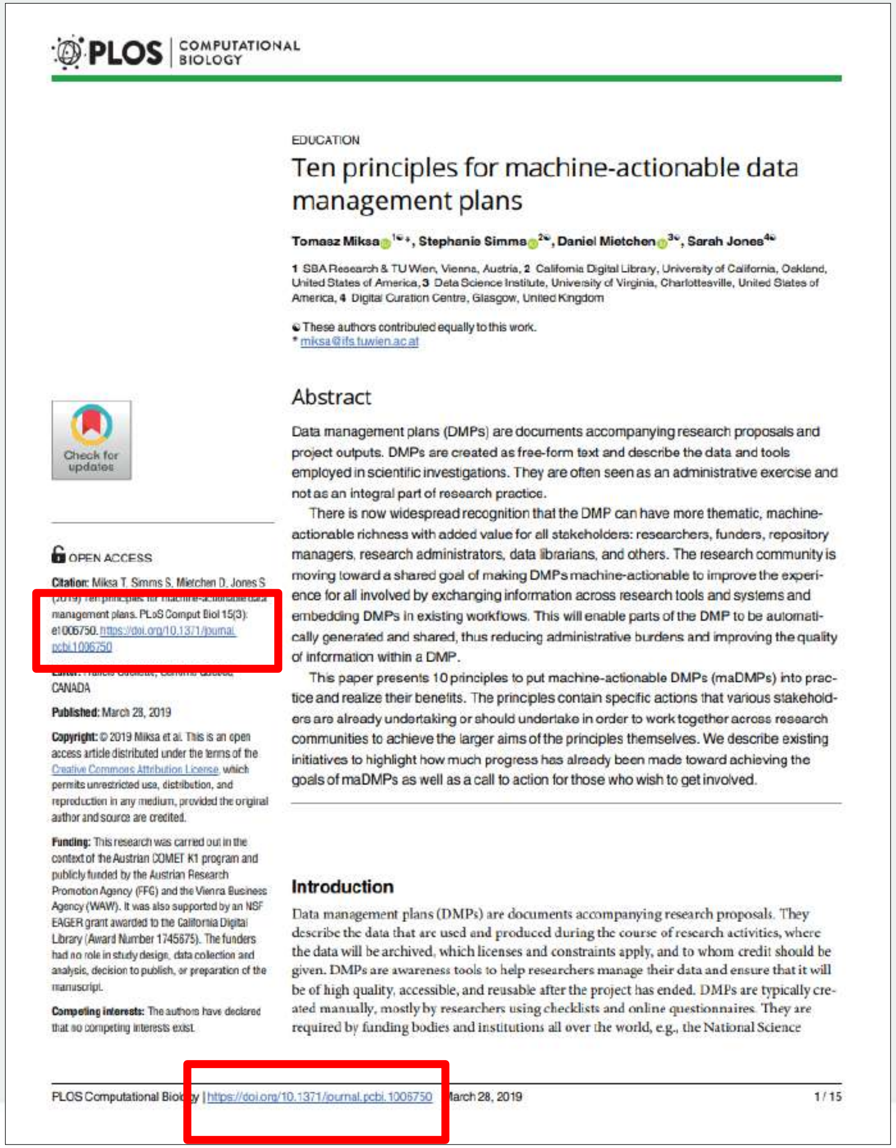


DOI – resolver service



https://www.doi.org/doi_handbook/3_Resolution.html

DOI example – assigned to publication



PLOS | COMPUTATIONAL BIOLOGY

EDUCATION

Ten principles for machine-actionable data management plans

Tomasz Miksa[1*], **Stephanie Simms**[2], **Daniel Mietchen**[3], **Sarah Jones**[4]

1 SBA Research & TU Wien, Vienna, Austria, 2 California Digital Library, University of California, Oakland, United States of America, 3 Data Science Institute, University of Virginia, Charlottesville, United States of America, 4 Digital Curation Centre, Glasgow, United Kingdom

These authors contributed equally to this work.
* miksa@ifs.tuwien.ac.at

Citation: Miksa T, Simms S, Mietchen D, Jones S (2019) Ten principles for machine-actionable data management plans. PLoS Comput Biol 15(3): e1006750. https://doi.org/10.1371/journal.pcbi.1006750

CANADA

Published: March 28, 2019

Funding: This research was carried out in the context of the Austrian COMET K1 program and publicly funded by the Austrian Research Promotion Agency (FFG) and the Vienna Business Agency (WAW). It was also supported by an NSF EAGER grant awarded to the California Digital Library (Award Number 1745675). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Data management plans (DMPs) are documents accompanying research proposals and project outputs. DMPs are created as free-form text and describe the data and tools employed in scientific investigations. They are often seen as an administrative exercise and not as an integral part of research practice.

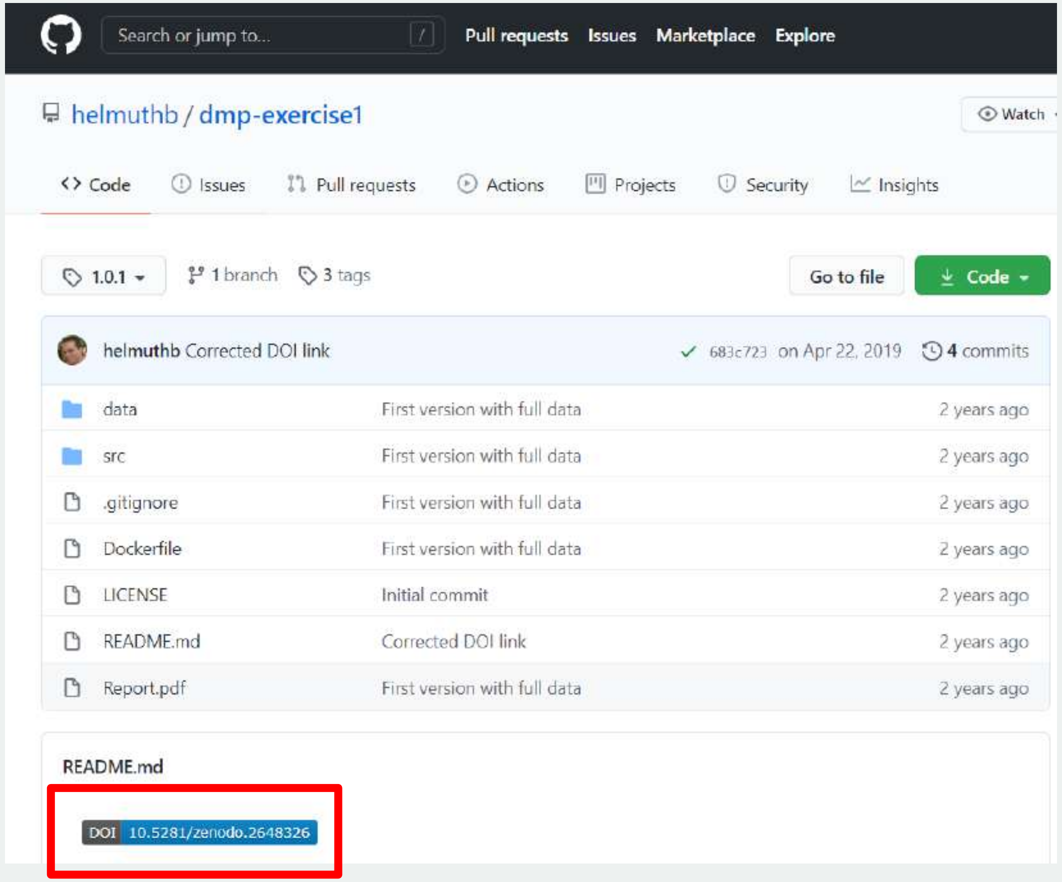
There is now widespread recognition that the DMP can have more thematic, machine-actionable richness with added value for all stakeholders: researchers, funders, repository managers, research administrators, data librarians, and others. The research community is moving toward a shared goal of making DMPs machine-actionable to improve the experience for all involved by exchanging information across research tools and systems and embedding DMPs in existing workflows. This will enable parts of the DMP to be automatically generated and shared, thus reducing administrative burdens and improving the quality of information within a DMP.

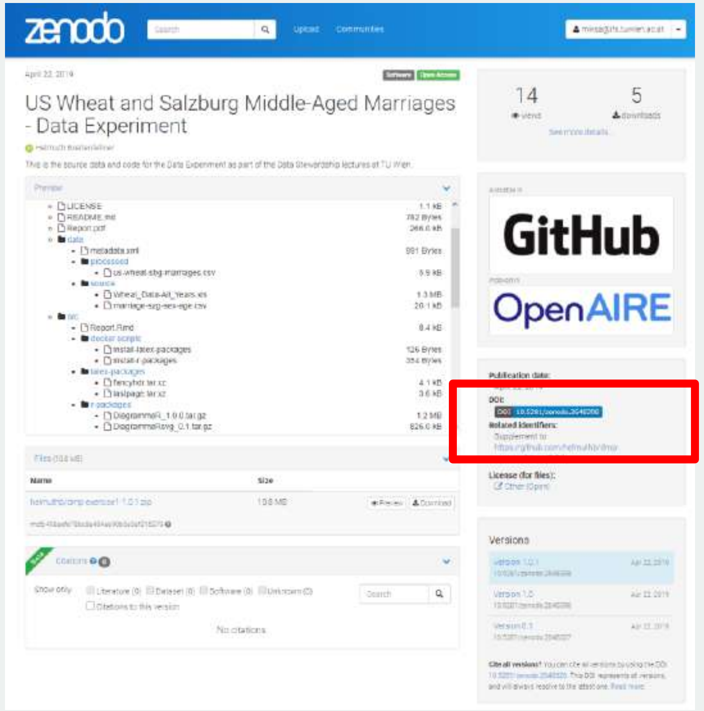
This paper presents 10 principles to put machine-actionable DMPs (maDMPs) into practice and realize their benefits. The principles contain specific actions that various stakeholders are already undertaking or should undertake in order to work together across research communities to achieve the larger aims of the principles themselves. We describe existing initiatives to highlight how much progress has already been made toward achieving the goals of maDMPs as well as a call to action for those who wish to get involved.

Introduction

Data management plans (DMPs) are documents accompanying research proposals. They describe the data that are used and produced during the course of research activities, where the data will be archived, which licenses and constraints apply, and to whom credit should be given. DMPs are awareness tools to help researchers manage their data and ensure that it will be of high quality, accessible, and reusable after the project has ended. DMPs are typically created manually, mostly by researchers using checklists and online questionnaires. They are required by funding bodies and institutions all over the world, e.g., the National Science

DOI example – assigned to code





DOI example - assigned to data

Search

Log in

November 27, 2020 | Version 1.0

Dataset

Embargoed

The Sentinel-1 Global Backscatter Model (S1GBM) - Mapping Earth's Land Surface with C-Band Microwaves

Bauer-Marschallinger, Bernhard [1]; Cao, Senmao [1, 2]; Navacchi, Claudio [1]; Freeman, Vahid [1, 3]; Reuß, Felix [1]; Geudtner, Dirk [4]; Rommen, Björn [4]; Vega, Francisco Ceba [4]; Snoeij, Paul [5]; Attema, Evert [4]; Reimer, Christoph [2]; Wagner, Wolfgang [1, 2] show affiliations

Description

This dataset was generated by the Remote Sensing Group of the TU Wien Department of Geodesy and Geoinformation (https://mrs.geo.tuwien.ac.at/), within a dedicated project by the European Space Agency (ESA).

With this dataset publication, we open up a new perspective on Earth's land surface, providing a normalised microwave backscatter map from spaceborne Synthetic Aperture Radar (SAR) observations. The Sentinel-1 Global Backscatter Model (S1GBM) describes Earth for the period 2016-17 by the mean C-band radar cross section in VV- and VH-polarization at a 10 m sampling, giving a high-quality impression on surface- structures and -patterns.

At TU Wien, we processed 0.5 million Sentinel-1 scenes totaling 1.1 PB and performed semi-automatic quality curation and backscatter harmonisation related to orbit geometry effects. The overall mosaic quality excels (the few) existing datasets, with minimised imprinting from orbit discontinuities and successful angle normalisation in large parts of the world. Supporting the designand verification of upcoming radar sensors, the obtained S1GBM data potentially also serve land cover classification and determination of vegetation and soil states, as well as water body mapping.

We invite developers from the broader user community to exploit this novel data resource and to integrate S1GBM parameters in models for various variables of land cover, soil composition, or vegetation structure.

Versions

Version 1.0
DOI: 10.48436/n2d1v-gqb91

Cite As

Bauer-Marschallinger, Bernhard et al. (2020). The Sentinel-1 Global Backscatter Model (S1GBM) - Mapping Earth's Land Surface with C-Band Microwaves (Version 1.0) [Dataset]. TU Data. https://doi.org/10.48436/n2d1v-gqb91

Persistent Identifiers

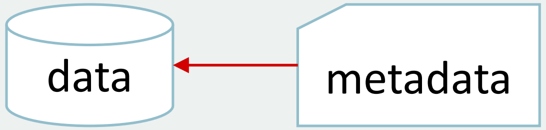
More on persistent identifiers (PIDs) in the other lectures

Note: DOI and ORCID are not the only PIDs in use!

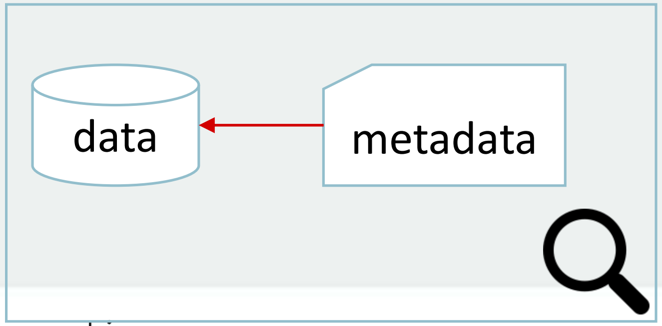
Findable (F2-F4)

F2. Data are described with rich metadata

F3. Metadata clearly and explicitly include the identifier of the data they describe

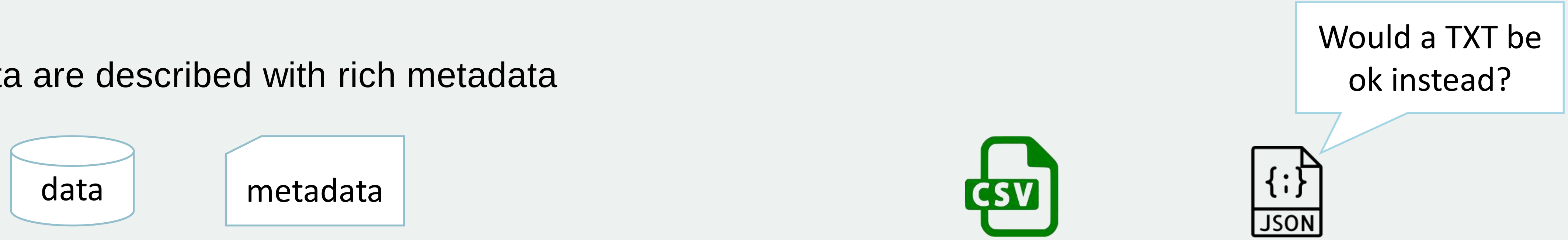


F4. (Meta)data are registered or indexed in a searchable resource

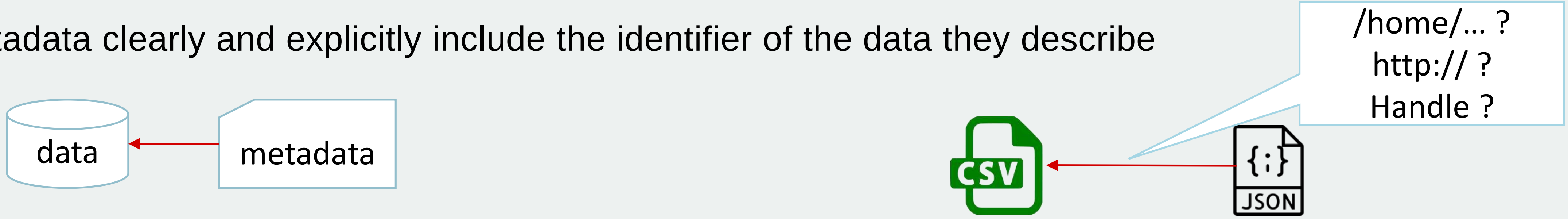


Findable

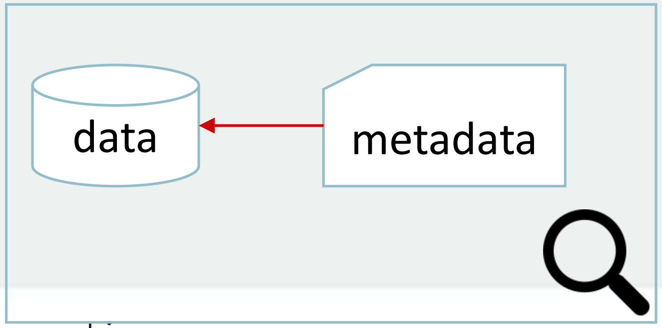
F2. Data are described with rich metadata



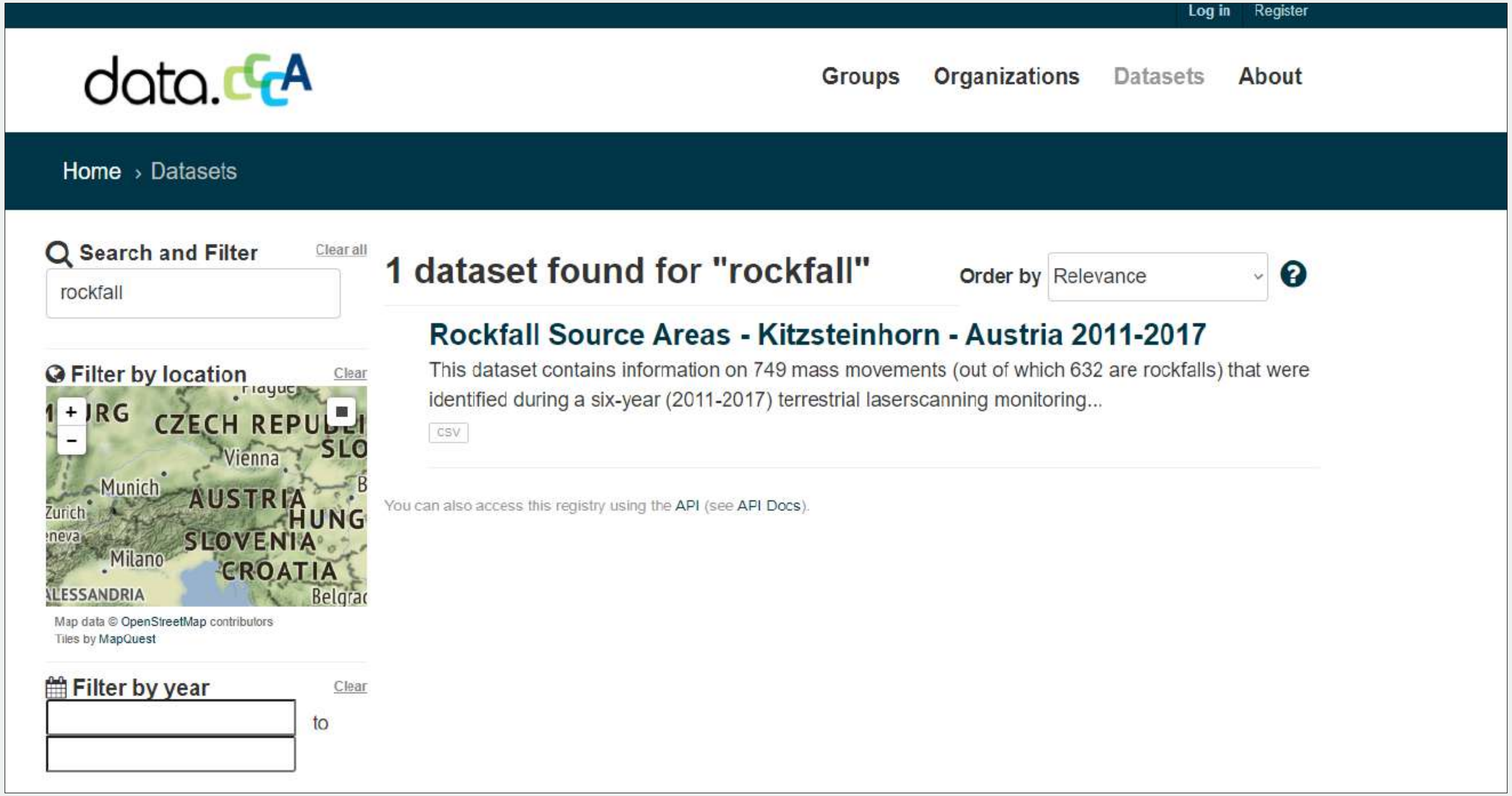
F3. Metadata clearly and explicitly include the identifier of the data they describe



F4. (Meta)data are registered or indexed in a searchable resource



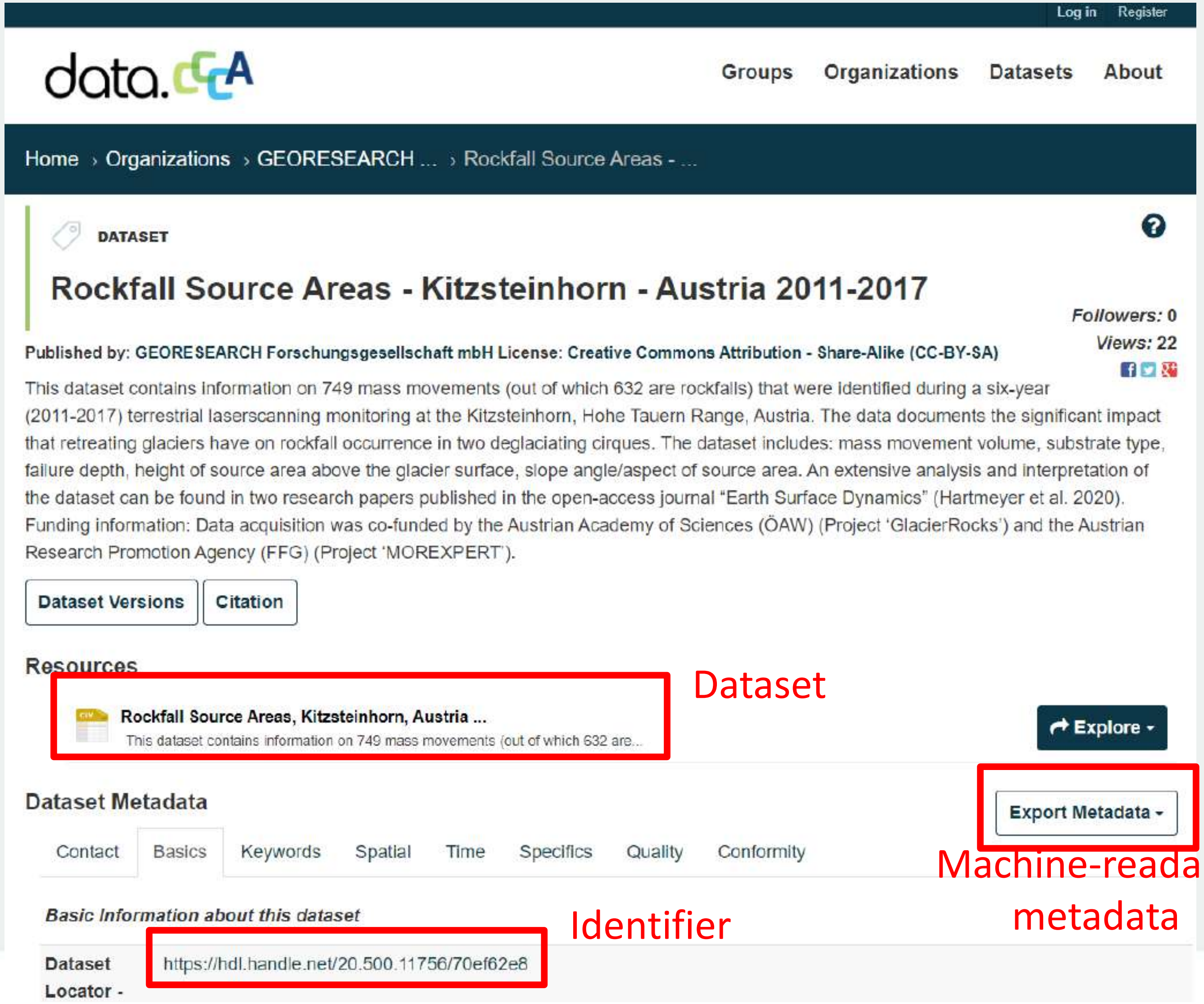
F4. (Meta)data are registered or indexed in a searchable resource



More on the F4 in the lecture on repositories.

https://data.ccca.ac.at

F3. Metadata clearly and explicitly include the identifier of the data they describe



https://hdl.handle.net/20.500.11756/70ef62e8

F2. Data are described with rich metadata

Resources

Rockfall Source Areas, Kitzsteinhorn, Austria ...
This dataset contains information on 749 mass movements (out of which 632 are...

Explore

Dataset Metadata

Export Metadata

Contact | Basics | Keywords | Spatial | Time | Specifics | Quality | Conformity

Basic Information about this dataset

Dataset Locator - URI	https://hdl.handle.net/20.500.11756/70ef62e8
Abstract	This dataset contains information on 749 mass movements (out of which 632 are rockfalls) that were identified during a six-year (2011-2017) terrestrial laserscanning monitoring at the Kitzsteinhorn, Hohe Tauern Range, Austria. The data documents the significant impact that retreating glaciers have on rockfall occurrence in two deglaciating cirques. The dataset includes: mass movement volume, substrate type, failure depth, height of source area above the glacier surface, slope angle/aspect of source area. An extensive analysis and interpretation of the dataset can be found in two research papers published in the open-access journal "Earth Surface Dynamics" (Hartmeyer et al. 2020). Funding information: Data acquisition was co-funded by the Austrian Academy of Sciences (ÖAW) (Project 'GlacierRocks') and the Austrian Research Promotion Agency (FFG) (Project 'MOREXPERT').
Metadata Language	English
License	cc-by-sa
Visibility	public
Use Limitation	no limitation

F2. Data are described with rich metadata

Resources

Rockfall Source Areas, Kitzsteinhorn, Austria ...
This dataset contains information on 749 mass movements (out of which 632 are...

Explore

Dataset Metadata

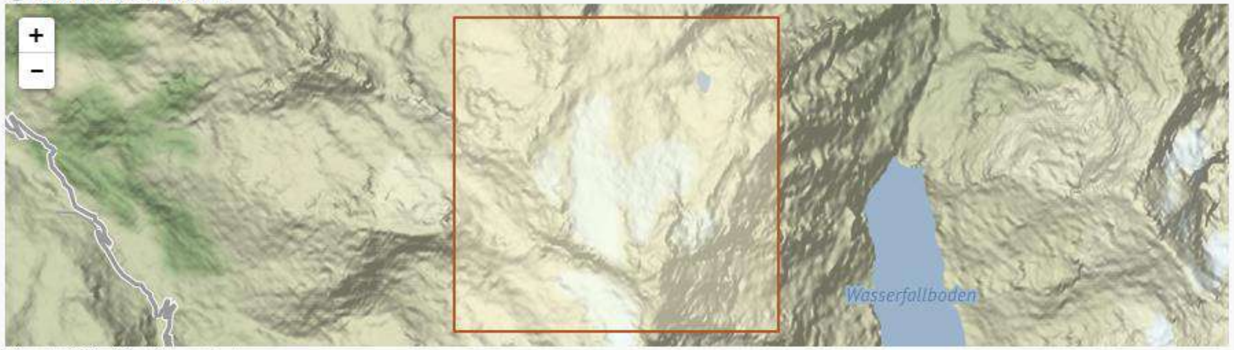
Export Metadata

Contact | Basics | Keywords | Spatial | Time | Specifics | Quality | Conformity

Geographic Aspects of the Resources

Polygon

Dataset extent



Map data © OpenStreetMap contributors
Tiles by MapQuest

Coverage Kitzsteinhorn, Hohe Tauern Range, Austria

Metadata - Chemistry

National Library of Medicine
National Center for Biotechnology Information

PubChem About Blog Submit Contact

COMPOUND SUMMARY

Water ethanol

PubChem CID	19096565
Structure	 2D 3D Find Similar Structures
Molecular Formula	$C_2H_8O_2$
Synonyms	ethanol water ethanol-water water ethanol water-ethanol EtOH water More...
Molecular Weight	64.08 g/mol
Parent Compound	CID 702 (Ethanol)
Component Compounds	CID 702 (Ethanol) CID 962 (Water)
Dates	Modify 2021-02-27 Create 2007-12-04

2 Names and Identifiers

2.1 Computed Descriptors

2.1.1 IUPAC Name

ethanol;hydrate

Computed by LexiChem 2.6.6 (PubChem release 2019.06.18)

▸ PubChem

2.1.2 InChI

InChI=1S/C2H6O.H2O/c1-2-3;/h3H,2H2,1H3;1H2

Computed by InChI 1.0.5 (PubChem release 2019.06.18)

▸ PubChem

2.1.3 InChI Key

IDGUHHHQCWSQLU-UHFFFAOYSA-N

Computed by InChI 1.0.5 (PubChem release 2019.06.18)

▸ PubChem

2.1.4 Canonical SMILES

CCO.O

Computed by OEChem 2.1.5 (PubChem release 2019.06.18)

▸ PubChem

2.2 Molecular Formula

C2H8O2

Computed by PubChem 2.1 (PubChem release 2019.06.18)

▸ PubChem

3 Chemical and Physical Properties

3.1 Computed Properties

Property Name	Property Value
Molecular Weight	64.08 g/mol
Hydrogen Bond Donor Count	2
Hydrogen Bond Acceptor Count	2
Rotatable Bond Count	0
Exact Mass	64.052429 g/mol
Monoisotopic Mass	64.052429 g/mol
Topological Polar Surface Area	21.2 Å^2
Heavy Atom Count	4
Formal Charge	0
Complexity	2.8
Isotope Atom Count	0
Defined Atom Stereocenter Count	0
Undefined Atom Stereocenter Count	0
Defined Bond Stereocenter Count	0
Undefined Bond Stereocenter Count	0
Covalently-Bonded Unit Count	2
Compound Is Canonicalized	Yes

▸ PubChem

https://pubchem.ncbi.nlm.nih.gov/compound/19096565

Accesible

A1. (Meta)data are retrievable by their identifier using a standardised communications protocol

- A1.1 The protocol is open, free, and universally implementable
- A1.2 The protocol allows for an authentication and authorisation procedure, where necessary

A2. Metadata are accessible, even when the data are no longer available

A1.1 The protocol is open, free, and universally implementable

“Anyone with a computer and an internet connection can access at least the metadata”

HTTP

- Open – specification of the protocol is known to everyone
- Free – no need to pay to “use Internet”

Proprietary protocols

- evade

OSI model

Layer	Name	Example protocols
7	Application Layer	HTTP, FTP, DNS, SNMP, Telnet
6	Presentation Layer	SSL, TLS
5	Session Layer	NetBIOS, PPTP
4	Transport Layer	TCP, UDP
3	Network Layer	IP, ARP, ICMP, IPSec
2	Data Link Layer	PPP, ATM, Ethernet
1	Physical Layer	Ethernet, USB, Bluetooth, IEEE802.11

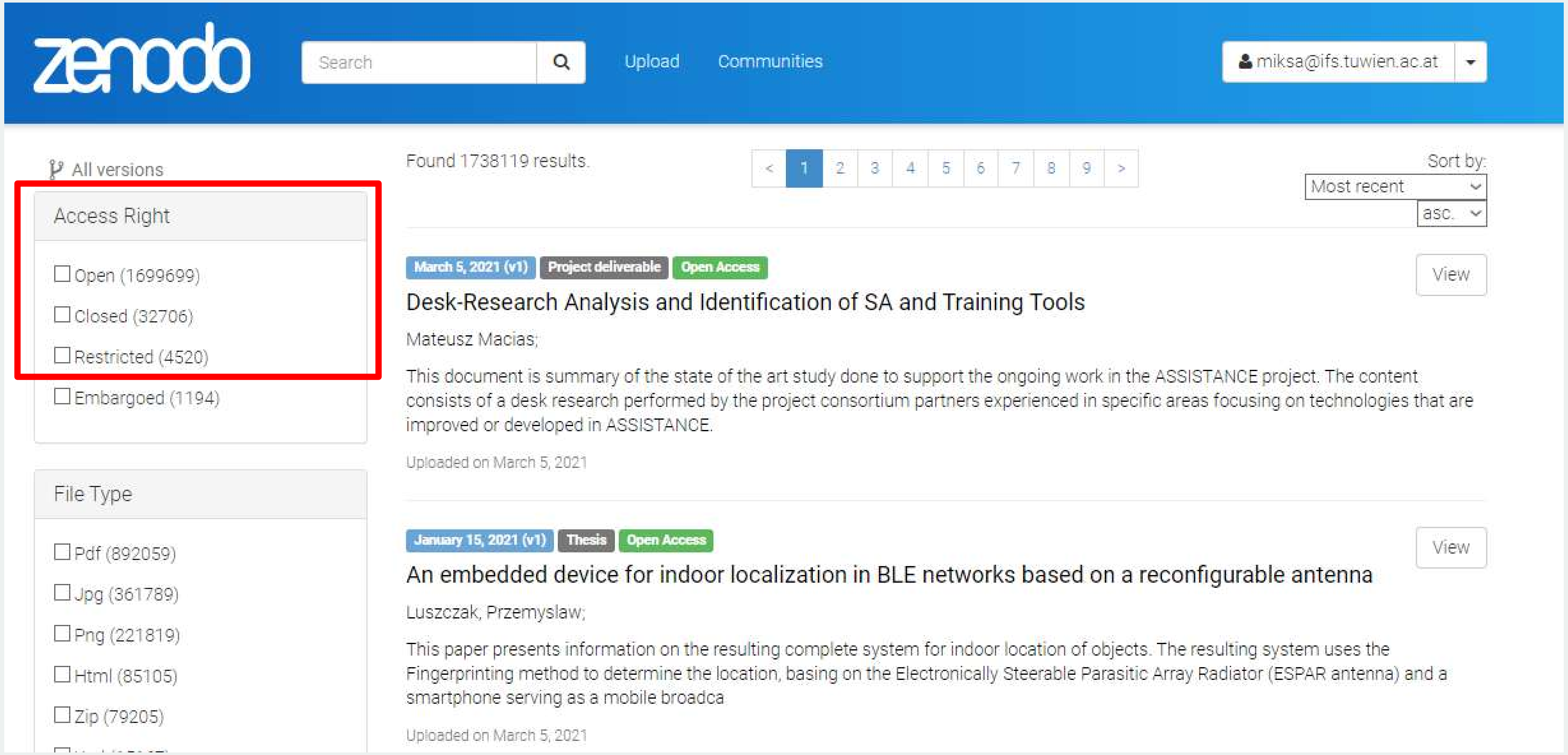
A1.2 protocol allows for authentication and authorisation

Protected and private data can be FAIR

Possible types of access

- Open – everyone has access
- Shared or restricted – only a selected/ invited group of people can access
- Closed – only the owner has access

Accessible - example



https://zenodo.org/search?page=1&size=20&q=

Accessible - example

zenodo Search Upload Communities

March 5, 2021

Dataset Restricted Access

Phase-contrast X-ray tomography of free-breathing murine lungs

Kian Shaker; Ilian Häggmark; Jakob Reichmann; Marie Arsenian-Henriksson; Hans M. Hertz

Full resolution phase-contrast X-ray tomographic dataset acquired of the lungs of a free-breathing mouse (NMRI nude mice, BomTac:NMRI-*Foxn*^{nu}, Taconic Biosciences, DK), weighing roughly 28 g.

Datasize: 3851x3951 pixels per slice, stack of 1700 slices, 16-bit, .tif

Voxel-size: 8.25x8.25x8.25 micrometer

A sample slice is available for download, prior to downloading the full dataset.

If you use the dataset, please cite:

"Phase-contrast X-ray tomography of free-breathing mice reveals the tracheobronchial tree", Kian Shaker, Ilian Häggmark, Jakob Reichmann, Marie Arsenian-Henriksson, and Hans M. Hertz, 2021 (under review)

Files

Restricted Access

You may request access to the files in this upload, provided that you fulfil the conditions below. The decision whether to grant/deny access is solely under the responsibility of the record owner.

Dataset is available upon request. Please state the purpose of your request and the intended usage of the data.

Request access...

Accessible (A1) - comments

FAIR data =/= open data!

- Common misconception stemming from A1.1
- FAIR data *can* be open data, but it has nothing to do with A1

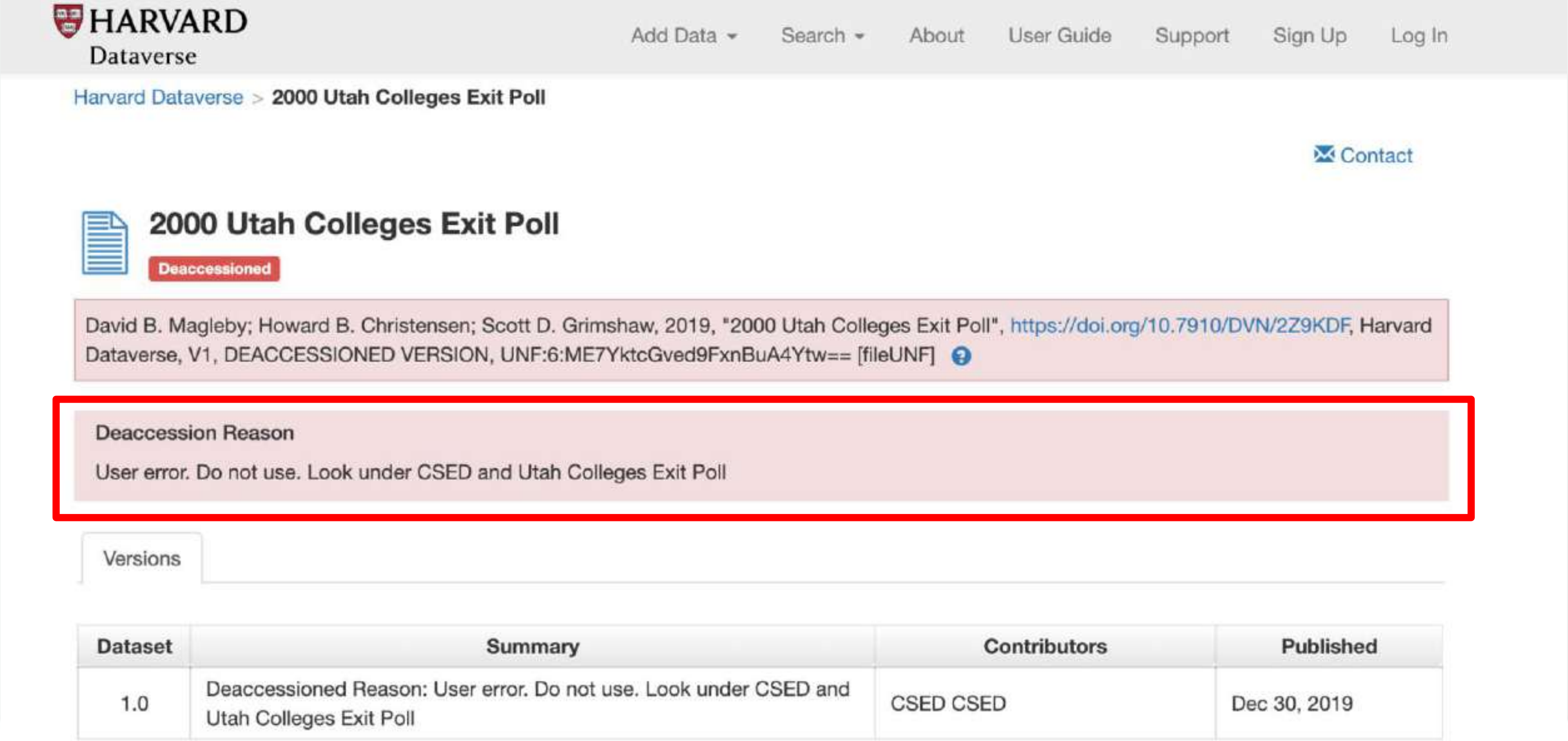
Access can/should be realized over APIs

- SPARQL endpoints
- HTTP APIs
- Client libraries

Access is not only “click to download”

Tombstone pages (A2)

Metadata is accessible, even when the data is no longer available



Interoperable

I1. (Meta)data use a formal, accessible, shared, and broadly applicable **language for knowledge representation**.

I2. (Meta)data use **vocabularies** that follow FAIR principles

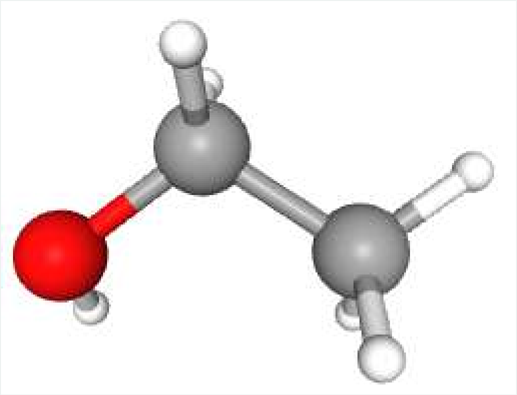
I3. (Meta)data include **qualified references** to other (meta)data

I1. (Meta)data use a formal, accessible, shared, and broadly applicable **language for knowledge representation**

“Data that should be readable for machines without the need for specialised or ad hoc algorithms, translators, or mappings”

Use:

- Common formats
  - RDF, JSON (+schema),
  - CSV (+ good README)
- Well defined/described data models
- Known representations
  - e.g. InChi Key: IDGUHHHQCWSQLU-UHFFFAOYSA-N



I2: (Meta)data use **vocabularies** that follow the FAIR principles

Help evade ambiguities

“My plane lands in London…” – where exactly?

County	ICAO	IATA	Airport name	Usage
Greater London	EGKB	BQH	*London Biggin Hill Airport*	Public
Greater London	EGML		Damyns Hall Aerodrome	Private
Greater London	EGLL	LHR	***Heathrow Airport***	Public
Greater London	EGWU	NHT	*RAF Northolt*	Military
Greater London	EGLC	LCY	***London City Airport***	Public
Greater London	EGLW		*London Heliport*	Public

Controlled vocabularies: IATA and ICAO

https://en.wikipedia.org/wiki/List_of_airports_in_the_United_Kingdom_and_the_British_Crown_Dependencies

Vocabularies

- Tag units of information to make search and retrieval eaiser
- No need to be an expert in car manufacturers to rent a car worldwide
- Example: IDAD – Intermediate category, 4/5 doors, automatic transmission, diesel engine, air-conditioning fitted

Category	Type	Trans / Driven wheels	Fuel / air-con
M: Mini	B: 2-3 Door	M: Manual (drive unspecified)	R: Unspecified Fuel With Air
N: Mini Elite	C: 2/4 Door	N: Manual 4WD	N: Unspecified Fuel Without Air
E: Economy	D: 4-5 Door	C: Manual AWD	D: Diesel Air
H: Economy Elite	W: Wagon/Estate	A: Auto (drive unspecified)	Q: Diesel No Air
C: Compact	V: Passenger Van	B: Auto 4WD	H: Hybrid Air
D: Compact Elite	L: Limousine	D: Auto AWD	I: Hybrid No Air
I: Intermediate	S: Sport		E: Electric Air
J: Intermediate Elite	T: Convertible		C: Electric No Air
S: Standard	F: SUV		L: LPG/Compressed Gas Air
R: Standard Elite	J: Open Air All Terrain		S: LPG/Compressed Gas No Air
F: Fullsize	X: Special		A: Hydrogen Air
G: Fullsize Elite	P: Pick up Regular Cab		B: Hydrogen No Air
P: Premium	Q: Pick up Extended Cab		M: Multi Fuel/Power Air
U: Premium Elite	Z: Special Offer Car		F: Multi Fuel/Power No Air
L: Luxury	E: Coupe		V: Petrol Air
W: Luxury Elite	M: Monospace		Z: Petrol No Air
O: Oversize	R: Recreational Vehicle		U: Ethanol Air
X: Special	H: Motor Home		X: Ethanol No Air
	Y: 2 Wheel Vehicle		
	N: Roadster		
	G: Crossover		
	K: Commercial Van/Truck		

https://en.wikipedia.org/wiki/Association_of_Car_Rental_Industry_Systems_Standards

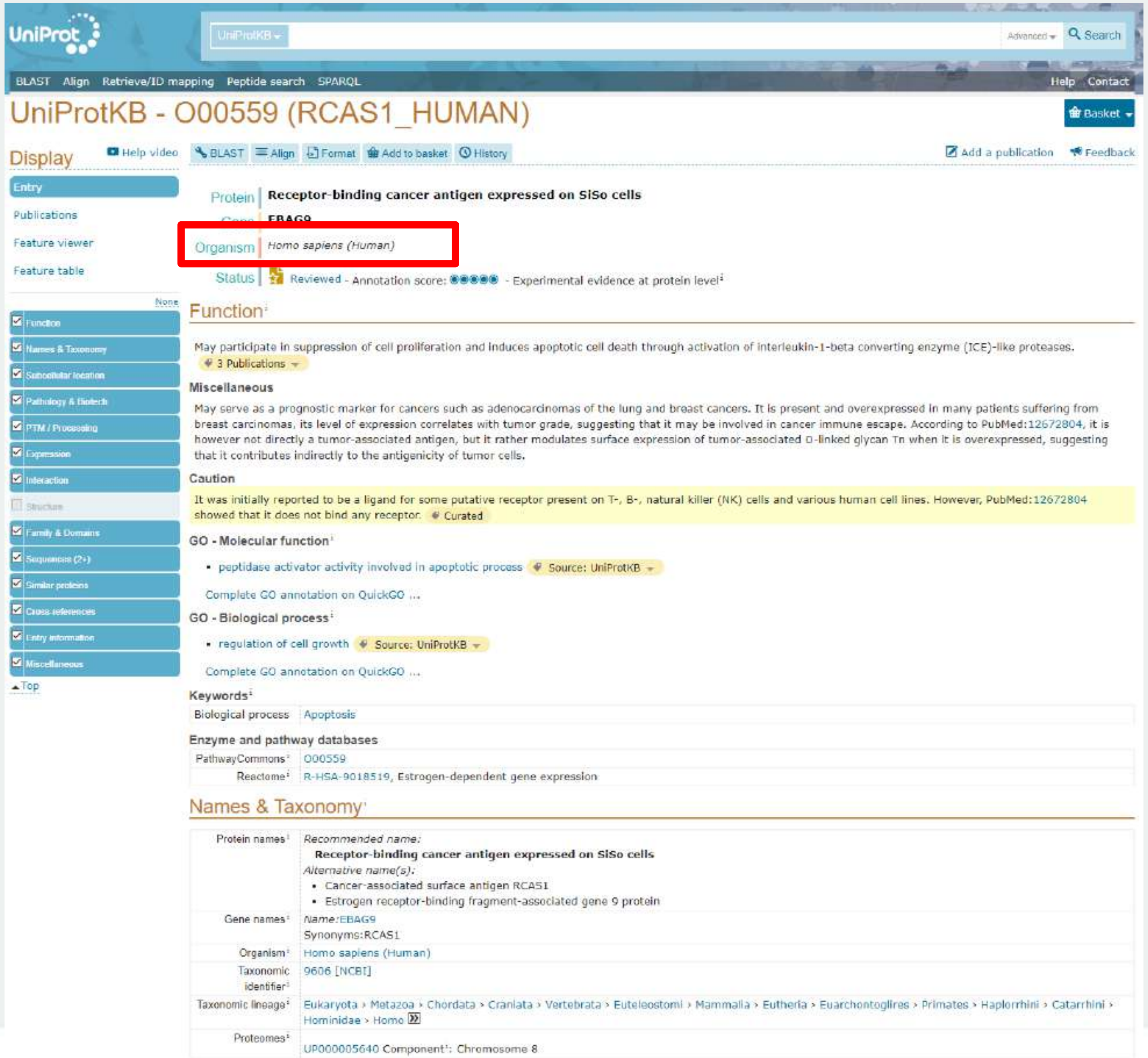
Vocabularies

Less time/money spent on data cleaning

- Different languages
- Spelling mistakes
- Abbreviations
- Capital letters

Vienna	**Beč** (Croatian, Serbian, older Bulgarian), **Beç** (older Turkish)*, **Bech** or **Vidnya** (Romani), **Bécs** (Hungarian)*, **Bin** / **Pin** - 빈 (Korean), **Dunaj** (Slovene)*, **Fienna** (Welsh), **Vedunia** (Celtic), **Vena** - Вена (Russian), **Vídeň** (Czech)*, **Viden'** / **Videň** (Ukrainian)*, **Viedeň** (Slovak), **Viên** (Vietnamese), **Viena** / **Vijena**/ **Виена** (Belarusian, Bulgarian, Macedonian), **Viena** (Catalan*, Lithuanian, Portuguese*, Romanian*, Spanish*, Tagalog*), **Vienna** (Italian)*, **Vienne** (French)*, **Viénni** - *Βιέννη* (Greek), **Vieno** (Esperanto), **Viin** (Estonian), **Vin** - ווין (Yiddish), **Vín** (Irish, Icelandic), **Vina** - וינה (Hebrew), **Vínarborg** (Icelandic variant), **Vindobona** (Latin), **Vīne** (Latvian)*, **Viyana** (Turkish)*, **Vjenë** (Albanian), **Vjenna** (Maltese), **Vyana** (Azeri), **Wean** (local Viennese, Austrian and Bavarian dialects)*, **Weiyena** - 維也納 (Chinese)*, **Wene** (Afrikaans), **Wenen** (Dutch)*, **Wiedeń** (Polish)*, **Wien** (Danish*, Finnish*, German*, Norwegian*, Swedish*), **Wīn** - ウィーン (Japanese)*, **Wina** (Indonesian), فيينا (Arabic), وین (Persian)

Vocabularies (I2)



https://www.uniprot.org/uniprot/O00559

Vocabularies (I2)

Each metadata field has its definition

Organism

Last modified April 10, 2018

This subsection of the Names and taxonomy section provides information on the name(s) of the organism that is the source of the protein sequence.

The organism designation consists of the Latin scientific name, usually composed of the genus and species names (the binomial system developed by Linnaeus), followed optionally by the English common name and a synonym.
Examples: Bacillus subtilis, Homo sapiens (Human), Cardamine pratensis (Cuckoo flower) (Alpine bitter cress)

The synonym can be a common name in English (or in Latin in the case of some historical legacy names).
Example: Radianthus magnifica (Magnificent sea anemone) (Heteractis magnifica)

In the case of viruses, the designation does not follow the binomial system. The English common name is used as the scientific name, sometimes followed by an acronym. When possible, viruses are named according to the nomenclature of the International Committee on Taxonomy of Viruses (ICTV).
Examples: Human immunodeficiency virus type 1 (isolate BRU/LAI group M subtype B) (HIV-1) , Influenza A virus (strain A/Aichi/2/1968 H3N2)

The organism name can differ from that given by the international nucleotide sequence databases for the same taxon. This is mainly due to our efforts in providing the most descriptive common names and synonyms to our users.

Note that the proteome for a given organism, when available, can be accessed through the proteomes page of our website.

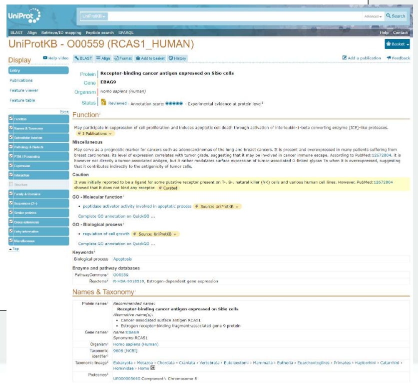
Related documents

Taxonomy

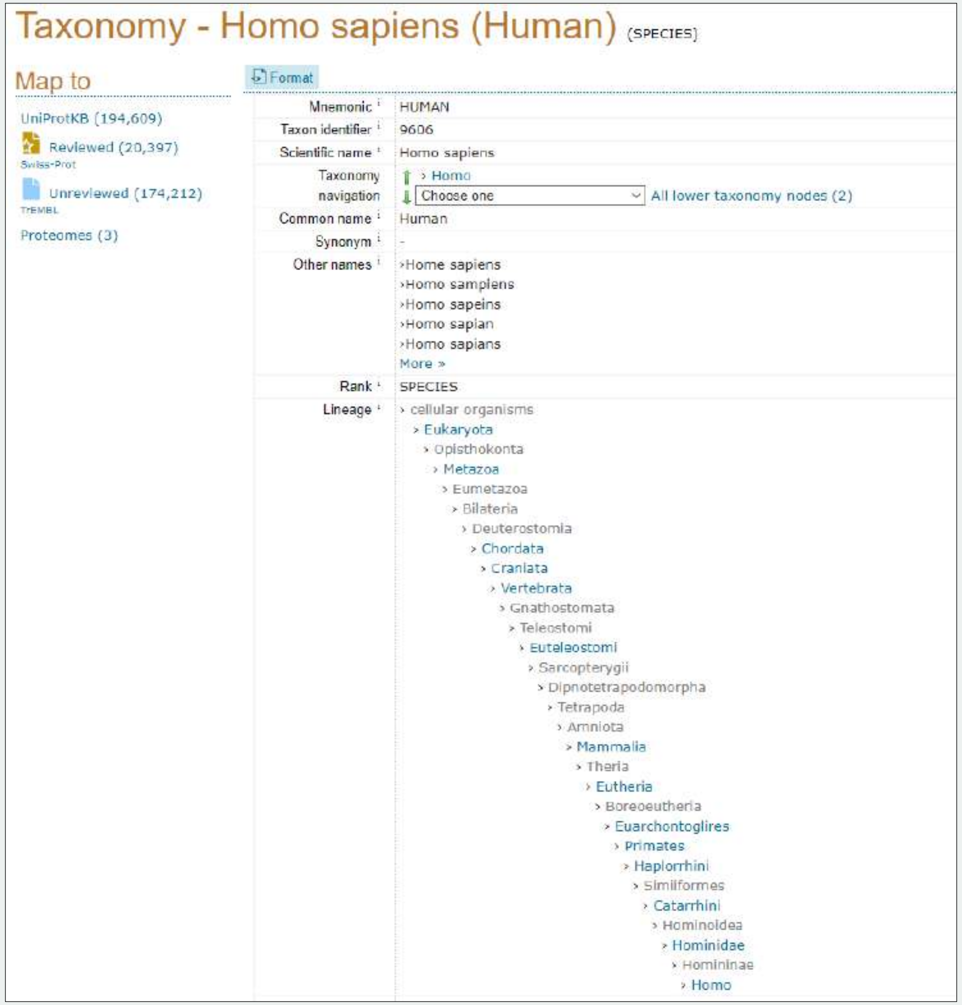
Controlled vocabulary of species

What are proteomes?

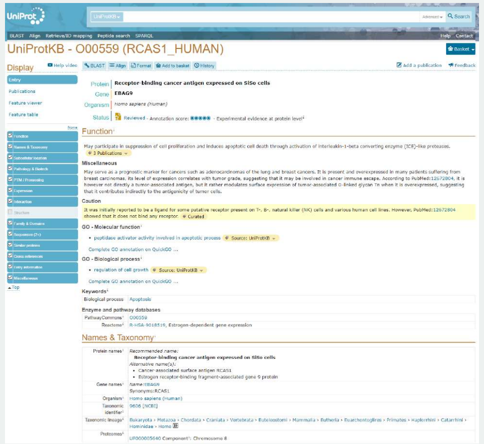
What are reference proteomes?



Vocabularies (I2)



Each metadata **value** comes from a controlled vocabulary – no free form answers.



Qualified References (I3)

Meaningul links to describe connections

- Dataset X was *derived from* dataset Y
- Dataset Y was *produced* using code Z

Standard relations define by Data Cite

- https://schema.datacite.org/meta/kernel-4.3/

Use persistent identifiers


```xml
<language>en</language>
<resourceType resourceTypeGeneral="Workflow">Software</resourceType>
<relatedIdentifiers>
  <relatedIdentifier relatedIdentifierType="DOI"
  relationType="IsReferencedBy">10.5072/2047-217X-1-1</relatedIdentifier>
  <relatedIdentifier relatedIdentifierType="DOI"
  relationType="Compiles">10.5072/100038</relatedIdentifier>
</relatedIdentifiers>
<sizes>
  <size>31 MB</size>
</sizes>
```


https://schema.datacite.org/meta/kernel-4.3/example/datacite-example-workflow-v4.xml

IsContinuedBy
Continues
IsDescribedBy
Describes
HasMetadata
IsMetadataFor
HasVersion
IsVersionOf
IsNewVersionOf
IsPreviousVersionOf
IsPartOf
HasPart
IsReferencedBy
References
IsDocumentedBy
Documents
IsCompiledBy
Compiles
IsVariantFormOf
IsOriginalFormOf
IsIdenticalTo
IsReviewedBy
Reviews
IsDerivedFrom
IsSourceOf
IsRequiredBy
Requires
IsObsoletedBy
Obsoletes

Qualified References (I3)

TU WIEN

Search

Log in

January 19, 2021 | Version 1.0

Dataset

Open Access

European Sentinel-1 Forest Type and Tree Cover Density Maps

Dostalova, Alena [1]; Cao, Senmao [1,2]; Wagner, Wolfgang [1,2] show affiliations

Description

This dataset was generated by the TU Wien Department of Geode

European Sentinel-1 forest type and tree cover density maps repre
Sentinel-1 C-Band Synthetic Aperture Radar (SAR) backscatter da
European continent with 10 m and 100 m sampling for forest type
derived using the method described in https://www.tandfonline.con

The forest type map shows the dominant forest type class (conifer
percentage of forest canopy cover within the 100 m pixel.

Please be referred to our peer-reviewed article at https://doi.org/10
assessment accross Europe.

Dataset Record

The forest type and tree cover density maps are sampled at 10 m
georeferenced to the Equi7Grid and divided into square tiles of 10
maps consist of 728 tiles over the European continent, with data v

The tiles' file-format is a LZW-compressed GeoTIFF holding 16-bit
and georeference. Compatibility with common geographic informat
libraries as GDAL is given.

In this repository, we provide each forest map as tiles, whereas tw
download below.

Code Availability

For the usage of the **Equi7Grid** we provide data and tools via the python package available on GitHub at https://github.com/TUW-GEO/Equi7Grid. More details on the grid reference can be found in https://www.sciencedirect.com/science/article/pii/S0098300414001629.

Acknowledgements

The computational results presented have been achieved using the Vienna Scientific Cluster (VSC).

Versions

Version 1.0
DOI: 10.48436/tkkfs-11b75

Cite As

Dostalova, Alena, Cao, Senmao & Wagner, Wolfgang. (2021). European Sentinel-1

Details

Licenses		
Resource type	Dataset	
Formats	application/x-geotiff	
Related identifiers	**isreferencedby**	10.3390/rs13030337 (doi)
	issupplementto	10.5281/zenodo.3515933 (doi)
		https://github.com/TUW-GEO/Equi7Grid (url)
	references	10.1080/01431161.2018.1479788 (doi)
		10.1016/j.cageo.2014.07.005 (doi)

Paper citing this dataset

Code

Paper describing the method to produce this dataset

https://researchdata.dl.hpc.tuwien.ac.at/records/tkkfs-11b75

Qualified References (I3)

README.rst

Equi7Grid

build passing | coverage 32% | pypi package 0.0.12 | docs passing

A python class for working with Equi7Grid - how to convert to - how to use the tiling system - etc.

It's a python package that handles the geometric and geographic operations of a gridded and tiled projection system. It was designed for data cubes ingesting satellite imagery and builds the basis for the Equi7Grid (see https://github.com/TUW-GEO/Equi7Grid).

A detailed documentation on the Equi7Grid definition is at:

~/docs/doc_files/

Overlays for visualisation in Google Earth can be found here:

~/docs/doc_files/google_earth_overlays/

Citation

DOI 10.5281/zenodo.1048530

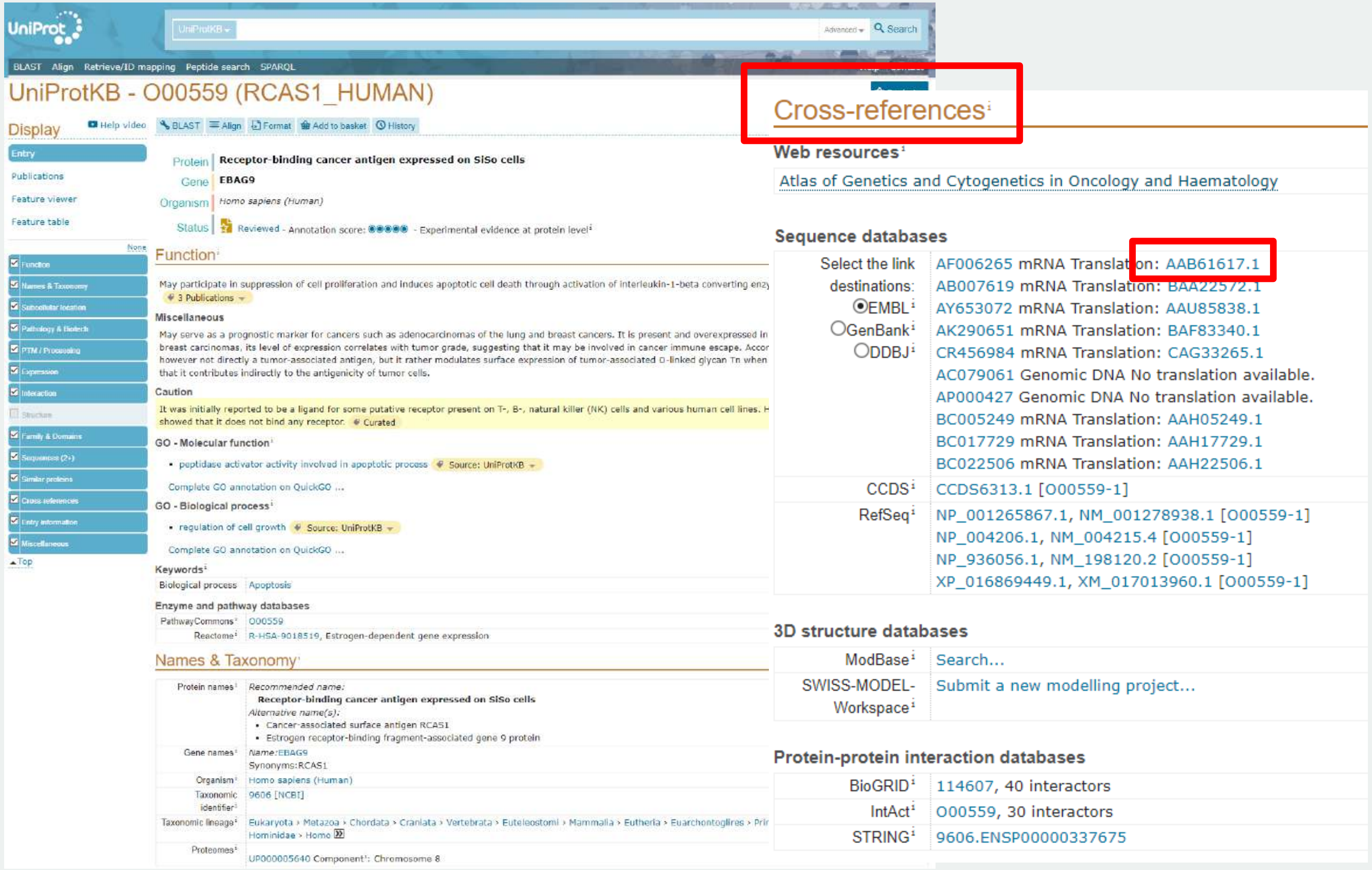
If you use the software in a publication then please cite it using the Zenodo DOI. Be aware that this badge links to the latest package version.

Please select your specific version at https://doi.org/10.5281/zenodo.1048530 to get the DOI of that version. You should normally always use the DOI for the specific version of your record in citations. This is to ensure that other researchers can access the exact research artefact you used for reproducibility.

You can find additional information regarding DOI versioning at http://help.zenodo.org/#versioning

https://github.com/TUW-GEO/Equi7Grid

Qualified References (I3)



Reusable

R1. (Meta)data are richly described with a plurality of accurate and relevant attributes

- R1.1. (Meta)data are released with a clear and accessible data usage **license**
- R1.2. (Meta)data are associated with detailed **provenance**
- R1.3. (Meta)data meet domain-relevant **community standards**

R1.1. (Meta)data are released with a clear and accessible data usage license

Public repository on GitHub

- May suggest that authors are willing to share code

No license

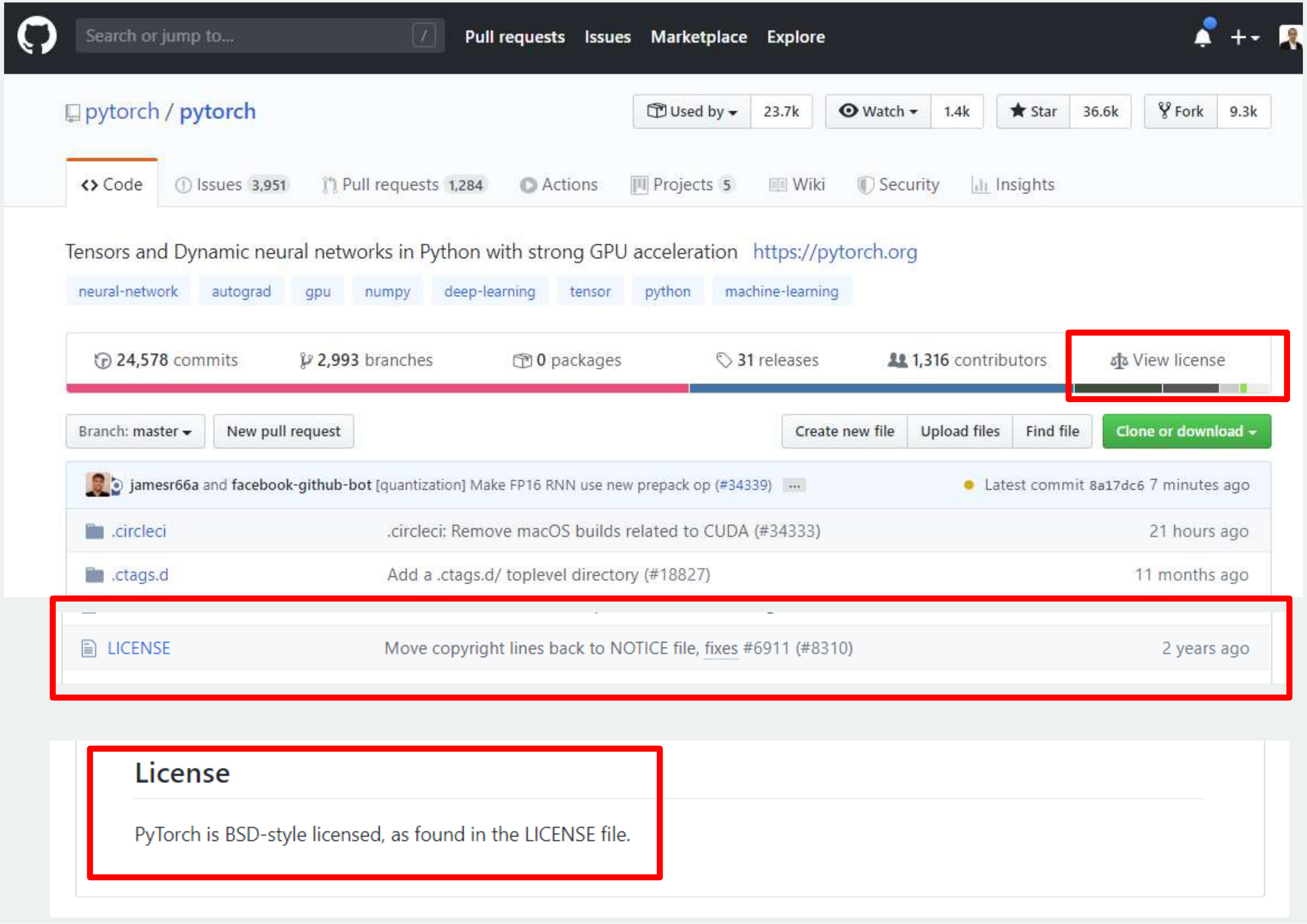
- no possibility for reuse
- can only by viewed (only because terms of use enforce that)

Code without a license is like an object in a museum

- You can watch and admire it, but you cannot touch it!

More on licenses in other lectures

License (R1.1)


Search or jump to...
Pull requests
Issues
Marketplace
Explore
pytorch / pytorch
Used by 23.7k
Watch 1.4k
Star 36.6k
Fork 9.3k
Code
Issues 3,951
Pull requests 1,284
Actions
Projects 5
Wiki
Security
Insights
Tensors and Dynamic neural networks in Python with strong GPU acceleration https://pytorch.org
neural-network autograd gpu numpy deep-learning tensor python machine-learning
24,578 commits
2,993 branches
0 packages
31 releases
1,316 contributors
View license
Branch: master
New pull request
Create new file
Upload files
Find file
Clone or download
jamesr66a and facebook-github-bot [quantization] Make FP16 RNN use new prepack op (#34339)
Latest commit 8a17dc6 7 minutes ago
.circleci
.circleci: Remove macOS builds related to CUDA (#34333)
21 hours ago
.ctags.d
Add a .ctags.d/ toplevel directory (#18827)
11 months ago
LICENSE
Move copyright lines back to NOTICE file, fixes #6911 (#8310)
2 years ago
License
PyTorch is BSD-style licensed, as found in the LICENSE file.

R1.2 (Meta)data are associated with detailed provenance

Provenance

- Describes origin of data
- Who? What? When? How?

Supports evaluation and can build trust in data

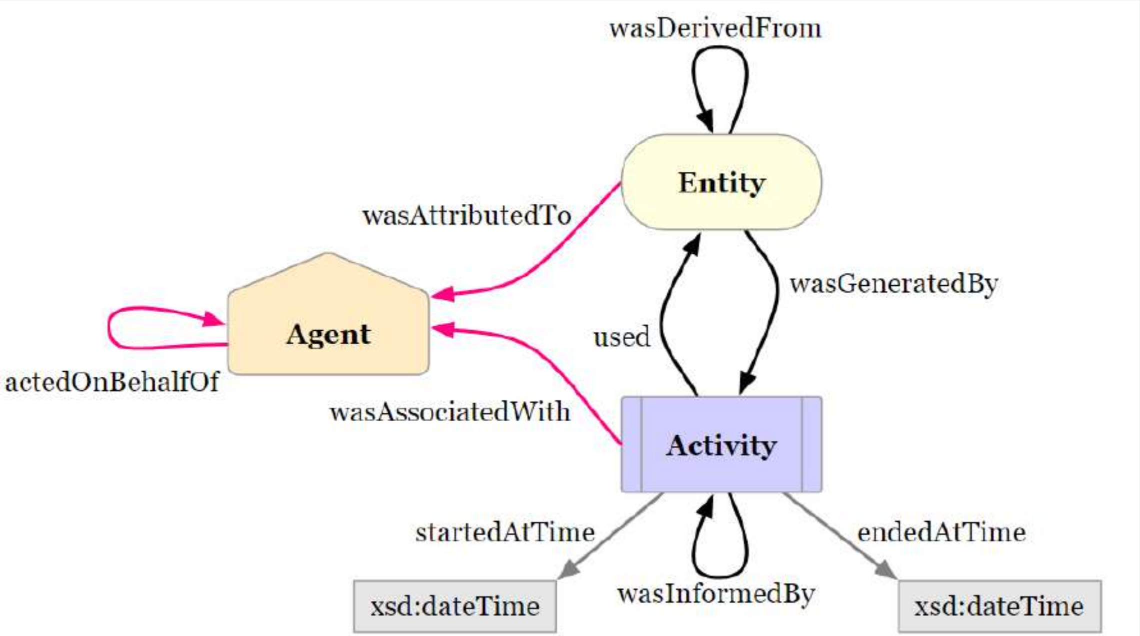
- ‘Officially, North Korea claims to have identified zero cases of COVID-19 inside its territory’

https://www.npr.org/sections/goatsandsoda/2020/02/20/807027901/north-korea-claims-zero-cases-of-coronavirus-infection-but-experts-are-skeptical?t=1615196582563

R1.2 (Meta)data are associated with detailed provenance

PROV-O: The PROV Ontology

- Machine-actionable way to express provenance

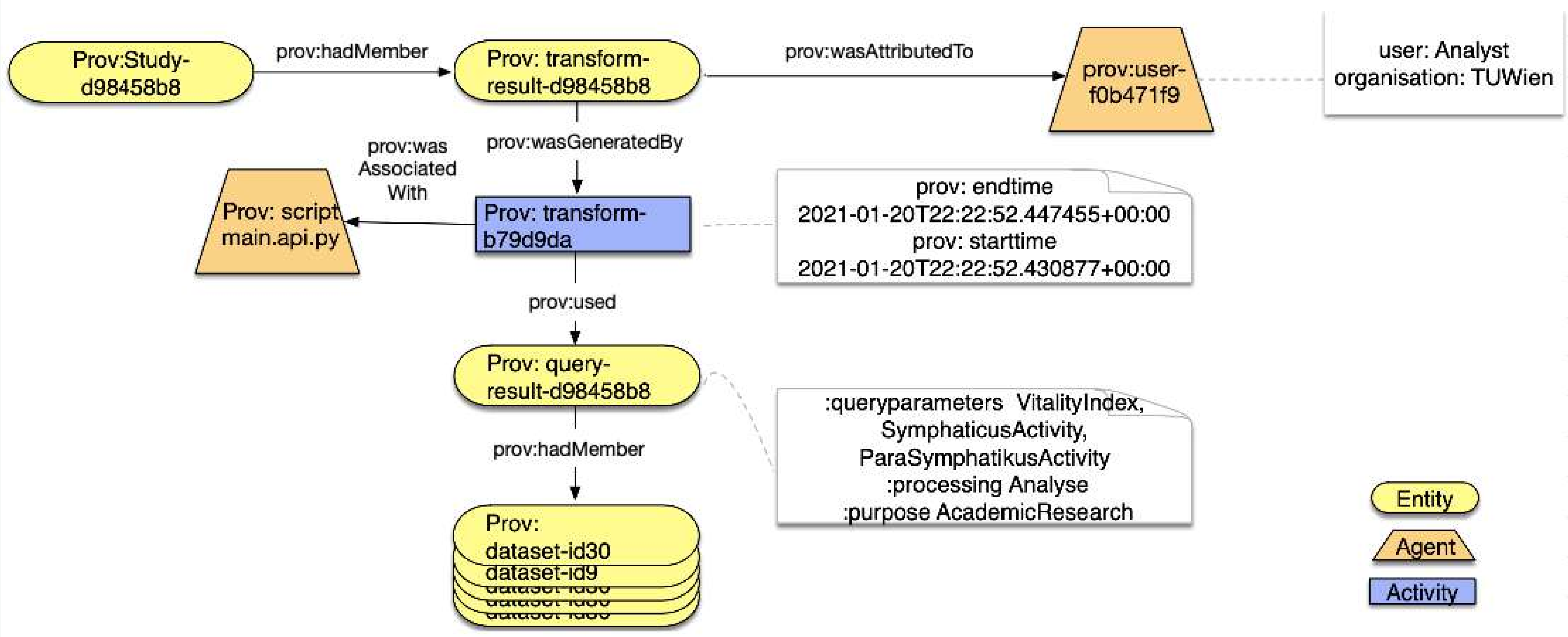


https://www.w3.org/TR/prov-o/

R1.2 (Meta)data are associated with detailed provenance

Example of PROV-O Instance

- ***Analyst** from TU Wien made a **study** in which data was **transformed**.*
- *To do so, a **script** (…) was executed to **query** data, using **parameters** (…) and following **datasets** were used (…).*



R1.3. (Meta)data meet domain-relevant community standards

Who is the “community” ?

What is the “standard” ?

- English vs other languages

Metadata

- Domain independent
  - e.g. Dublin Core
- Domain specific
  - e.g. EXIF for images

Sometimes no common standard exist

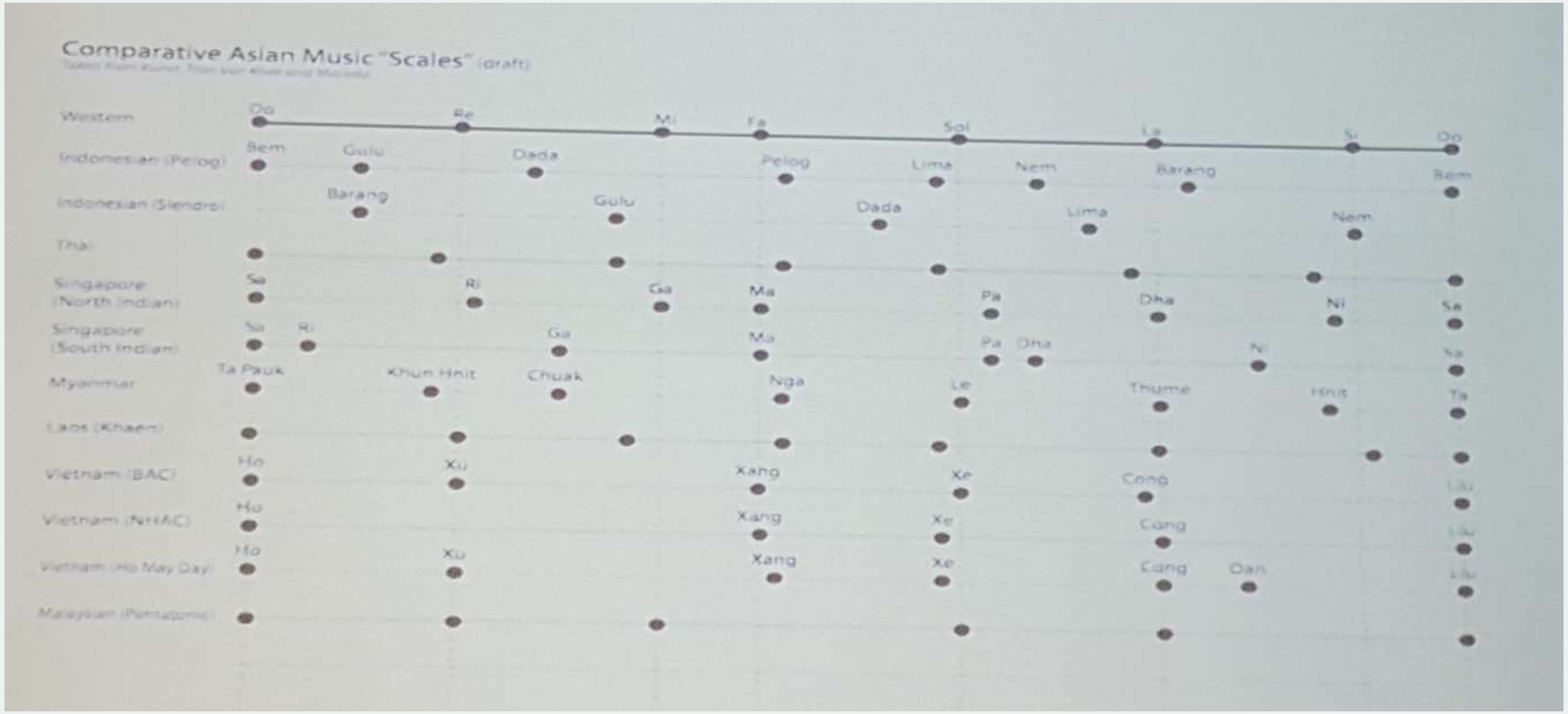
- Good documentation and README

There is no universal guideline – it always depends!

R1.3. (Meta)data meet domain-relevant community standards

Do Re Mi Fa Sol La Si Do

- Does not have to be a standard for everyone!

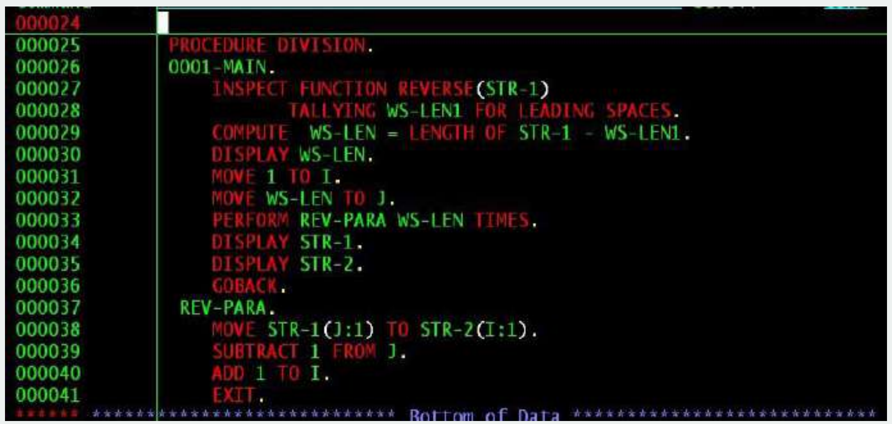


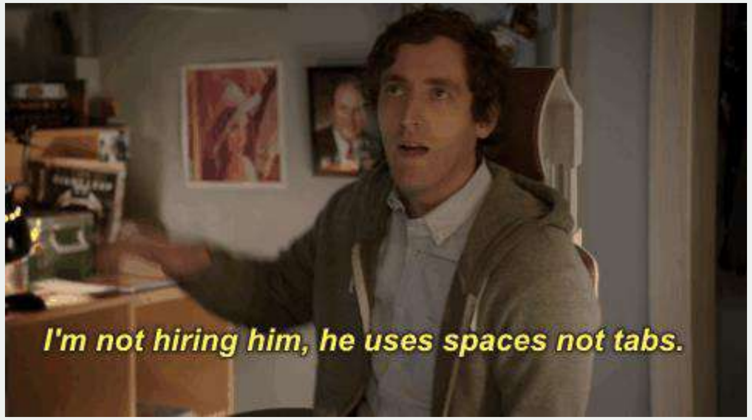
R1.3. (Meta)data meet domain-relevant community standards

Follow standards and <u>domain specific </u>conventions

Examples

- Sharing COBOL code
  - with Data Science students?
  - with mainframe operators?
- Obfuscated code/data





R1.3. (Meta)data meet domain-relevant community standards

Good documentation supports reuse

- Removes ambiguities (especially where there are no common controlled vocabularies or others standards)

Example

- Confirmed cases of COVID-19: testing date vs reporting date

Indicators	Definition
Tests	Cumulative number of tests carried out for SARS-CoV-2, from 27 February 2020 up to and including the reporting date. Responsible for data consolidation: Office of the respective federal state government (Land), data status: morning of the reporting day.
Laboratory-confirmed cases	Cumulative number of laboratory-confirmed cases of SARS-CoV-2 infection (sum of "Active cases", "Recovered cases" and "Deceased cases") with laboratory diagnosis date since 27.02.2020 up to and including the reporting date.
Active cases	Cumulative number of laboratory-confirmed cases of SARS-CoV-2 infection with laboratory diagnosis date from 27.02.2020 up to and including the reporting date, which have not been classified as "recovered" or "deceased" on the reporting date.
Recovered cases	Cumulative number of laboratory-confirmed cases of SARS-CoV-2 infection with laboratory diagnosis date from 27.02.2020 up to and including the report date, which are classified as "recovered" on the report date. Definition of "recovered" (since 9 July): in the case of home care, 10-day home isolation after the onset of symptoms or laboratory diagnosis; in case of severe disease progression, the earliest 10 days after onset of symptoms, at least 48 hours without symptoms AND the following result by RT-PCR according to the Charité protocol: no nucleic acid detection of beta-coronavirus SARS-CoV-2 or nucleic acid detection of beta-coronavirus SARS-CoV-2 at a Ct value of more than 30. Further details can be found in the recommendation for the release of COVID-19 cases, recommendation for the release of COVID-19 cases from isolation.
Deceased cases	Cumulative number of laboratory-confirmed cases of SARS-CoV-2 infection with a laboratory diagnosis date from 27.02.2020 up to and including the report date, which are classified as "deceased" on the report date.

https://covid19-dashboard.ages.at/basisinfo_en.html?l=en

FAIR ASSESSMENT

FAIR Assessment

Still pretty much work in progress

Most approaches

- Try to quantify FAIRness
- Based on self-assessment
  - https://ardc.edu.au/resources/working-with-data/fair-data/fair-self-assessment-tool/
  - https://fairaware.dans.knaw.nl

FAIR Aware

Australian Research Data Commons

FAIR Aware – self-assessment example

FAIR questions

FINDABLE

1. Are you aware that a dataset should be assigned a globally unique persistent and resolvable identifier when deposited with a data repository? — Yes / No
2. Are you aware that when you deposit a dataset with a repository, you will need to provide some details (known as discovery metadata) in order to make the data findable, understandable and reusable to others? — Yes / No
3. Are you aware that the repository providing access to your dataset should make the metadata describing your datasets available in a format readable by machines as well as humans? — Yes / No

ACCESSIBLE

4. Are you aware that access to your dataset may need to be controlled and that metadata should include licence information under which the data can be reused? — Yes / No
5. Are you aware that metadata should remain available over time, even if the data is no longer accessible? — Yes / No

INTEROPERABLE

6. Are you aware that the metadata describing your datasets should use semantic vocabularies? — Yes / No

REUSABLE

7. Are you aware that provenance information about the collection and/or generation of data should be included in the metadata? — Yes / No
8. Are you aware that metadata describing your data should follow the specifications of a community-endorsed standard? — Yes / No
9. Are you aware that data should be deposited preferably in a file format that is open – to support reuse – and supported by the repository for long-term preservation? — Yes / No
10. Are you aware that maintaining your dataset FAIR over time requires professional data curation and preservation? — Yes / No

Findable

Does the dataset have any identifiers assigned?	No identifier
Is the dataset identifier included in all metadata records/files describing the data?	No
How is the data described with metadata?	The data is not described
What type of repository or registry is the metadata record in?	The data is not described in any repository

Accessible

How accessible is the data?	No access to data or metadata
Is the data available online without requiring specialised protocols or tools once access has been approved?	No access to data
Will the metadata record be available even if the data is no longer available?	Unsure

Interoperable

What (file) format(s) is the data available in?	Mostly in a proprietary format
What best describes the types of vocabularies/ontologies/tagging schemas used to define the data elements?	Data elements not described
How is the metadata linked to other data and metadata (to enhance context and clearly indicate relationships)?	There are no links to other metadata

Reusable

Which of the following best describes the license/usage rights attached to the data?	No license
How much provenance information has been captured to facilitate data reuse?	No provenance information is recorded

RDA FAIR Data Maturity Model Specification and Guidelines Recommendation

17 minimum viable metrics to systematically measure the extent to which **research data objects** are FAIR

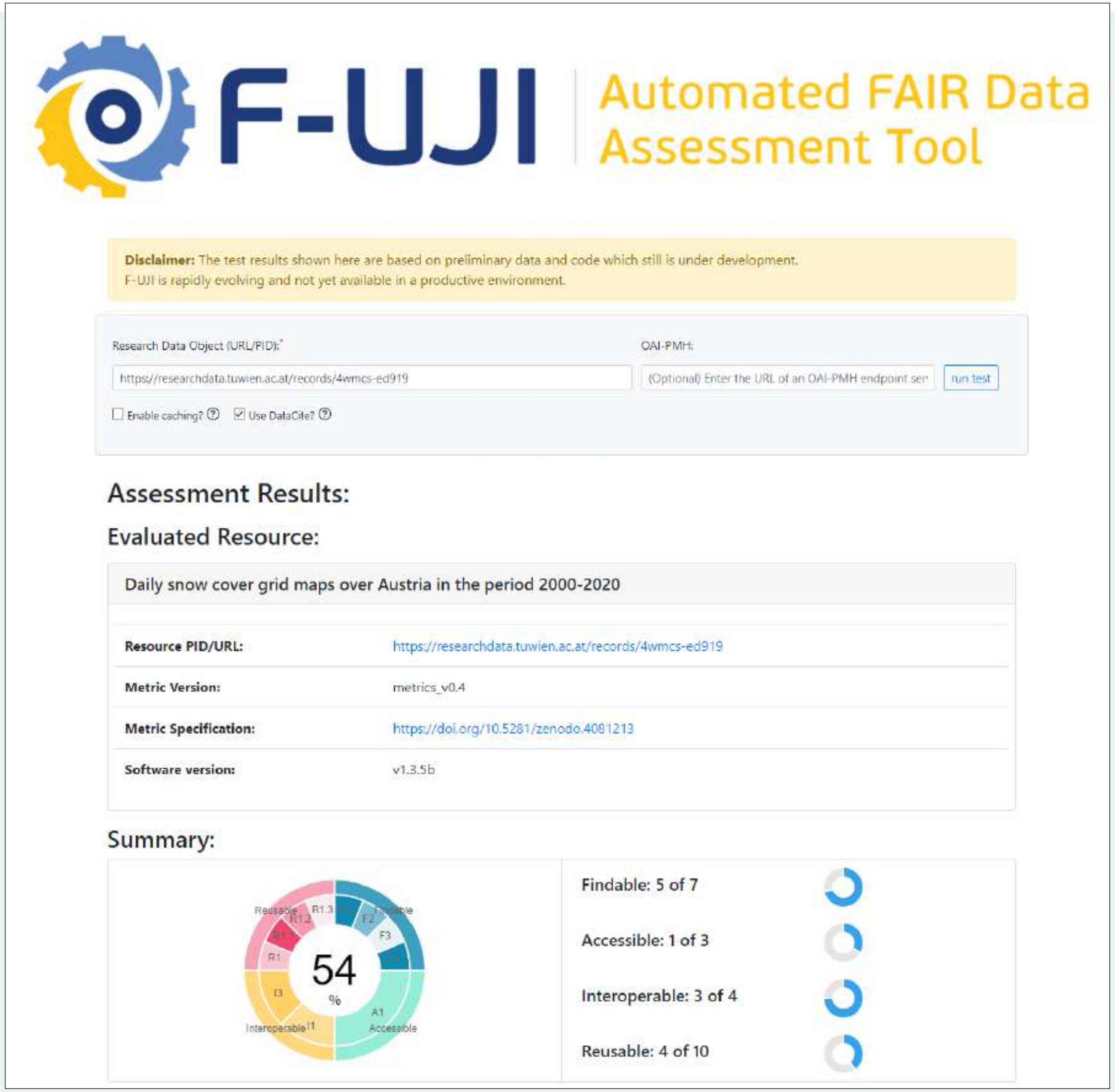
Not systems as a whole!

2.1 Globally Unique Identifier

FIELD	DESCRIPTION
Metric Identifier	FsF-F1-01D
Metric Name	Data is assigned a globally unique identifier.
Description	A data object may be assigned with a globally unique identifier such that it can be referenced unambiguously by humans or machines. Globally unique means an identifier should be associated with only one resource at any time. Examples of unique identifiers of data are Internationalized Resource Identifier (IRI)[16], Uniform Resource Identifier (URI) such as URL and URN, Digital Object Identifier (DOI), the Handle System, identifiers.org, w3id.org and Archival Resource Key (ARK). A data repository may assign a globally unique identifier to your data or metadata when you publish and make it available through its curation service.
FAIR Principle	F1. (Meta) data are assigned globally unique and persistent identifiers
CoreTrustSeal Alignment	R13. The repository enables users to discover the data and refer to them in a persistent way through proper citation
ASSESSMENT	
Requirement(s)	• Data identifier (IRI, URL) • List of globally unique identifier schemes
Method	Check if the identifier is specified based on a globally unique identifier scheme.
COMMENTS	
Related Resources: • Identifiers compiled by FAIRsharing, https://fairsharing.org/standards/?q=&selected_facets=type_exact:identifier%20schema • A list of Uniform Resource Identifier (URI) schemes, available in different formats, https://www.iana.org/assignments/uri-schemes/uri-schemes.xhtml#uri-schemes-1 • Uniform Resource Identifier (URI) Generic Syntax (RFC 3986), https://tools.ietf.org/html/rfc3986	

F-UJI Automated FAIR Data Assessment Tool

http://doi.org/10.5281/zenodo.4081213 (Published: 12 October 2020)



Code still under development!

https://www.f-uji.net/index.php

Report:

Findable

- FsF-F1-01D - Data is assigned a globally unique identifier. ✓
- FsF-F1-02D - Data is assigned a persistent identifier. ✓
- FsF-F2-01M - Metadata includes descriptive core elements (creator, title, data identifier, publisher, publication date, summary and keywords) to support data findability. ✓
- FsF-F3-01M - Metadata includes the identifier of the data it describes. ✗
- FsF-F4-01M - Metadata is offered in such a way that it can be retrieved programmatically. ✓

Accessible

- FsF-A1-01M - Metadata contains access level and access conditions of the data. ✗
- FsF-A1-03D - Data is accessible through a standardized communication protocol. ✗
- FsF-A1-02M - Metadata is accessible through a standardized communication protocol. ✓

Interoperable

- FsF-I1-01M - Metadata is represented using a formal knowledge representation language. ✓
- FsF-I1-02M - Metadata uses semantic resources ✗
- FsF-I3-01M - Metadata includes links between the data and its related entities. ✓

Reusable

- FsF-R1-01MD - Metadata specifies the content of the data. ✗
- FsF-R1.1-01M - Metadata includes license information under which data can be reused. ✓
- FsF-R1.2-01M - Metadata includes provenance information about data creation or generation. ✓
- FsF-R1.3-01M - Metadata follows a standard recommended by the target research community of the data. ✓
- FsF-R1.3-02D - Data is available in a file format recommended by the target research community. ✗

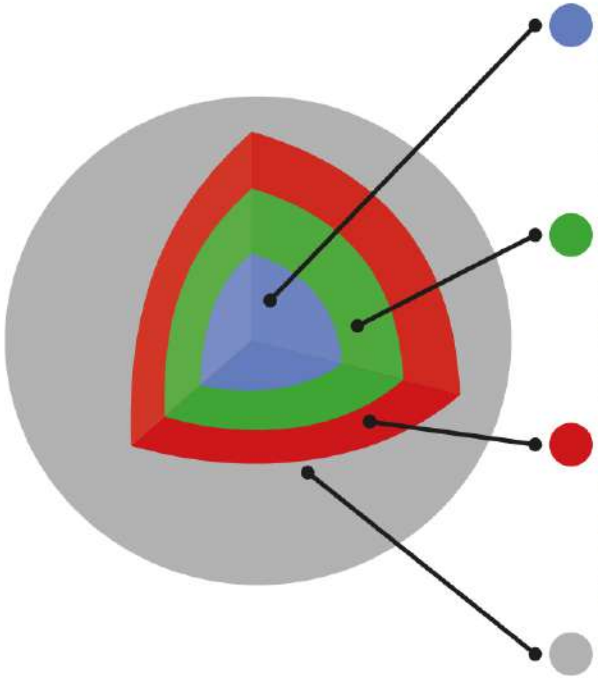
Metric tests:

Test:	Test name:	Result:
FsF-R1-01MD-1	Minimal information about available data content is given in metadata	✓
FsF-R1-01MD-1a	Resource type (e.g. dataset) is given in metadata	✓
FsF-R1-01MD-1b	Information about data content (e.g. links) is given in metadata	✗
FsF-R1-01MD-2	Verifiable data descriptors (file info, measured variables or observation types) are specified in metadata	✗
FsF-R1-01MD-2a	File size and type information are specified in metadata	✗
FsF-R1-01MD-2b	Measured variables or observation types are specified in metadata	✗
FsF-R1-01MD-3	Data content matches file type and size specified in metadata	✗
FsF-R1-01MD-4	Data content matches measured variables or observation types specified in metadata	✗

Debug:

Level:	Message:
INFO	Object landing page accessible status -: True
SUCCESS	Resource type specified -: dataset
WARNING	NO data object content available/accessible to perform file descriptors (type and size) tests
WARNING	NO measured variables found in metadata, skip 'measured_variable' test.
WARNING	Measured variables given in metadata do not match data object content

FAIR Digital Object



DIGITAL OBJECT
Data, code and other research outputs
At its most basic level, data or code is a bitstream or binary sequence. For this to have meaning and to be FAIR, it needs to be represented in standard formats and be accompanied by Persistent Identifiers (PIDs), metadata and documentation. These layers of meaning enrich the object and enable reuse.

IDENTIFIERS
Persistent and unique (PIDs)
Digital Objects should be assigned a unique and persistent identifier such as a DOI or URN. This enables stable links to the object and supports citation and reuse to be tracked. Identifiers should also be applied to other related concepts such as the data authors (ORCIDs), projects (RAIDs), funders and associated research resources (RRIDs).

STANDARDS & CODE
Open, documented formats
Digital Objects should be represented in common and ideally open file formats. This enables others to reuse them as the format is in widespread use and software is available to read the files. Open and well-documented formats are easier to preserve. Data also need to be accompanied by the code use to process and analyse the data.

METADATA
Contextual documentation
In order for Digital Objects to be assessable and reusable, they should be accompanied by sufficient metadata and documentation. Basic metadata will enable data discovery, but much richer information and provenance is required to understand how, why, when and by whom the objects were created. To enable the broadest reuse, they should be accompanied by a plurality of relevant attributes and a clear and accessible usage license.

Figure 8. A model for FAIR Digital Objects, noting the elements that need to be in place for data to be Findable, Accessible, Interoperable and Reusable

Turning FAIR into reality https://op.europa.eu/s/oM5N

RO-Crate as an example of a FAIR Data Object

Currently, no commonly accepted FAIR Data Object

RO-Crate

- packages **research artefacts** along with their **metadata** in a **machine readable** way
- based on Schema.org annotations in JSON-LD
- https://www.researchobject.org/ro-crate/

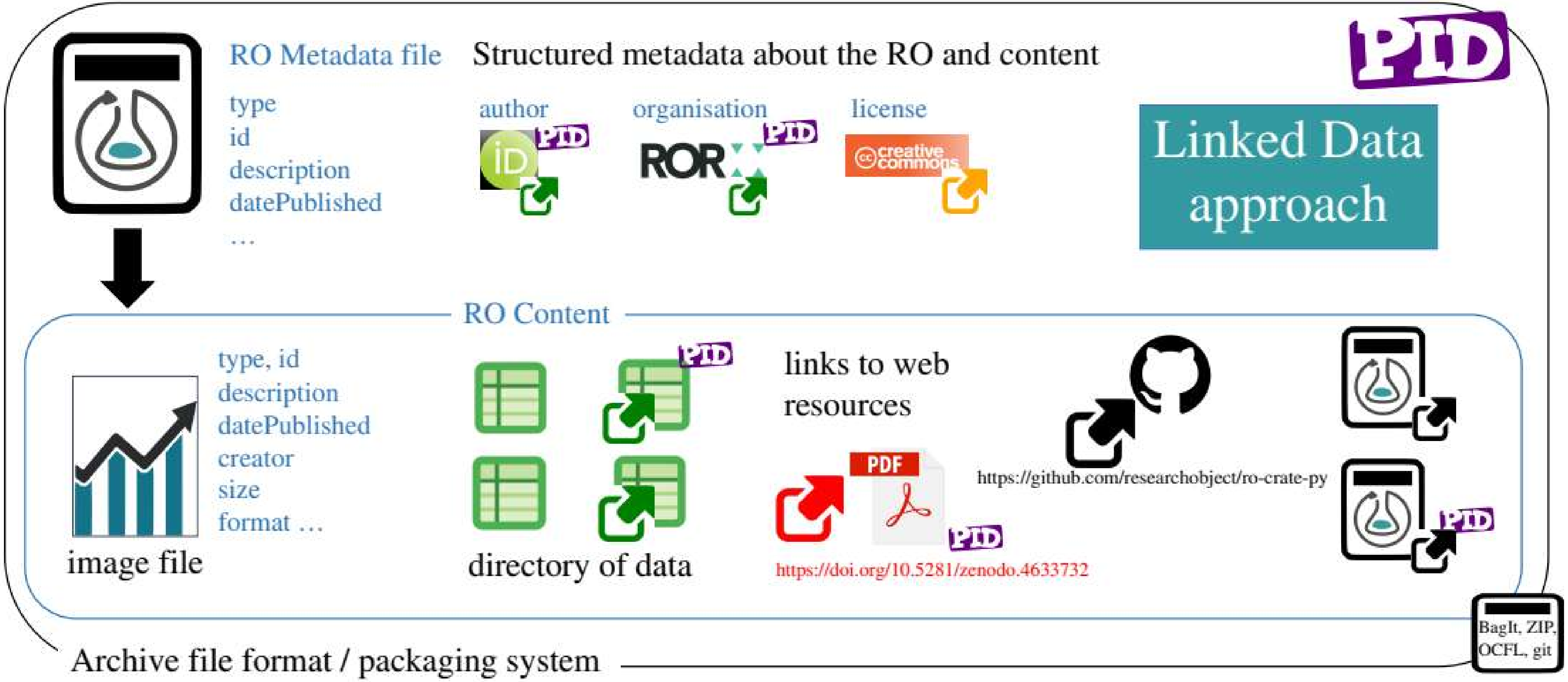


Fig. 1. **Conceptual overview of RO-Crate**. A *Persistent Identifier* (PID) [25] points to a *Research Object* (RO), which may be archived using different packaging approaches like BagIt [26], OCFL [27], git or ZIP. The RO is described within a *RO-Crate Metadata File*, providing identifiers for *authors* using ORCID, *organisations* using Research Organization Registry (ROR) [28] and licences such as Creative Commons using SPDX identifiers. The *RO-Crate content* is further described with additional metadata following a Linked Data approach. Data can be embedded files and directories, as well as links to external Web resources, PIDs and nested RO-Crates.

```json
{ "@context": "https://w3id.org/ro/crate/1.1/context",
  "@graph": [
      { "@id": "ro-crate-metadata.json",
        "@type": "CreativeWork",
        "conformsTo": {"@id": "https://w3id.org/ro/crate/1.1"},
        "about": {"@id": "./"}
      },
      { "@id": "./",
        "@type": "Dataset",
        "name": "A simplified RO-Crate",
        "author": {"@id": "#alice"},
        "license": {"@id": "https://spdx.org/licenses/CC-BY-4.0"},
        "datePublished": "2021-11-02T16:04:43Z",
        "hasPart": [
          {"@id": "survey-responses-2019.csv"},
          {"@id": "https://example.com/pics/5707039334816454031_o.jpg"}
        ]
      },
      { "@id": "survey-responses-2019.csv",
        "@type": "File",
        "about": {"@id": "https://example.com/pics/5707039334816454031_o.jp
        "author": {"@id": "#alice"}
      },
      { "@id": "https://example.com/pics/5707039334816454031_o.jpg",
        "@type": ["File", "ImageObject"],
        "contentLocation": {"@id": "http://sws.geonames.org/8152662/"},
        "author": {"@id": "https://orcid.org/0000-0002-1825-0097"}
      },
      { "@id": "#alice",
        "@type": "Person",
        "name": "Alice"
      },
      { "@id": "https://orcid.org/0000-0002-1825-0097",
        "@type": "Person",
        "name": "Josiah Carberry"
      },
      { "@id": "http://sws.geonames.org/8152662/",
        "@type": "Place",
        "name": "Catalina Park"
      },
      { "@id": "https://spdx.org/licenses/CC-BY-4.0",
        "@type": "CreativeWork",
        "name": "Creative Commons Attribution 4.0"
      }
  ]
}
```



```json
{ "@id": "./",
  "@type": "Dataset",
  "name": "A simplified RO-Crate",
  "author": {"@id": "#alice"},
  "license": {"@id": "https://spdx.org/licenses/CC-BY-4.0"},
  "datePublished": "2021-11-02T16:04:43Z",
  "hasPart": [
    {"@id": "survey-responses-2019.csv"},
    {"@id": "https://example.com/pics/5707039334816454031_o.jpg"}
  ]
},
{ "@id": "survey-responses-2019.csv",
  "@type": "File",
  "about": {"@id": "https://example.com/pics/5707039334816454031_o.jp
  "author": {"@id": "#alice"}
},
```


https://www.researchobject.org/2021-packaging-research-artefacts-with-ro-crate/manuscript.html

SUMMARY

You should know and be able to explain

- Why we need FAIR principles
- Differences between specific principles
  - Provide your own examples
- Relation between FAIR principles, machine-actionability and open data
- How to apply FAIR principles in practice

Read and watch more about FAIR

Principles explained (by their authors)

- https://www.nature.com/articles/sdata201618
- https://www.go-fair.org/fair-principles/

Watch (why FAIR matters)

- https://vimeo.com/143245835

Related papers

- ‘FAIR vs Open’ https://insights.uksg.org/articles/10.1629/uksg.468/

FAIR cookbook

- https://faircookbook.elixir-europe.org/content/home.html

Let's Make Our Data FAIR! Webinar for GO-FAIR

- https://www.youtube.com/watch?v=dEV2HnraqaI

FAIR underlies European Open Science Cloud

- https://eosc-launch.eu/declaration/

Work with us: TUW Data Services

- Topics:
  - FAIR, PIDs, Data repositories, virtual research environments, Jupyter, Kubernetes, earth observation,…
- Skills
  - ‘Coding and setting up things’
- Part time job
  - First contract for 5-6 months
  - Up to 20 hours/week
- Details to be discussed directly with you
- Contact
  - tomasz.miksa@tuwien.ac.at

Machine-actionable Data Management Plans

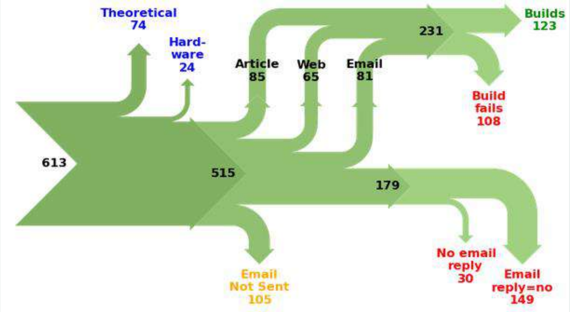
Tomasz Miksa

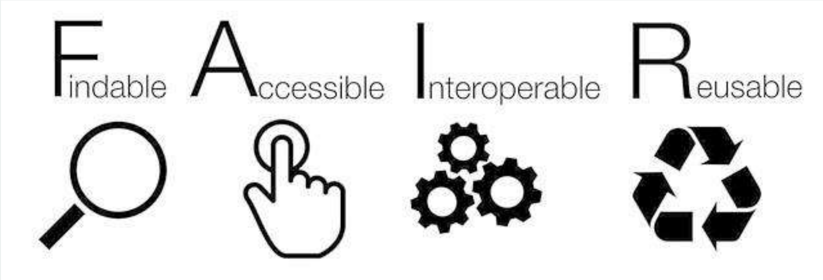
SBA Research & TU Wien

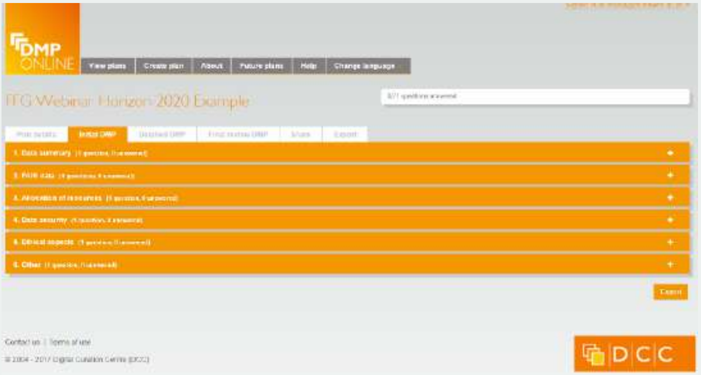
tmiksa@sba-research.org

TU WIEN Informatics

Previous lecture on DMPs







Data Management Plans (DMPs) currently

	Data Officer	*Who is responsible for the data management and the DMP of the project (name/email address)?*
I	**Data Characteristics**	
I.1	**Description of the data**	*What kinds of data/source code will be generated or reused (type, format, volume)?* *How will the research data be generated and which methods will be used?* *How will you structure the data and handle versioning?* *Who is the target audience?*
II	**Documentation and Metadata**	
II.1	**Metadata standards**	*What metadata standards (if any) will be in use and why? (see Digital Curation Centre)*
II.2	**Documentation of data**	*What information is needed for the data to be findable, accessible, interoperable and re-usable (FAIR) in the future?* *Is the data machine-readable?* *How are you planning to document this information?*
II.3	**Data quality control**	*What quality assurance processes will you adopt?* *How will the consistency and quality of data collection be controlled and documented? (This may include processes such as repeat samples or measurements, standardised data capture, peer review of data or representation with controlled vocabularies.)*
III	**Data Availability and Storage**	
III.1	**Data sharing strategy**	*How and when will the data be shared and made accessible?* *What repository will you be using?* *What persistent identifier will be used?*
III.2	**Data storage strategy**	*What data are to be preserved for the long-term, and what data will not be stored? How and where will the data be stored and backed up during the research?* *How and where will the data be stored after the project ends?* *For how long will the data be stored?* *Are there any costs that need to be covered for storage?* *At what point during or after the project will the data be stored?* *Are there any technical barriers to making the research data fully or partially accessible?*

https://www.fwf.ac.at/fileadmin/files/Dokumente/Open_Access/FWF_DMPTemplate_e.pdf

Data Management Plans (DMPs)

manually created text documents

considered as bureaucracy

created too late

vague

depend on human factor

- scrupulousness
- awareness

Data Management Plans

How to discover these tools?

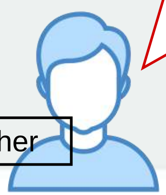
Which one do I need to use?

Why do I have to provide the same information again?

Why haven't they consulted us before?

Who is going to pay for this?

We don't have enough people for that!



Researcher

Stakeholders

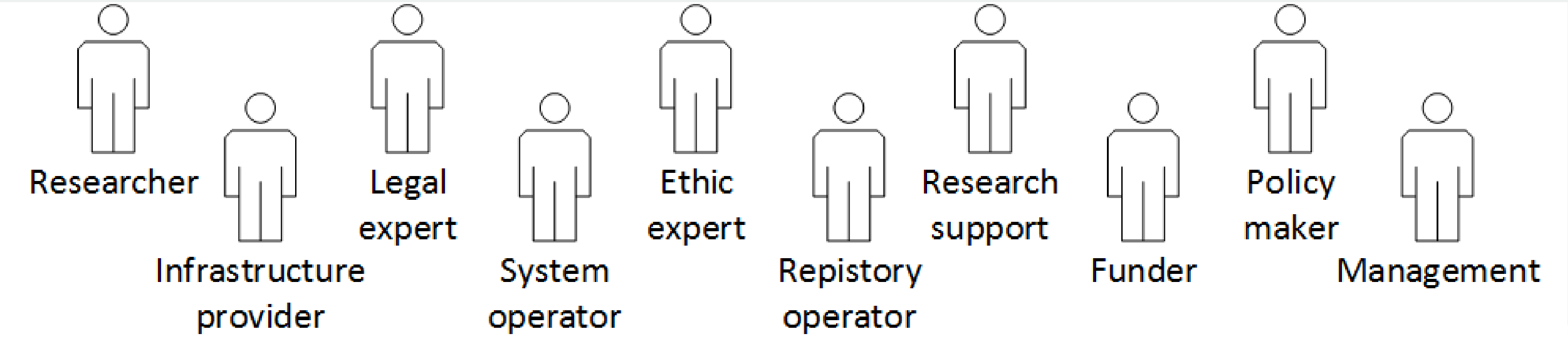
Research data lifecycle

Stakeholders involved in research data management

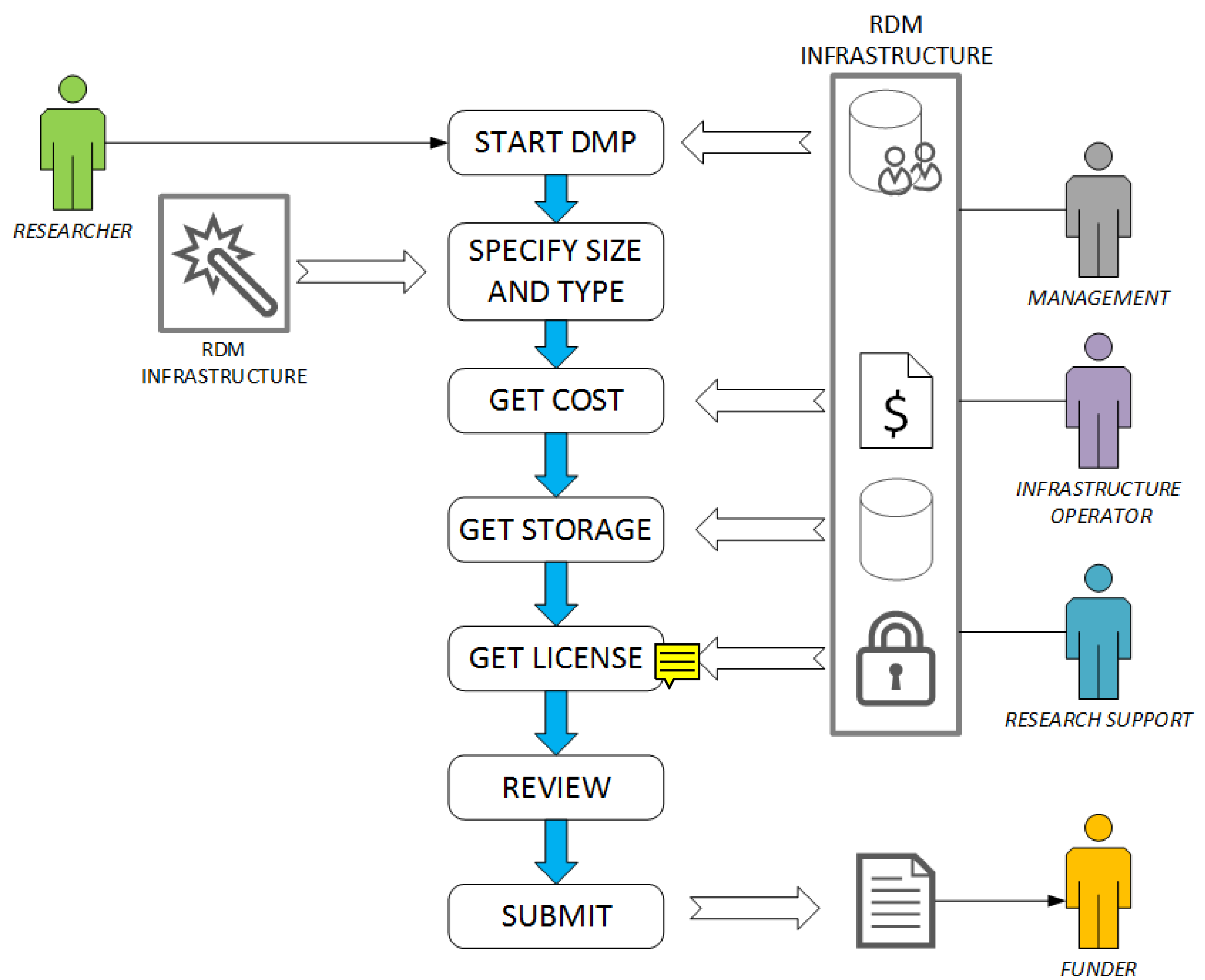
- require information at certain stages
- can provide information if requested at a proper stage

Many problems can be avoided when

- timing is right
- information flow is ensured



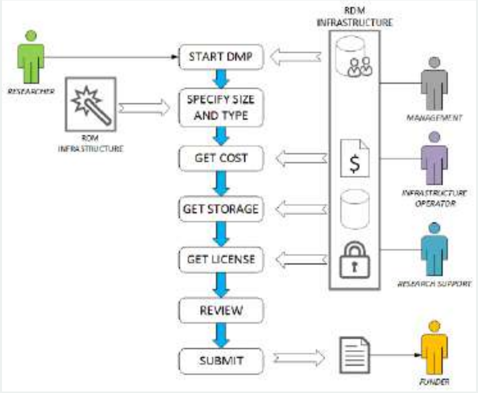
Automated Data Management Workflow



Automated Data Management Workflow

Requires

- common data model
  - to exchange information
- well-defined RDM workflows
  - Who? What? When? How?
- data management infrastructure
  - systems and services implementing workflows



Example

•Current DMPs – model questionnaires

```
<administrative_data>
    <question>Who will be the Principle Investigator?</question>
    <answer>The PI will be John Smith from our university.</answer>
</administrative_data>
```

•Machine-actionable DMPs – model information

```
"dc:creator":[ {
        "foaf:name":"John Smith",
        "@id":"orcid.org/0000-1111-2222-3333",
        "foaf:mbox":"mailto:jsmith@tuwien.ac.at",
        "madmp:institution":" AT-Vienna-University-of-Technology"
 } ],
```

Example

•Currently available – not very useful

```
<administrative_data>
    <question>Who will be the Principle Investigator?</question>
    <answer>The PI will be John Smith from our university.</answer>
</administrative_dat
```

Reuse existing standards, e.g. Dublin Core, PREMIS, etc.

•Machine-action

```
"dc:creator":[ {
        "foaf:name":"John Smith",
        "@id":"orcid.org/0000-1111-2222-3333",
        "foaf:mbox":"mailto:jsmith@tuwien.ac.at",
        "madmp:institution":"AT-Vienna-University-of-Technology"
 } ],
```

Example

- Currently available – not very useful

```
<administrative_data>
    <question>Who will be the Principle Investigator?</question>
    <answer>The PI will be John Smith from our university.</answer>
</administrative_data>
```

- Machine-actionable DMP

```
"dc:creator":[ {
        "foaf:name":"John Smith",
        "@id":"orcid.org/0000-1111-2222-3333",
        "foaf:mbox":"mailto:jsmith@tuwien.ac.at",
        "madmp:institution":"AT-Vienna-University-of-Technology"
 } ],
```

Use PIDs whenever possible, e.g. ORCID

Example

- Currently available – not very useful

```
<administrative_data>
    <question>Who will be the Principle Investigator?</question>
    <answer>The PI will be John Smith from our university.</answer>
</administrative_data>
```

- Machine-actionable DMP

```
"dc:creator":[ {
        "foaf:name":"John Smith",
        "@id":"orcid.org/0000-1111-2222-3333",
        "foaf:mbox":"mailto:jsmith@tuwien.ac.at",
        "madmp:institution":"AT-Vienna-University-of-Technology"
 } ],
```

Use controlled vocabularies

Example

- Currently available – not very useful

```
<administrative_data>
    <question>Who will be the Principle Investigator?</question>
    <answer>The PI will be John Smith from our university.</answer>
</administrative_data>
```

- Machine-actionable DMP

```
"dc:creator":[ {
        "foaf:name":"John Smith",
        "@id":"orcid.org/0000-1111-2222-3333",
        "foaf:mbox":"mailto:jsmith@tuwien.ac.at",
        "madmp:institution":"AT-Vienna-University-of-Technology"
 } ],
```

Develop own concepts and vocabularies only when needed

What is RDA

Research Data Alliance

- community-driven organization
- 6,000 members from 130 countries
- different stakeholders

Plenary meetings

Interest Groups (IGs)

- Active DMPs

Working Groups (WGs)

- DMP Common Standards

https://rd-alliance.org/

DMP Common Standards WG

Launched in October 2017

Result of a consultation made by Active DMPs IG

Focus on machine-actionable DMPs

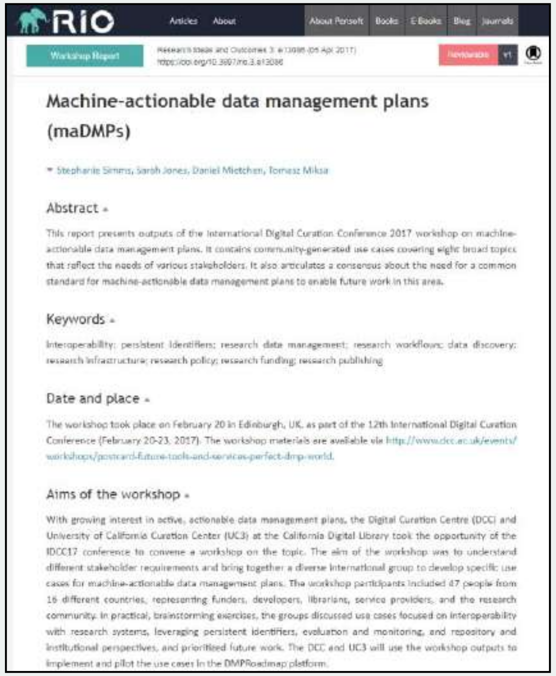
280 members from all continents

DMP tool owners are part of it

DMP ONLINE

RDMO

DSW
DATA STEWARDSHIP WIZARD



RIO

Articles About About Pensoft Books E-Books Blog Journals

Workshop Report

Machine-actionable data management plans (maDMPs)

Stephanie Simms, Sarah Jones, Daniel Mietchen, Tomasz Miksa

Abstract

This report presents outputs of the International Digital Curation Conference 2017 workshop on machine-actionable data management plans. It contains community-generated use cases covering eight broad topics that reflect the needs of various stakeholders. It also articulates a consensus about the need for a common standard for machine-actionable data management plans to enable future work in this area.

Keywords

Interoperability; persistent identifiers; research data management; research workflows; data discovery; research infrastructure; research policy; research funding; research publishing

Date and place

The workshop took place on February 20 in Edinburgh, UK, as part of the 12th International Digital Curation Conference (February 20-23, 2017). The workshop materials are available via http://www.dcc.ac.uk/events/workshops/postcard-future-tools-and-services-perfect-dmp-world.

Aims of the workshop

With growing interest in active, actionable data management plans, the Digital Curation Centre (DCC) and University of California Curation Center (UC3) at the California Digital Library took the opportunity of the IDCC17 conference to convene a workshop on the topic. The aim of the workshop was to understand different stakeholder requirements and bring together a diverse international group to develop specific use cases for machine-actionable data management plans. The workshop participants included 47 people from 16 different countries, representing funders, developers, librarians, service providers, and the research community. In practical, brainstorming exercises, the groups discussed use cases focused on interoperability with research systems, leveraging persistent identifiers, evaluation and monitoring, and repository and institutional perspectives, and prioritized future work. The DCC and UC3 will use the workshop outputs to implement and pilot the use cases in the DMPRoadmap platform.

https://doi.org/10.3897/rio.3.e13086

RESEARCH DATA ALLIANCE

O&A Members 61

Active Organisational & Affiliate members

MEMBERSHIP Members: 11604

Becoming a member of RDA is simple and open to both individuals and organizations

Register now

RDA Groups WG & IGs: 97

Discover what RDA Working and Interest Groups and all other Groups are up to and find out how to join them. Explore Groups

ABOUT RDA · GET INVOLVED · GROUPS · RECOMMENDATIONS & OUTPUTS · RDA FOR DISCIPLINES · PLENARIES & EVENTS · NEWS & MEDIA

Home » Working and Interest Groups » Working Group » DMP Common Standards WG

WG

DMP Common Standards WG

Taxonomy:

Posts · Create Wiki index · Events · Repository · Outputs · Case Statements · Plenaries · Members

create new content

Group Status: WGs Maintaining deliverables (maintenance group)

Join Group

Status: Recognised & Endorsed
Chair (s): Paul Walk, Peter Neish, Tomasz Miksa
Secretariat Liaison: enquiries[at]rd-alliance.org
TAB Liaison: Isabelle Perseil

The need for establishing this working group was articulated during the 9th plenary meeting in Barcelona during the Active DMPs IG session. The discussion was framed by a white paper by Simms et al. on machine-actionable data management plans (DMPs). The white paper is based on outputs from the IDCC workshop held in Edinburgh in 2017 that gathered almost 50 participants from Africa, America, Australia, and Europe. It describes eight community use cases which articulate consensus about the need for a common standard for machine-actionable DMPs (where machine actionable is defined as "information that is structured in a consistent way so that machines, or computers, can be programmed against the structure")

The specific focus of this working group is on developing common information model and specifying access mechanisms that make DMPs machine-actionable. The outputs of this working group will help in making systems interoperable and will allow for automatic exchange, integration, and validation of information provided in DMPs, for example, by checking whether a provided PID links to an existing dataset, if hashes of files match to their provenance traces, or whether a license was specified. The common information models are NOT intended to be prescriptive templates or questionnaires, but to provide re-usable ways of representing machine-actionable information on themes covered by DMPs.

https://www.rd-alliance.org/groups/dmp-common-standards-wg

SCOPING MADMPS

Scoping maDMPs by DMP Common Standards WG

1st consultation

2nd consultation

Proof of concept tools

BPMN processes

Model development

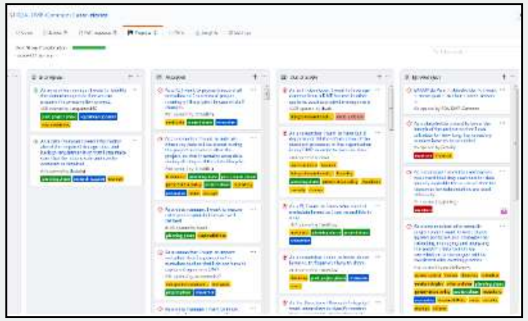
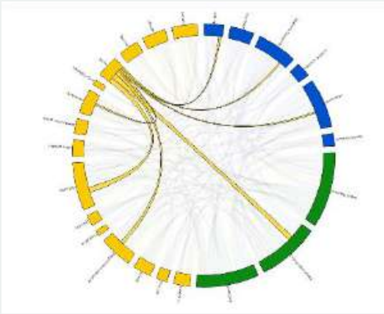
1st consultation – user stories

Goals

- identify stakeholders at each lifecycle stage
  - define which information they **provide**
  - define which information they **expect**

As a <stakeholder>, I want <goal> so that <reason >.

*As a **researcher**, I want to **inform repository operator** on the amount of data in the planning phase, so that they provide **information on costs**.*

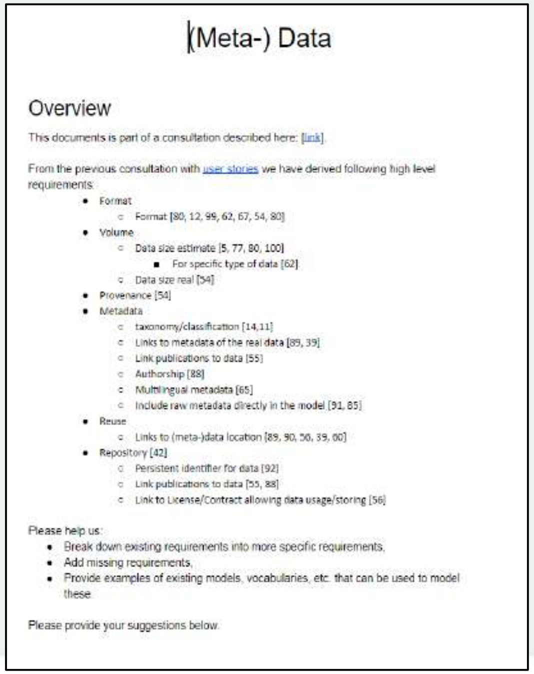




http://ifs.tuwien.ac.at/~miksa/papers/2018-iPres-maDMPs.pdf

2nd consultation – existing models

2nd consultation goes deep

- how do we model specific requirements
  - which specific fields are needed?
  - which models exist?



(Meta-) Data

Overview

This documents is part of a consultation described here: [link].

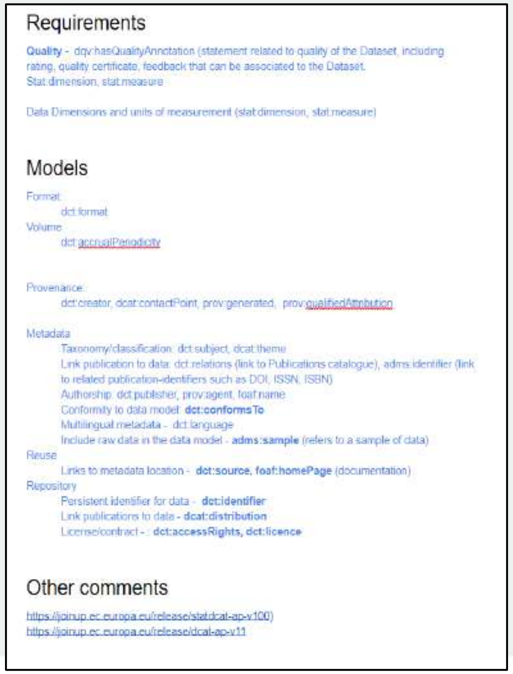
From the previous consultation with user stories we have derived following high level requirements:

- Format
  - Format [80, 12, 99, 62, 67, 54, 80]
- Volume
  - Data size estimate [5, 77, 80, 100]
    - For specific type of data [62]
  - Data size real [54]
- Provenance [54]
- Metadata
  - taxonomy/classification [14,11]
  - Links to metadata of the real data [89, 39]
  - Link publications to data [55]
  - Authorship [88]
  - Multilingual metadata [65]
  - Include raw metadata directly in the model [91, 85]
- Reuse
  - Links to (meta-)data location [89, 90, 56, 39, 60]
- Repository [42]
  - Persistent identifier for data [92]
  - Link publications to data [55, 88]
  - Link to License/Contract allowing data usage/storing [56]

Please help us:

- Break down existing requirements into more specific requirements,
- Add missing requirements,
- Provide examples of existing models, vocabularies, etc. that can be used to model these

Please provide your suggestions below.



Requirements

Quality - dqv:hasQualityAnnotation (statement related to quality of the Dataset, including rating, quality certificate, feedback that can be associated to the Dataset.
Stat:dimension, stat:measure

Data Dimensions and units of measurement (stat:dimension, stat:measure)

Models

Format:
dct:format
Volume:
dct:accrualPeriodicity

Provenance:
dct:creator, dcat:contactPoint, prov:generated, prov:qualifiedAttribution

Metadata
Taxonomy/classification: dct:subject, dcat:theme
Link publication to data: dct:relations (link to Publications catalogue), adms:identifier (link to related publication-identifiers such as DOI, ISSN, ISBN)
Authorship: dct:publisher, prov:agent, foaf:name
Conformity to data model: **dct:conformsTo**
Multilingual metadata - dct:language
Include raw data in the data model - **adms:sample** (refers to a sample of data)
Reuse
Links to metadata location - **dct:source, foaf:homePage** (documentation)
Repository
Persistent identifier for data - **dct:identifier**
Link publications to data - **dcat:distribution**
License/contract - : **dct:accessRights, dct:licence**

Other comments

https://joinup.ec.europa.eu/release/statdcat-ap-v100)
https://joinup.ec.europa.eu/release/dcat-ap-v11

Proof of concept tools

Requirements

- Provide minimum input
- Import as much as possible from existing systems to help in creating maDMPs

Tools available as Docker on GitHub

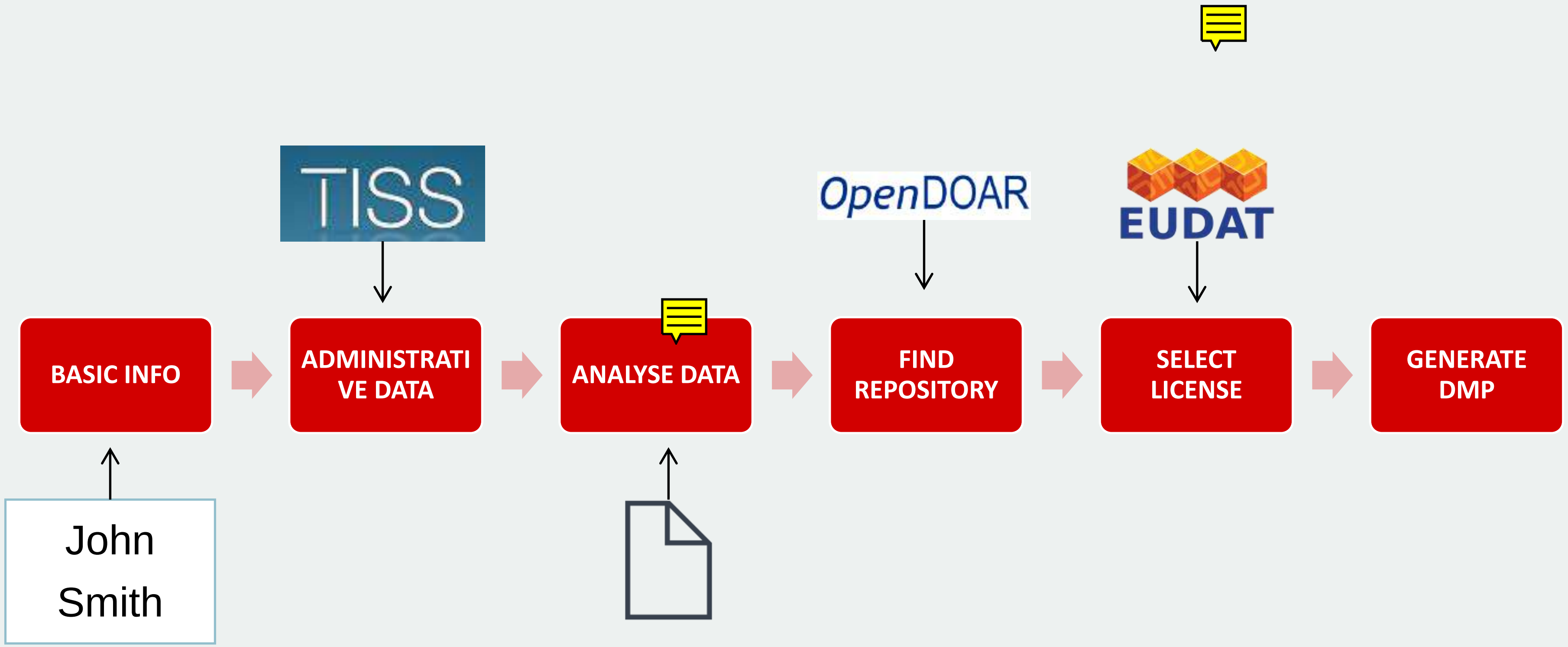
- https://github.com/TomMiksa/DMPGenerator
- https://github.com/TomMiksa/digital_preservation_ex_1_2
- https://github.com/TomMiksa/tu-dpue-lab2-ss18
- https://github.com/TomMiksa/DigitalPreservation_2
- https://github.com/TomMiksa/digitalpreservation-dmp-generator
- https://github.com/TomMiksa/DMPlanner

Example of a landing page for maDMPs

- https://oblassers.github.io/fair-data-science/
- https://github.com/oblassers/fair-data-science

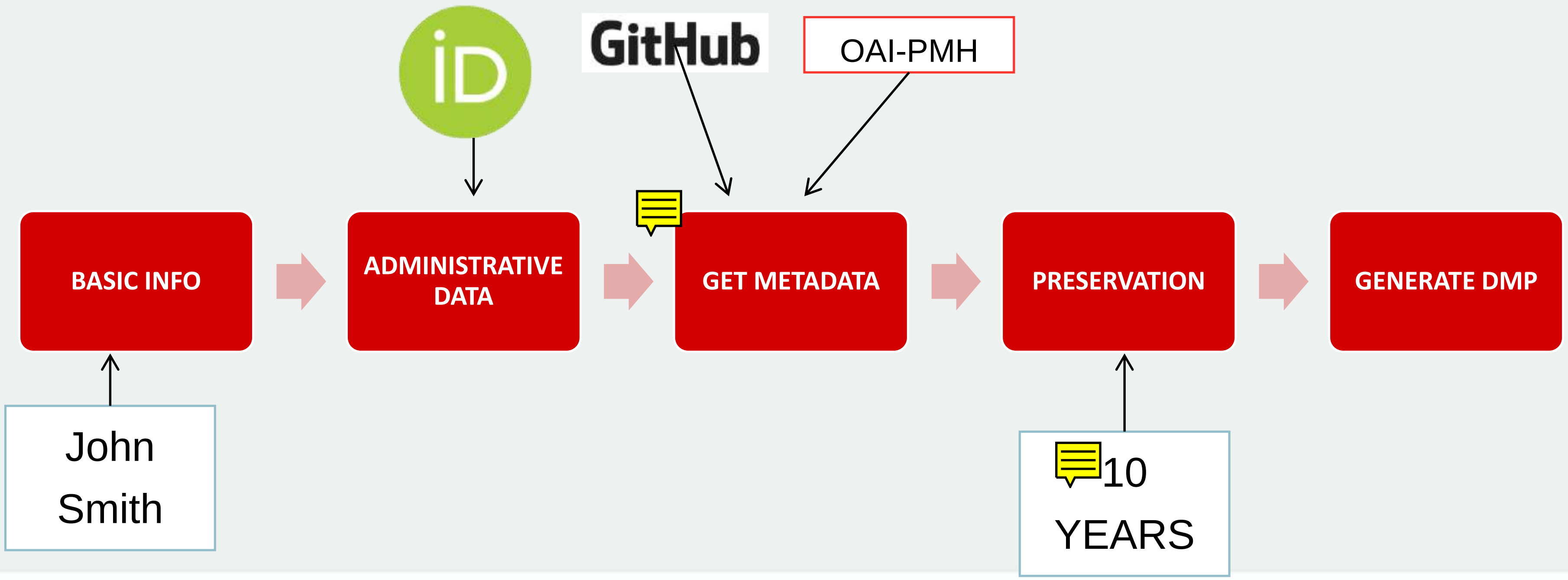
Planning phase

- Goal: get **estimations** and **recommendations** (which are feasible to implement later)



Project and Post-project phases

- Goal: **update** DMP with **real** information by **re-using** (linking) information provided elsewhere



Name

Please provide your full name.

Search ORCiD

full_name Tomasz Miksa

orcid 0000-0002-4929-7875

current_employment_name SBA Research

Resources

Add as many Github repositories or OAI-PMH compliant DOIs as you like.*

Add Resource

Zenodo Ten Simple Rules For Machine-Actionable Data Management Plans (Preprint)

documentation | Remove

Github TomMiksa/DMPlanner | software | Remove

Preservation Time

Choose how many years the data for each group should be kept.

Software: 10 years

Documentation: 20 years

TUW DMP

A Data Management Plan created using DMPlanner.

Creator

Name: Tomasz Miksa
ORCiD: 0000-0002-4929-7875
Current Work: SBA Research

How will you manage copyright and Intellectual Property Rights (IPR) issues?

The software which was created in the course of the project has the license restrictions "MIT License".

Which data are of long-term value and should be retained, shared, and/or preserved?

In this project especially the documentation, as well as the software has a long-term value and should at least be as long preserved as the targeted preservation time specifies. The targeted preservation time for the documentation is 20 years. The targeted preservation time for the software is 10 years.

What is the long-term preservation plan for the dataset?

One of the main strategies of the long-term preservation plan is the use of public accessible repositories to save the components of the project. The documentation resource "Ten Simple Rules For Machine-Actionable Data Management Plans (Preprint)" is hosted on Zenodo. The software resource "DMPlanner" is hosted on Github.

How will you share the data?

The data will be primarily shared through the public repositories listed above. This way the data is openly accessible and findable, as well as searchable. The data is available at the repositories as of this moment.

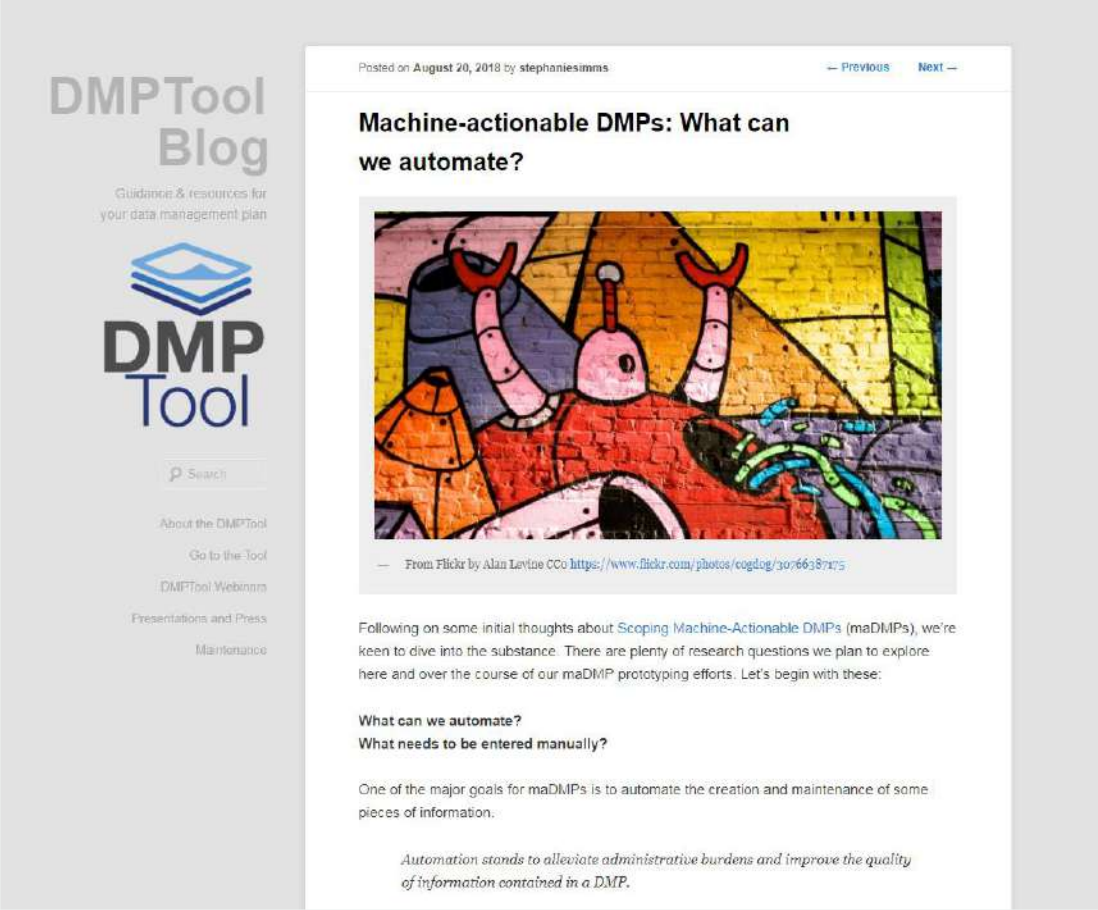
Are any restrictions on data sharing required?

The restrictions on data sharing are composed of the used licenses together with the long-term preservation plan. With this in mind the following restrictions for the resources of the project apply. The documentation resource "Ten Simple Rules For Machine-Actionable Data Management Plans (Preprint)" will be hosted on Zenodo for at least 20 years. The software resource "DMPlanner" will be hosted on Github for at least 10 years.

Who will be responsible for data management?

The creator of this data management plan is Tomasz Miksa. Therefore Tomasz Miksa is also the reference person for possible reviews and revisions regarding this data management plan in the future. Unless amended Tomasz Miksa is additionally responsible for the adherence to the plan.

Machine-actionable DMPs: What can we automate?



DMPTool Blog

Guidance & resources for your data management plan

Search

About the DMPTool

Go to the Tool

DMPTool Webinars

Presentations and Press

Maintenance

Posted on August 20, 2018 by stephaniesimms

← Previous Next →

Machine-actionable DMPs: What can we automate?

— From Flickr by Alan Levine CC0 https://www.flickr.com/photos/cogdog/30766387175

Following on some initial thoughts about Scoping Machine-Actionable DMPs (maDMPs), we're keen to dive into the substance. There are plenty of research questions we plan to explore here and over the course of our maDMP prototyping efforts. Let's begin with these:

What can we automate?
What needs to be entered manually?

One of the major goals for maDMPs is to automate the creation and maintenance of some pieces of information.

> *Automation stands to alleviate administrative burdens and improve the quality of information contained in a DMP.*

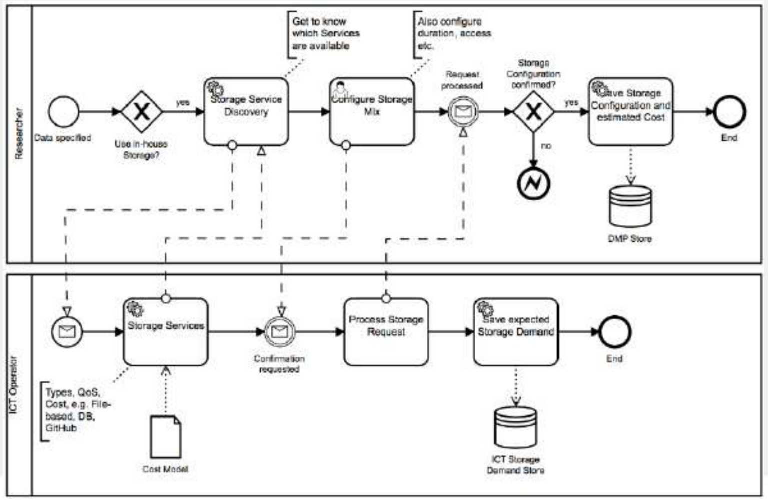
https://blog.dmptool.org/2018/08/20/machine-actionable-dmps-what-can-we-automate/

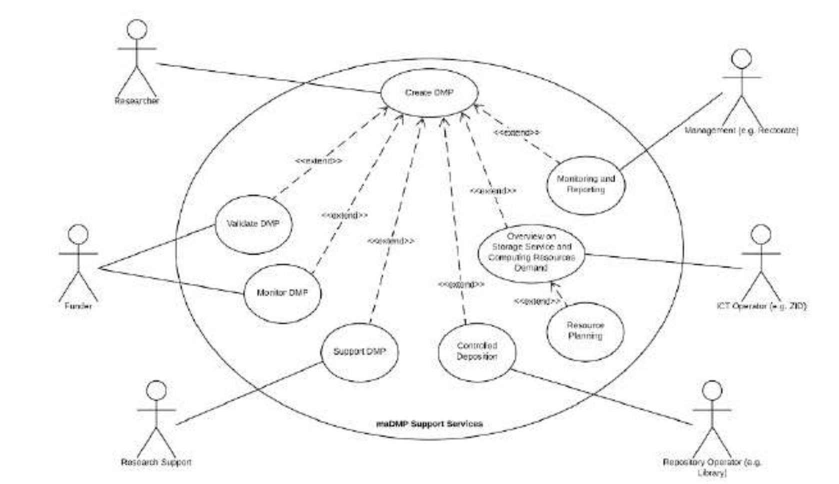
https://blog.dmptool.org/2019/10/14/whats-new-with-our-machine-actionable-dmp-work/

Processes

Processes help identify

- **tasks** performed by stakeholders
  - e.g. ICT operator provide costs of storage
- **systems** needed to be put in place
  - e.g. maDMP repository or costing service
- **concepts** to be developed or agreed
  - e.g. cost model for storage





Processes

Useful in deploying maDMPs

Allow us to narrow down focus

- common model does not contain business logic
  - e.g. cost estimation is done by a service that provides a value
- common model is an information carrier
  - tools, services, processes make maDMPs *machine-actionable*

http://doi.org/10.5281/zenodo.2607556

Scoping maDMPs - summary

1st consultation (user stories) went broad

- to define scope of maDMPs

2nd consultation went deep

- to identify models for specific requirements

Proof of concept tools

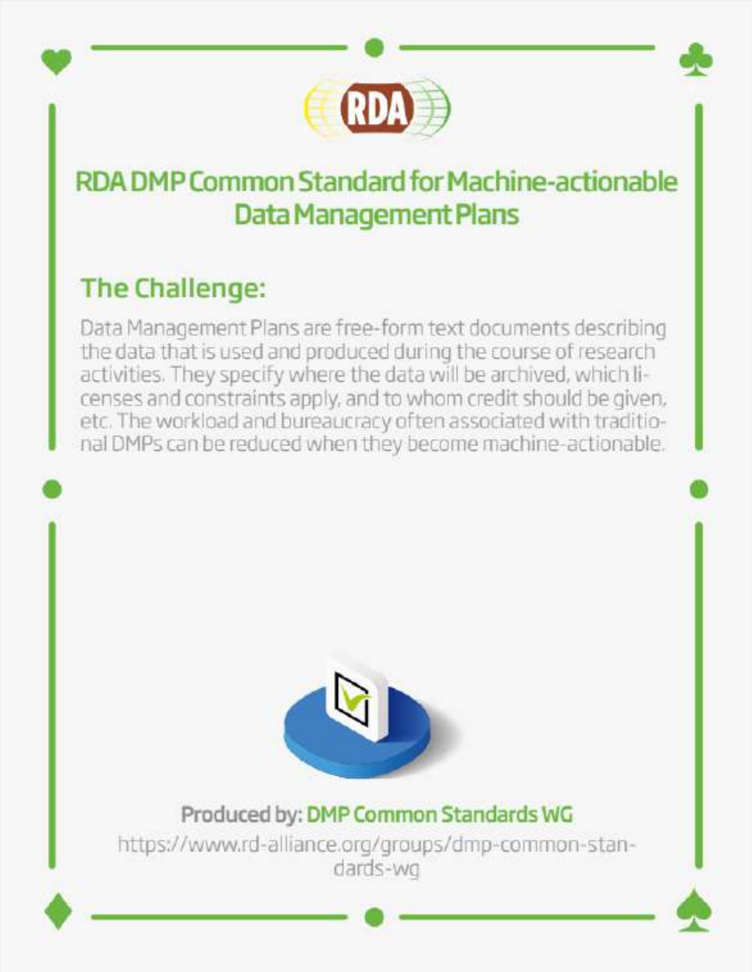
- to demonstrate how model can be used to automate tasks

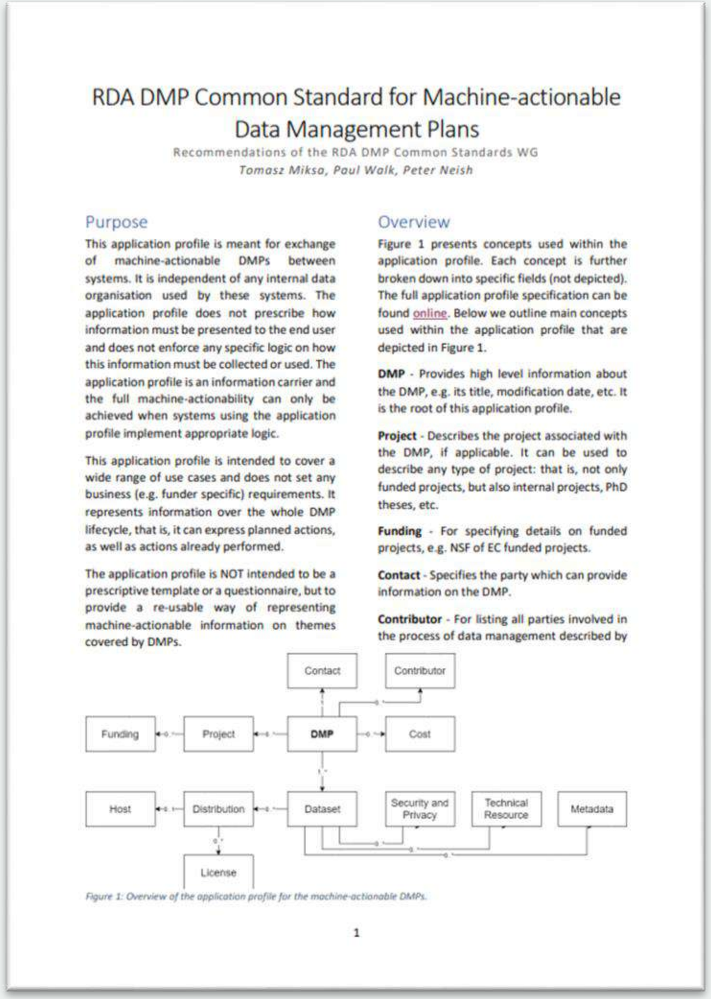
BPMN processes

- to identify systems and stakeholders involved

Model development

Official RDA Recommendation on maDMPs





RDA DMP Common Standard for Machine-actionable Data Management Plans

Recommendations of the RDA DMP Common Standards WG

Tomasz Miksa, Paul Walk, Peter Neish

Purpose

This application profile is meant for exchange of machine-actionable DMPs between systems. It is independent of any internal data organisation used by these systems. The application profile does not prescribe how information must be presented to the end user and does not enforce any specific logic on how this information must be collected or used. The application profile is an information carrier and the full machine-actionability can only be achieved when systems using the application profile implement appropriate logic.

This application profile is intended to cover a wide range of use cases and does not set any business (e.g. funder specific) requirements. It represents information over the whole DMP lifecycle, that is, it can express planned actions, as well as actions already performed.

The application profile is NOT intended to be a prescriptive template or a questionnaire, but to provide a re-usable way of representing machine-actionable information on themes covered by DMPs.

Overview

Figure 1 presents concepts used within the application profile. Each concept is further broken down into specific fields (not depicted). The full application profile specification can be found online. Below we outline main concepts used within the application profile that are depicted in Figure 1.

DMP - Provides high level information about the DMP, e.g. its title, modification date, etc. It is the root of this application profile.

Project - Describes the project associated with the DMP, if applicable. It can be used to describe any type of project: that is, not only funded projects, but also internal projects, PhD theses, etc.

Funding - For specifying details on funded projects, e.g. NSF of EC funded projects.

Contact - Specifies the party which can provide information on the DMP.

Contributor - For listing all parties involved in the process of data management described by

Figure 1: Overview of the application profile for the machine-actionable DMPs.

1

Miksa, T., Walk, P., & Neish, P. (2020). RDA DMP Common Standard for Machine-actionable Data Management Plans. https://doi.org/10.15497/rda00039

Pending adoptions (selected)

DMP ONLINE

DMP Tool

DSW

F1000 Research

Haplo Repository

ARGOS

RDMO

FAIR DATA AUSTRIA

DMP OPIDoR

NSD NORWEGIAN CENTRE FOR RESEARCH DATA

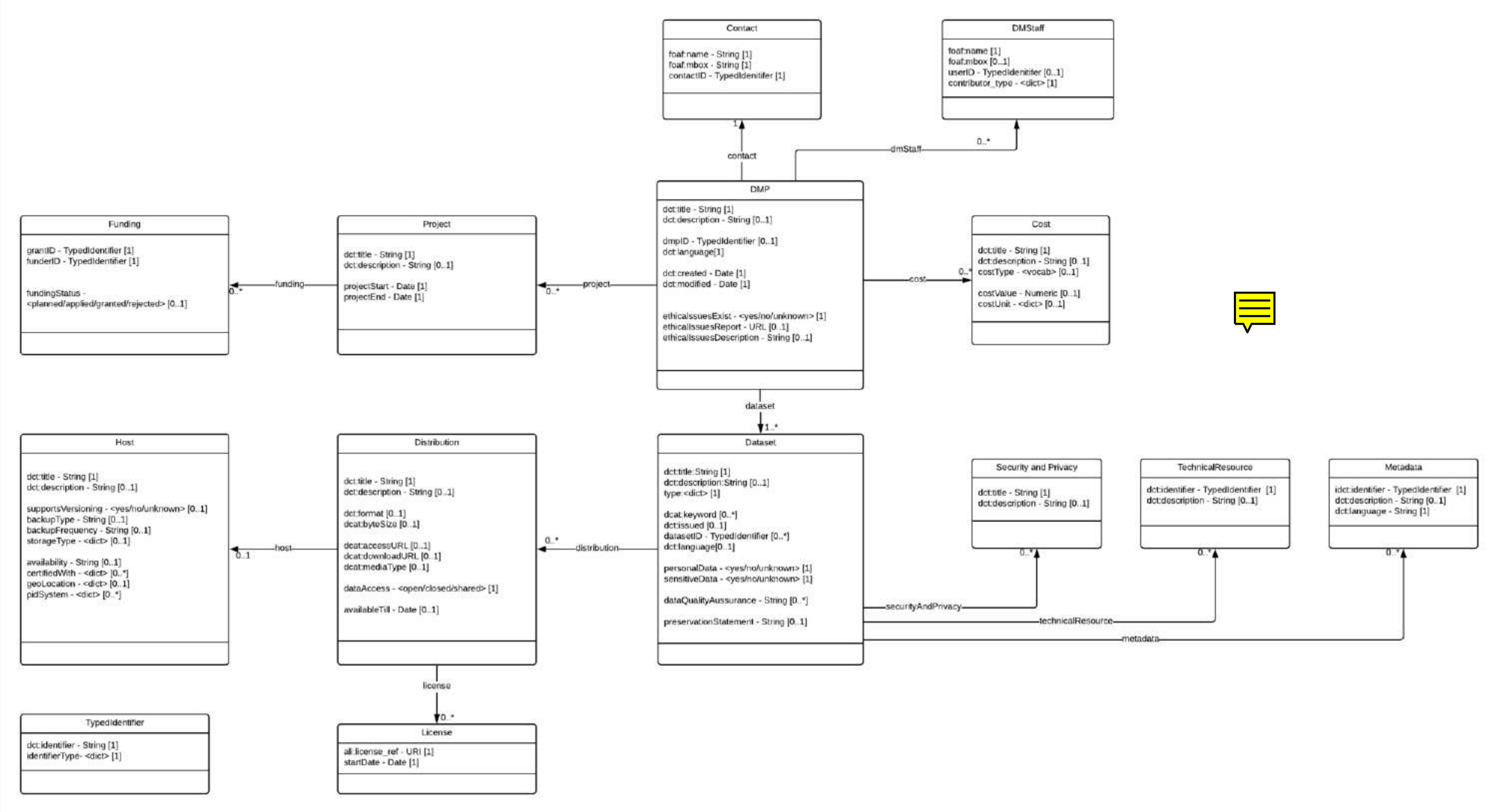
Webinar available

https://www.rd-alliance.org/rda-working-groups-solutions-dmp-recording-and-slides-webinar-now-available-0

Part 2

COMMON STANDARD FOR MADMPS

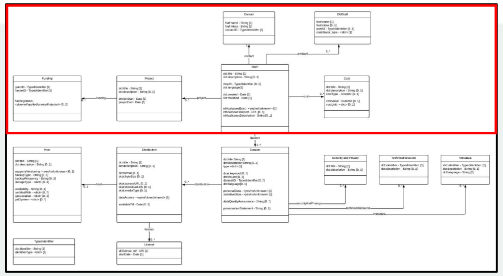
Common standard for maDMPs

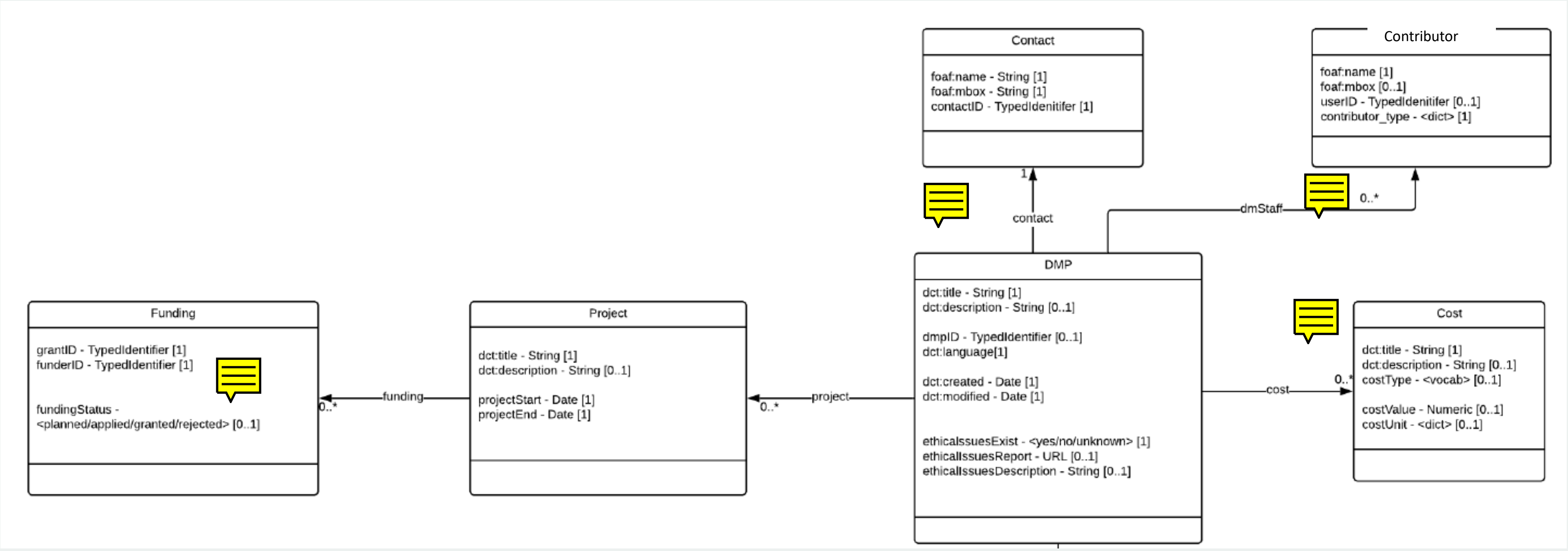


Note: Diagram depicts draft version. Details evolved! Use official release.

https://github.com/RDA-DMP-Common/RDA-DMP-Common-Standard

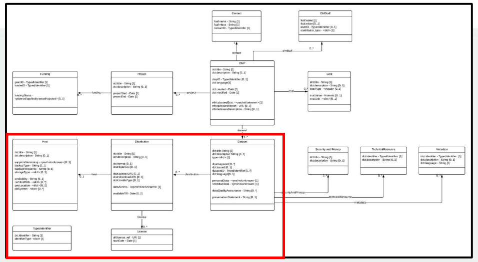
Common standard for maDMPs

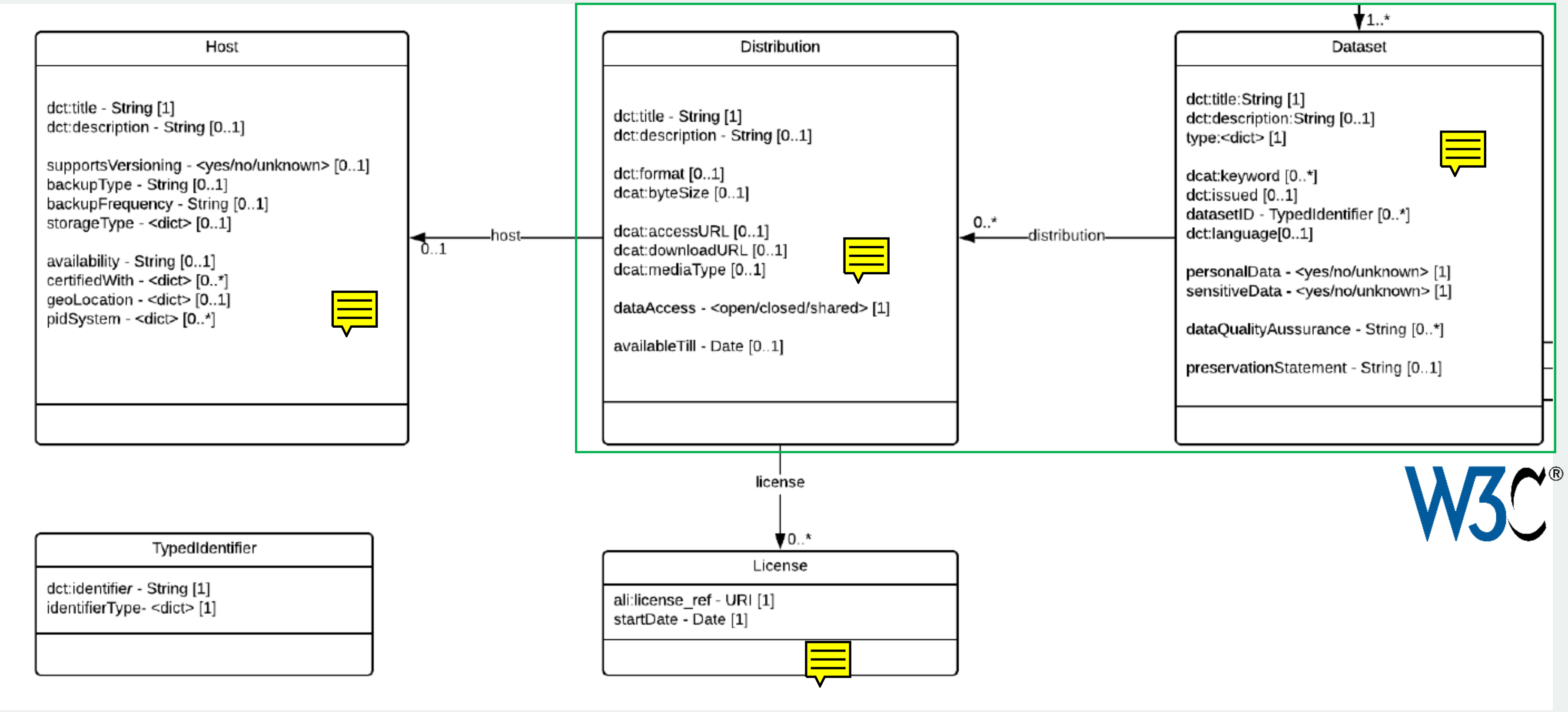




Note: Diagram depicts draft version. Details evolved! Use official release.

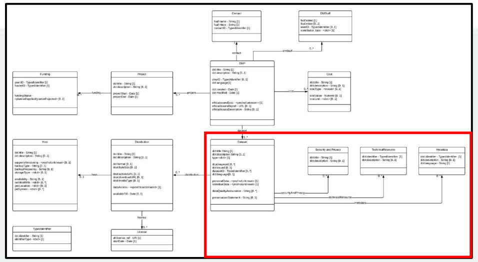
Common standard for maDMPs

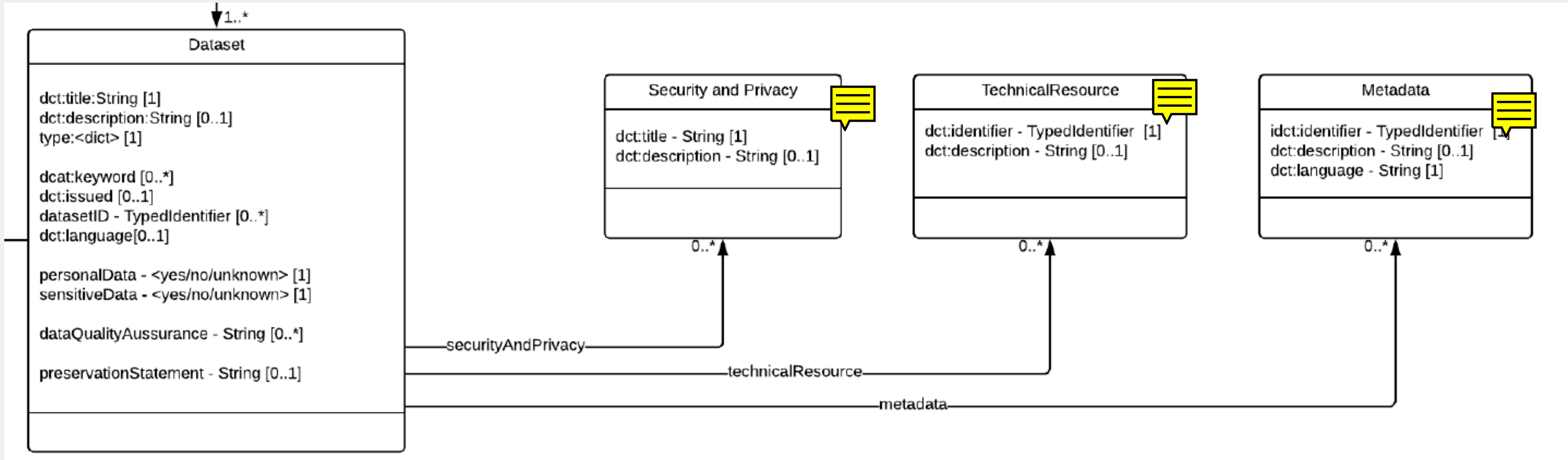




Note: Diagram depicts draft version. Details evolved! Use official release.

Common standard for maDMPs





Note: Diagram depicts draft version. Details evolved! Use official release.

Standard - documentation

Properties in 'contact'

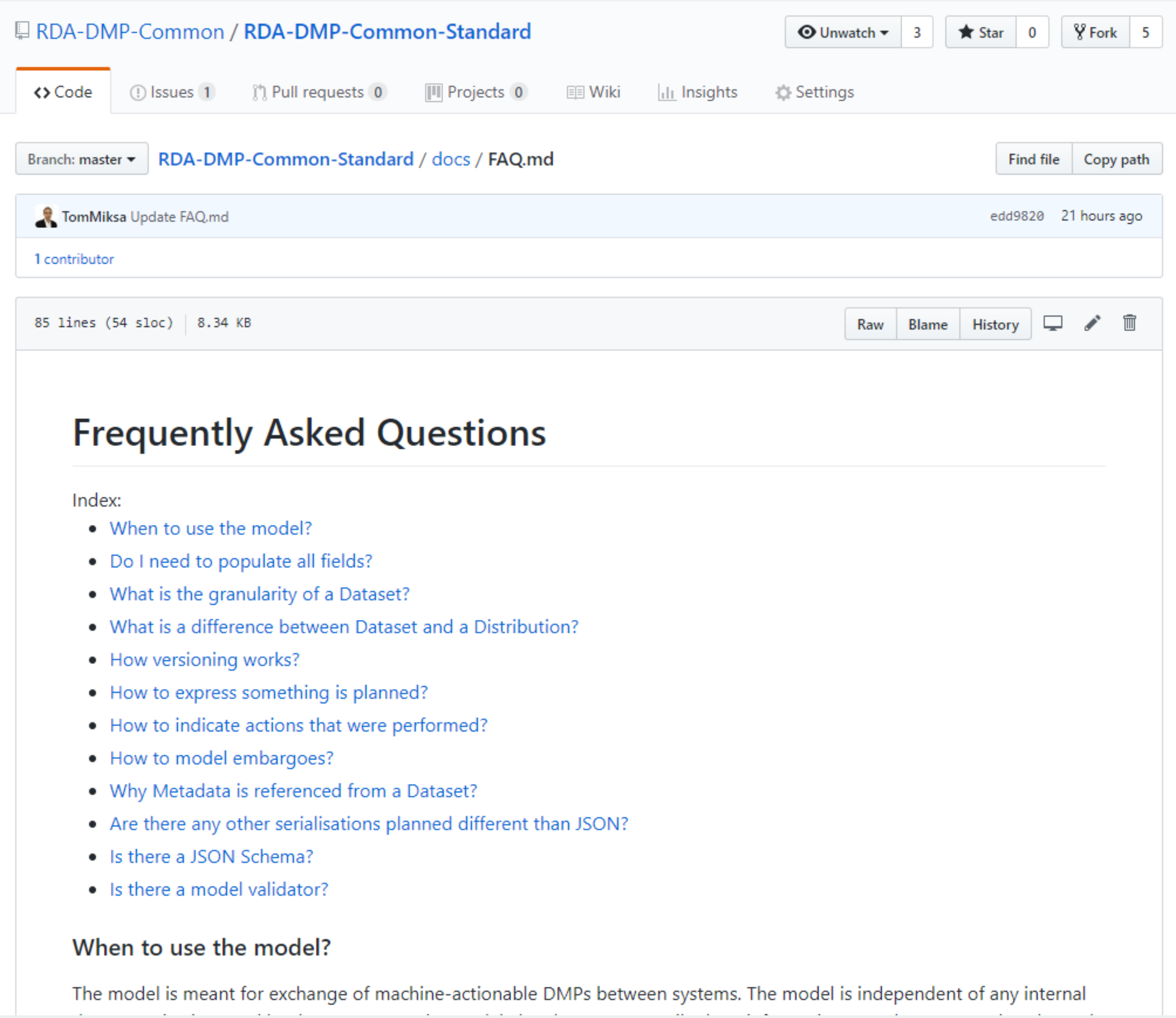
Name	Description	Data Type	Cardinality	Example Value
contact_id	Identifier for a contact person	String	Exactly One	http://orcid.org/0000-0000-0000-0000
mail	E-mail address	String	Exactly One	cc@example.com
name	Name of the contact person	String	Exactly One	Charlie Chaplin

Properties in 'cost'

Name	Description	Data Type	Cardinality	Example Value
currency_code	Allowed values defined by ISO 4217.	Term from Controlled Vocabulary	Zero or One	EUR
description	Description	String	Zero or One	Costs for maintaining....
title	Title	String	Exactly One	Storage and backup
type	Type	Term from Controlled Vocabulary	Zero or One	
value	Value	Number	Zero or One	1000

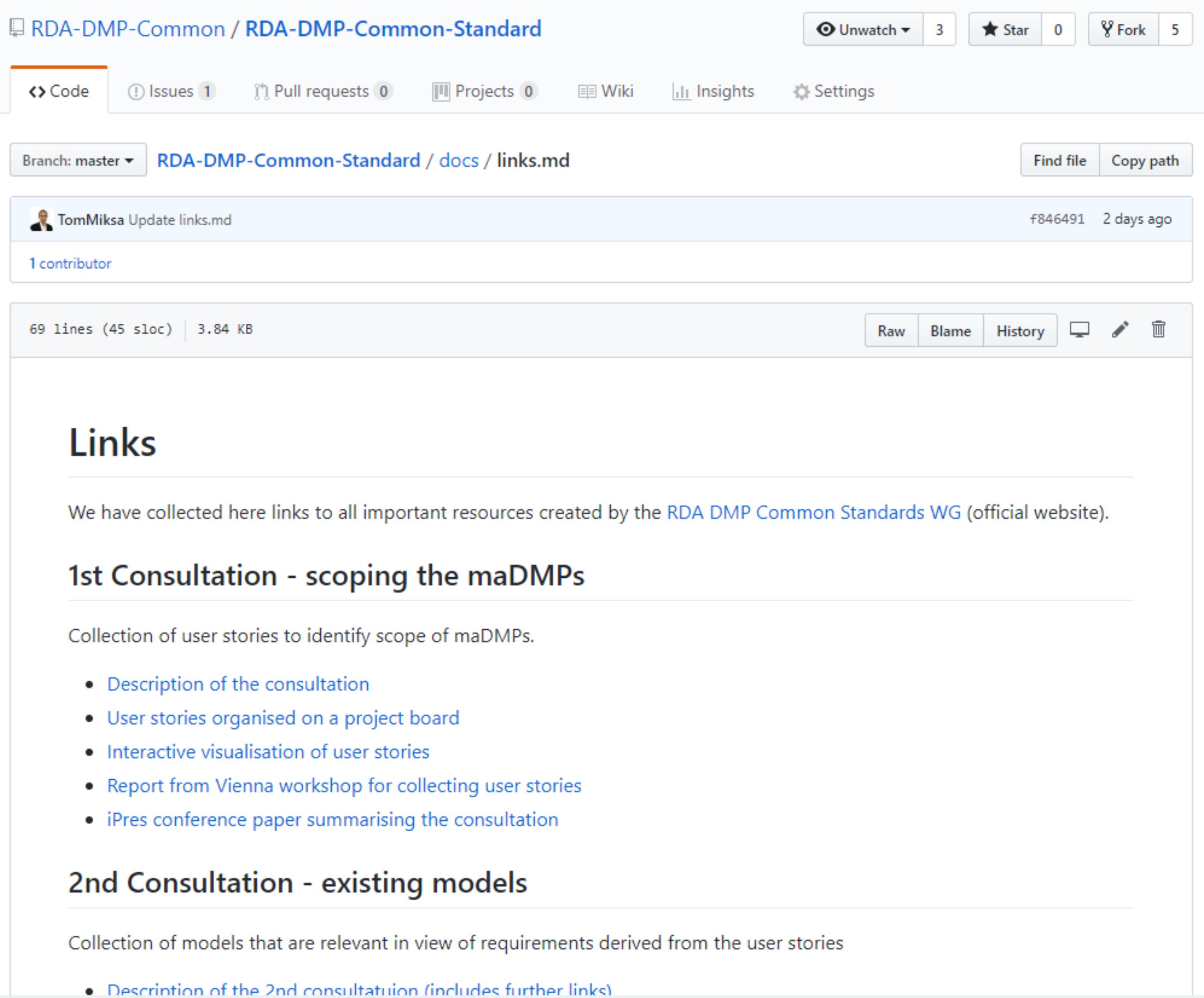
https://github.com/RDA-DMP-Common/RDA-DMP-Common-Standard/blob/master/docs/index.md

Standard – FAQ



https://github.com/RDA-DMP-Common/RDA-DMP-Common-Standard/blob/master/docs/FAQ.md

Standard – useful links



https://github.com/RDA-DMP-Common/RDA-DMP-Common-Standard/blob/master/docs/links.md

Standard – JSON examples

master · RDA-DMP-Common-Standard / examples / JSON / Go to file Add file ···

SotosTsepe Closes #32 … e10966f on Nov 7, 2020 History

..		
JSON-schema	Closes #32	4 months ago
ex1-header-fundedProject.json	Updated examples	13 months ago
ex2-dataset-planned.json	Remove tab from ex2-dataset-planned.json	9 months ago
ex3-dataset-finished.json	Updated examples	13 months ago
ex4-dataset-embargo.json	Updated examples	13 months ago
ex5-dataset-planned-host.json	Updated examples	13 months ago
ex6-dataset-closed.json	Updated examples	13 months ago
ex7-dataset-many.json	Updated examples	13 months ago
ex8-dmp-minimal-content.json	Updated examples	13 months ago
ex9-dmp-long.json	New example	13 months ago

https://github.com/RDA-DMP-Common/RDA-DMP-Common-Standard/tree/master/examples/JSON

DMP and Project – JSON example

40 lines (34 sloc) | 825 Bytes   Raw Blame History

```
{
        "DMP": {
                "title": "Funded DMP",
                "description": "Example of a DMP header for a funded project.",

                "created": "2019-02-22T13:20:15.5",
                "modified": "2019-02-22T15:10:56.9",
                "contact": {
                        "name": "First Last",
                        "mbox": "test@test",
                        "contactID": {
                                "identifier": "https://orcid.org/0000-0002-4929-7875",
                                "identifierType": "HTTP-ORCID"
                        }
                },
                "ethicalIssuesExist": "false",

                "project": {
                        "title": "Making maDMPs awesome",
                        "projectStart": "2017-01-01",
                        "projectEnd": "2020-12-31",

                        "funding": {
                                "funderID": {
                                        "identifier": "501100002428",
                                        "identifierType": "FUNDREF"
                                },
                                "grantID": {
                                        "identifier": "1234567-AT",
                                        "identifierType": "custom"
                                },
                                "fundingStatus": "granted"
                        }
                },

                "dataset" : {}

        }
}
```

https://github.com/RDA-DMP-Common/RDA-DMP-Common-Standard/blob/master/examples/JSON/ex1-header-fundedProject.json

DMP and Project – JSON example

40 lines (34 sloc) | 825 Bytes

```json
{
        "DMP": {
                "title": "Funded DMP",
                "description": "Example of a DMP header for a funded project.",

                "created": "2019-02-22T13:20:15.5",
                "modified": "2019-02-22T15:10:56.9",
                "contact": {
                        "name": "First Last",
                        "mbox": "test@test",
                        "contactID": {
                                "identifier": "https://orcid.org/0000-0002-4929-7875",
                                "identifierType": "HTTP-ORCID"
                        }
                },
                "ethicalIssuesExist": "false",
```

https://github.com/RDA-DMP-Common/RDA-DMP-Common-Standard/blob/master/examples/JSON/ex1-header-fundedProject.json

DMP and Project – JSON example

```
                "project": {
                        "title": "Making maDMPs awesome",
                        "projectStart": "2017-01-01",
                        "projectEnd": "2020-12-31",

                        "funding": {
                                "funderID": {
                                        "identifier": "501100002428",
                                        "identifierType": "FUNDREF"
                                },
                                "grantID": {
                                        "identifier": "1234567-AT",
                                        "identifierType": "custom"
                                },
                                "fundingStatus": "granted"
                        }
                },

                "dataset" : {}

        }
}
```

https://github.com/RDA-DMP-Common/RDA-DMP-Common-Standard/blob/master/examples/JSON/ex1-header-fundedProject.json

Standard assumptions – relaxed constraints

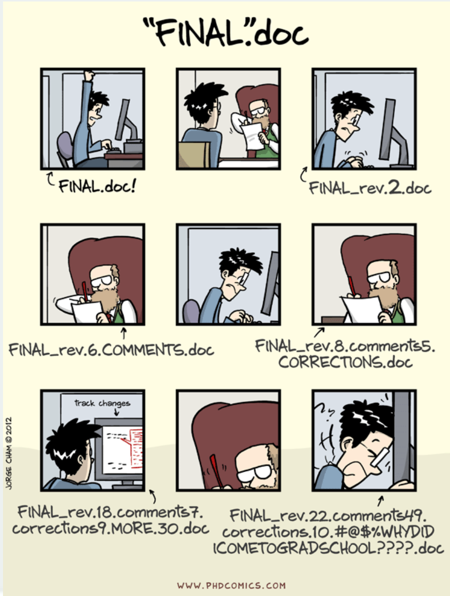
Model must be applicable in different settings

- relaxed constraints within the model
  - e.g. DMP **can** relate to a project [0..*]
- constraints introduced at the ‘business level’
  - tool implementing the model
  - e.g. DMP **must** relate to a project [1..*]
- DMP instances are still compatible

Standard assumptions - interoperability

Model will be pre-dominantly used to exchange information between systems

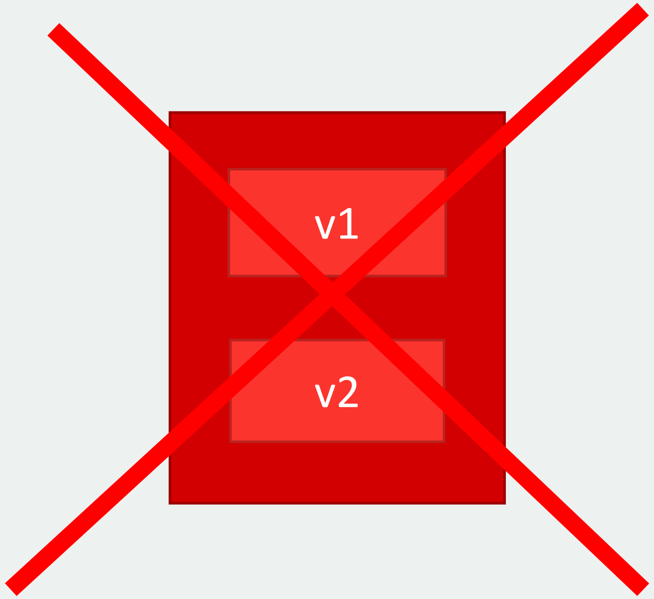
- internal representation of information in a DMP tool may differ (physical model)
  - e.g. database may have a different schema
- No ‘meta-fields’ about DMP
  - e.g. no DMP state field ‘final’



Standard assumptions - versioning

DMP versioning done by systems using the model

- model provides fields allowing to identify DMP version
- model does not track connections between versions



Standard assumptions – evolving information

Model expresses ‘certainty’ of provided information

- to support different phases of DMPs

Example

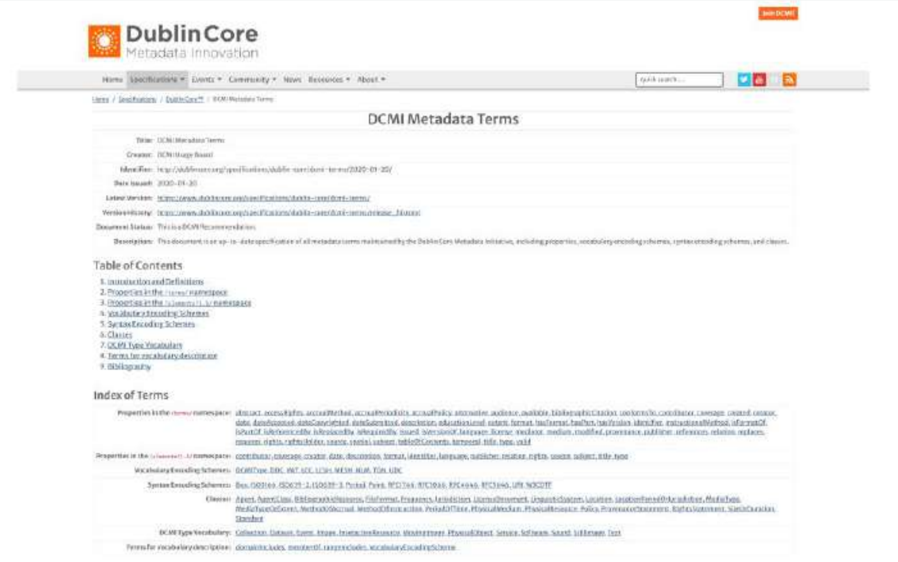
- Source code will be issued on 2019-06-30 (planned) in ‘some-repo’.
- There will be an embargo period till 2019-12-31.
- Later on the source code will be available on a CC-BY license.

```
"DMP": {
    "modified": "2019-02-22T13:20:15.5“
    "dataset": {
        "title": "Source Code",
        "issued": "2019-06-30",
        "distribution": {
            "accessURL": "http://some-repo...",
            "license": {
                "license_ref": "https://creativecommons.org/licenses/by/4.0/",
                "startDate": "2019-12-31
}}}}
```

Standard – reused standards

id	label	uri
ali	Access License and Indicators	http://www.niso.org/schemas/ali/1.0/
dces	Dublin Core Element Set	http://purl.org/dc/elements/1.1/
dct	DCMI Metadata Terms	http://purl.org/dc/terms/
foaf	Friend of a Friend (FOAF)	http://xmlns.com/foaf/0.1/
dcat	DCAT	https://www.w3.org/TR/vocab-dcat/
datacite	Data Cite	https://schema.datacite.org
cerif	Cerif	https://www.eurocris.org/ontologies/cerif/1.3/index.html#currencyCode
coar	COAR	http://vocabularies.coar-repositories.org/pubby/resource_type.html
iso6391	ISO 6391-1	Two letter country code
iso4217	ISO 4217	Currency code

Dublin Core



Core set of element for describing *resources*

- digital resources (video, images, web pages, etc.)
- physical *(books, etc.)*

Examples

- Contributor – "An entity responsible for making contributions to the resource".
- Creator – "An entity primarily responsible for making the resource".
- Date – "A point or period of time associated with an event in the lifecycle of the resource".
- Description – "An account of the resource".
- Format – "The file format, physical medium, or dimensions of the resource".
- Identifier – "An unambiguous reference to the resource within a given context".
- Language – "A language of the resource".
- Type – "The nature or genre of the resource".

https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

Term Name: title	More details
URI	http://purl.org/dc/terms/title
Label	Title
Definition	A name given to the resource.
Type of Term	Property
Has Range	http://www.w3.org/2000/01/rdf-schema#Literal
Subproperty of	• Title (http://purl.org/dc/elements/1.1/title)

Any string

Term Name: type	More details
URI	http://purl.org/dc/terms/type
Label	Type
Definition	The nature or genre of the resource.
Comment	Recommended practice is to use a controlled vocabulary such as the DCMI Type Vocabulary [DCMI-TYPE]. To describe the file format, physical medium, or dimensions of the resource, use the property Format.
Type of Term	Property
Subproperty of	• Type (http://purl.org/dc/elements/1.1/type)

Dublin Core

Index of Terms

Properties in the /terms/ namespace:	abstract, accessRights, accrualMethod, accrualPeriodicity, accrualPolicy, alternative, audience, available, bibliographicCitation, conformsTo, contributor, coverage, created, creator, date, dateAccepted, dateCopyrighted, dateSubmitted, description, educationLevel, extent, format, hasFormat, hasPart, hasVersion, identifier, instructionalMethod, isFormatOf, isPartOf, isReferencedBy, isReplacedBy, isRequiredBy, issued, isVersionOf, language, license, mediator, medium, modified, provenance, publisher, references, relation, replaces, requires, rights, rightsHolder, source, spatial, subject, tableOfContents, temporal, title, type, valid
Properties in the /elements/1.1/ namespace:	contributor, coverage, creator, date, description, format, identifier, language, publisher, relation, rights, source, subject, title, type
Vocabulary Encoding Schemes:	DCMIType, DDC, IMT, LCC, LCSH, MESH, NLM, TGN, UDC
Syntax Encoding Schemes:	Box, ISO3166, ISO639-2, ISO639-3, Period, Point, RFC1766, RFC3066, RFC4646, RFC5646, URI, W3CDTF
Classes:	Agent, AgentClass, BibliographicResource, FileFormat, Frequency, Jurisdiction, LicenseDocument, LinguisticSystem, Location, LocationPeriodOrJurisdiction, MediaType, MediaTypeOrExtent, MethodOfAccrual, MethodOfInstruction, PeriodOfTime, PhysicalMedium, PhysicalResource, Policy, ProvenanceStatement, RightsStatement, SizeOrDuration, Standard
DCMI Type Vocabulary:	Collection, Dataset, Event, Image, InteractiveResource, MovingImage, PhysicalObject, Service, Software, Sound, StillImage, Text
Terms for vocabulary description:	domainIncludes, memberOf, rangeIncludes, VocabularyEncodingScheme

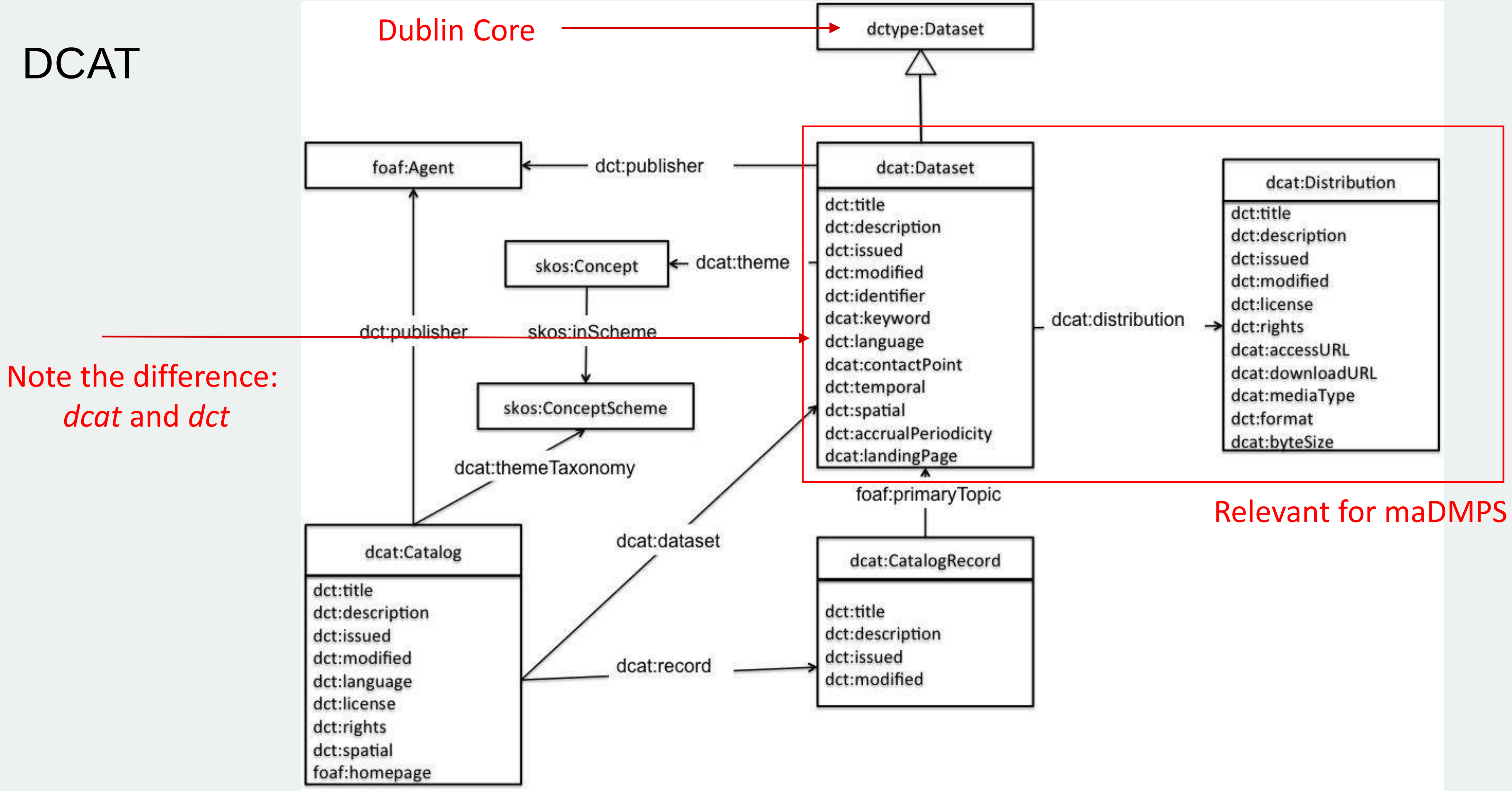
Term Name: Dataset	
URI	http://purl.org/dc/dcmitype/Dataset
Label	Dataset
Definition	Data encoded in a defined structure.
Comment	Examples include lists, tables, and databases. A dataset may be useful for direct machine processing.
Type of Term	Class
Member Of:	http://purl.org/dc/terms/DCMIType

https://www.dublincore.org

DCAT

- DCAT is an RDF vocabulary designed to facilitate interoperability between data catalogs published on the Web.
  - Open data portals in Europe did not have the same schema originally
    - https://data.gov.uk/
  - Potential use case: support federated search
- No serialization enforced
- Reuses other vocabularies as well, e.g. Dublin Core

DCAT

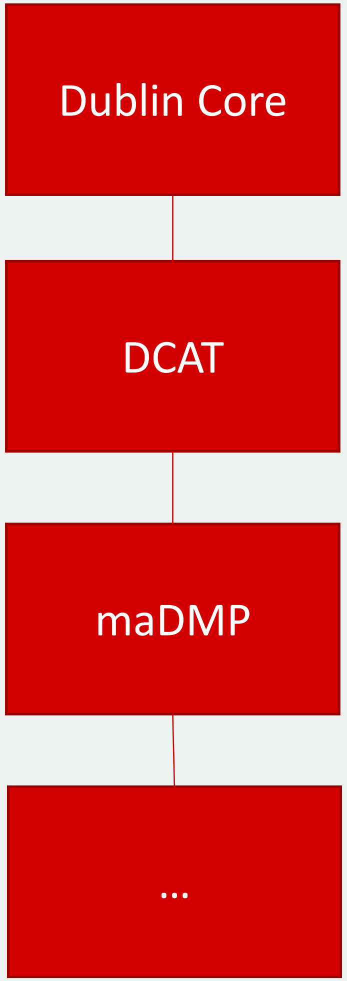


https://www.w3.org/TR/vocab-dcat-1/

(newer version exists, adds more concepts, does NOT make this one obsolete)

Application Profile

- Adds additional constraints
- Does NOT break compliance
- For example
  - Cardinality constraints (MAY -> MUST)
  - Sub-classes
  - Use further vocabularies
- For this reason DCAT has most of the fields optional
- DCAT itself uses Dublin Core in this way
- maDMP recommendation does the same
  - Reuses classes
  - Sets constraints
  - Adds new terms



10 PRINCIPLES FOR MADMPS

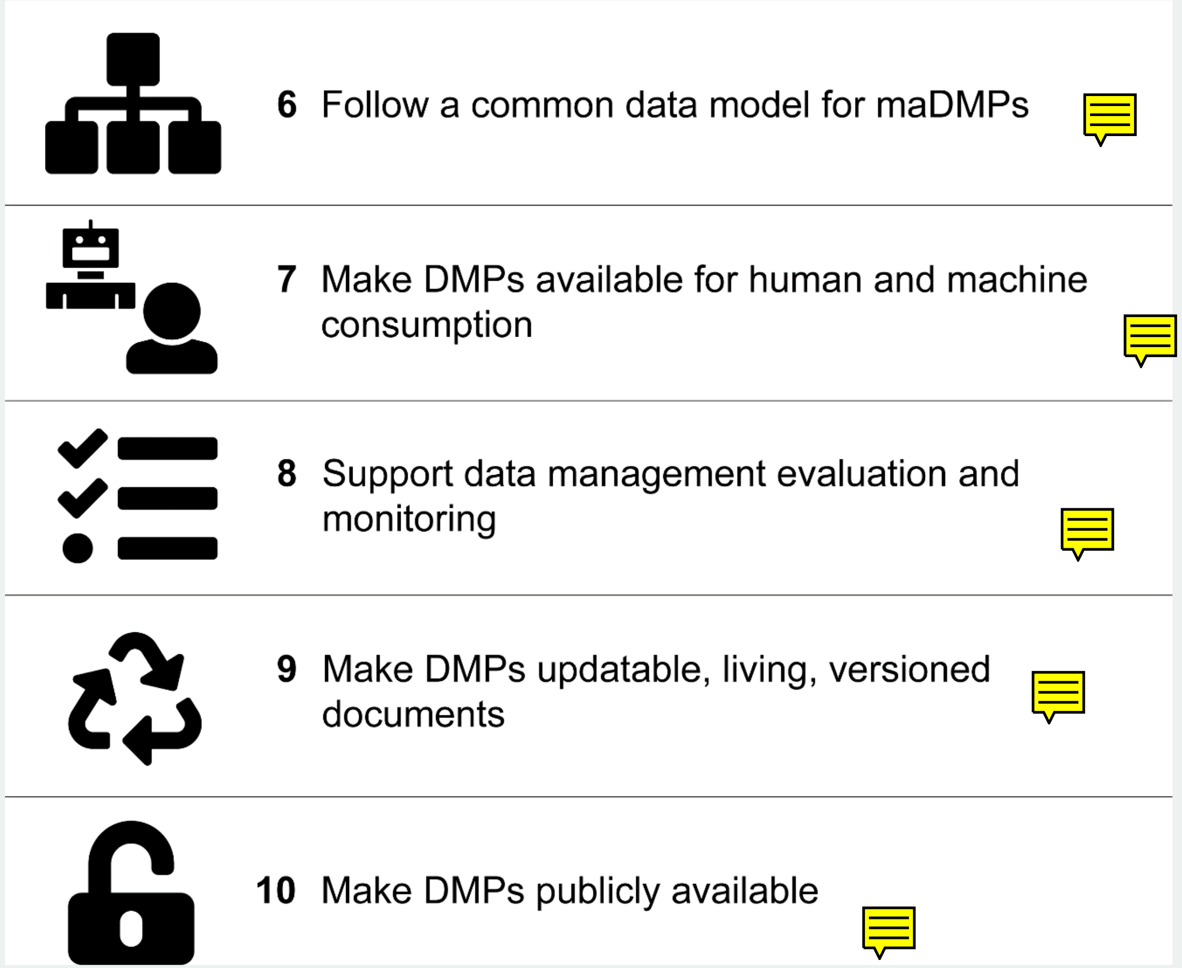
10 principles for maDMPs

1. Integrate DMPs with the workflows of all stakeholders in the research data ecosystem
2. Allow automated systems to act on behalf of stakeholders
3. Make policies (also) for machines, not just for people
4. Describe—for both machines and humans—the components of the data management ecosystem
5. Use PIDs and controlled vocabularies

Miksa T, Simms S, Mietchen D, Jones S (2019) Ten principles for machine-actionable data management plans. PLoS Comput Biol 15(3): e1006750. https://doi.org/10.1371/journal.pcbi.1006750

This is important…just saying…

10 principles for maDMPs



Miksa T, Simms S, Mietchen D, Jones S (2019) Ten principles for machine-actionable data management plans. PLoS Comput Biol 15(3): e1006750. https://doi.org/10.1371/journal.pcbi.1006750

This is important…just saying…

FAIR Data Austria

RDM INFRASTRUCTURE

RDM Infrastructure @ TU Wien

DMPs are not for funders only

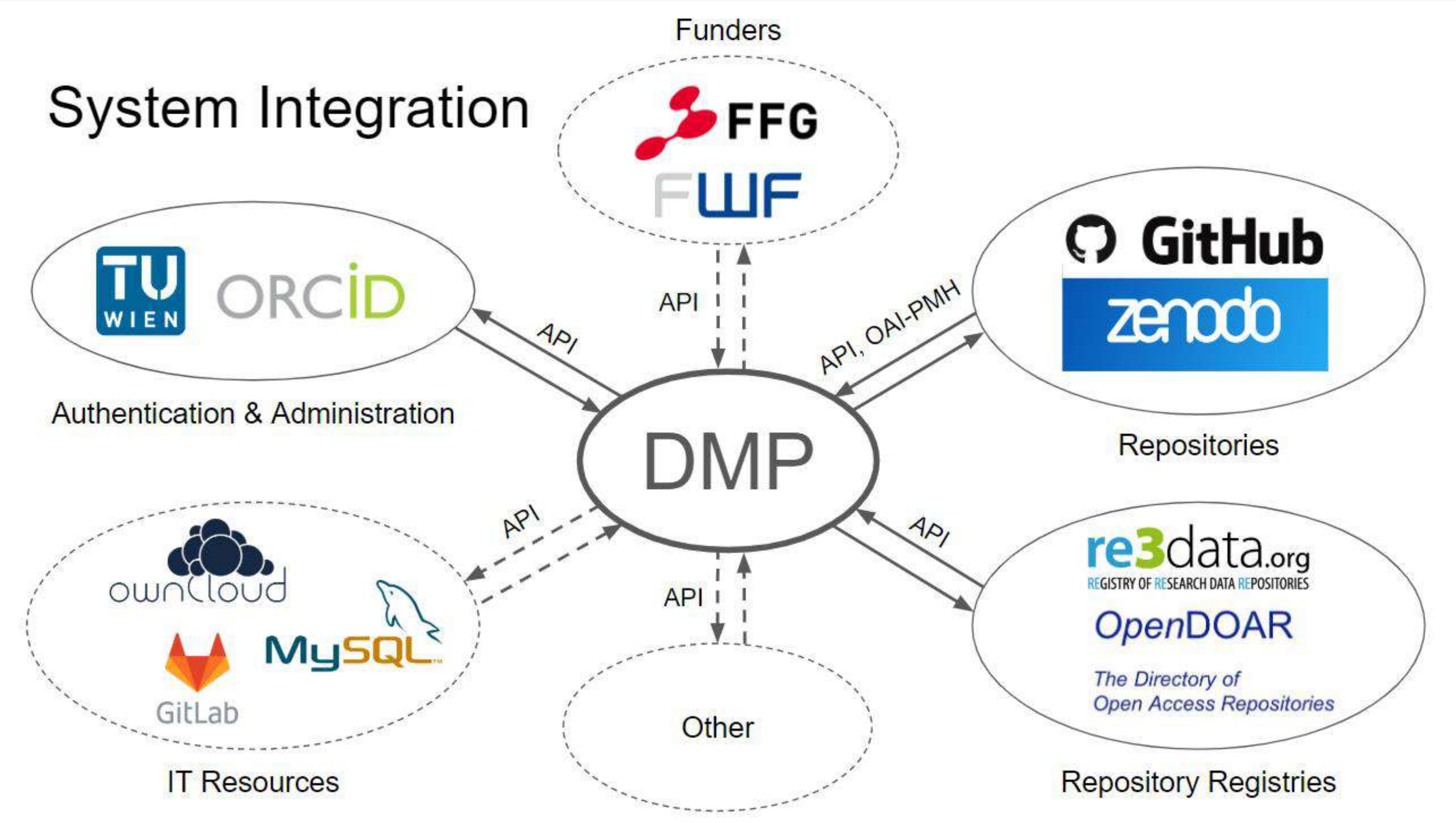
DMPs must also create benefits for researchers

- Less work
- Automation of tasks
- Reuse of information

DMPs are the ‘glue’ between different systems

- Automate
  - getting data in
  - getting data out

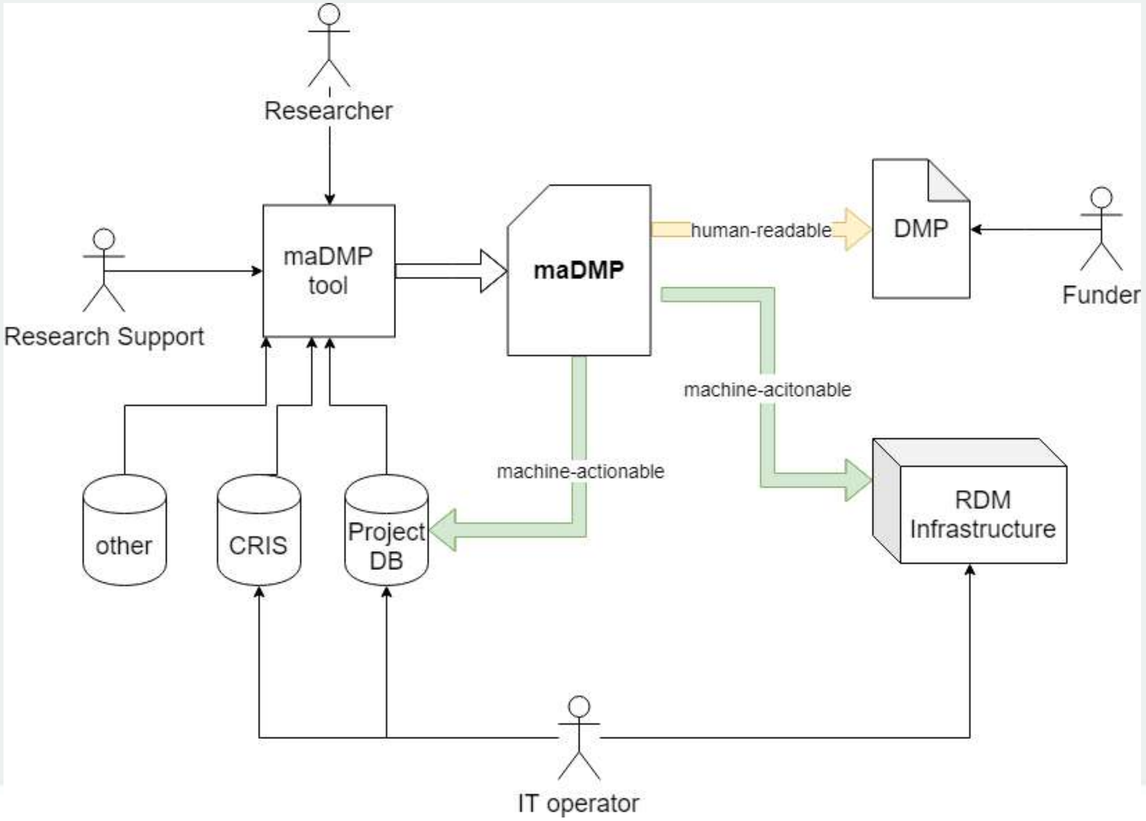
RDM Infrastructure @ TU Wien



maDMPs and RDM infrastructure

once-only principle

- do not ask researchers same questions in different places



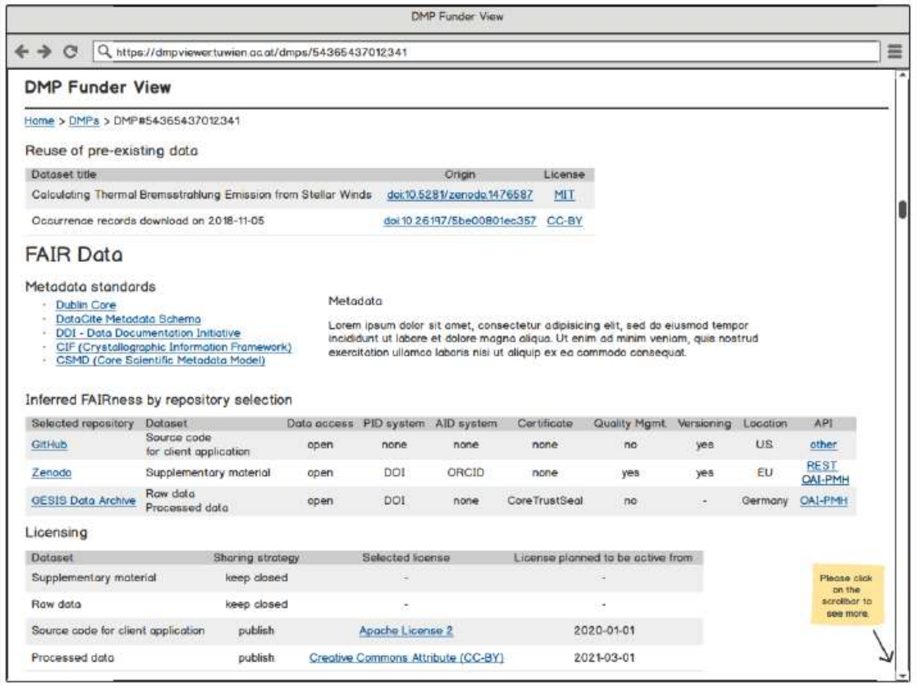
Mock-ups

Requirements collection

- Community
- TU Wien interviews

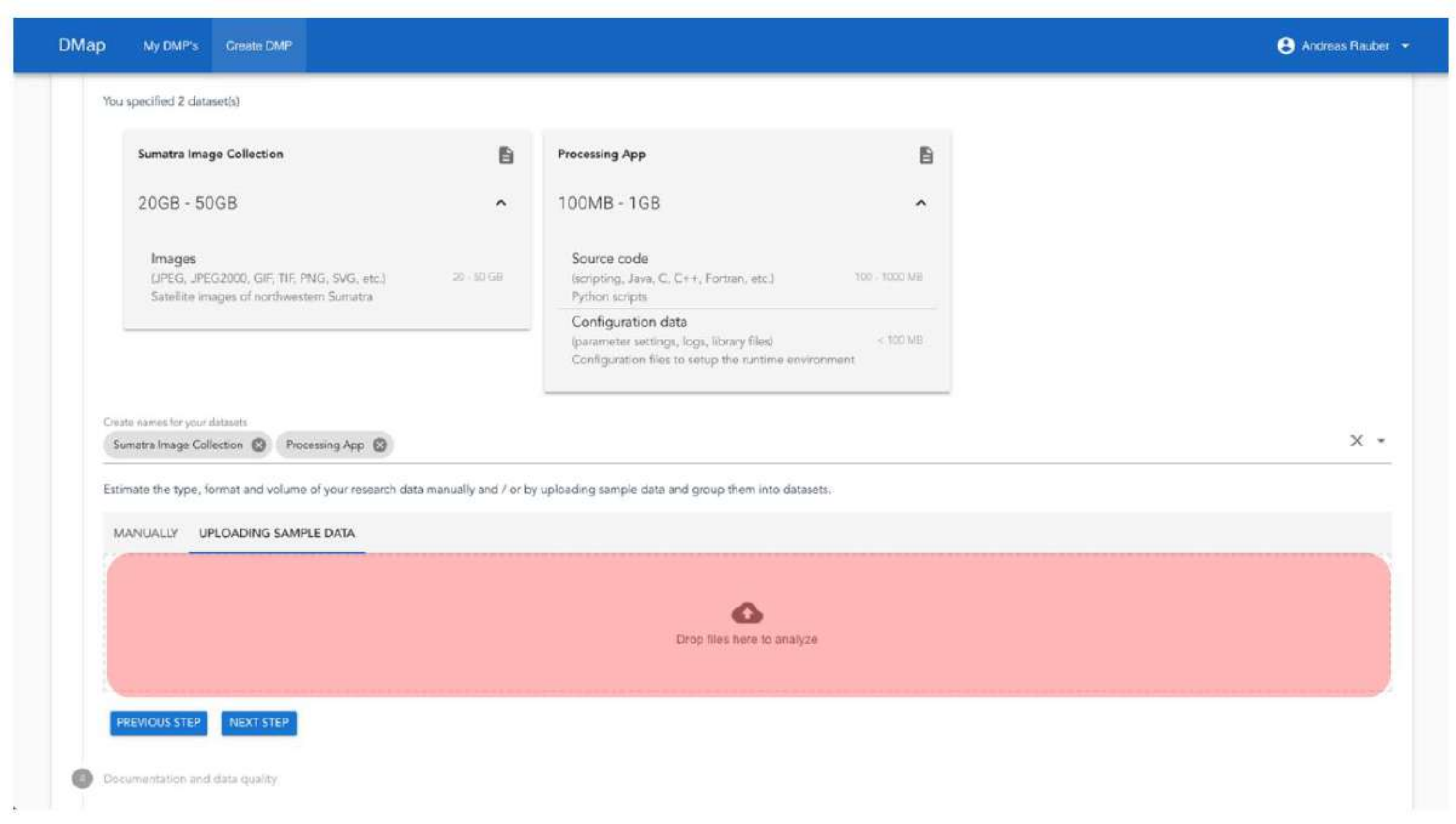
To be used for MVP

Implemented in DMap



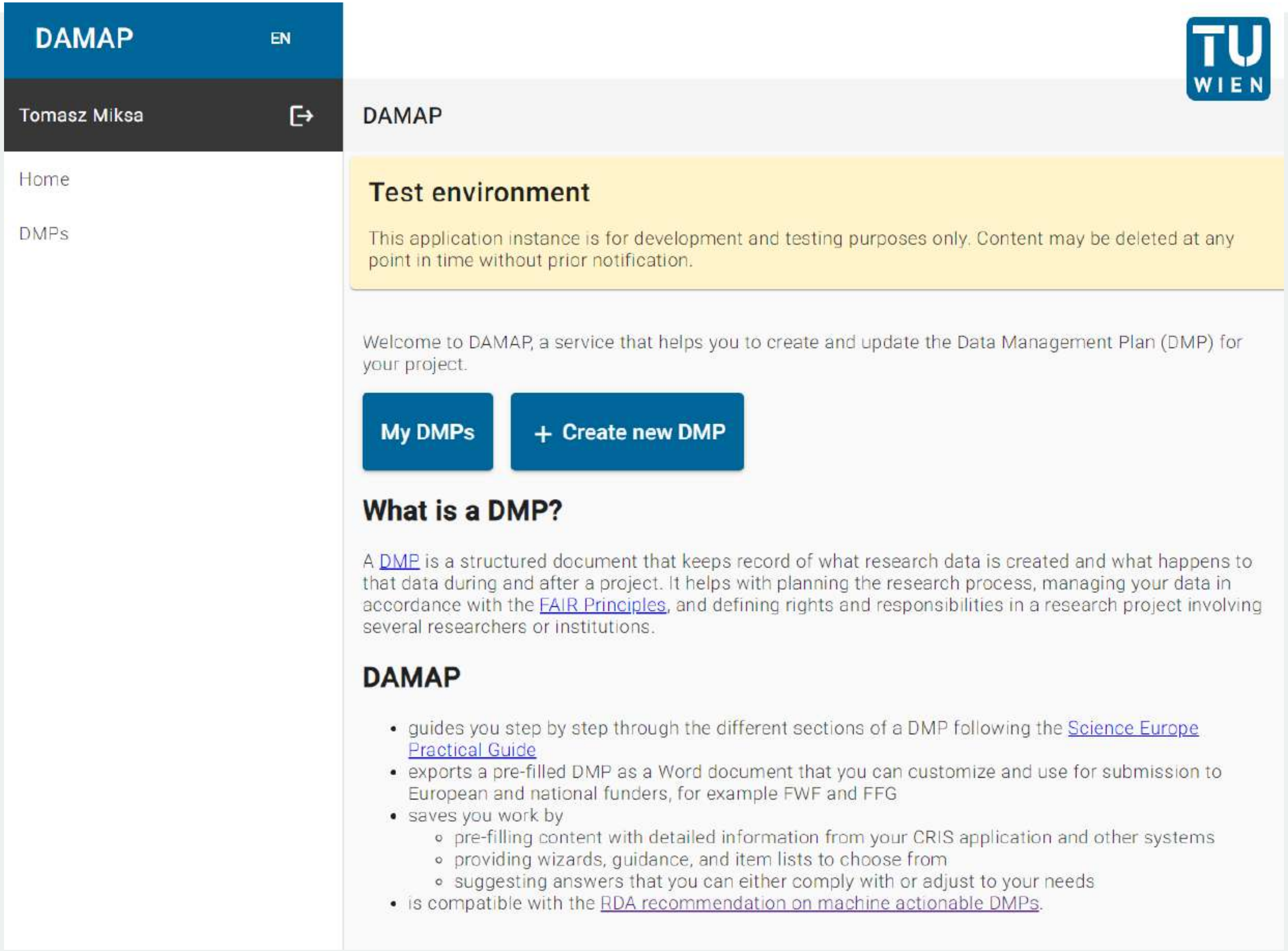
https://oblassers.github.io/dmap-mockups/

DMap



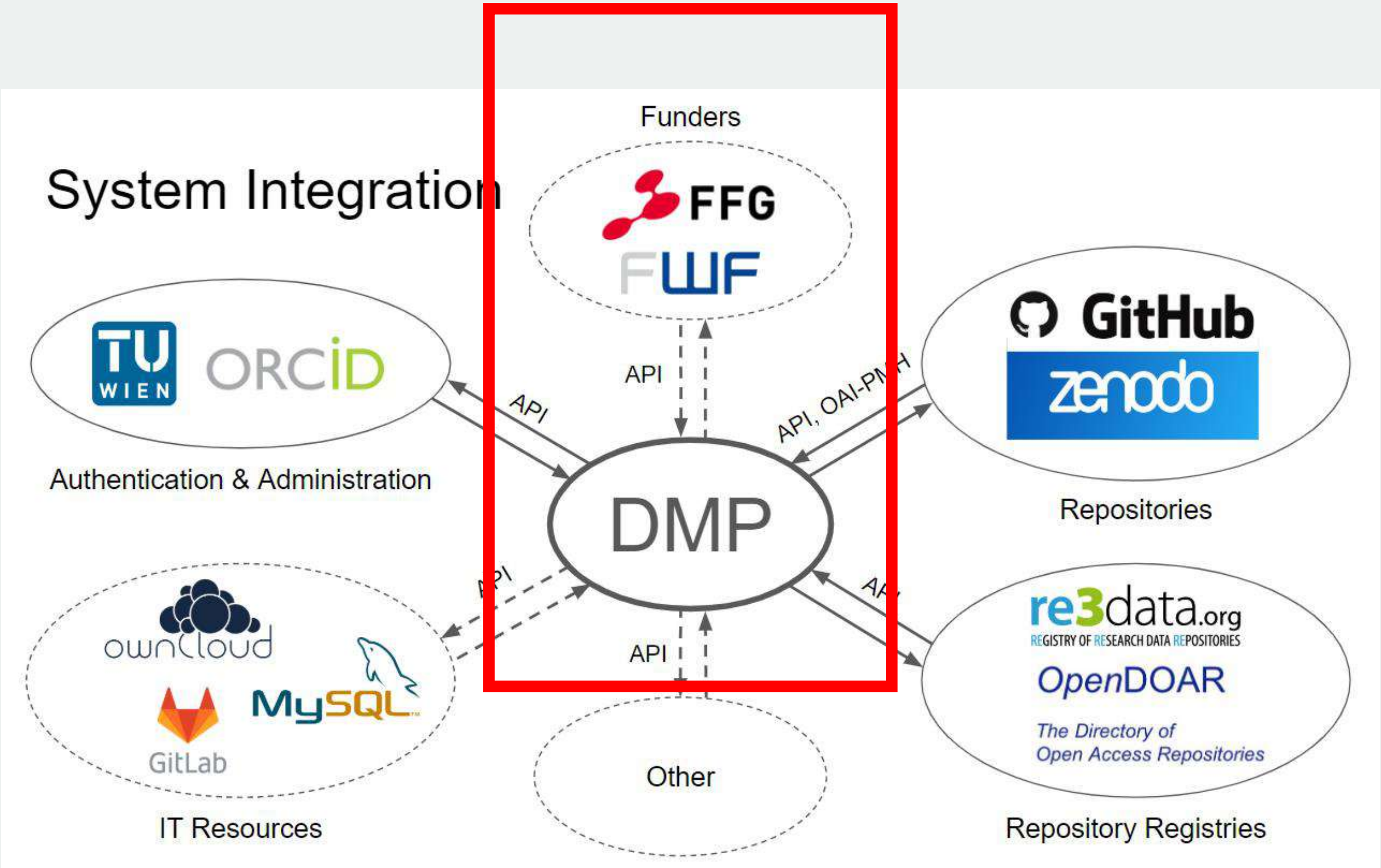
https://www.rd-alliance.org/system/files/documents/2019-RDA-DMAP-Oblasser.pdf

DAMAP

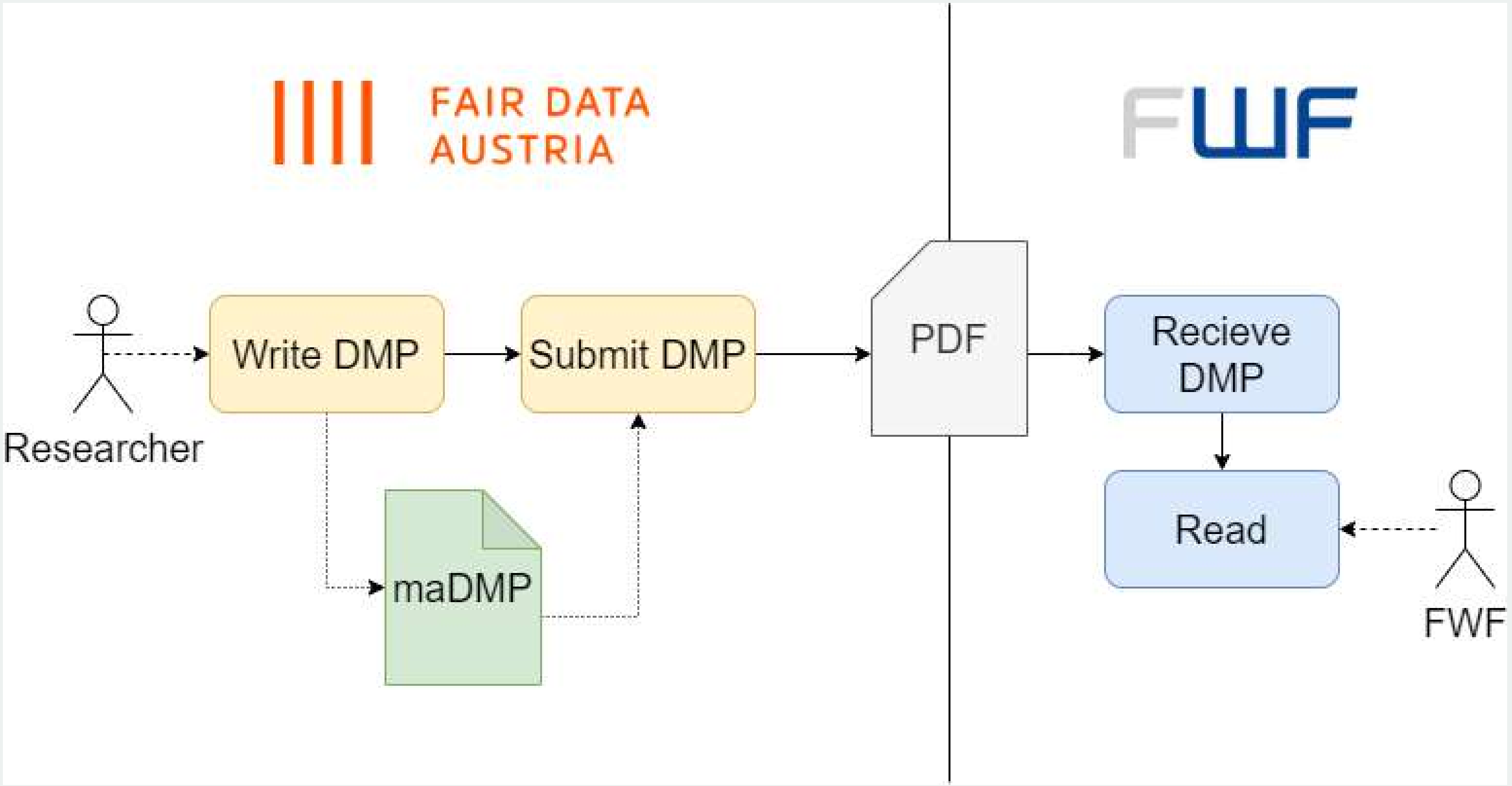


https://damap-qa.apps.dev.csd.tuwien.ac.at

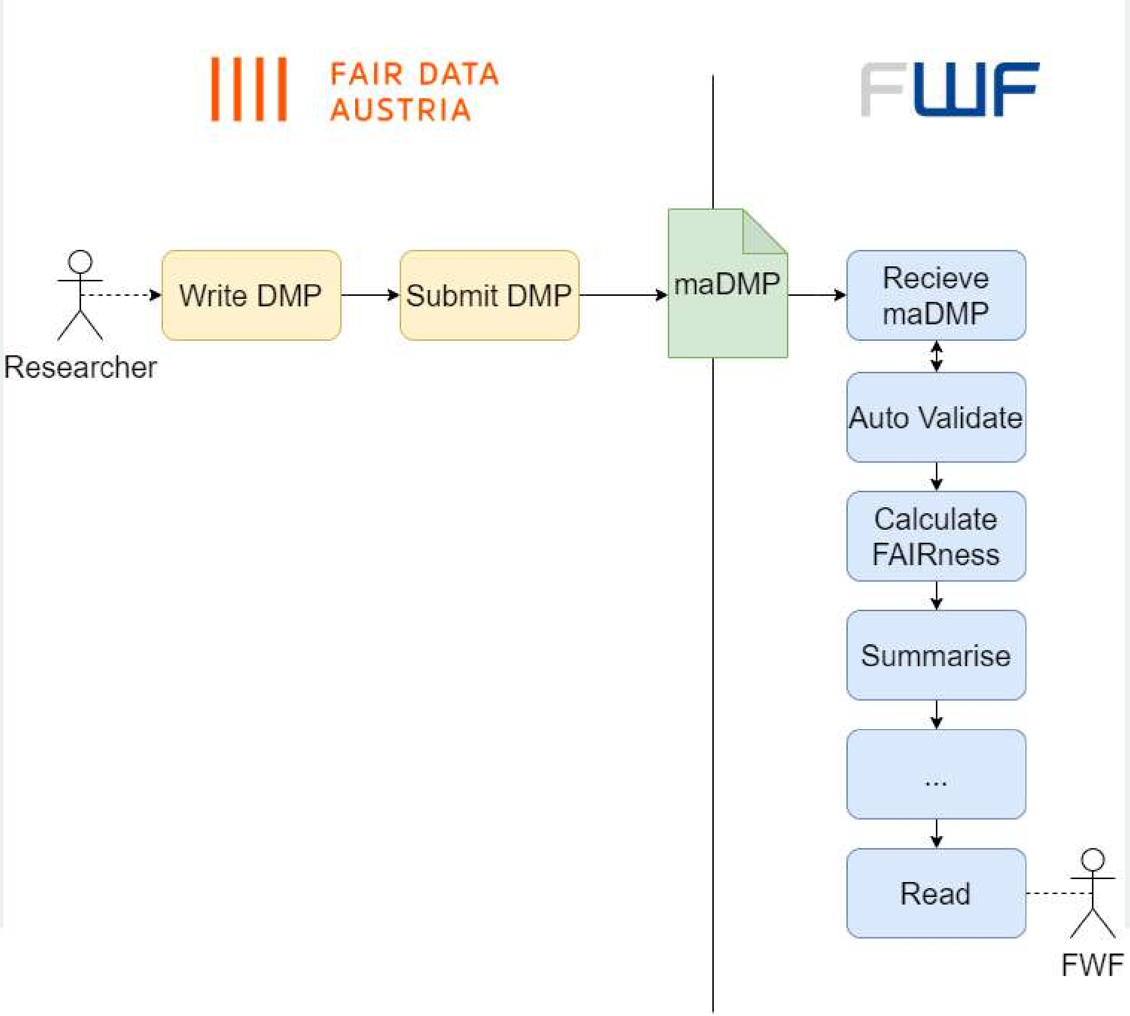
Next steps: Funder integration



Potential scenarios - mixed



Potential scenarios - full



SUMMARY

You should know

Why traditional DMPs are not enough

What is required to make DMPs machine-actionable

What are the related standards and how application profiles work

How common standard for maDMPs works

10 principles for maDMPs

Publications

- Tomasz Miksa, Simon Oblasser, and Andreas Rauber**. Automating research data management using machine-actionable data management plans**. ACM Transactions on Management Information Systems, 13(2), dec 2021.

- Tomasz Miksa, Paul Walk, Peter Neish, Simon Oblasser, Hollydawn Murray, Tom Renner, Marie-Christine Jacquemot-Perbal, João Cardoso, Trond Kvamme, Maria Praetzellis, Marek Suchánek, Rob Hooft, Benjamin Faure, Hanne Moa, Adil Hasan, and Sarah Jones. **Application profile for machine-actionable data management plans**. CODATA Data Science Journal, 20(1):32, October 2021

- João Cardoso, Leyla Jael Castro, and Tomasz Miksa**. Interconnecting Systems Using Machine-Actionable Data Management Plans Hackathon Report**. Data Science Journal, 20, 2021.

- Tomasz Miksa, Maroua Jaoua, and Ghaith Arfaoui. **Research Object Crates and Machine-actionable Data Management Plans**. In DaMaLOS - First Workshop on Data and Research Objects Management for Linked Open Science : Co-located at the International Semantic Web Conference ISWC 2020. PUBLISSO, November 2020.

- Simon Oblasser, Tomasz Miksa, Asanobu Kitamoto: **Finding a Repository with the Help of Machine-Actionable DMPs: Opportunities and Challenges**. IDCC 2020

- Tomasz Miksa, Stephanie Simms, Daniel Mietchen, Sarah Jones (2019) **Ten principles for machine-actionable data management plans**. PLOS Computational Biology 15(3): e1006750.

- Tomasz Miksa, Peter Neish, Paul Walk, Andreas Rauber: **Defining requirements for machine-actionable Data Management Plans**. iPres 2018

- Tomasz Miksa, João Cardoso, José Luis Borbinha: **Framing the scope of the common data model for machine-actionable Data Management Plans**. BigData 2018: 2733-2742

A Database Repository for Sensitive Data

Martin Weise

Information and Software Engineering
Technische Universität Wien

May 9th, 2022

Introduction

About me

This is my first lecture!

- PreDoc Researcher @IFS
- MSc in *Software Engineering & Internet Computing* 2022 (TU Wien)
- BSc in *Software & Information Engineering* 2019 (TU Wien)

Introduction

Problem Description

Digital repositories frequently struggle to make databases available in their collection:

1. **Separation of concerns**: data stewards receive a database without semantic understanding of the data, curation activities start already poorly
2. **Project phases**: deployed at research unit level, maintained by researchers and handover to IT staff once finished
3. **FAIR queries**: preserving databases, FAIR principles become a challenge
4. **Reproducability**: re-execution of a (persistently identified) query to reproduce research ouput is often times neglected, preservation is used in a archiving context
5. **Versioning**: monolithic records, large administrative overhead through lack of machine-readable interfaces

Digital repositories also struggle to make sensitive data available in their collection:

1. **Control**: maintaining protection over sensitive data
2. **Open science**: requires data sharing, conflicts with allowing access to third parties
3. **Data sharing**: give up control of the data, typically once this happens the data is gone

Definition: Sensitive Data [1]

Any ideas?

Introduction

Problem Description

Digital repositories also struggle to make sensitive data available in their collection:

1. **Control**: maintaining protection over sensitive data
2. **Open science**: requires data sharing, conflicts with allowing access to third parties
3. **Data sharing**: give up control of the data, typically once this happens the data is gone

Definition: Sensitive Data [1]

"[...] any information that is protected against unwarranted disclosure. Protection of data may be required for legal or ethical reasons, for issues pertaining to personal privacy, or for proprietary considerations."

Introduction

What is Data Visiting?

- Data stays under the control of the owner
  - Who can access
  - Over which period
  - Which subset of data
  - Answer which research question(s) and activities
- Allow consumers (i.e. analysts, machine learning algorithms) to come to the data
  - Ensure proper usage of data (=minimize risk)
  - Allow maximum usage of data
- Closely monitoring processes and interaction with the data during these visits
- Safe-guards to prevent accidental leakage and intentional data breaches

Introduction

Example: Intentional Data Breaches

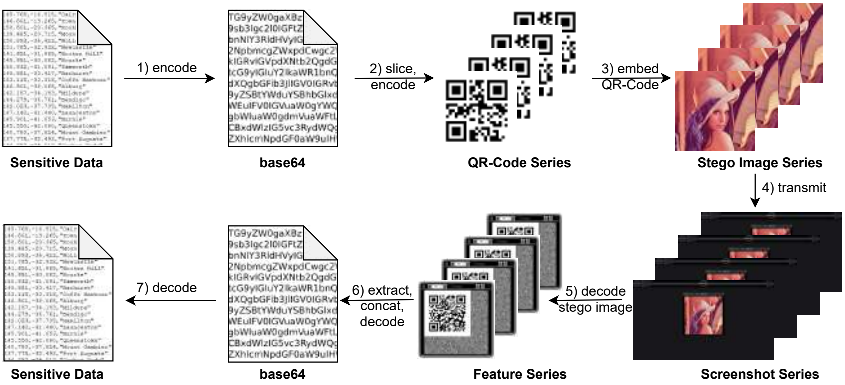


Figure: QR-Code Optical Covert Channel [2]

(Overview)

- *Infrastructure*: Openstack, Docker
- *Databases*: MariaDB, Postgres
- *Automatization*: Ansible, Python, Bash
- *Identity management*: FreeIPA
- *Retrieval*: ElasticSearch
- *Data visiting*: OpenVPN, TigerVNC

What makes this a repository?

We use a similar stack as most web shops nowadays, how does this differ?

Introduction

What makes this a repository

Technical Measures

Safeguarding data against technical threats and design flaws, provide a highly-controlled virtual research environment

Introduction

What makes this a repository

Technical Measures

Safeguarding data against technical threats and design flaws, provide a highly-controlled virtual research environment

Organizational Measures

Identification of researchers, research questions and guide all involved people through well-defined processes on need-to-know basis

Introduction

What makes this a repository

Technical Measures

Safeguarding data against technical threats and design flaws, provide a highly-controlled virtual research environment

Organizational Measures

Identification of researchers, research questions and guide all involved people through well-defined processes on need-to-know basis

Legal Measures

Build trust with data providers, provide legal framework to discourage information sharing, provide foundation for legal steps after (accidental) leaks

Technical Measures

Example: Homomorphic Encryption

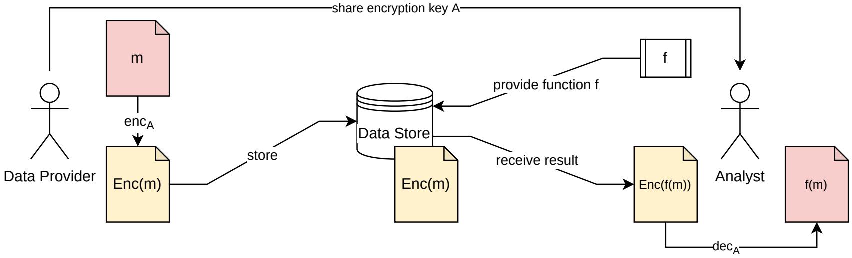


Figure: Simplified homomorphic encryption scheme with a *untrusted* Data Store

The Data Store is now hidden from the Analyst

Is this sufficient for most researchers?

Technical Measures

Example: Homomorphic Encryption

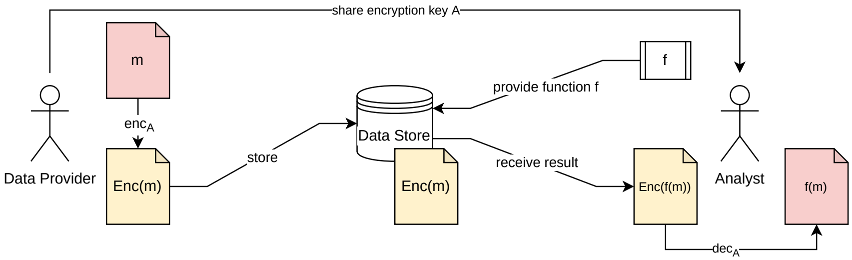


Figure: Simplified homomorphic encryption scheme with a *untrusted* Data Store

The Data Store is now hidden from the Analyst

Still not sufficient for exploratory/interactive analyses types required in many settings!

Technical Measures

Example: Secure Enclaves

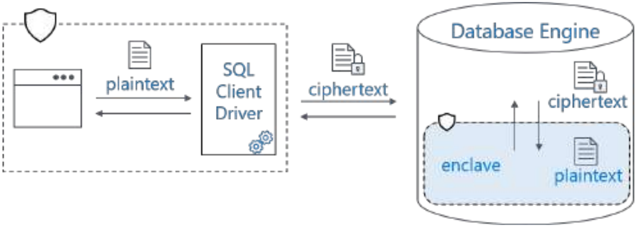


Figure: Permanent encryption of sensitive data in Azure SQL Database [3]

Secure enclave computation on untrusted hardware

Provide hardware isolation and memory encryption, isolating application code and data from untrusted hardware. Many manufacturers of CPUs offer secure enclave support for consumer electronics.

Technical Measures

Data Visiting: Air-Gap Isolation

"Five safe" dimensions [4] to maximize public value while also protecting personal rights of individuals and allowing flexible solutions:

- *Safe projects*: management decisions regarding appropriateness of the usage of the data through auditability and review processes.
- *Safe people*: identify individuals that access the sensitive data and require them to sign legally binding terms of use.
- *Safe data*: ensure appropriate data de-identification and access capabilities with respect to the research questions formulated.
- *Safe settings*: address the necessity of security and transparency to achieve trust with the public and data owners.
- *Safe outputs*: ensure only approved, aggregated research results can be exported.

Technical Measures

Data Visiting: Air-Gap Isolation

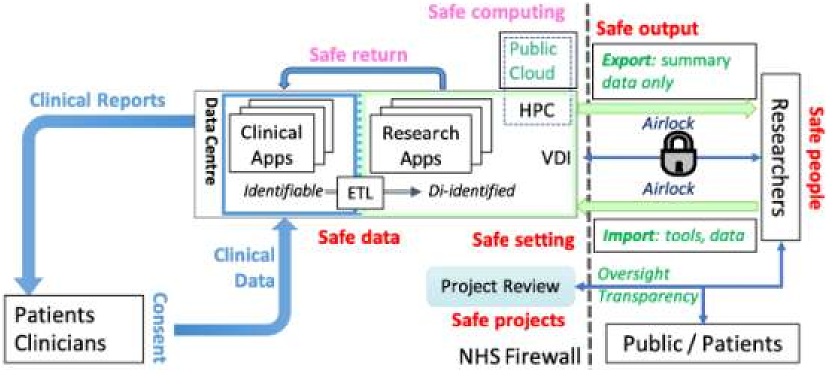


Figure: Schematic overview example of a TRE environment [5]

Technical Measures

Data Visiting: Air-Gap Isolation

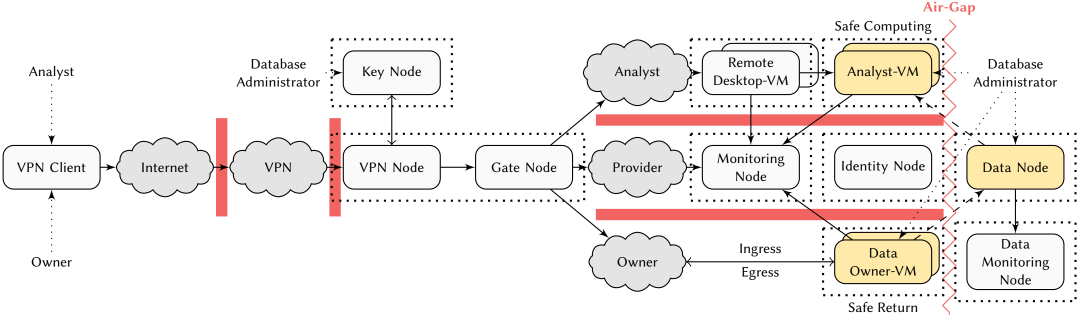


Figure: Our open-source technical architecture specification of OSSDIP [6]

ID	Sensor	Temp
1	A	23.1
2	B	25.8

(a) Original Table

ID	Sensor	Temp
1	A	22.1
2	B	25.8

(b) Corrected Table

Figure: Update operation on a table, creating a new version at a different timestamp

Why do we need data versioning? Why not use snaphots to cite them?

Any ideas?

Technical Measures

Data Versioning

ID	Sensor	Temp
1	A	23.1
2	B	25.8

(a) Original Table

ID	Sensor	Temp
1	A	22.1
2	B	25.8

(b) Corrected Table

Figure: Update operation on a table, creating a new version at a different timestamp

Why do we need data versioning? Why not use snaphots to cite them?

Full copies pollute the database, selection of the "right" snapshot difficult knowing only the query, challenges to archive every single snapshot, ...

Data Versioning (cont.)

ID	Sensor	Temp	Valid From	Valid To
1	A	23.1	t1	
2	B	25.8	t2	

(a) Original Table

ID	Sensor	Temp	Valid From	Valid To
1	A	23.1	t1	t3
2	B	25.8	t2	
1	A	22.1	t3	

(b) Corrected Table

Figure: Implementation of data versioning in MariaDB (v10.5)

Technical Measures

Accessing Sensitive Data

Machine-readability and -actionability through APIs:

- *Hypertext Transfer Protocol* (HTTP) API
  - REST constraints: client-server architecture, stateless, uniform resource identification
  - Service discovery: 8 services and 14 endpoints visible in the gateway
  - Allow any researcher to (re-)execute view-only queries [7]
- *Advanced Message Queue Protocol* (AMQP) API
  - Time-series tuple insert (i.e. from sensors)
  - One exchange per database, one queue per table, handover to HTTP API
- *Java Database Connectivity* (JDBC) API
  - Direct access for database experts to maintain the database

Interaction with external systems:

- DOI: DataCite schema to persistently identify a query and *mint a DOI* (soon)

Organizational Measures

Ingress Sensitive Data into the System

Data Owner wants to import a dataset into the infrastructure:

- Must sign a *data processing agreement*, receive training material, etc.
- Provide dataset (and metdata) via the GUI, receive an account
- Deposit into Data Owner-VM
- Database Administrator briefly connects Data Node and copies dataset, restores air-gap

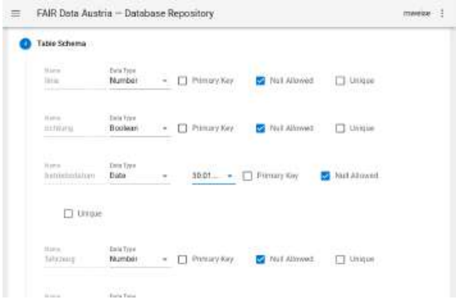


Figure: Provide dataset metadata

Organizational Measures

Ingress Sensitive Data into the System

Add metadata to the table schema:

- For now, only units of measurement [8] (centimeter, kilogramm, etc.)
- Assign a unit to each column (where applicable)

Why is this important?

Any ideas?

Figure: Assign a unit of measurement

Organizational Measures

Ingress Sensitive Data into the System

Add metadata to the table schema:

- For now, only units of measurement [8] (centimeter, kilogramm, etc.)
- Assign a unit to each column (where applicable)

Why is this important?

We assign specific semantic, machine-readable concept to a column that allows conversion to other concepts, e.g. 100 [°C] = 212 [°F]

Figure: Assign a unit of measurement

Organizational Measures

Ingress Sensitive Data into the System

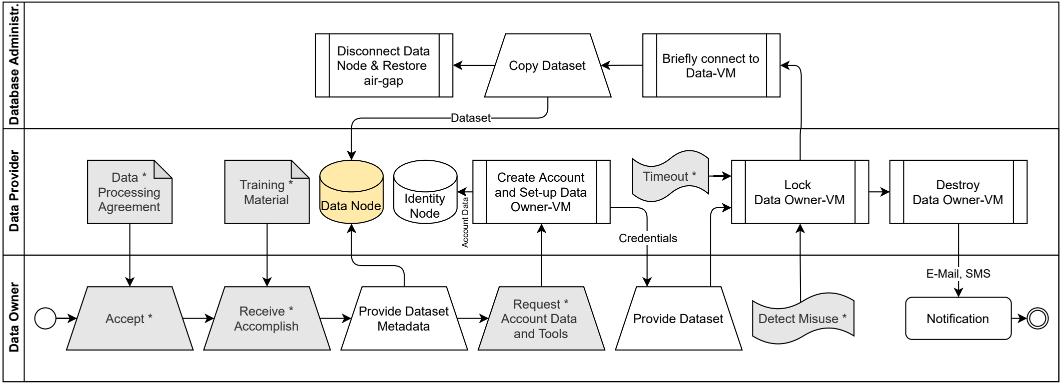


Figure: Complete data ingress process [6]

Organizational Measures

Requesting Access to Sensitive Data

1. *Findability*
   - *F1*: Assign a PID to the query that creates a subset (for the Analyst)
   - *F2*: Add unit of measurement, require metadata on data upfront
   - *F3*: Link PID of data along with the metadata
   - *F4*: Metadata is searchable by ElasticSearch
2. *Accessibility*
   - *A1*: open and authentication via HTTPS, AMQP, JDBC
   - *A2*: metadata always available in metadata database
3. *Interoperability*
   - *I1*: OWL/RDF concepts for units of measurement
   - *I2*: no metadata on the services yet, in development [9]
   - *I3*: metadata is interlinked in the metadata database
4. *Reuseability*
   - *R1*: open-source license, describe the data (currently only free form)

FAIR Sensitive Data

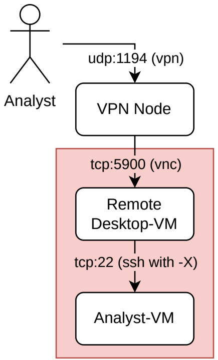
Even though the data is not open, the dataset is still FAIR!

Organizational Measures

Requesting Access to Sensitive Data

Analyst wants to access a sensitive dataset:

- Send personal identification data along with proposal
  - Required sensitive data (metadata is known)
  - Required tools to analyze data
- Extract the required subset via query (and store this query in the query store with timestamp, etc.)
- Potentially apply fingerprinting to the subset [10]
- Provision of Analyst-VM with the imported subset and Remote Desktop-VM



Organizational Measures

Requesting Access to Sensitive Data

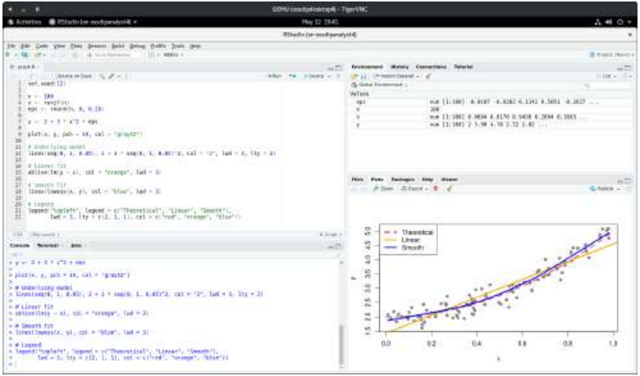


Figure: Work with the data with pre-approved tools via VPN+VNC

Organizational Measures

Requesting Access to Sensitive Data

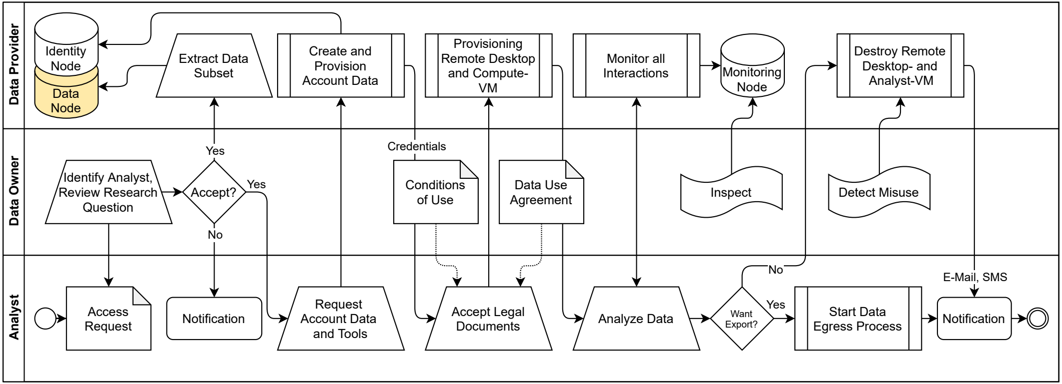


Figure: Complete data access process [6]

Organizational Measures

Subset Generation

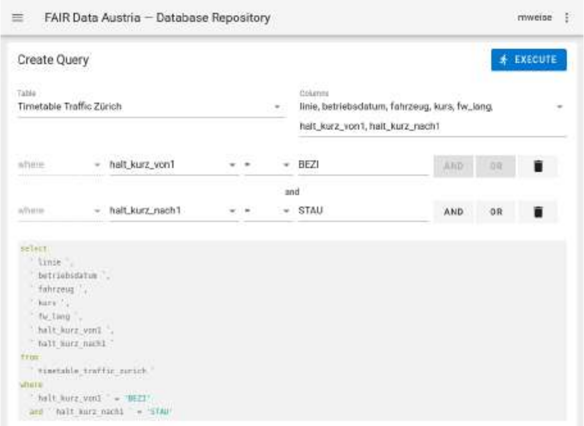


Figure: Lines from town hall (“Bezirksgebäude”) to Staudenbühl

Organizational Measures

Persistent Identification of Subsets

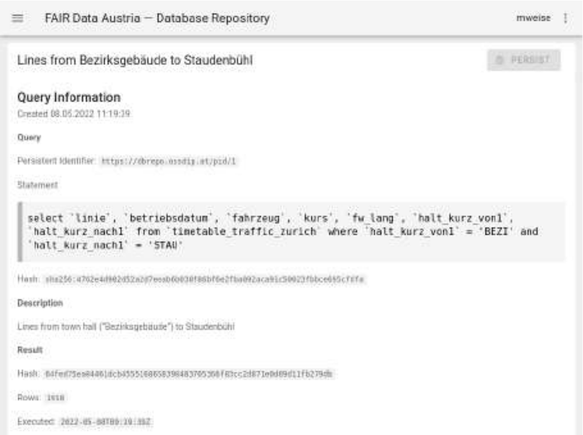


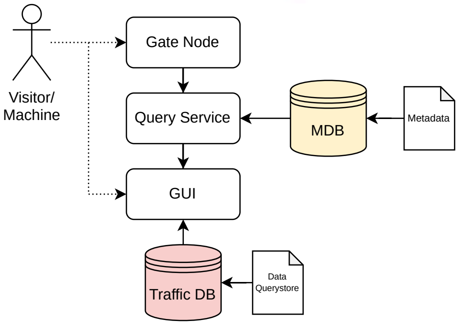
Figure: Query from the query store, persisted in the metadata database

Organizational Measures

Persistent Identification of Subsets

Stored query:

- Query ("raw" and normalized)
- Timestamp of creation
- Timestamp of execution
- Result hash
- Creator



How does a stored query differ from a persisted query?

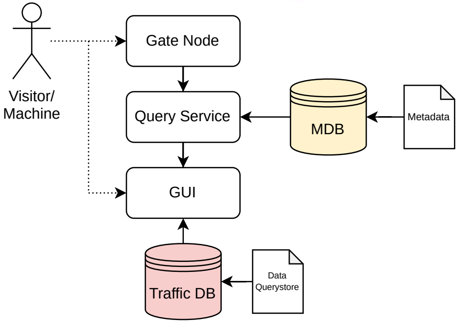
Any thoughts?

Organizational Measures

Persistent Identification of Subsets

Stored query:

- Query (“raw” and normalized)
- Timestamp of creation
- Timestamp of execution
- Result hash
- Creator



How does a stored query differ from a persisted query?

Persisted queries store metadata also in the metadata database. When the data is not available anymore, the metadata is still available!

Legal Measures

Agreements

Overall goal: prevent (unauthorized) leaks of sensitive data, what do we need legally?

- Trusted computing infrastructure for Data Owners
- Conditions of use for Analysts
  - Prohibition of data download
  - Prohibition of de-anonymization
  - Non-Disclosure Agreement (NDA)
  - Agreement to extensive monitoring
- Information in case of unauthorized leaks
  - Personal identifiable information
  - Evidence of misuse (compare to agreed research questions)
  - Evidence of relationship between leaked dataset and provided dataset (i.e. through fingerprinting)

Conclusion

TECHNISCHE UNIVERSITÄT WIEN

	Traditional Repository[1]	**Database Repository**
Representation	File storage	Database engine
Versioning	Snapshots	Temporal tables
Identification	Record	Query/Subset
Storage	Filesystem	Database
Concerns	Researcher	Database expert
Use-case	Deposit	Continous work

Table: Key differences between a traditional repository a database repositor

[1] e.g. Invenio, Dataverse, Figshare, Mendeley Data, Open Science Framework

Conclusion

A Database Repository for Sensitive Data

Important concepts for the exam:

1. Sensitive data and data visiting
2. Three aspects of sensitive data repositories
3. Data versioning
4. Stored and persisted queries
5. FAIR sensitive data

Conclusion

Learn more about these projects:

- Database Repository
  https://dbrepo-docs.ossdip.at
- Secure Data Infrastructure / Data Visiting
  https://ossdip.at

Talk to me after the lecture or drop me a mail martin.weise@tuwien.ac.at!

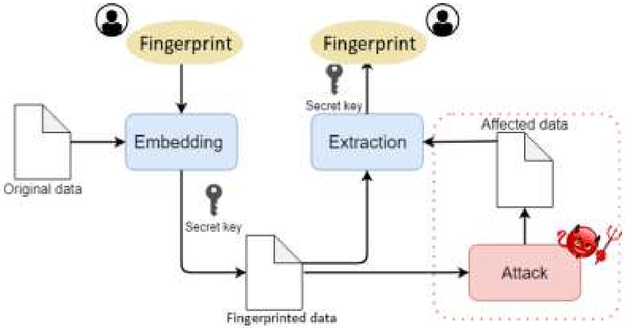


Figure: General fingerprinting scheme [10]

Backup Slides

Marking Ownership (cont.)

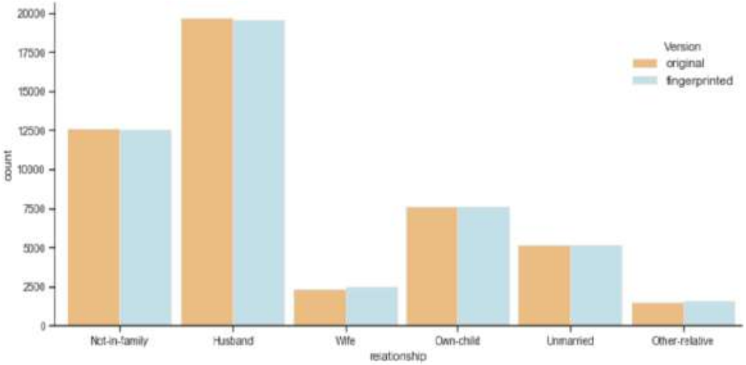


Figure: Distribution of a categorical attribute before and after fingerprinting [10]

Bibliography I

TECHNISCHE UNIVERSITÄT WIEN

- CSC - IT Center for Science, "Definition of sensitive data," [Online]. URL: https://research.csc.fi/definition-of-sensitive-data, accessed 2022-05-07.
- M. Weise, "A QR-Code Optical Covert Channel in an Air-Gapped Secure Data Infrastructure," Master's thesis, Technische Universität Wien, 2022.
- J. Szymaszek, "Always Encrypted with Secure Enclaves in Azure SQL Database Preview," [Online]. URL: https://techcommunity.microsoft.com/t5/azure-sql-blog/always-encrypted-with-secure-enclaves-in-azure-sql-database/ba-p/2051544, 2021, accessed 2022-05-07.
- T. Desai, F. Ritchie, and R. Welpton, "Five Safes: Designing data access for research," 2016. doi: 10.13140/RG.2.1.3661.1604 Economics Working Paper Series 1601.

Bibliography II

- UK Health Data Research Alliance and NHSX, "Building Trusted Research Environments - Principles and Best Practices; Towards TRE Ecosystems," 2021. doi: 10.5281/zenodo.5767586
- M. Weise, F. Kovacevic, N. Popper, and A. Rauber, "Open Source Secure Data Infrastructure and Processes Supporting Data Visiting," *Data Science Journal*, vol. 21, no. 1, p. 4, 2022. doi: 10.5334/dsj-2022-004
- A. Trisovic, P. Durbin, T. Schlatter *et al.*, "Advancing Computational Reproducibility in the Dataverse Data Repository Platform," in *Proceedings of the 3rd International Workshop on Practical Reproducible Evaluation of Computer Systems*, New York, NY, USA, 2020. doi: 10.1145/3391800.3398173 p. 15–20.
- H. Rijgersberg, "Ontology of Units of Measure," https://github.com/HajoRijgersberg/OM, 2022, accessed 2022-05-08.

Bibliography III

- L. O. Bonino, K. Burger, and R. Kaliyaperumal, "FAIR Data Point," [Online]. URL: https://specs.fairdatapoint.org/#repositorymetadata, 2021, accessed 2022-05-08, version 1.0.
- T. Sarcevic and R. Mayer, "Fingerprinting Relational Data Sets," [Poster]. [Online]. URL: https://www.sba-research.org/wp-content/uploads/2020/09/MLDM_Fingerprinting-Relational-Data-Sets_WEB.pdf, 2020, accessed 2022-05-07.

Repositories I:
TU Data[1] and its design

[1] The name finding process is still ongoing… Let’s hope it won’t be TURD again

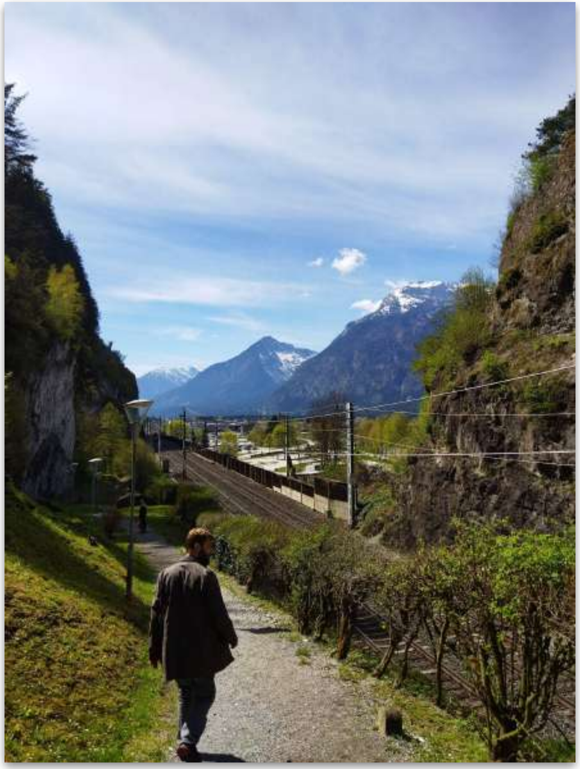
About Me

Maximilian Moser

- From the Tyrolean Alps
- Studied Software Engineering at TU Wien
- Not a lecturer, but a DevOp for TU Data

TU WIEN

mmoser@tuwien ~ $



What is a Research Data Repository?

- Central place for researchers to deposit and share their data
  - No more Google Drives and institute servers…
- For *all kinds* of research data
  - Stores datasets along with *metadata*
- Helps with FAIR-ification of the datasets
  - Provides visibility and access restrictions
- Ensures that data stays available
  - Archival and preservation

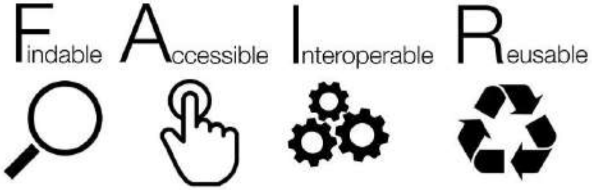
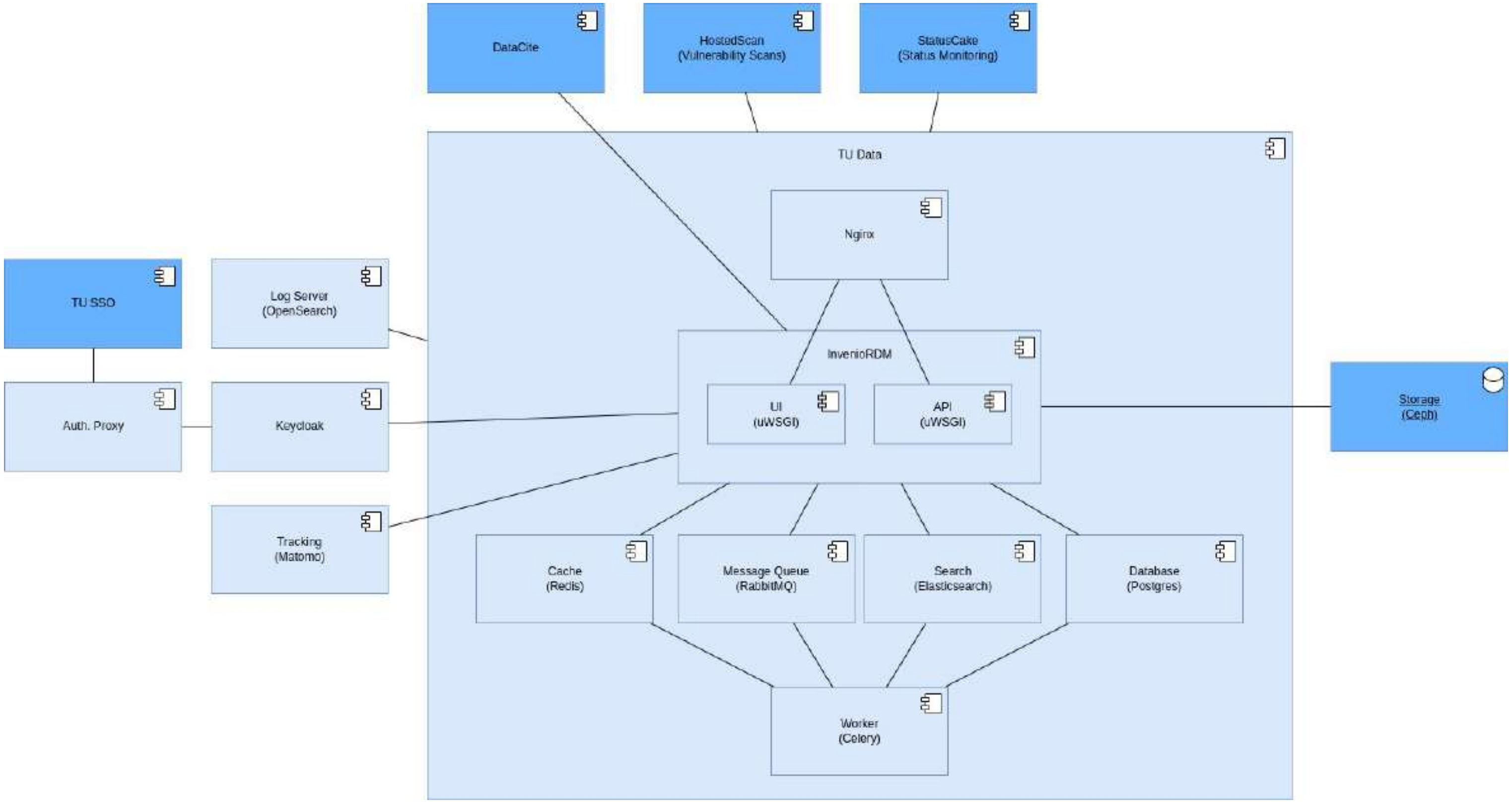


Image by [SangyaPundir]

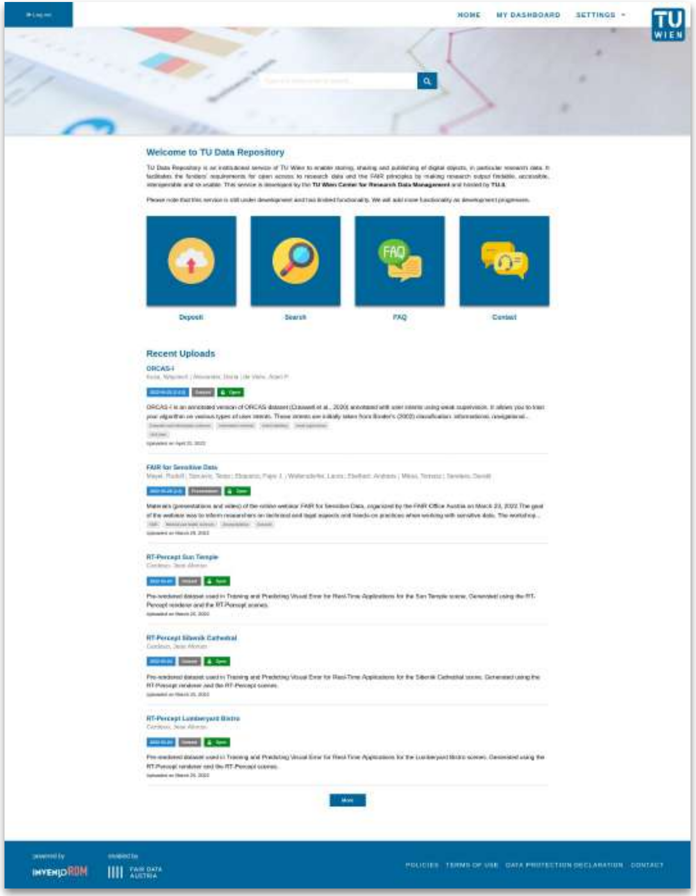
What is InvenioRDM / TU Data?

- File-based research data repository
  - Developed by CERN and partners: https://inveniosoftware.org/products/rdm/
- Flask Application
  - Frontend: React, Semantic UI
  - Backend: Python
- Highly customizable and extensible
  - TU Data is a themed and customized variant of InvenioRDM
- Still under development

Architecture Overview

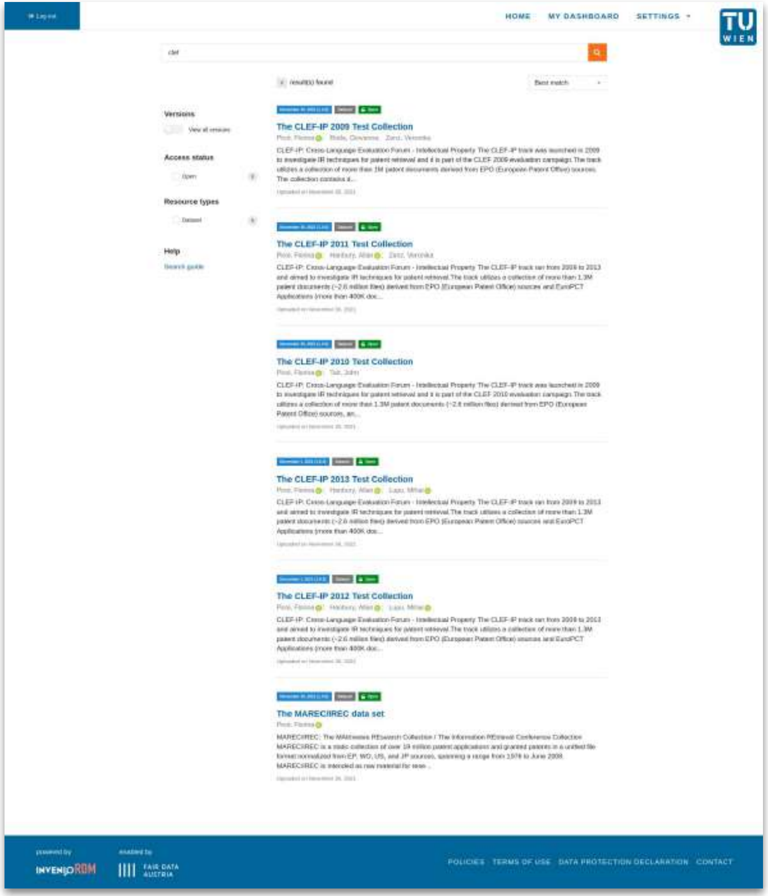


Frontpage

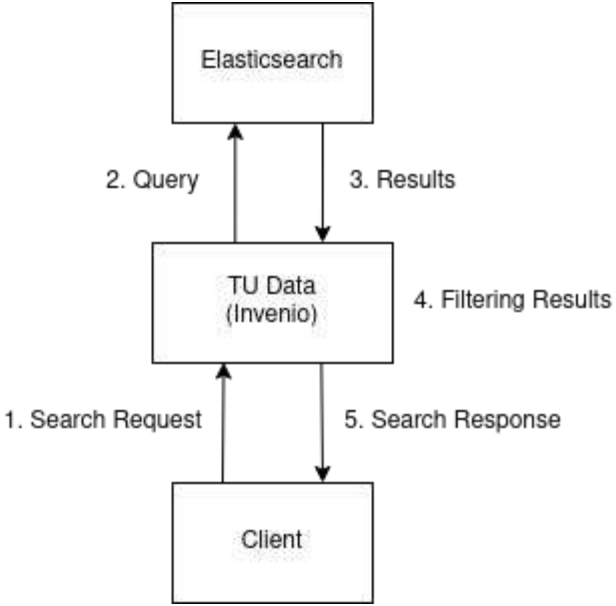


https://researchdata.tuwien.ac.at

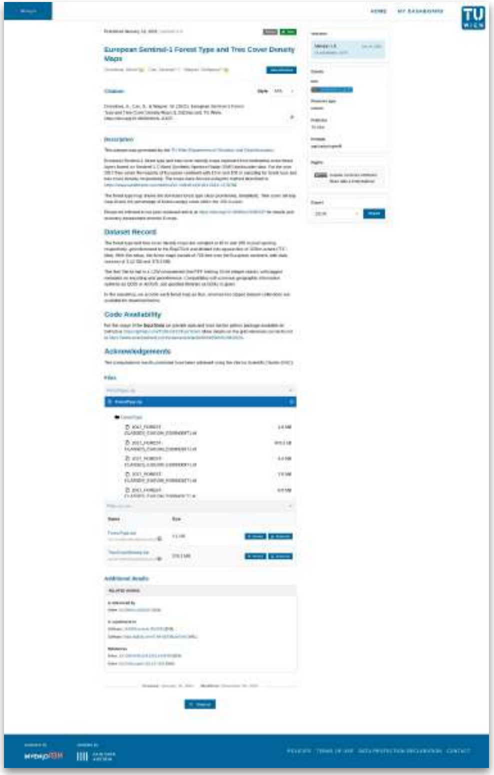
Search


HOME
MY DASHBOARD
SETTINGS
TU WIEN
Versions
Access status
Resource types
Help
Search guide
The CLEF-IP 2009 Test Collection
The CLEF-IP 2011 Test Collection
The CLEF-IP 2010 Test Collection
The CLEF-IP 2013 Test Collection
The CLEF-IP 2012 Test Collection
The MAREC/IREC data set
POLICIES
TERMS OF USE
DATA PROTECTION DECLARATION
CONTACT
FAIR DATA AUSTRIA

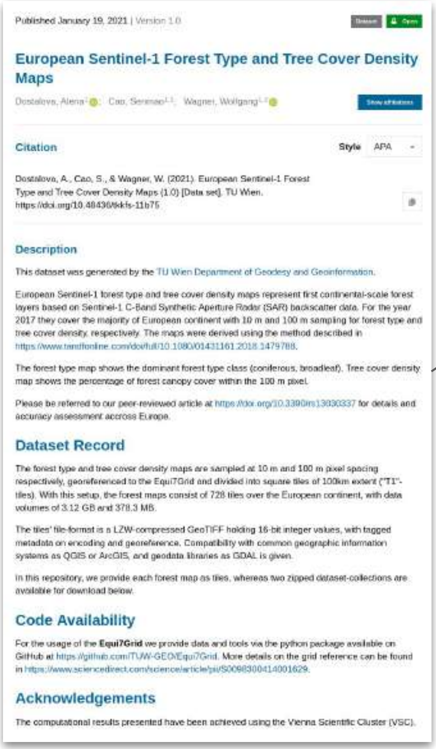
Search - Workflow



Record Landing Page



Record Landing Page



Published January 19, 2021 | Version 1.0

Dataset | Open

European Sentinel-1 Forest Type and Tree Cover Density Maps

Dostalova, Alena[1] Cao, Senmao[1,3] Wagner, Wolfgang[1,2]

Show affiliations

Citation

Style APA

Dostalova, A., Cao, S., & Wagner, W. (2021). European Sentinel-1 Forest Type and Tree Cover Density Maps (1.0) [Data set]. TU Wien. https://doi.org/10.48436/kkfs-11b75

Description

This dataset was generated by the TU Wien Department of Geodesy and Geoinformation.

European Sentinel-1 forest type and tree cover density maps represent first continental-scale forest layers based on Sentinel-1 C-Band Synthetic Aperture Radar (SAR) backscatter data. For the year 2017 they cover the majority of European continent with 10 m and 100 m sampling for forest type and tree cover density, respectively. The maps were derived using the method described in https://www.tandfonline.com/doi/full/10.1080/01431161.2018.1479788.

The forest type map shows the dominant forest type class (coniferous, broadleaf). Tree cover density map shows the percentage of forest canopy cover within the 100 m pixel.

Please be referred to our peer-reviewed article at https://doi.org/10.3390/rs13030337 for details and accuracy assessment accross Europe.

Dataset Record

The forest type and tree cover density maps are sampled at 10 m and 100 m pixel spacing respectively, georeferenced to the Equi7Grid and divided into square tiles of 100km extent ("T1"-tiles). With this setup, the forest maps consist of 728 tiles over the European continent, with data volumes of 3.12 GB and 378.3 MB.

The tiles' file-format is a LZW-compressed GeoTIFF holding 16-bit integer values, with tagged metadata on encoding and georeference. Compatibility with common geographic information systems as QGIS or ArcGIS, and geodata libraries as GDAL is given.

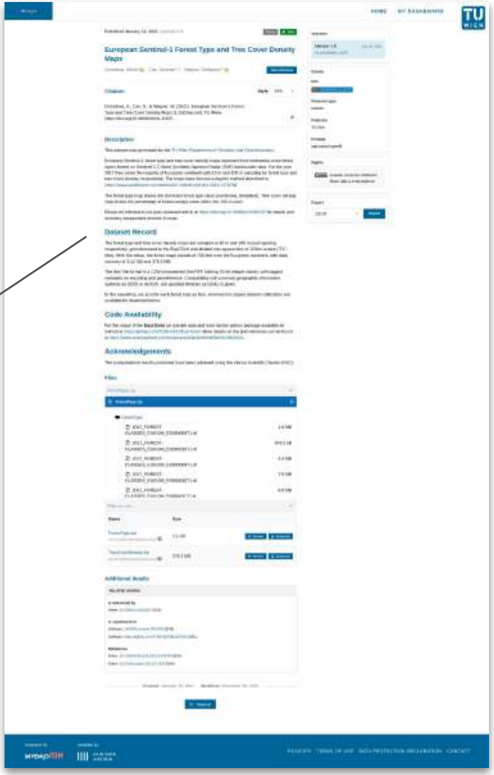
In this repository, we provide each forest map as tiles, whereas two zipped dataset-collections are available for download below.

Code Availability

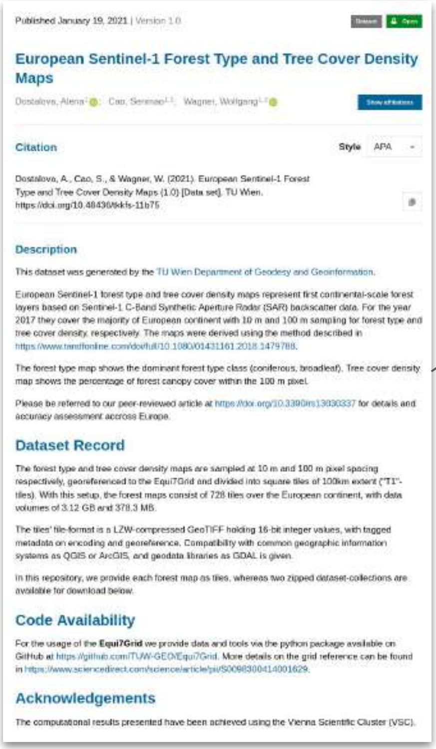
For the usage of the **Equi7Grid** we provide data and tools via the python package available on GitHub at https://github.com/TUW-GEO/Equi7Grid. More details on the grid reference can be found in https://www.sciencedirect.com/science/article/pii/S0098300414001629.

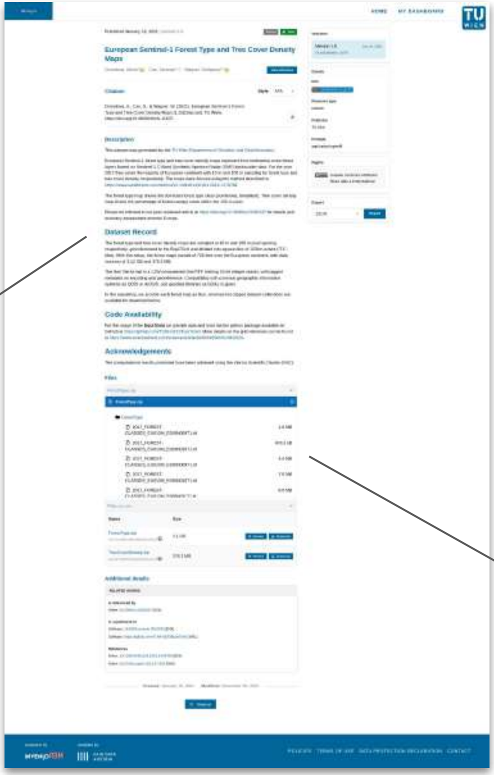
Acknowledgements

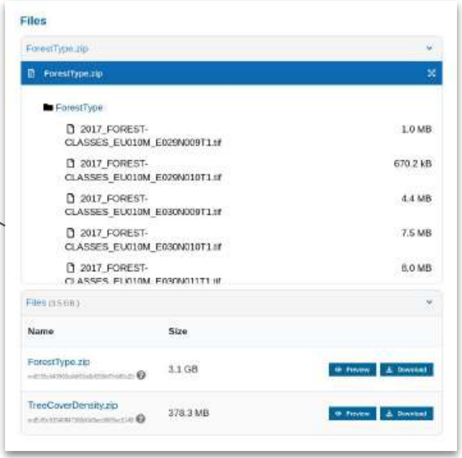
The computational results presented have been achieved using the Vienna Scientific Cluster (VSC).



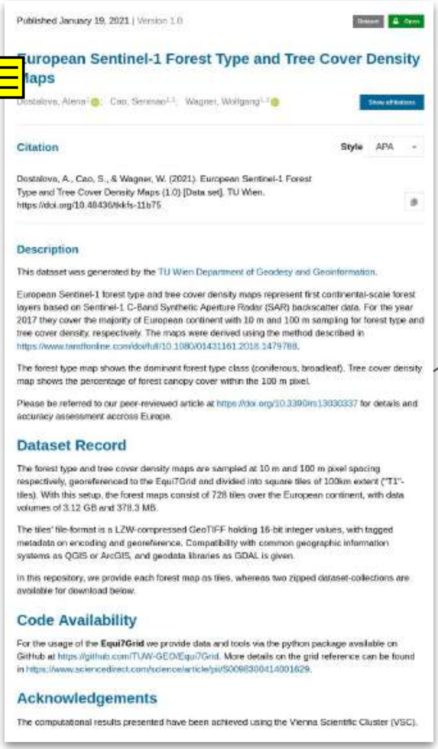
Record Landing Page

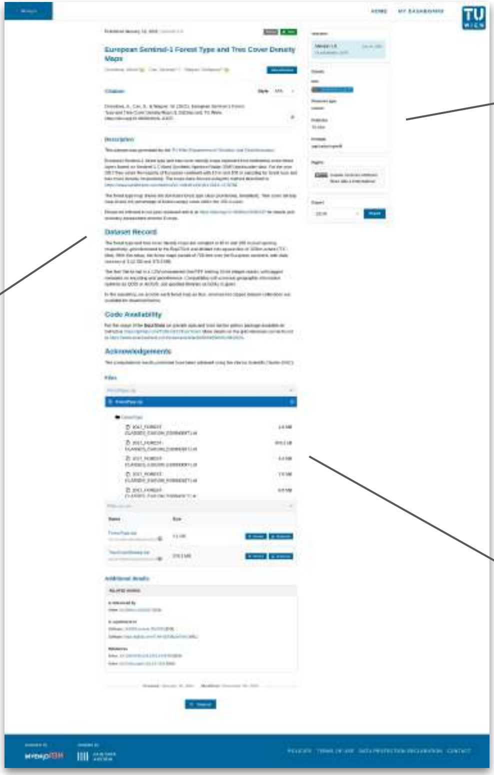


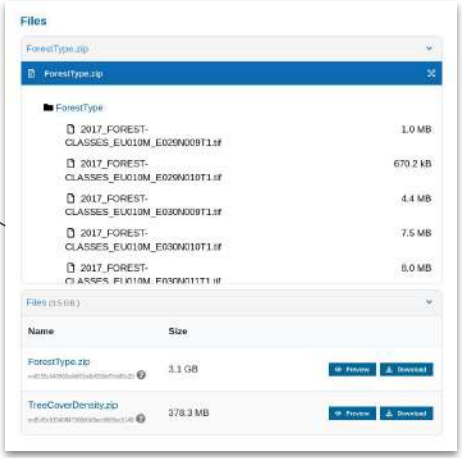




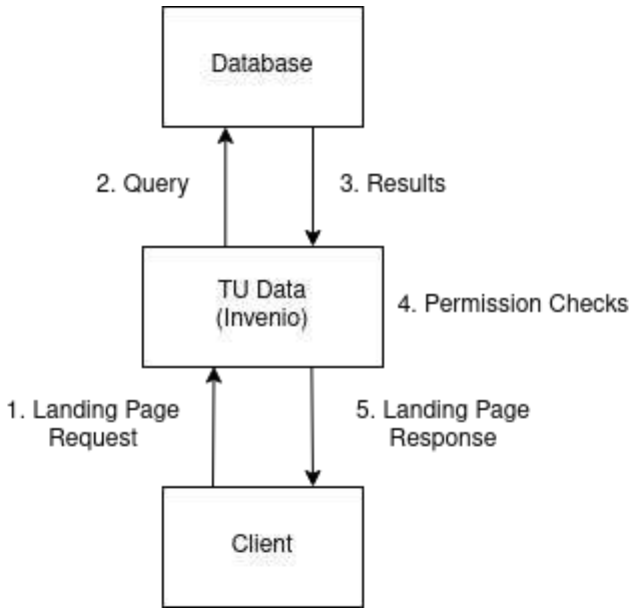
Record Landing Page







Record Landing Page - Workflow

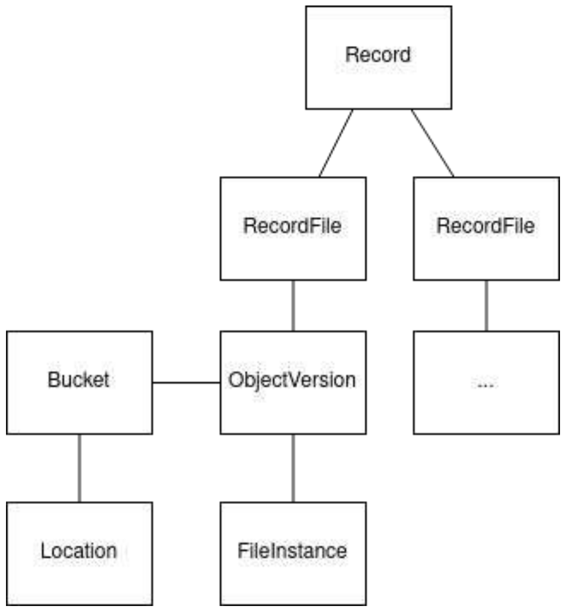


Records: Requirements

- Metadata

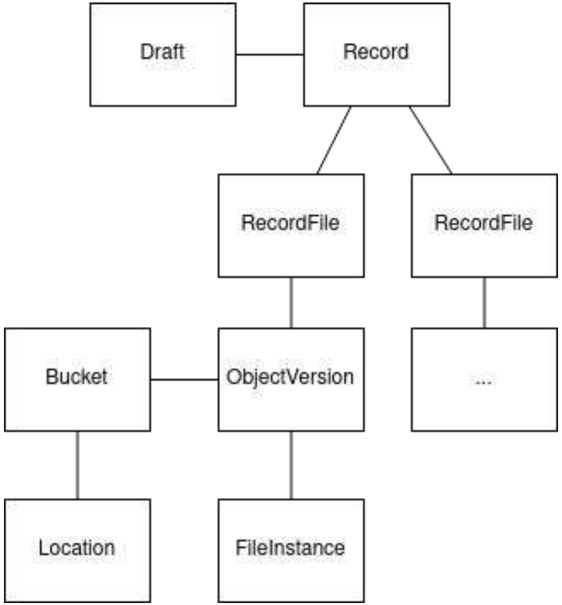
Records: Requirements

- Metadata
- Attached files



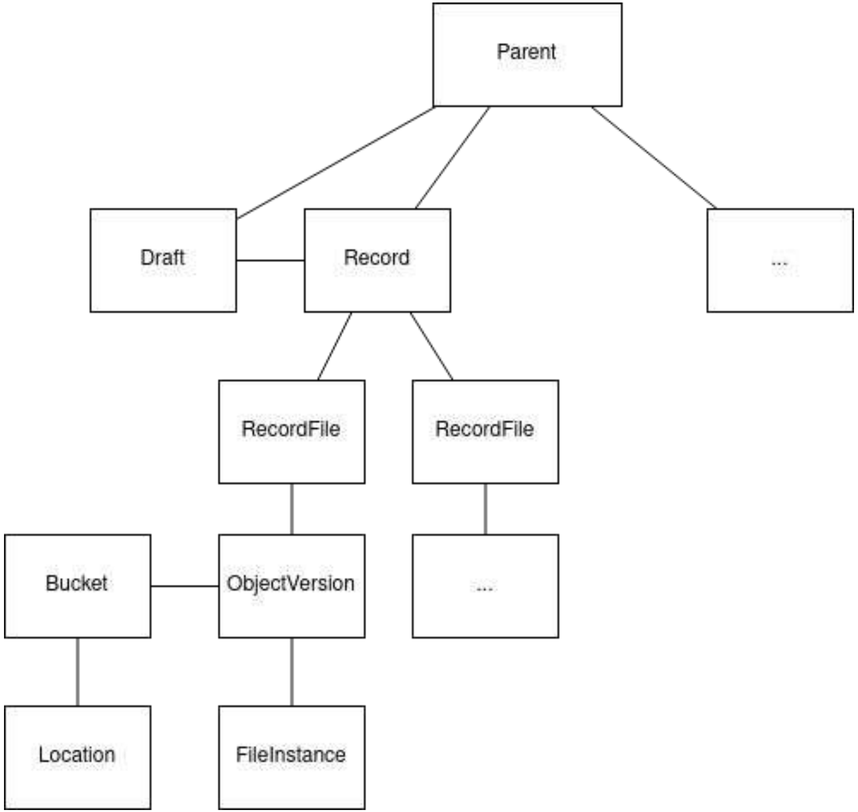
Records: Requirements

- Metadata
- Attached files
- Editable drafts



Records: Requirements

- Metadata
- Attached files
- Editable drafts
- Record versioning
  - PIDs (e.g. DOIs) are only meaningful if the content they point to is fixed
  - What if you found a mistake in your uploaded dataset?



Records: Metadata

- What does it look like (which standard)?
- How's it stored (format)?
- Validation
- Bonus for the Invenio framework:

  It's a library, not an application – keep it flexible!

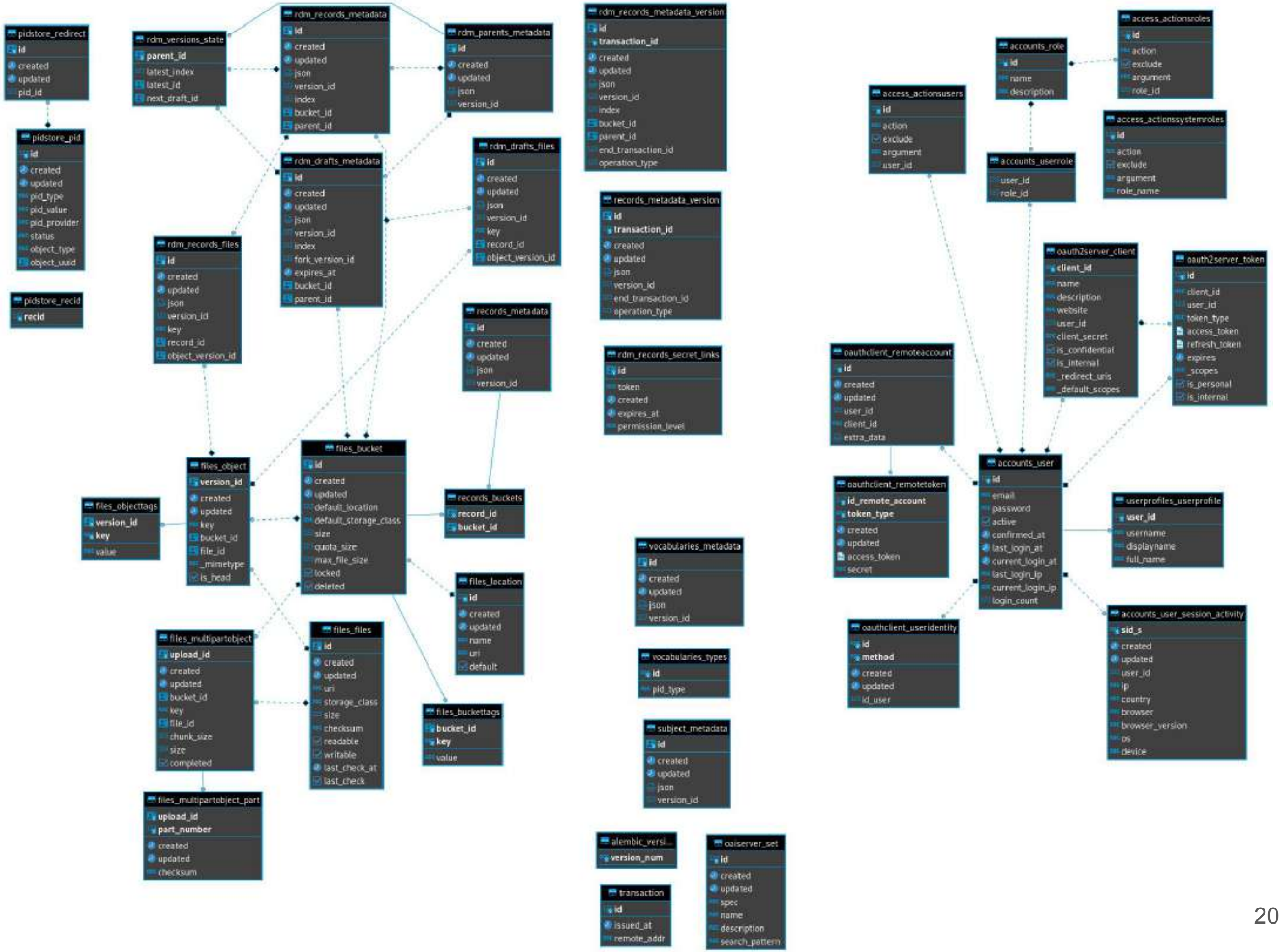
Image by [DataCite]

{json}

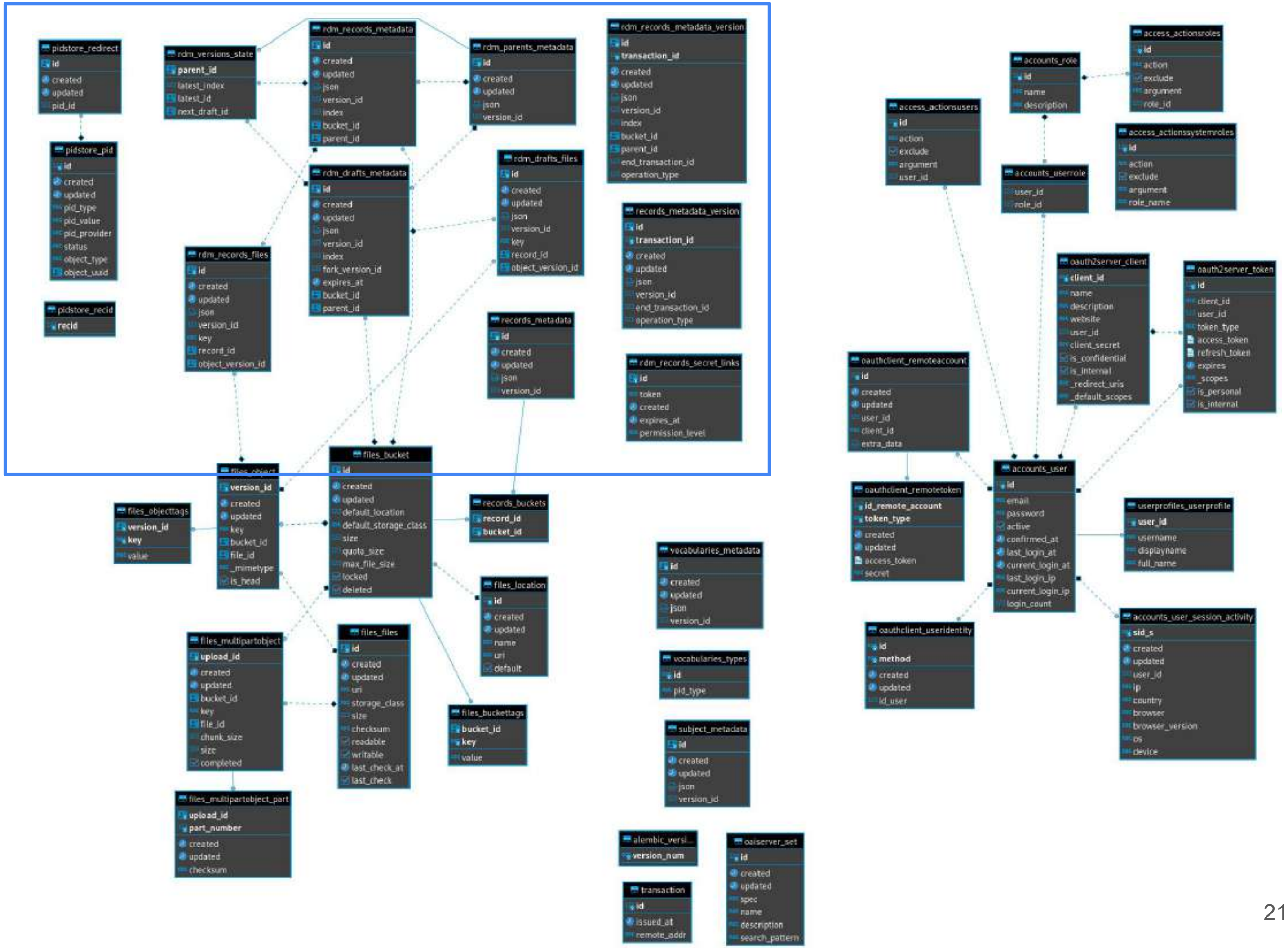
Image by [marshmallow-code]

Question: How would you store records?

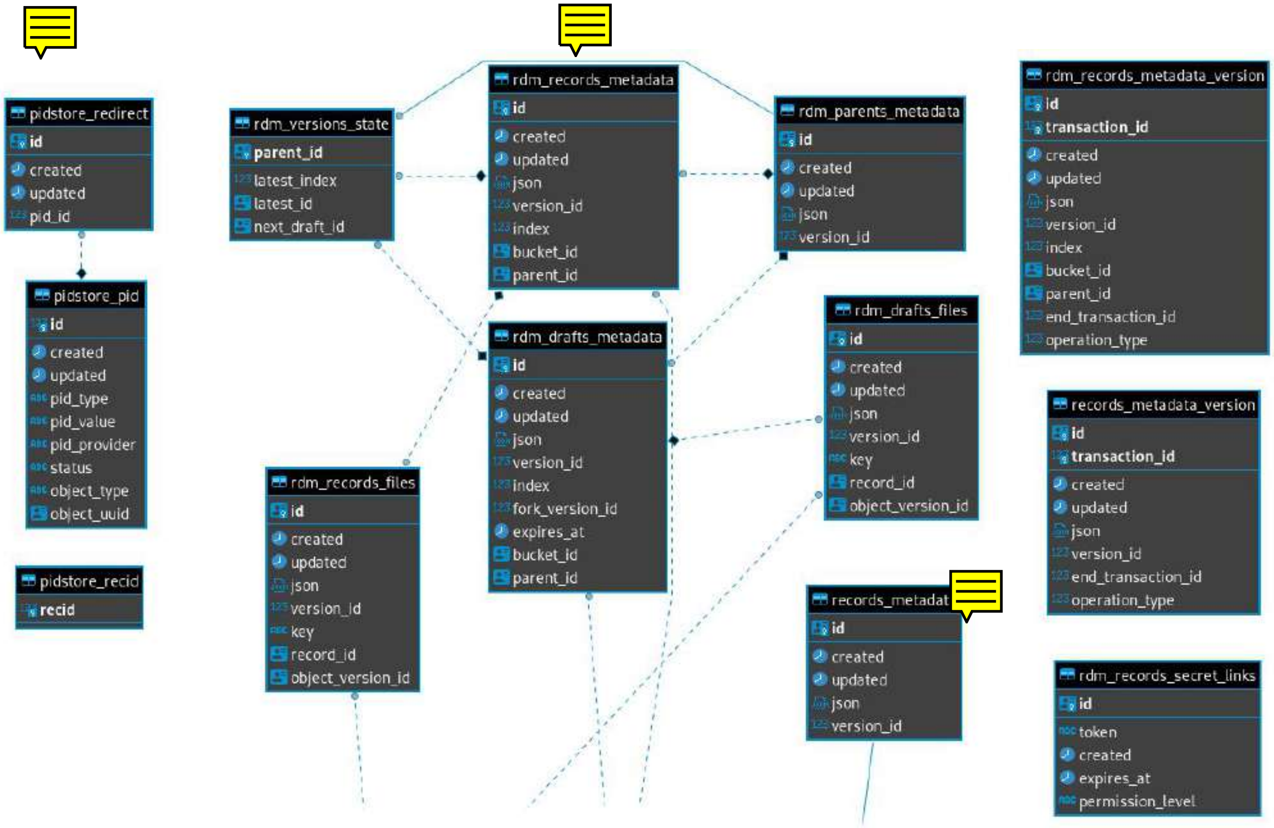
The Database



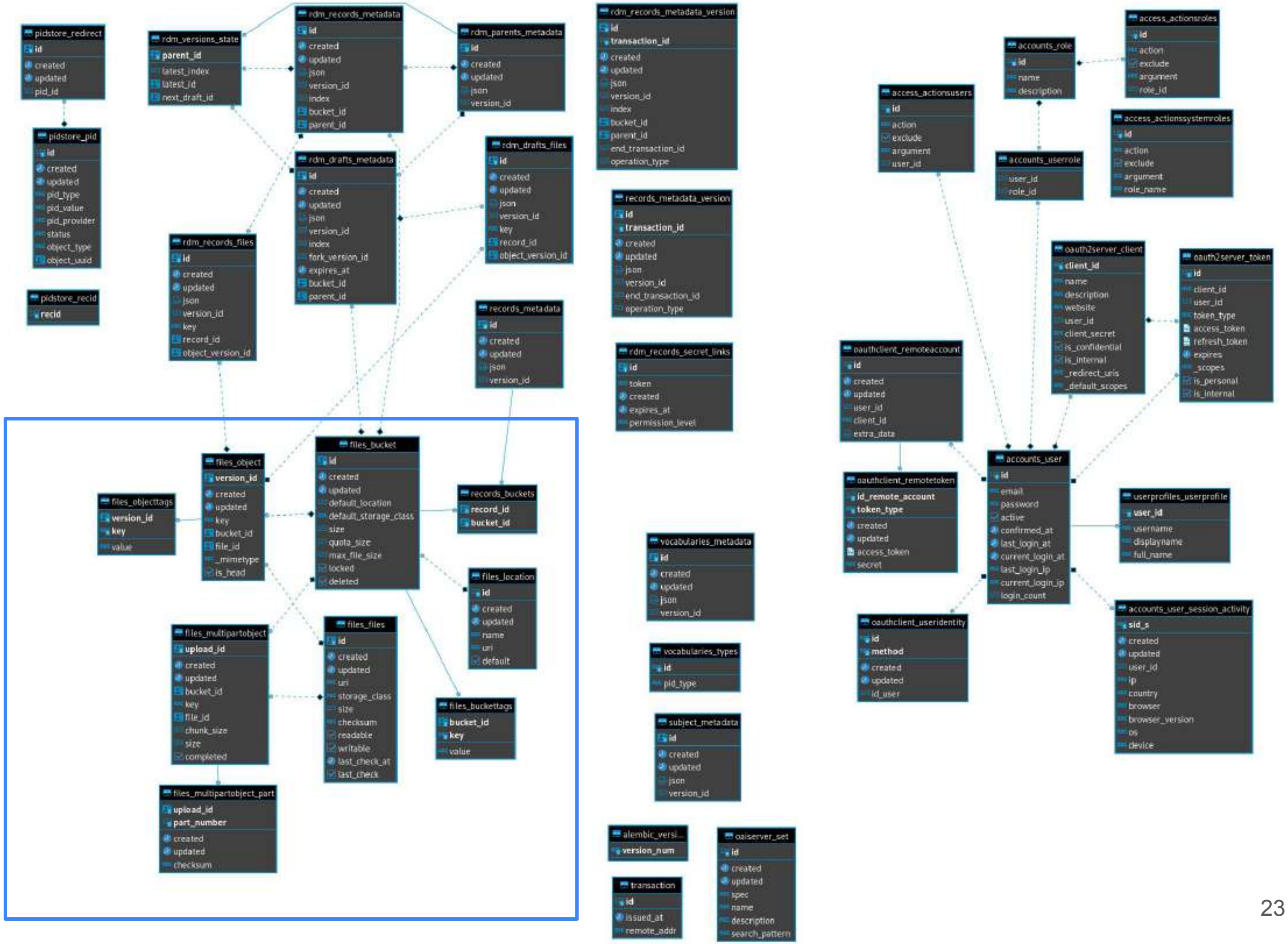
The Database (Records)



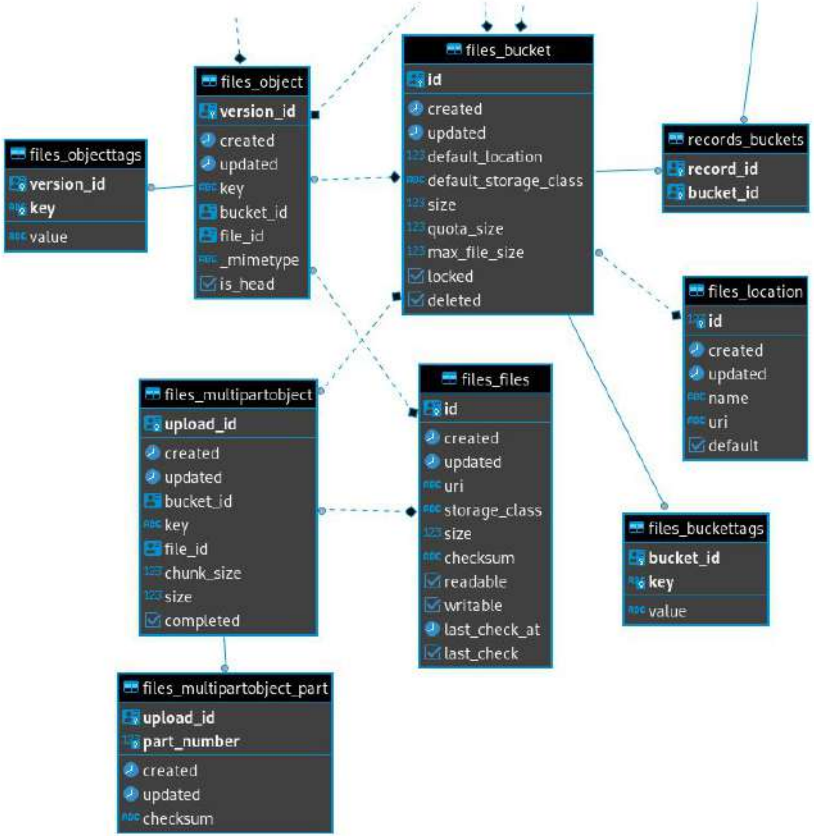
The Database (Records)



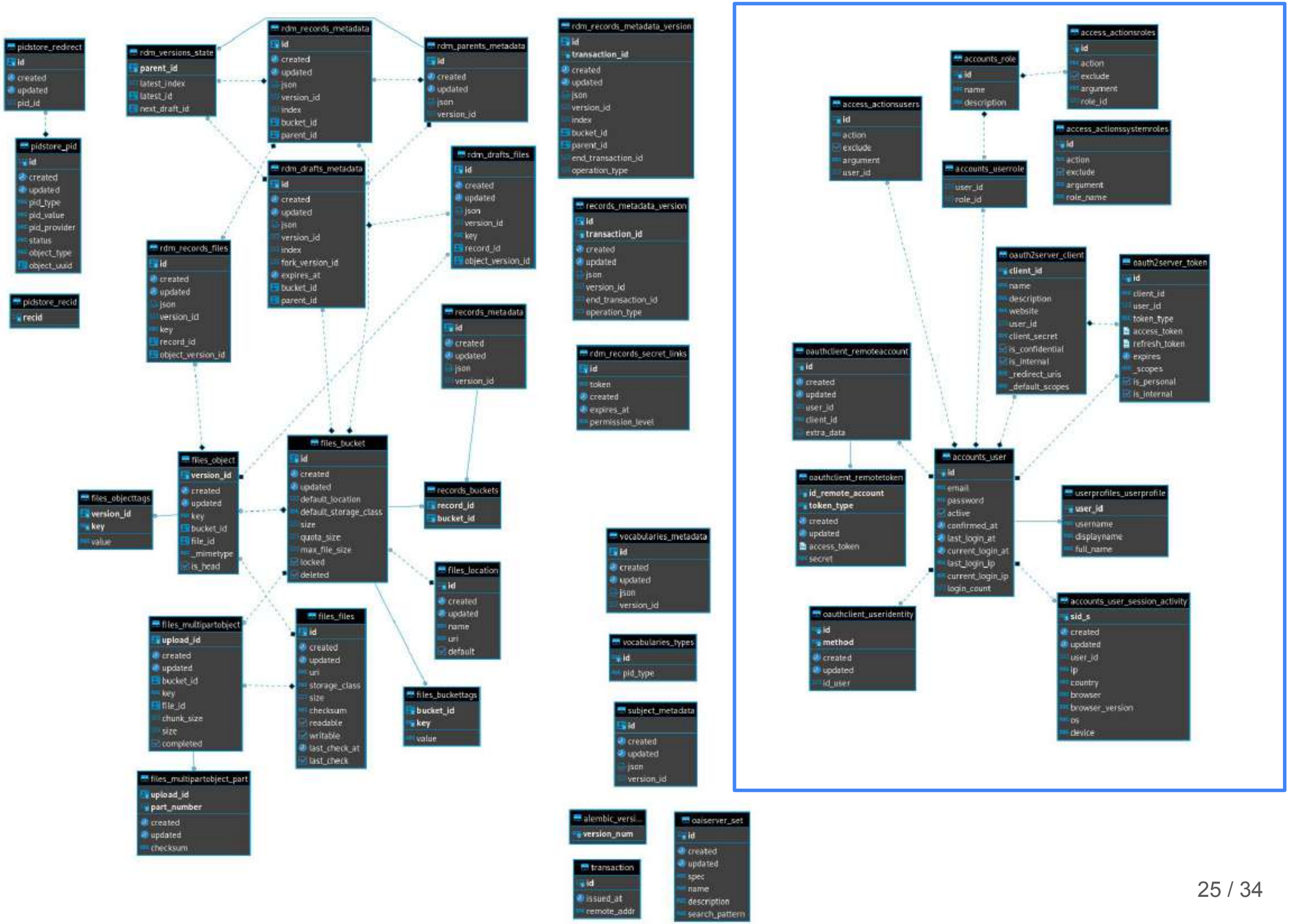
The Database (Files)



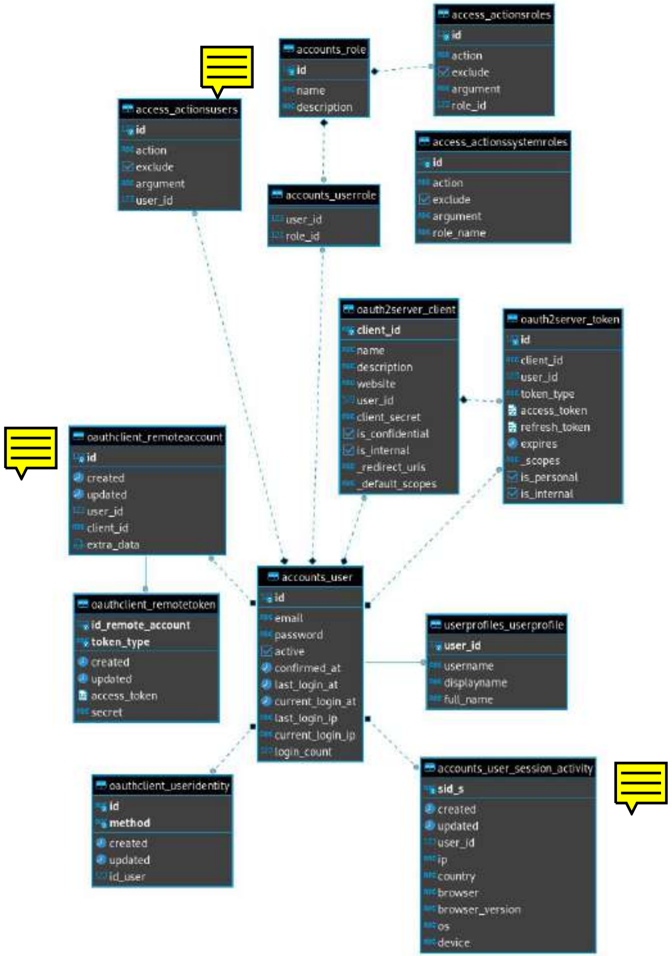
The Database (Files)



The Database (Users)

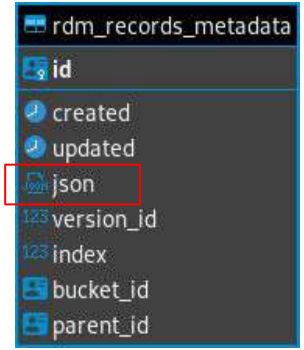


The Database (Users)

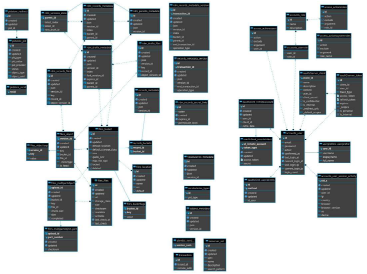


Record Metadata: Why a JSON field instead of ORM?

- A single table with a JSON field
  - Nested structures in JSON
    - No “cluttering” with tables
  - No accidental partial updates



vs.



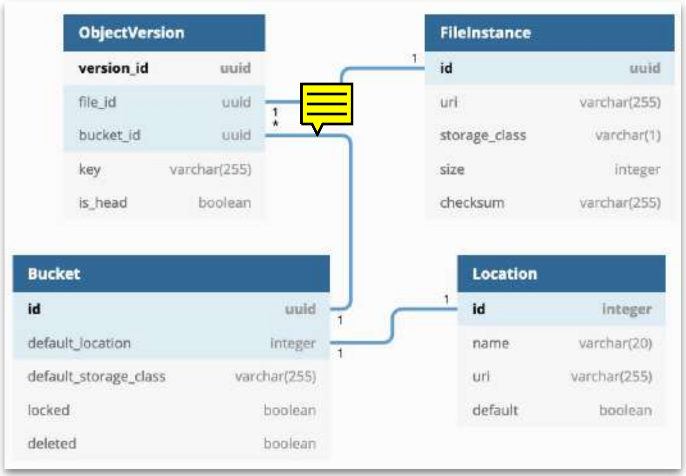
- Fits nicely with ES indexing

- Why not MongoDB then?
  - RDBMS still has benefits (ACID, battle-tested, PK lookups are fast, …)

- Ultimately, just a design decision

Records: File Deposits

- Where to store them (local, remote)?
- Make it scalable (like S3: buckets)
- Fixity checks (checksums)
- Efficiency (e.g. duplicates)
- More information about files in Invenio: https://invenio-files-rest.readthedocs.io/en/latest/overview.html

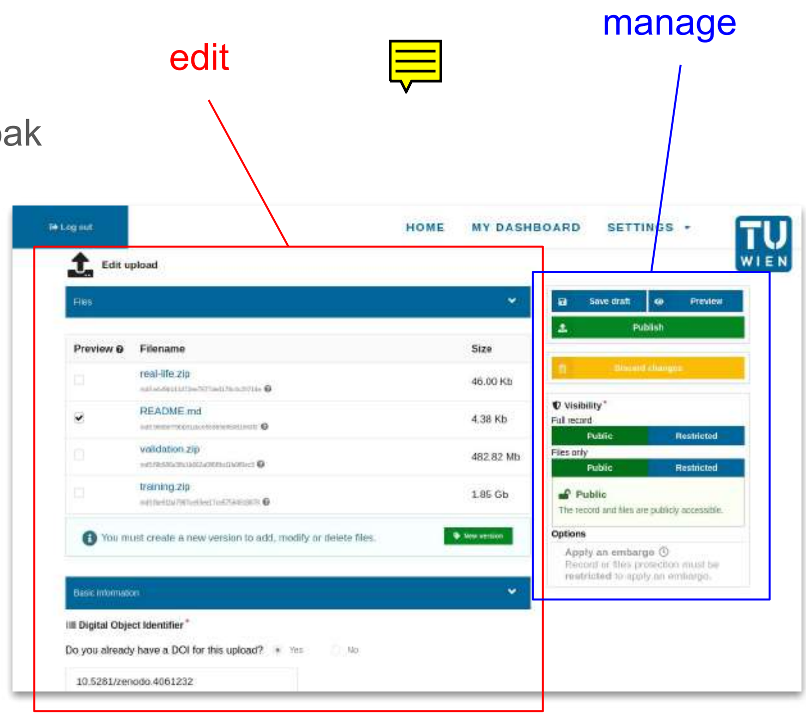


Records: Search

- Map the records to Elasticsearch
  - May have influence on the metadata design (e.g. nested fields)
  - Alternatively, add translation layer => increased complexity
- Try to avoid expensive database queries
  - Keep everything relevant in ES
- Filter out some results (e.g. closed-access records)
- It's not a database!
  - Everything can be recreated (re-indexed) from DB at any time

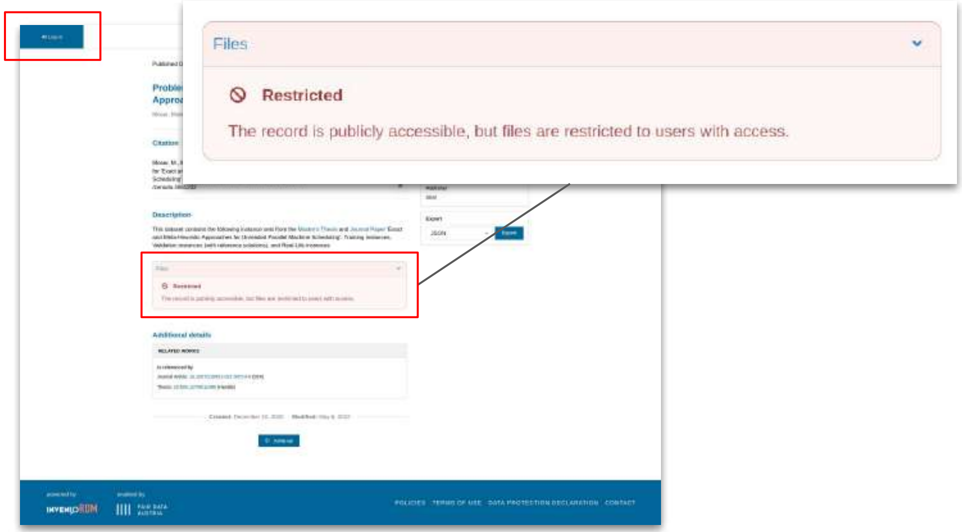
Access Restrictions

- Authentication
  - Local login vs OAuth2 (resp. OpenID Connect)
    - Get other services integrated, e.g. Keycloak => integrates TU SSO
  - Sessions for browsers, tokens for cURL
- Authorization
  - Different permissions for different actions
    - read, read_files, create, edit, manage, ...
  - Role-based access (e.g. role “trusted-user”)
  - Share-by-Link (like Google Docs)

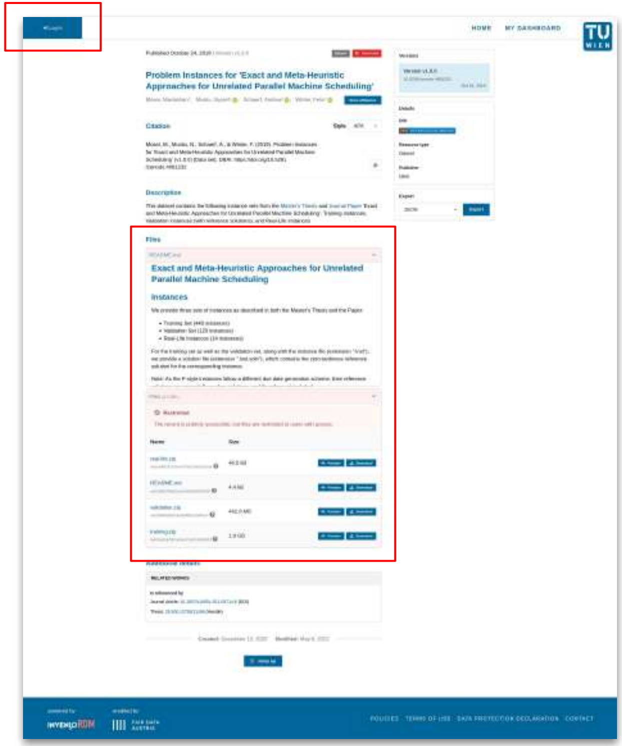


Share by Link

Can be used to share access to restricted datasets:



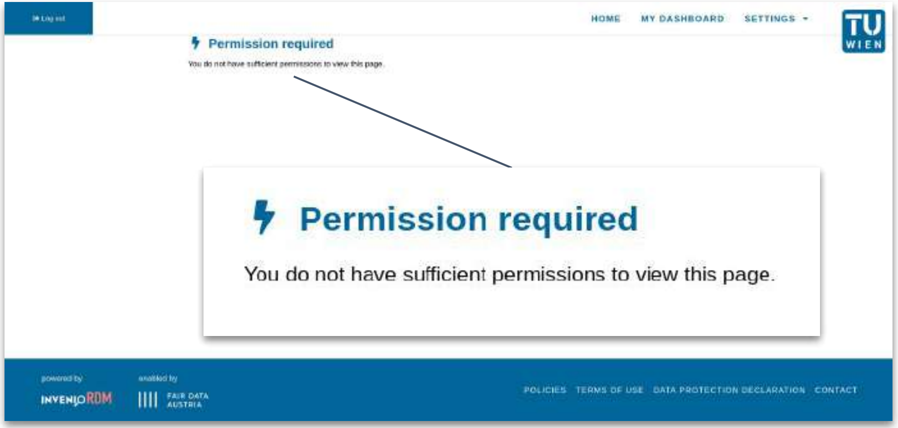
Normal Access



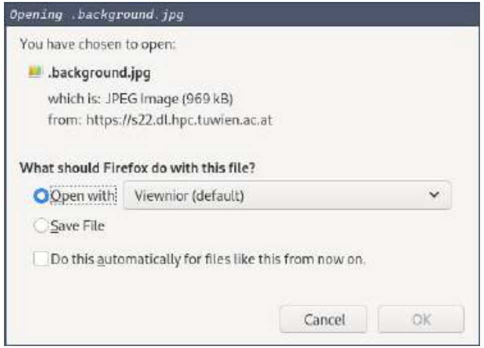
Access via Secret Link

Share by Link

Can also be used to enable double-blind peer reviews:



Metadata cannot be accessed...



… but files can be downloaded

Some Challenges for DevOps

- (Automatic) Deployment
- SSL certificates
- Secrets management
- Migrations (versions, environments)
- Failure detection, recovery, and resistance
- Identifying and fixing bottlenecks and other issues
- Backups
- Licensing
- Life cycles of software
  - Software hits EOL
  - Packages are abandoned
  - New versions introduce breaking changes
- Customizations
  - Custom logic (e.g. for permissions and integration with other services)
  - Custom styling to fit the corporate design
- Etc.

Thanks!

Play with our staging instance:
https://test.researchdata.tuwien.ac.at

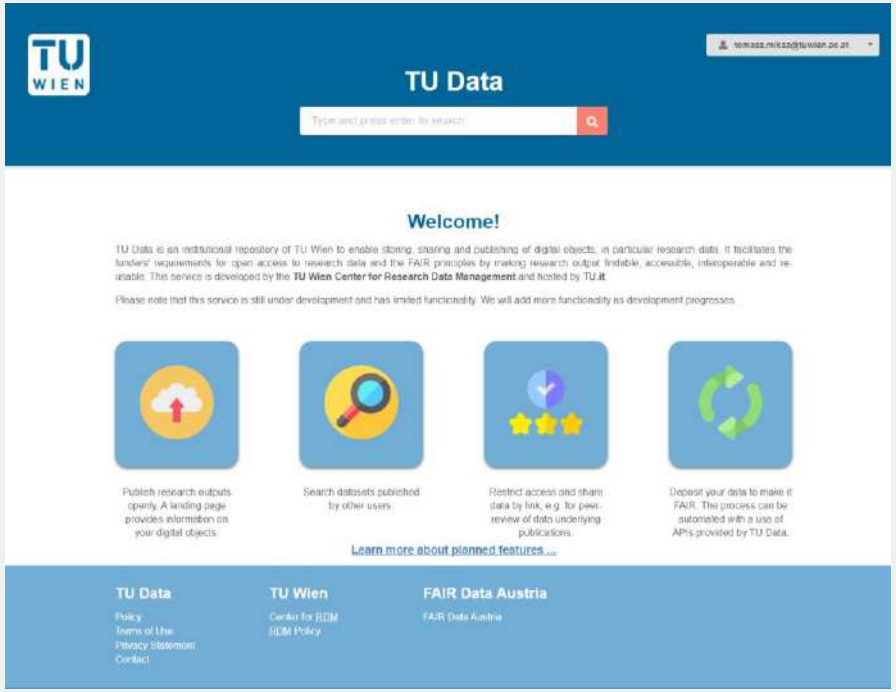
And send bug reports to:
maximilian.moser@tuwien.ac.at

More documentation:
Invenio RDM: https://inveniordm.docs.cern.ch/
Invenio Framework: https://invenio.readthedocs.io/en/latest/

Repository systems 2

Tomasz Miksa

Previous lecture



External visibility

Many institutions only tick off a box ‘we have a repository’

‘Having a repository’ is not enough

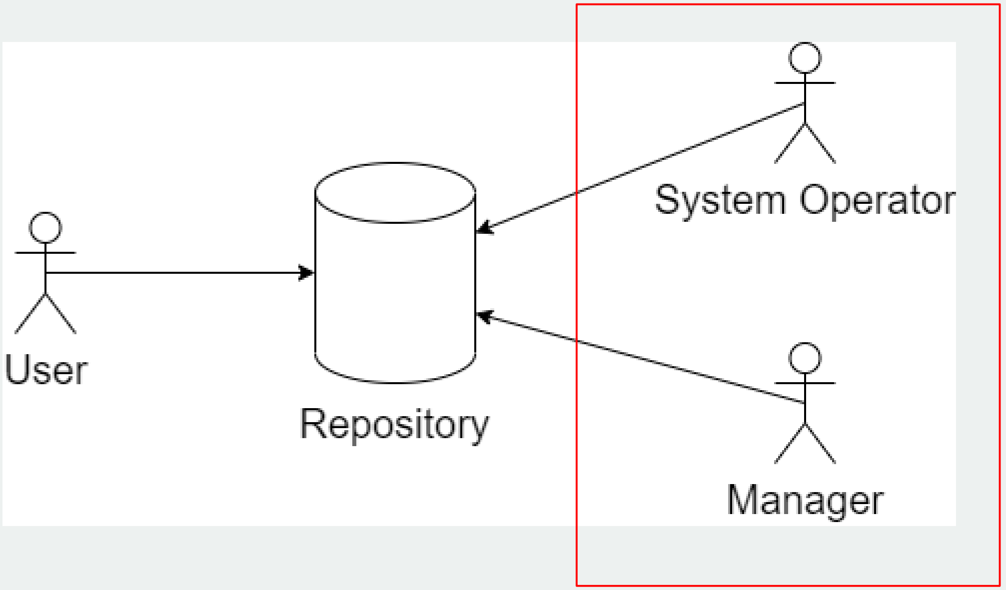
- Contents must be discoverable and FAIR
  - Integration with hubs
  - Metadata following standards and machine-actionability

Cargo cult

Agenda

How to make repository contents visible?

- DOI registration bodies
- Interoperability protocols
- Scholix
- DCAT
- Schema.org
- Repository registries



FAIR - Findable

Main focus today:

- **F4. (Meta)data are registered or indexed in a searchable resource**
- Repository itself is a ‘searchable resource’, but this is not enough

Repositories support also other principles

- e.g. handle access, assign identifiers, etc.
- Not discussed today (and you should know this by now anyway)

DOI registration bodies

DataCite

- DOI registration body
  - Handle based identifier
  - Metadata
  - Provides a range of identifiers to repositories
- Supported by public sector, e.g. DCC, CERN, ANDS
- National desk established at the TU Wien

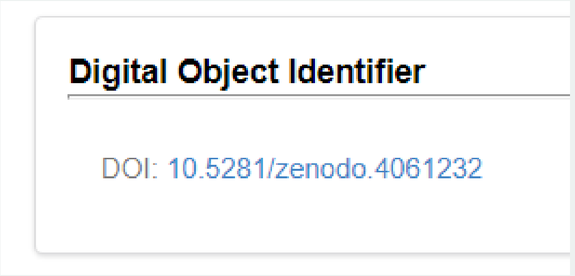
Crossref

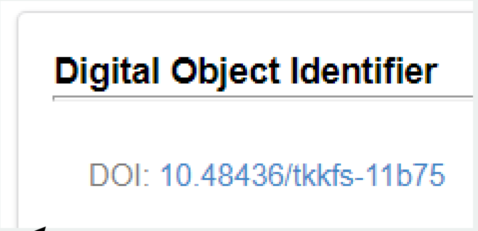
- DOI registration body
- Supported by private sector, e.g. Elsevier

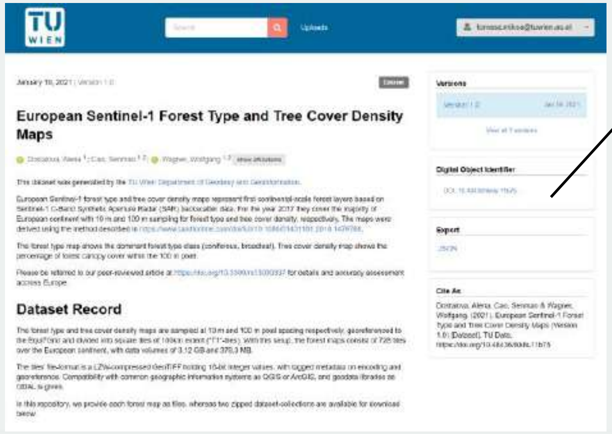
DOI – repository side

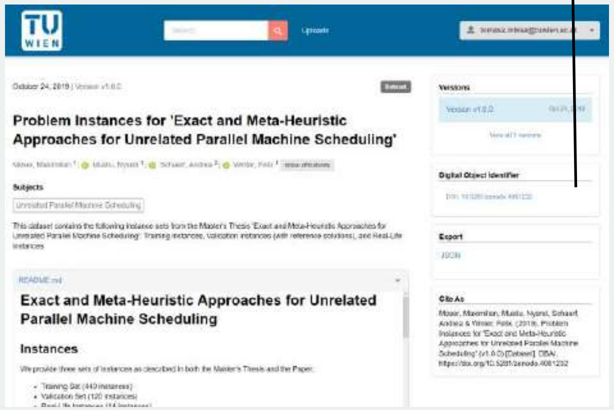
Repository must get an account first to be able to mint DOIs

- **Prefix**: 10.48436
- Suffix: tkkfs-11b75

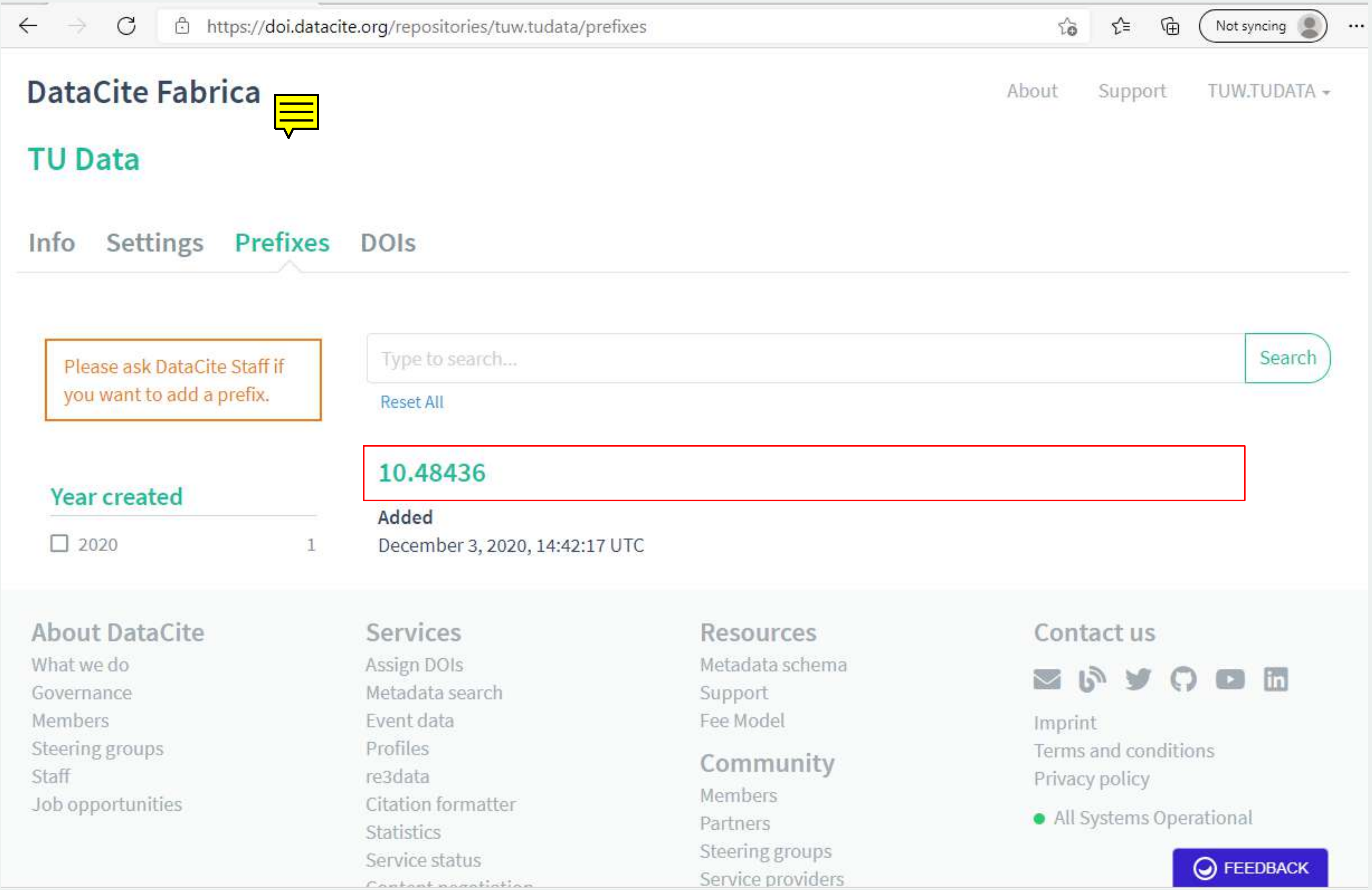








DOI – repository prefix



DOI – metadata

Repository must provide minimal metadata

- https://schema.datacite.org/meta/kernel-4.3/

Table 1: DataCite Mandatory Properties

ID	Property	Obligation
1	Identifier (with mandatory type sub-property)	M
2	Creator (with optional given name, family name, name identifier and affiliation sub-properties)	M
3	Title (with optional type sub-properties)	M
4	Publisher	M
5	PublicationYear	M
10	ResourceType (with mandatory general type description sub-property)	M

Table 2: DataCite Recommended and Optional Properties

ID	Property	Obligation
6	Subject (with scheme sub-property)	R
7	Contributor (with optional given name, family name, name identifier and affiliation sub-properties)	R
8	Date (with type sub-property)	R
9	Language	O
11	AlternateIdentifier (with type sub-property)	O
12	RelatedIdentifier (with type and relation type sub-properties)	R
13	Size	O
14	Format	O
15	Version	O
16	Rights	O
17	Description (with type sub-property)	R
18	GeoLocation (with point, box and polygon sub-properties)	R
19	FundingReference (with name, identifier, and award related sub-properties)	O

DataCite Fabrica

About Support TUW.TUDATA

TU Data

Info Settings Prefixes DOIs

Create DOI (Form)

More information about DOI registration via form can be found on our Support Website. Required properties are marked with a red asterix.

Required Properties

*** DOI** A globally unique string that identifies the resource and can't be changed.

10.48436 z0fz-p653

Click the circle icon for a new random suffix, or the cross icon to delete the random suffix and enter a value manually.

*** State** The state determines whether a DOI is registered and findable. Once in Registered or Findable state, a DOI can't be set back to Draft state. More ...

- (•) Draft only visible in Fabrica, DOI can be deleted
- ( ) Registered registered with the DOI Resolver
- ( ) Findable registered with the DOI Resolver and indexed in DataCite Search

*** URL** The location of the landing page with more information about the resource.

URL

Should be a https URL – within the allowed domain(s) of your repository if domain restrictions are enabled in the repository settings. Http and ftp are also supported.

*** Creators** The main researchers or organizations involved in producing the resource, in priority order.

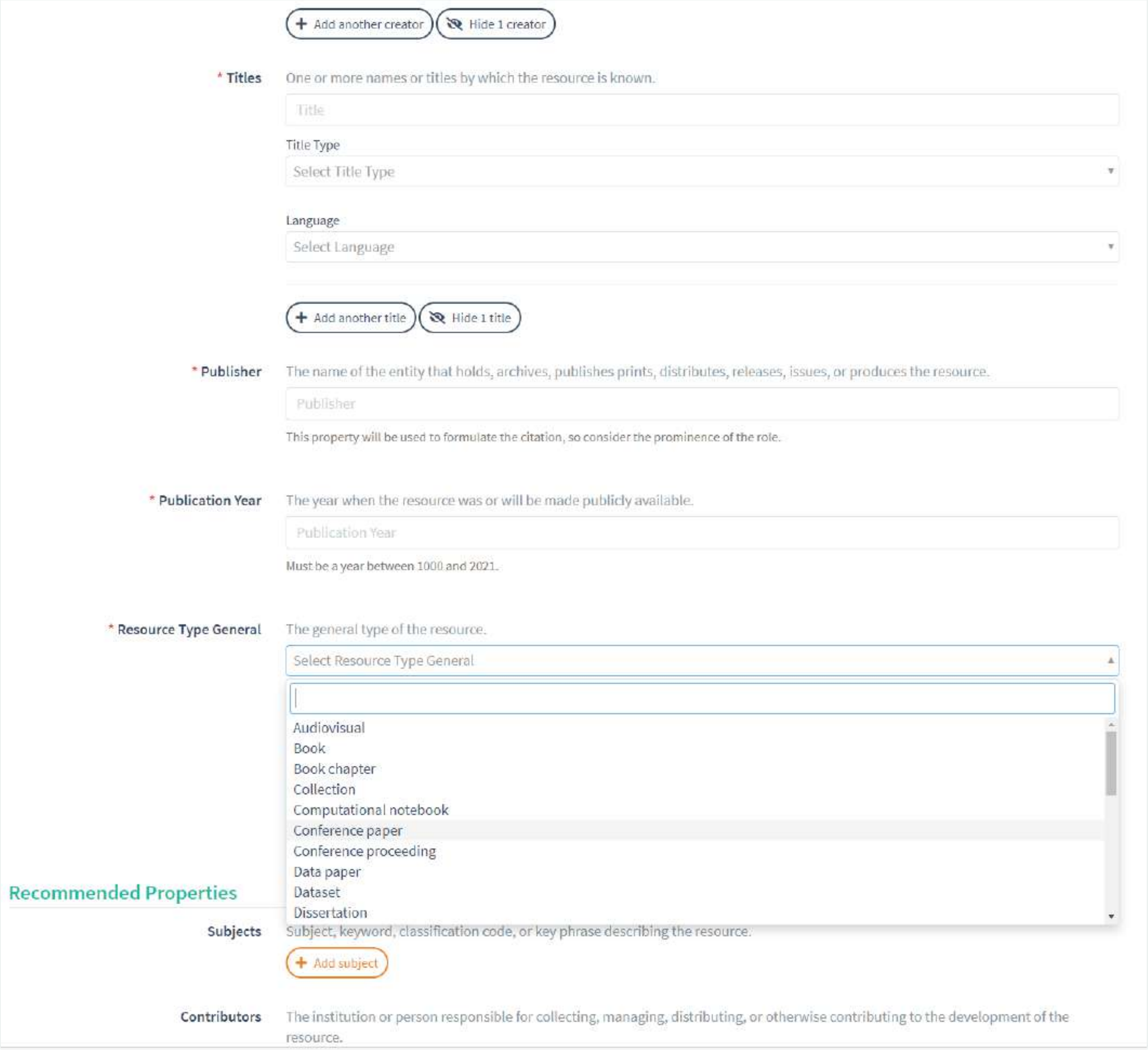
Name Identifier

Name Identifier

Uniquely identifies an individual or legal entity, according to various schemas, e.g. ORCID, ROR or ISNI. Use name identifier expressed as URL. The Given Name, Family Name and Name will automatically be filled out for ORCID and ROR identifiers.

Add another name identifier

( ) Person ( ) Organization (•) Unknown


+ Add another creator
Hide 1 creator
* Titles
One or more names or titles by which the resource is known.
Title
Title Type
Select Title Type
Language
Select Language
+ Add another title
Hide 1 title
* Publisher
The name of the entity that holds, archives, publishes prints, distributes, releases, issues, or produces the resource.
Publisher
This property will be used to formulate the citation, so consider the prominence of the role.
* Publication Year
The year when the resource was or will be made publicly available.
Publication Year
Must be a year between 1000 and 2021.
* Resource Type General
The general type of the resource.
Select Resource Type General
Audiovisual
Book
Book chapter
Collection
Computational notebook
Conference paper
Conference proceeding
Data paper
Dataset
Dissertation
Recommended Properties
Subjects
Subject, keyword, classification code, or key phrase describing the resource.
+ Add subject
Contributors
The institution or person responsible for collecting, managing, distributing, or otherwise contributing to the development of the resource.

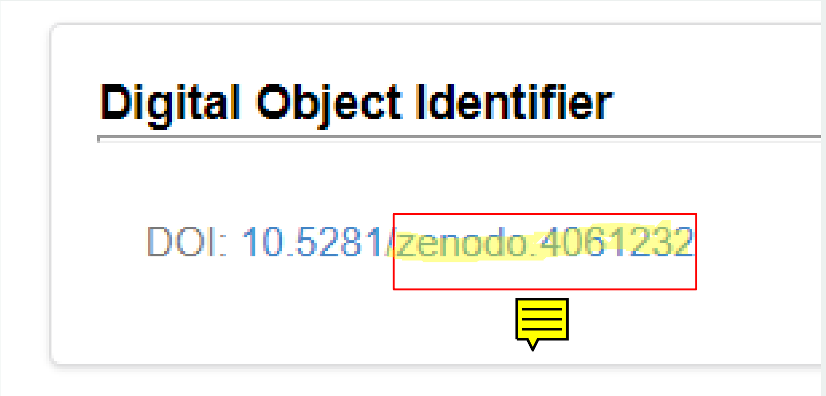
DOI – suffix

https://help.zenodo.org

- Why do you include “zenodo” in the DOI?
- Why don’t the DOIs have a version number suffix like “.v1”?

Cool DOIs

- https://blog.datacite.org/cool-dois/

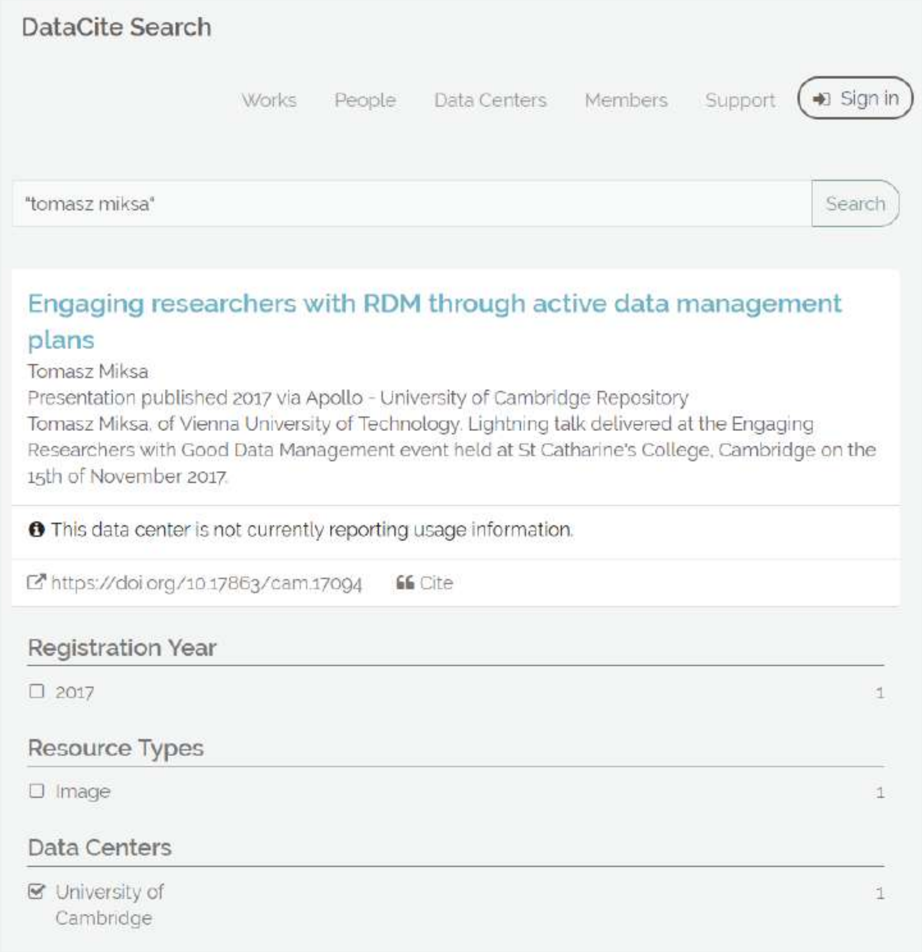


DOI registration bodies - example

Example shows

- Document located at Cambridge Repository
- Document has a DOI
- DOI was minted by Data Cite
- Data Cite has metadata about each DOI

Data Cite provides access to metadata registry



https://search.datacite.org

OAI-PMH - Open Archives Initiative Protocol for Metadata Harvesting

- query to discover repository contents
- only for metadata
- not for depositing

SWORD - Simple Web-service Offering Repository Deposit

- deposit to multiple repositories at once
- deposit by third party systems (e.g. lab equipment)

OAI-PMH

Data remains within a repository

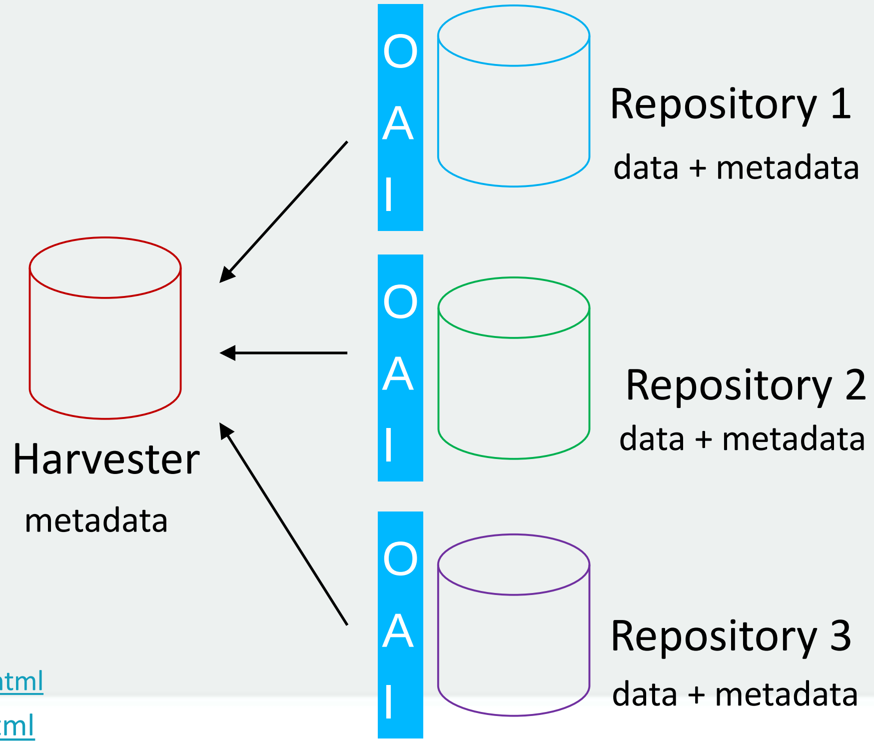
Harvester aggregates metadata

Useful to aggregate data

- e.g. for domain, country, institution

Dublin Core by default

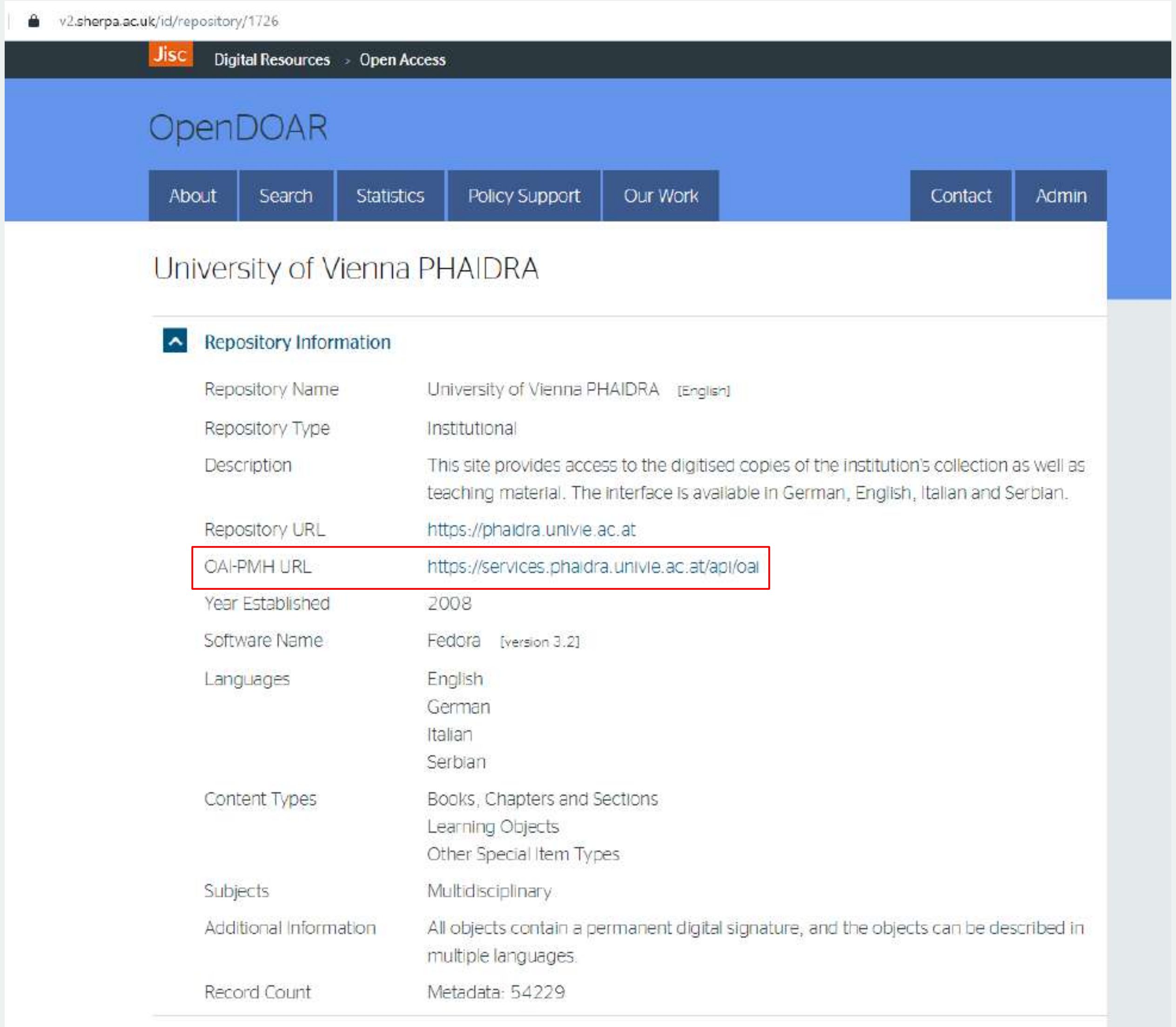
- Other Metadata standards can be added
  - OpenAIRE requires DataCite



https://guidelines.openaire.eu/en/latest/data/use_of_oai_pmh.html

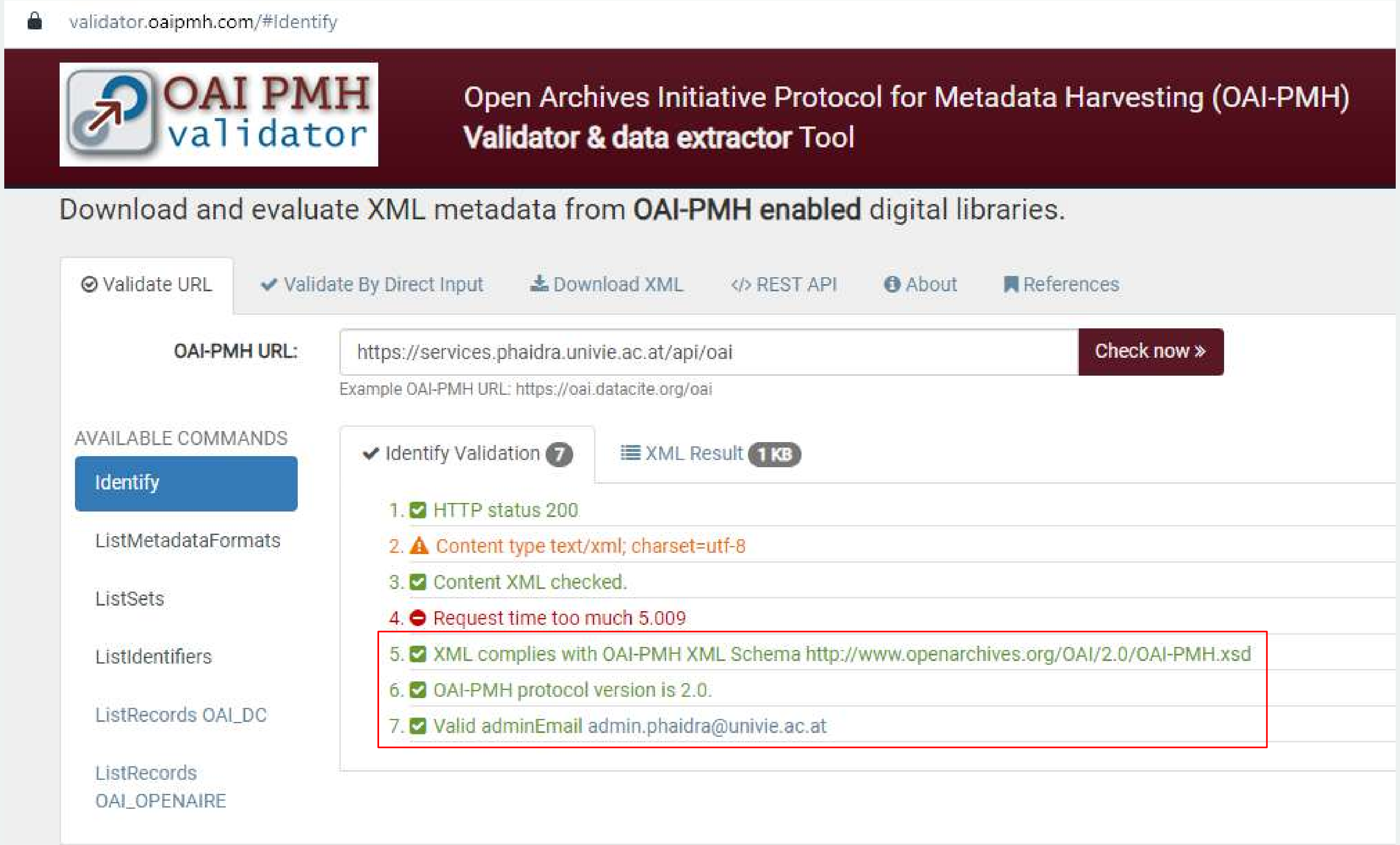
http://www.openarchives.org/OAI/openarchivesprotocol.html

OAI-PMH in OpenDOAR (registry)



https://v2.sherpa.ac.uk/id/repository/1726?template=opendoar

OAI-PMH



https://validator.oaipmh.com/

OAI-PMH

OAI-PMH URL: https://services.phaidra.univie.ac.at/api/oai
Check now »

Example OAI-PMH URL: https://oai.datacite.org/oai

AVAILABLE COMMANDS

- Identify
- ListMetadataFormats
- ListSets
- ListIdentifiers
- ListRecords OAI_DC
- ListRecords OAI_OPENAIRE

ListSets Validation 48 | XML Result 40 KB | ListSets help

1. HTTP status 200
2. Content type text/xml; charset=utf-8
3. Content XML checked.
4. Request time is 0.43 sec
5. XML complies with OAI-PMH XML Schema http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd
6. Found set "Nationale und Studienkataloge der Juridischen Fakultät" with setSpec "o688263".
7. Found set "Nachlass von Erwin Schrödinger (1887-1961)" with setSpec "europeana_a1051".
8. Found set "Nachlass Maximilian Hell: Aufzeichnungen des berühmten Wiener Astronomen" with setSpec "o314596".
9. Found set "iPRES 2019 - 16th International Conference on Preservation of Digital Objects: iPRES 2019" with setSpec "ipres2019".
10. Found set "iPRES 2017 - Proceedings of the 14th Conference on Preservation of Digital Objects" with setSpec "ipres2017".
11. Found set "iPRES 2012 - Proceedings of the 9th International Conference on Preservation of Digital Objects: iPRES 2012 - Digital Curation Institute, iSchool, Toronto" with setSpec "ipres2012".
12. Found set "Nachlass von Hans Thirring (1888-1976)" with setSpec "europeana_a1052".
13. Found set "DiFaB, DIgitales ForschungsArchiv Byzanz" with setSpec "difab".
14. Found set "Mündliche Geschichtsüberlieferung: Österreichische Exilpublizisten im Widerstand gegen den Nationalsozialismus – Ein Oral Video History-Projekt 1988-1996" with setSpec "o868526".
15. Found set "Archiv der Universität Wien, digitale Objekte" with setSpec "archivunivie".
16. Found set "iPRES 2005 - 2nd International Conference on Preservation of Digital Objects: iPRES 2005 - Göttingen" with setSpec "ipres2005".

OpenAIRE

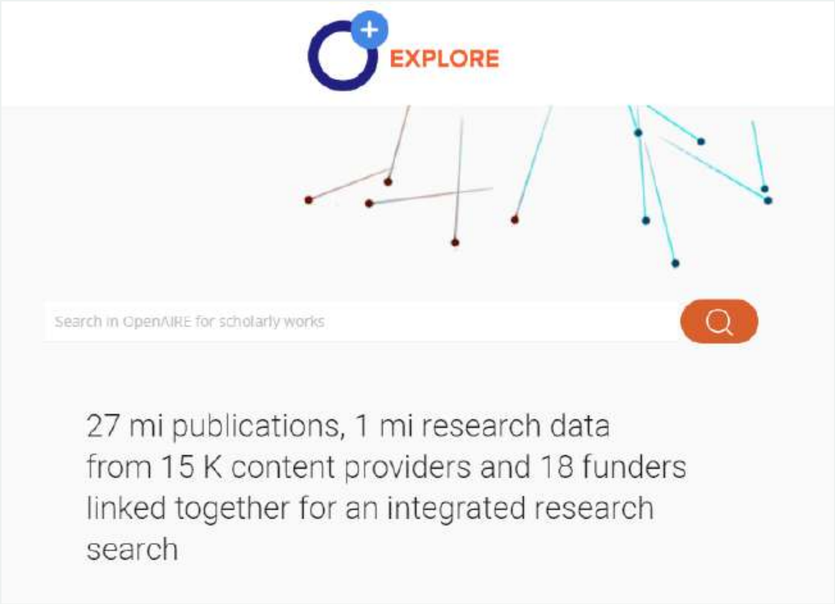
Open Access Infrastructure for Research in Europe

- Launched 2009 by European Commission
- Promotes Open Access

Network of experts

Technical infrastructure

- Harvest research output
  - Data, publications
  - Link it
  - Monitor
- Zenodo



https://explore.openaire.eu

OpenAIRE

Check what kind of sources are indexed and how metadata is collected

- https://www.openaire.eu/aggregation-and-content-provision-workflows

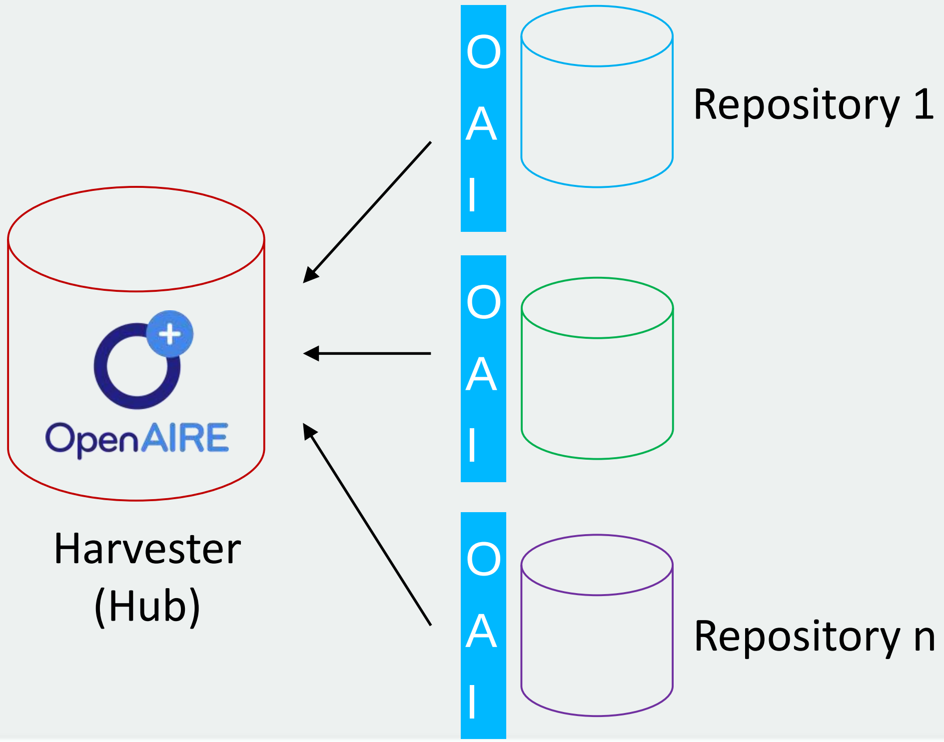
What else openAIRE does

- https://www.openaire.eu/faqs

OpenAIRE - compliance

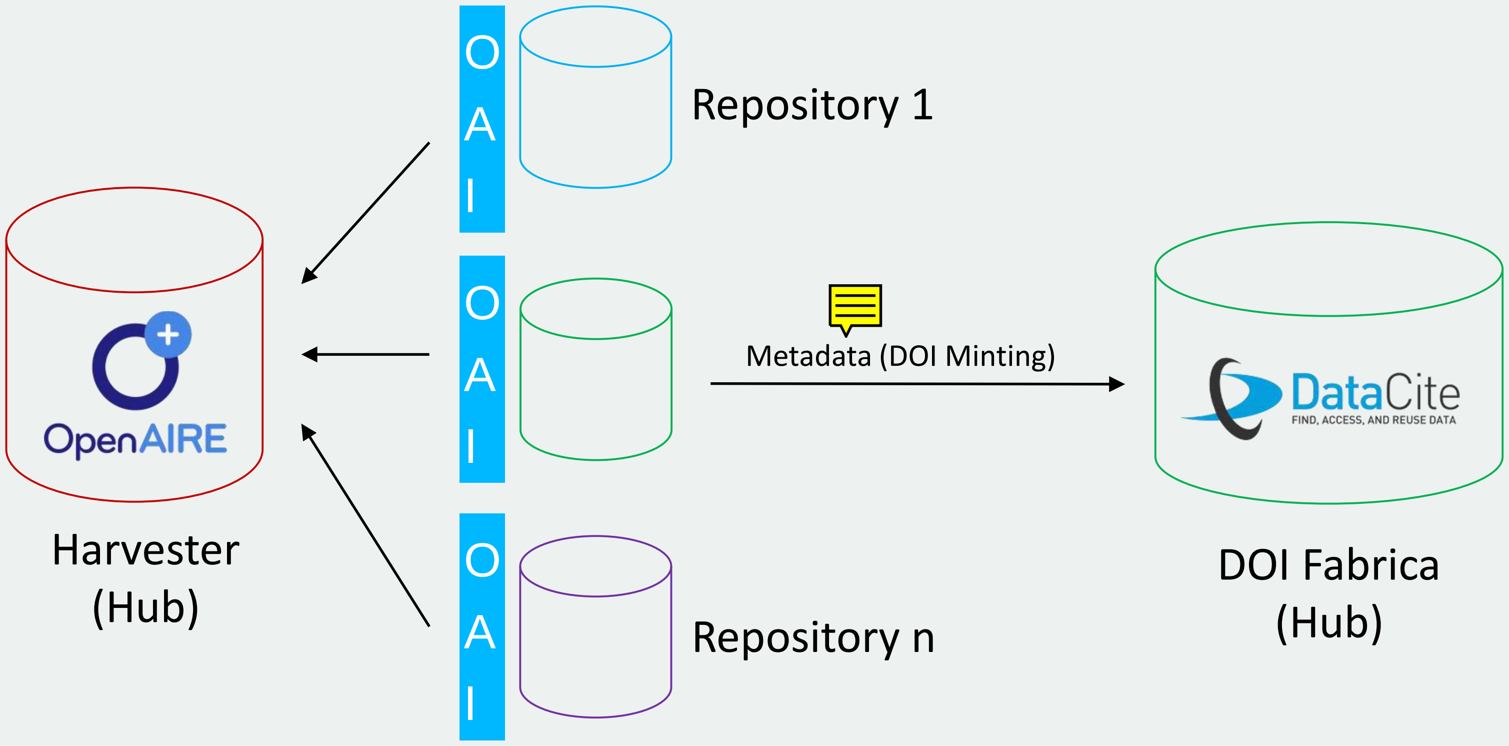
Repository must have an OAI-PMH endpoint

Metadata in DataCite format



https://guidelines.openaire.eu/en/latest/

OpenAIRE AND/OR DataCite?



OpenAIRE AND/OR DataCite?

Not every repository

- Must participate in OpenAIRE
  - lack of OAI-PMH
  - not located in Europe
- Must mint DOIs
  - Handles, ARKs, etc.

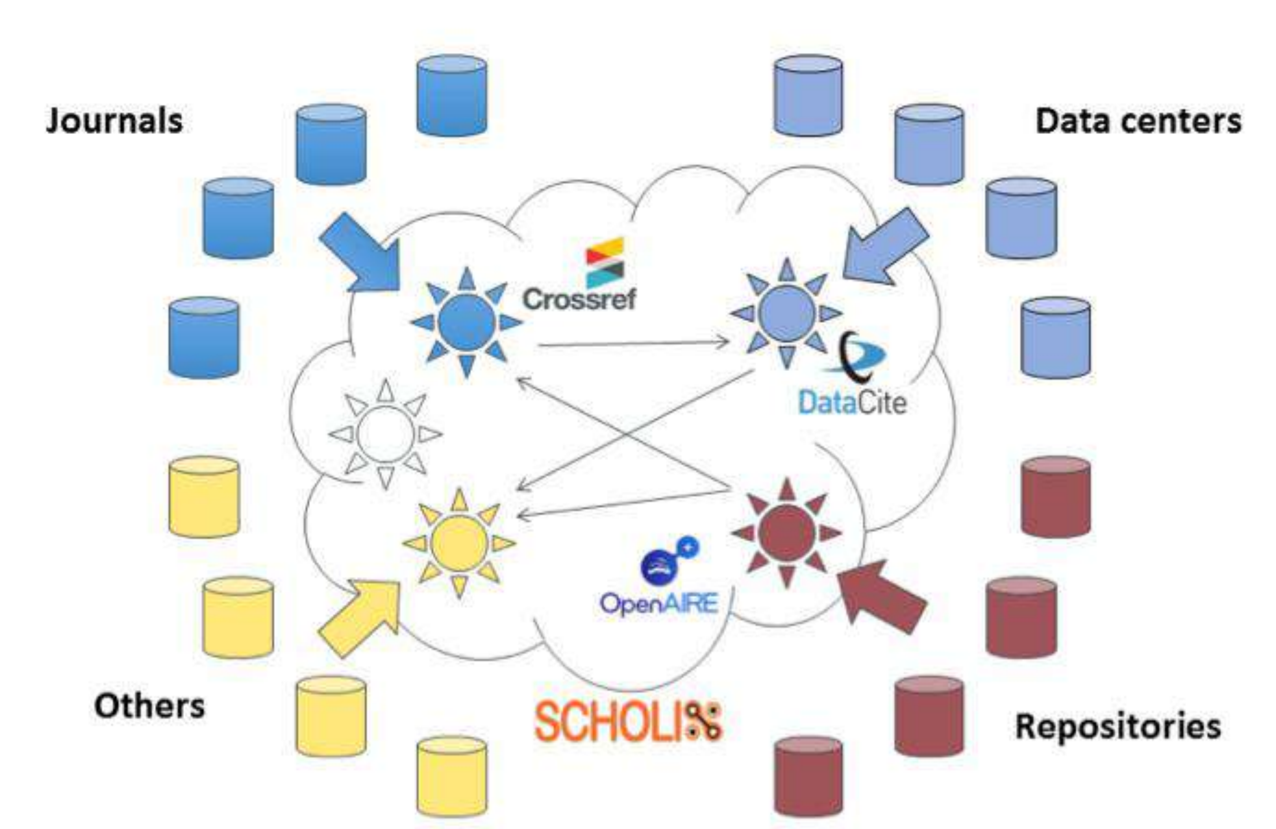
Everyone must decide

Eventually it theoretically makes little difference…

Scholix

Goal

- exchange links between publications and data



http://www.dlib.org/dlib/january17/burton/01burton.html

Scholix

At the core of the conceptual model is the *link* between two *objects*.

Main focus: literature and data.

- theoretically also: software, algorithms, models, protocols, tweets, comments, and so on.
- practically: not for the time being the focus

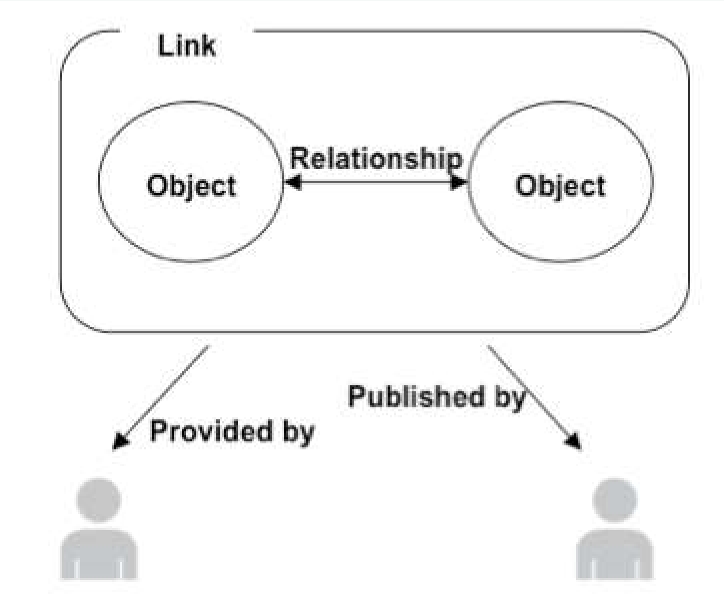
To become a contributor

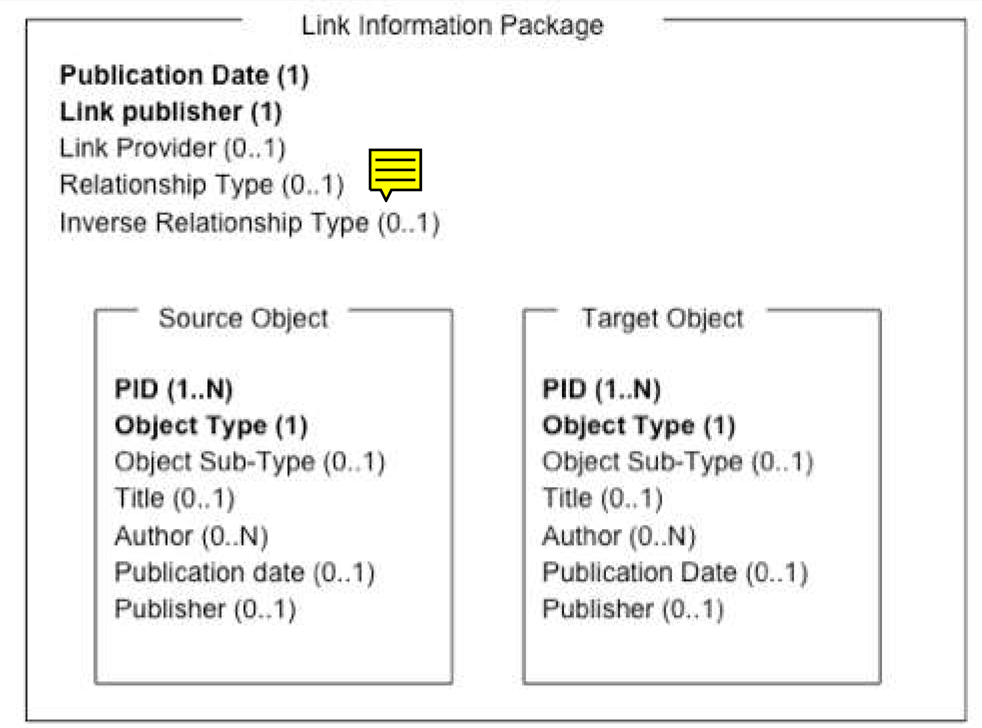
- Feed your data-literature link information to an existing Scholix hub using your existing community standards
  - e.g. OpenAIRE or DataCite registries

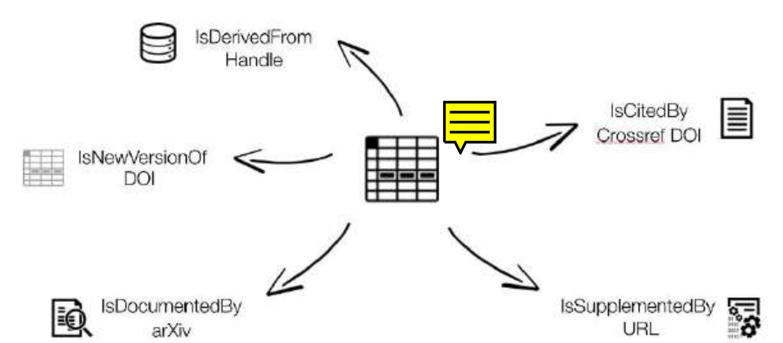
To retrieve links

- http://api.scholexplorer.openaire.eu/v2/ui/

Scholix







http://api.scholexplorer.openaire.eu/v2/ui/

Scholix was developed at RDA


WORLD DATA SYSTEM
ICSU
RDA

An open, universal literature-data cross-linking service

The Challenge:

Sharing information about the links between the literature and research data.

What is the solution?

Building on pre-existing components and international initiatives, the RDA/WDS Publishing Data Services Working Group is one of the drivers behind the "Data Literature Interlinking Service" (DLI), developed in a synergy with OpenAIRE & PANGAEA. DLI is aimed at improving visibility, discoverability, re-use and reproducibility by bringing 2M+ existing article/data links together, normalize them using a common schema, and exposing the full set as an open service.

Produced by: RDA/WDS Publishing Data Services WG
https://rd-alliance.org/groups/rdawds-publishing-data-services-wg.html

What is the impact?

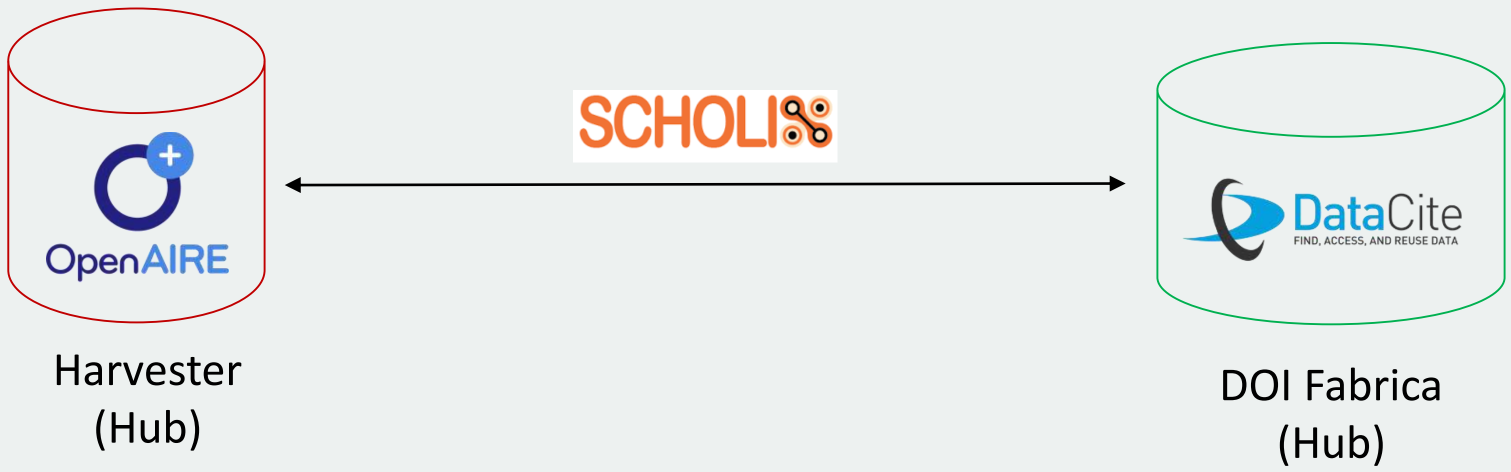
Accessing and using literature-data links at large scale in an efficient and reliable way allows different stakeholders in the data publishing landscape to improve their services, increase data discoverability and usability.

Data centres will be able to assess much better how often their data is used in the literature, and present their users with links to relevant publications.

Find out more about the RDA/WDS Publishing Data Services WG Recommendation

https://www.rd-alliance.org/groups/rdawds-scholarly-link-exchange-scholix-wg

Scholix


OpenAIRE
SCHOLIX
DataCite
FIND, ACCESS, AND REUSE DATA
Harvester
(Hub)
DOI Fabrica
(Hub)

W3C Data Cataolog Vocabulary (DCAT)

- RDF vocabulary for interoperability between data catalogues
- decentralized publishing
- facilitates federated dataset search

Relaxed constraints

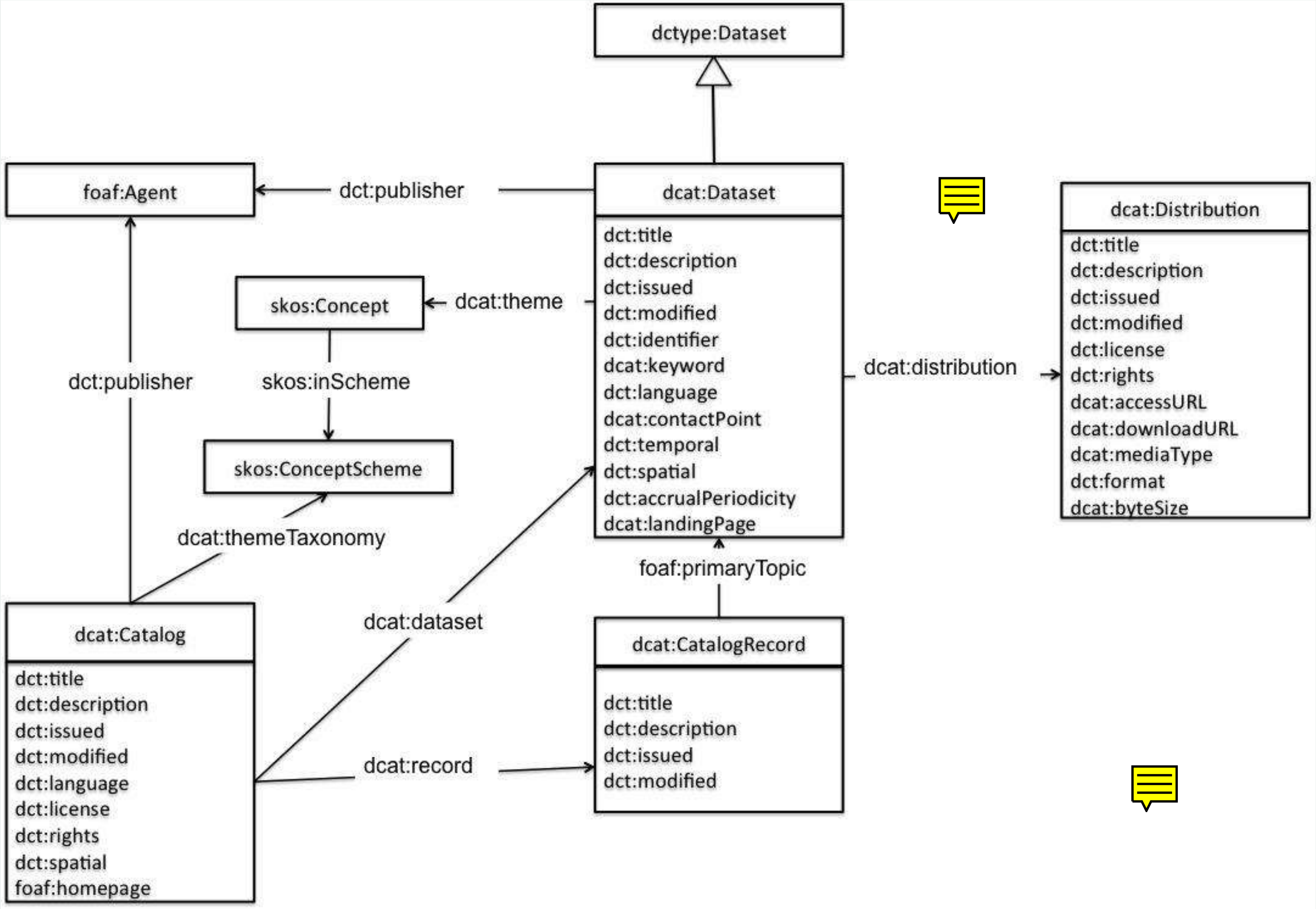
- Most fields are optional

No specific deployment method

- RDF via SPARQL, embedded in HTML, serialised to RDF/XML or Turtle, etc.

Mostly used in open governmental data repositories

DCAT



Specification and examples: https://www.w3.org/TR/vocab-dcat/

DCAT Application Profile

DCAT profile is a specification that adds additional constraints

Application Profile for open data in Europe

- https://joinup.ec.europa.eu/solution/dcat-application-profile-data-portals-europe/about

4.3. Dataset

4.3.1. Mandatory properties for Dataset

Property	URI	Range	Usage note	Card
description	dct:description	rdfs:Literal	This property contains a free-text account of the Dataset. This property can be repeated for parallel language versions of the description.	1..n
title	dct:title	rdfs:Literal	This property contains a name given to the Dataset. This property can be repeated for parallel language versions of the name.	1..n

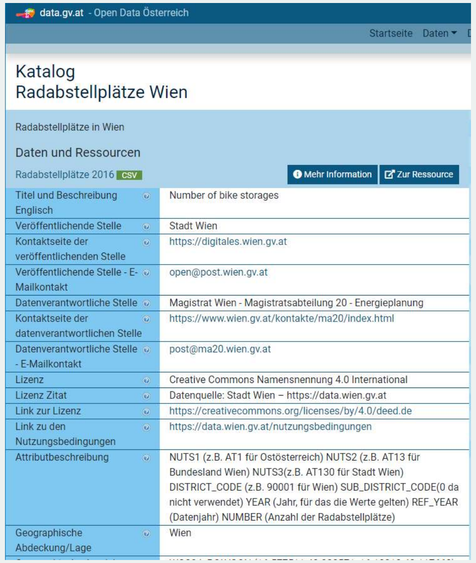
4.3.2. Recommended properties for Dataset

Property	URI	Range	Usage note	Card
contact point	dcat:contactPoint	vcard:Kind	This property contains contact information that can be used for sending comments about the Dataset.	0..n
dataset distribution	dcat:distribution	dcat:Distribution	This property links the Dataset to an available Distribution.	0..n
keyword/ tag	dcat:keyword	rdfs:Literal	This property contains a keyword or tag describing the Dataset.	0..n
publisher	dct:publisher	foaf:Agent	This property refers to an entity (organisation) responsible for making the Dataset available.	0..1
theme/ category	dcat:theme, subproperty of dct:subject	skos:Concept	This property refers to a category of the Dataset. A Dataset may be associated with multiple themes.	0..n

DCAT - example

Find a dataset and .rdf/xml/ttl/n3/jsonld

- https://www.data.gv.at/katalog/dataset/af8e02b6-1e03-4464-a69b-8533d8703ffe
- https://www.data.gv.at/katalog/dataset/af8e02b6-1e03-4464-a69b-8533d8703ffe**.ttl**



DCAT - example

```
<dcat:Dataset rdf:about="https://www.data.gv.at/dataset/af8e02b6-1e03-4464-a69b-8533d8703ffe">
    <dct:title>Radabstellplätze Wien</dct:title>
    <dct:description>Radabstellplätze in Wien</dct:description>
    <dcat:keyword>fahrrad</dcat:keyword>
    <dcat:keyword>räder</dcat:keyword>
    <dcat:keyword>verkehr</dcat:keyword>
    <dcat:keyword>fahrräder</dcat:keyword>
    <dcat:distribution>
      <dcat:Distribution rdf:about="https://www.data.gv.at/dataset/af8e02b6-1e03...">
        <dct:title>Radabstellplätze 2016</dct:title>
        <dct:format>CSV</dct:format>
        <dcat:accessURL rdf:resource="https://www.wien.gv.at/gogv/l9radabstellplaetze2016"/>
      </dcat:Distribution>
    </dcat:distribution>
  </dcat:Dataset>
```

SPARQL endpoint - example

EU Open Data Portal

Retrieve all the resources from a dataset with a title that contains specific words (eg.'Vienna')

SPARQL

You can search for the metadata stored in the EU Open Data Portal triple store by using the SPARQL endpoint query editor below.

Namespaces *

```
PREFIX dcat: <http://www.w3.org/ns/dcat#>
PREFIX odp:  <http://data.europa.eu/euodp/ontologies/ec-odp#>
PREFIX dc: <http://purl.org/dc/terms/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
```

SPARQL query *

```
SELECT  ?DatasetTitle  ?Publisher ?ResourceDescription WHERE { graph ?g  {?DatasetURI a dcat:Dataset;
dc:publisher ?Publisher; dc:title ?DatasetTitle; dcat:distribution ?Resource. ?Resource dc:description ?
ResourceDescription. FILTER(regex(?DatasetTitle,"Vienna", "i")) } }      LIMIT 10
```

https://data.europa.eu/euodp/en/linked-data

SPARQL endpoint - example

EU Open Data Portal

DatasetTitle	Publisher	ResourceDescription
"Glossary City of Vienna"	http://publications.europa.eu/resource/authority/corporate-body/CNECT	"Data archive containing files in the following formats: application/xml"
"University of Vienna Termbanks"	http://publications.europa.eu/resource/authority/corporate-body/CNECT	"Data archive containing files in the following formats: application/xml"
"Audioguide for the Military History Museum in Vienna"	http://publications.europa.eu/resource/authority/corporate-body/CNECT	"Data archive containing files in the following formats: MS-Word doc"

https://data.europa.eu/euodp/en/linked-data

Schema.org

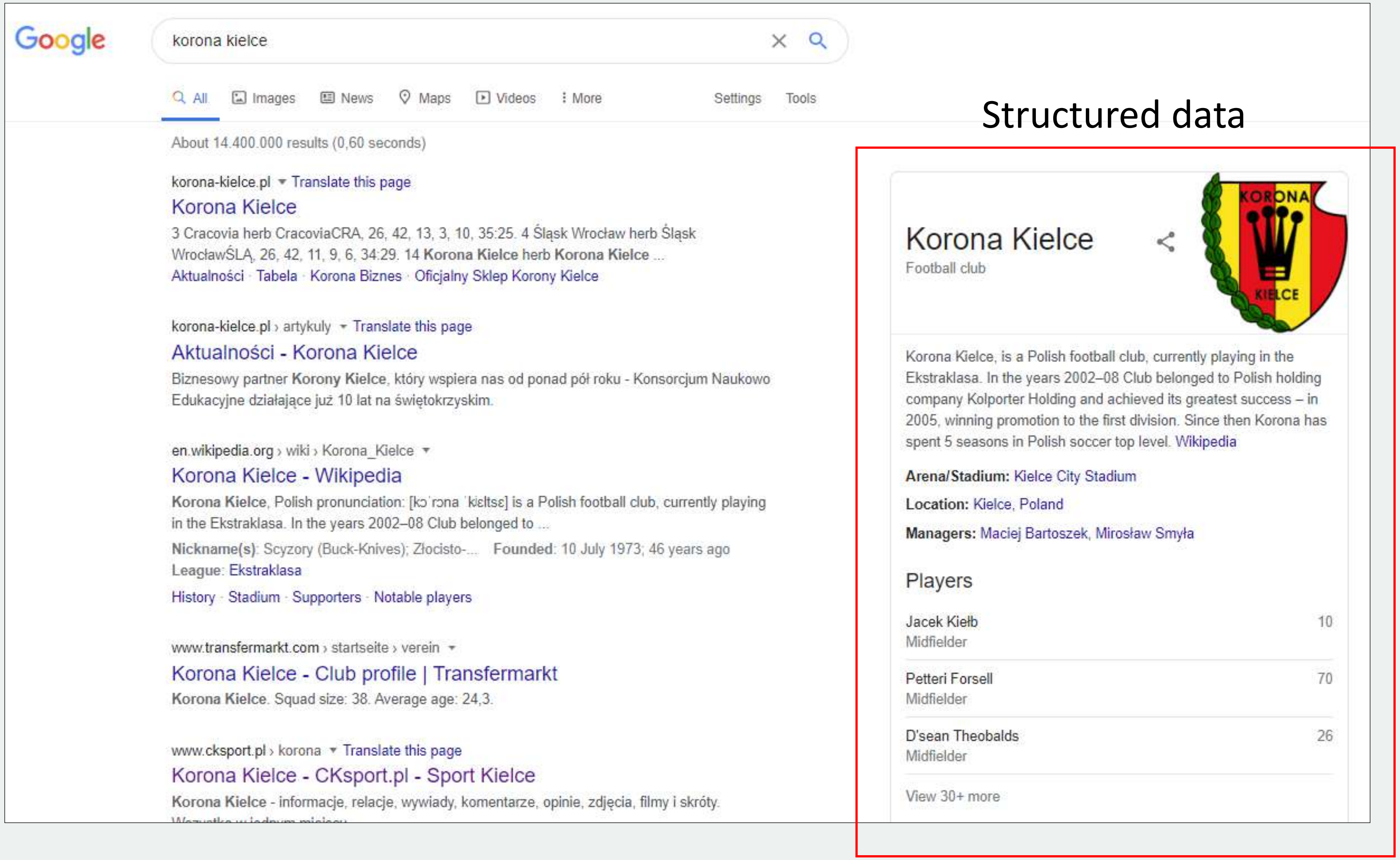
Started by Google, Microsoft, Yahoo, and Yandex to help with indexing web pages for search

Schema.org metadata can be embedded using microdata, RDFa or JSON-LD

Commonly used types

- Creative works: CreativeWork, Book, Movie, MusicRecording, Recipe, TVSeries
- Embedded non-text objects: AudioObject, ImageObject, VideoObject
- Event
- Organization
- Person
- Place, LocalBusiness, Restaurant ...
- Product, Offer, AggregateOffer
- Review, AggregateRating

Schema.org


Google
korona kielce
All
Images
News
Maps
Videos
More
Settings
Tools
Structured data
About 14.400.000 results (0,60 seconds)
korona-kielce.pl ▾ Translate this page
Korona Kielce
3 Cracovia herb CracoviaCRA, 26, 42, 13, 3, 10, 35:25. 4 Śląsk Wrocław herb Śląsk WrocławŚLĄ, 26, 42, 11, 9, 6, 34:29. 14 Korona Kielce herb Korona Kielce ...
Aktualności · Tabela · Korona Biznes · Oficjalny Sklep Korony Kielce
korona-kielce.pl › artykuly ▾ Translate this page
Aktualności - Korona Kielce
Biznesowy partner Korony Kielce, który wspiera nas od ponad pół roku - Konsorcjum Naukowo Edukacyjne działające już 10 lat na świętokrzyskim.
en.wikipedia.org › wiki › Korona_Kielce ▾
Korona Kielce - Wikipedia
Korona Kielce, Polish pronunciation: [kɔˈrɔna ˈkʲɛltsɛ] is a Polish football club, currently playing in the Ekstraklasa. In the years 2002–08 Club belonged to ...
Nickname(s): Scyzory (Buck-Knives); Złocisto-... Founded: 10 July 1973; 46 years ago
League: Ekstraklasa
History · Stadium · Supporters · Notable players
www.transfermarkt.com › startseite › verein ▾
Korona Kielce - Club profile | Transfermarkt
Korona Kielce. Squad size: 38. Average age: 24,3.
www.cksport.pl › korona ▾ Translate this page
Korona Kielce - CKsport.pl - Sport Kielce
Korona Kielce - informacje, relacje, wywiady, komentarze, opinie, zdjęcia, filmy i skróty.
Korona Kielce
Football club
KORONA
KIELCE
Korona Kielce, is a Polish football club, currently playing in the Ekstraklasa. In the years 2002–08 Club belonged to Polish holding company Kolporter Holding and achieved its greatest success – in 2005, winning promotion to the first division. Since then Korona has spent 5 seasons in Polish soccer top level. Wikipedia
Arena/Stadium: Kielce City Stadium
Location: Kielce, Poland
Managers: Maciej Bartoszek, Mirosław Smyła
Players
Jacek Kiełb 10
Midfielder
Petteri Forsell 70
Midfielder
D'sean Theobalds 26
Midfielder
View 30+ more

Google Dataset Search

schema.org:Dataset

- based on W3C DCAT

Full definition

- https://schema.org/Dataset

Google has an *application profile*

- https://developers.google.com/search/docs/data-types/dataset

Required properties	
`description`	Text A short summary describing a dataset.
`name`	Text A descriptive name of a dataset. For example, "Snow depth in Northern Hemisphere".

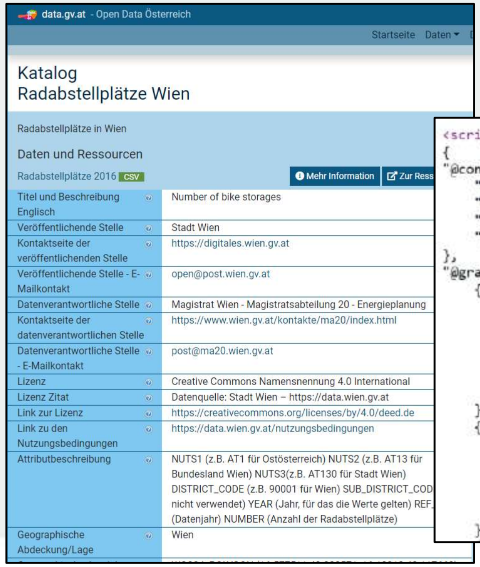
Recommended properties

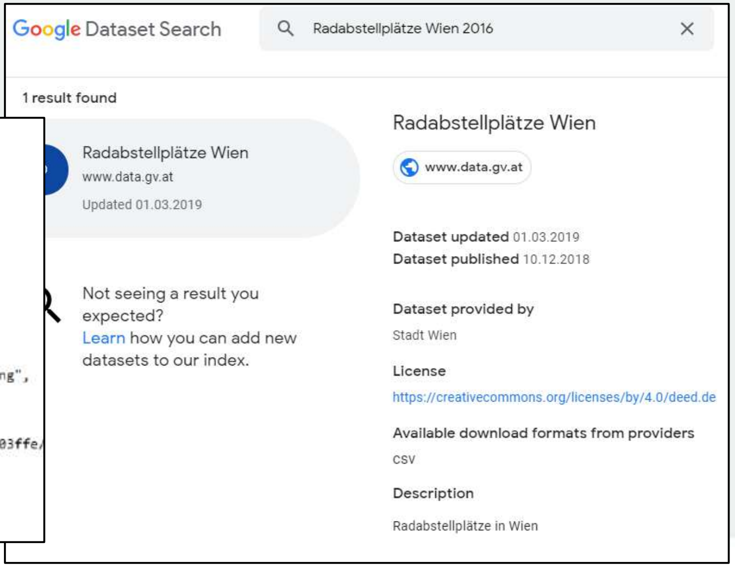
Schema.org - example

View page source of any dataset at data.gv.at

- Navigate to <script> section

Search in Google Dataset Search for ‘Radabstellplätze Wien 2016’





data.gv.at - Open Data Österreich

Startseite Daten

Katalog
Radabstellplätze Wien

Radabstellplätze in Wien

Daten und Ressourcen

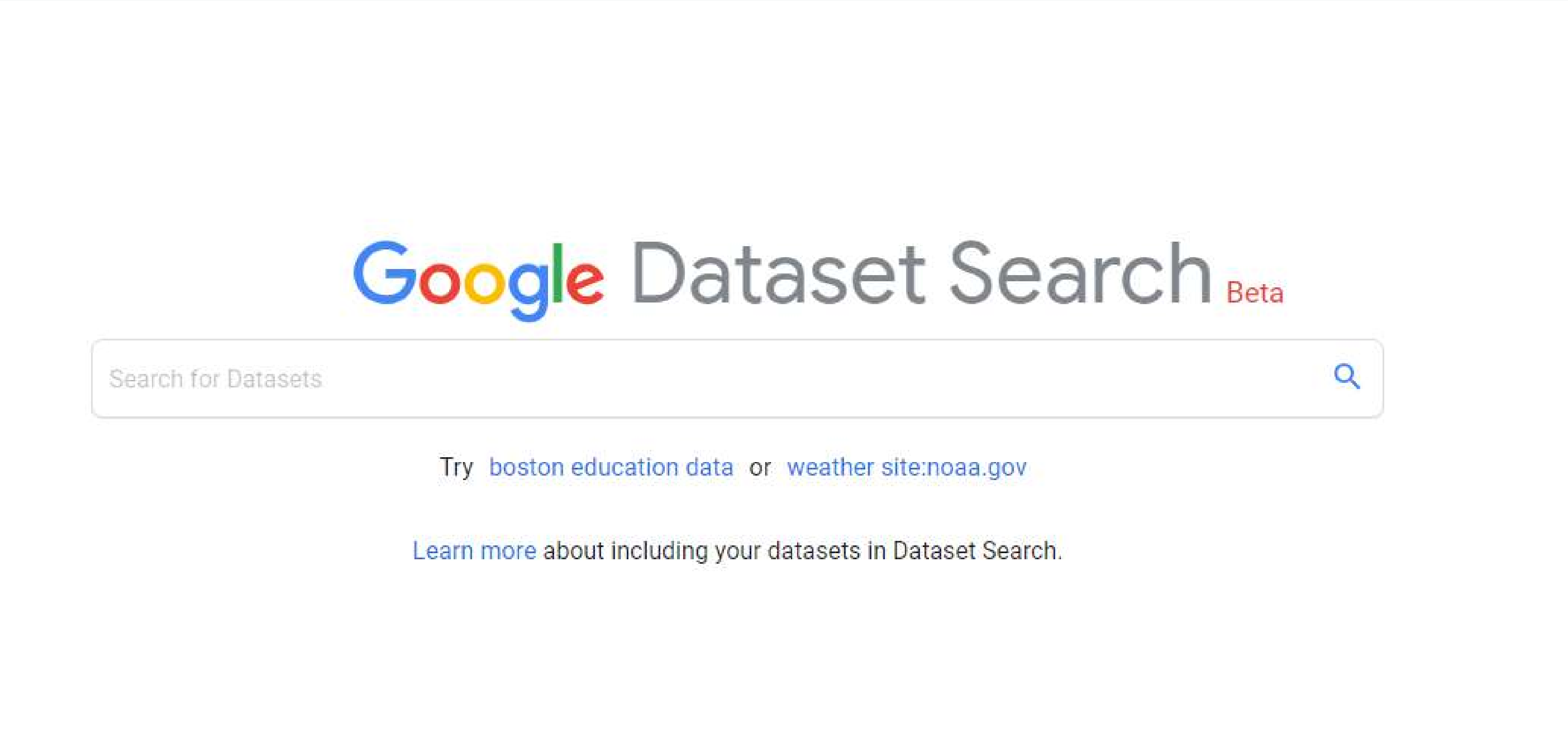
Radabstellplätze 2016 CSV

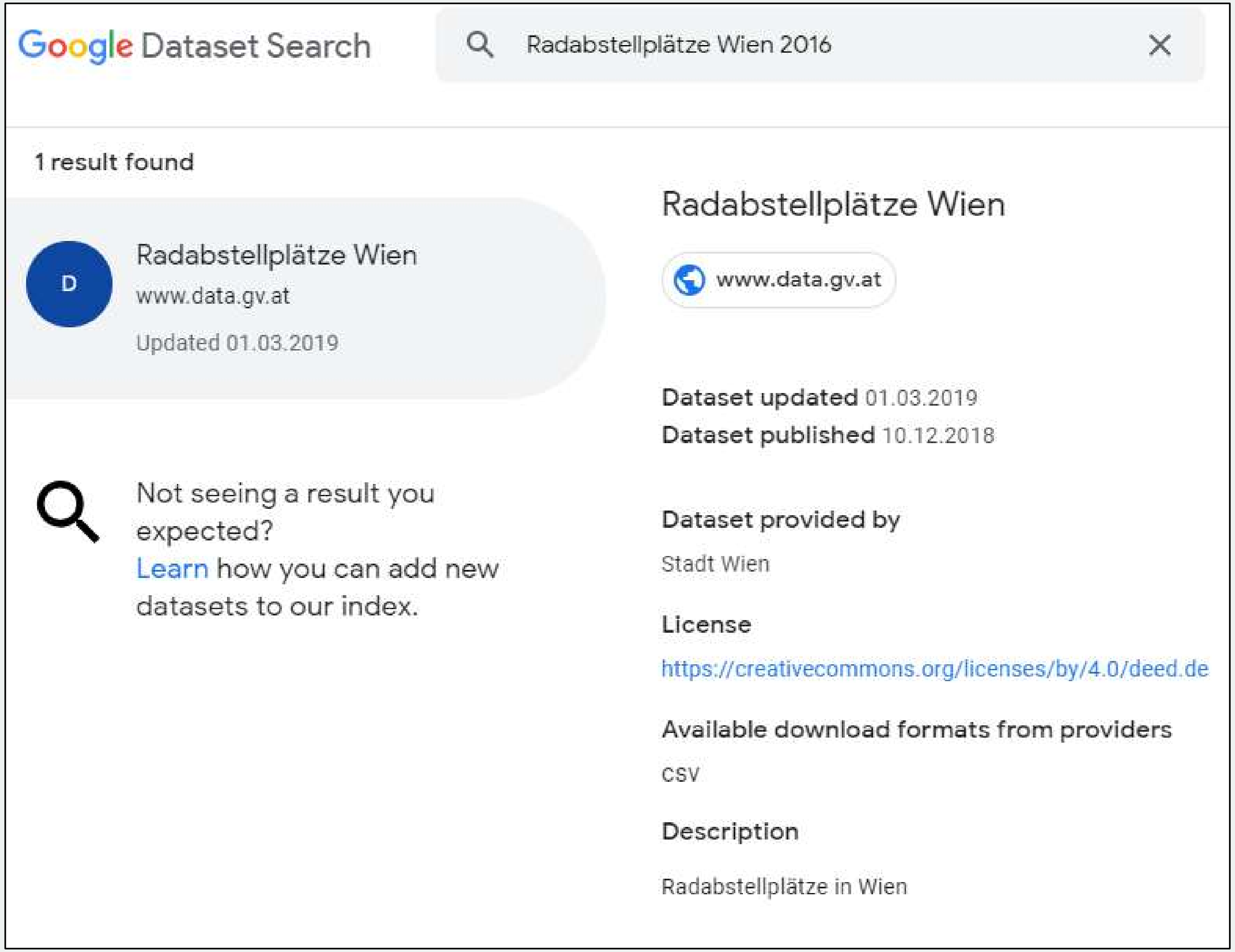
Mehr Information | Zur Ressource

Titel und Beschreibung Englisch	Number of bike storages
Veröffentlichende Stelle	Stadt Wien
Kontaktseite der veröffentlichenden Stelle	https://digitales.wien.gv.at
Veröffentlichende Stelle - E-Mailkontakt	open@post.wien.gv.at
Datenverantwortliche Stelle	Magistrat Wien - Magistratsabteilung 20 - Energieplanung
Kontaktseite der datenverantwortlichen Stelle	https://www.wien.gv.at/kontakte/ma20/index.html
Datenverantwortliche Stelle - E-Mailkontakt	post@ma20.wien.gv.at
Lizenz	Creative Commons Namensnennung 4.0 International
Lizenz Zitat	Datenquelle: Stadt Wien – https://data.wien.gv.at
Link zur Lizenz	https://creativecommons.org/licenses/by/4.0/deed.de
Link zu den Nutzungsbedingungen	https://data.wien.gv.at/nutzungsbedingungen
Attributbeschreibung	NUTS1 (z.B. AT1 für Ostösterreich) NUTS2 (z.B. AT13 für Bundesland Wien) NUTS3(z.B. AT130 für Stadt Wien) DISTRICT_CODE (z.B. 90001 für Wien) SUB_DISTRICT_CODE(0 da nicht verwendet) YEAR (Jahr, für das die Werte gelten) REF_YEAR (Datenjahr) NUMBER (Anzahl der Radabstellplätze)
Geographische Abdeckung/Lage	Wien

```
<script type="application/ld+json">
{
"@context": {
    "rdf": "http://www.w3.org/1999/02/22-rdf-syntax-ns#",
    "rdfs": "http://www.w3.org/2000/01/rdf-schema#",
    "schema": "http://schema.org/",
    "xsd": "http://www.w3.org/2001/XMLSchema#"
},
"@graph": [
    {
        "@id": "_:N7d7157714cf1492a944372bc2c224f3f",
        "@type": "schema:ContactPoint",
        "schema:contactType": "customer service",
        "schema:email": "open@post.wien.gv.at ",
        "schema:name": "Magistrat Wien - Magistratsabteilung 20 - Energieplanung",
        "schema:url": "https://www.data.gv.at"
    },
```

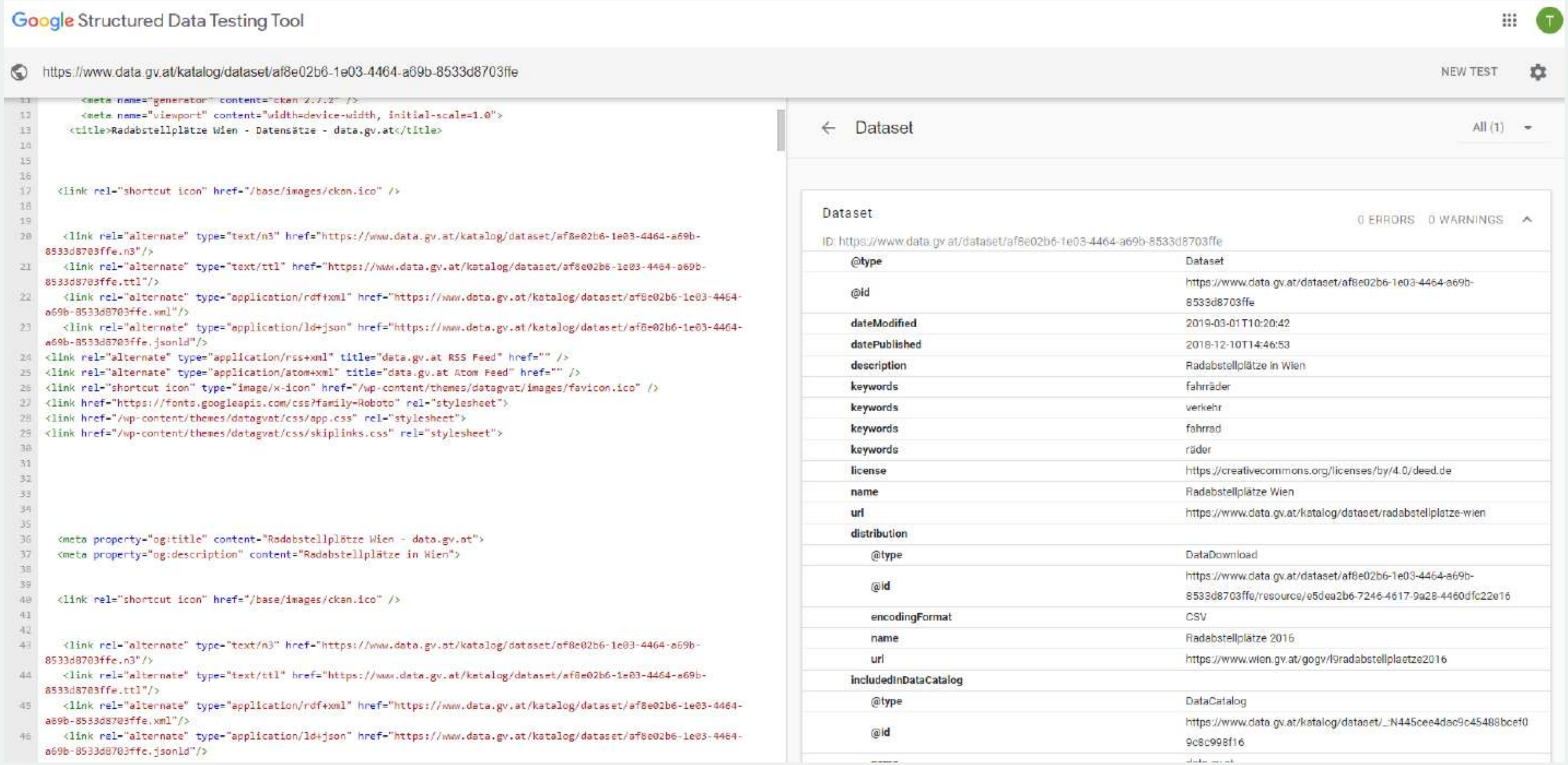
```
    "@id": "https://www.data.gv.at/dataset/af8e02b6-1e03-4464-a69b-8533d8703ffe",
    "@type": "schema:Dataset",
    "schema:dateModified": "2019-03-01T10:20:42.981483",
    "schema:datePublished": "2018-12-10T14:46:53.400738",
    "schema:description": "Radabstellpl\u00e4tze in Wien\r\n",
    "schema:distribution": {
        "@id": "https://www.data.gv.at/dataset/af8e02b6-1e03-4464-a69b-8533d8703ffe/resour
    },
    "schema:includedInDataCatalog": {
        "@id": "_:N5611dfe394574bbf9e32004749146 54a"
```


Google Dataset Search Beta
Search for Datasets
Try boston education data or weather site:noaa.gov
Learn more about including your datasets in Dataset Search.



https://toolbox.google.com/datasetsearch

Structured Data Testing Tool



https://search.google.com/structured-data/testing-tool

Structured Data Testing Tool

Dataset

0 ERRORS 0 WARNINGS

ID: https://www.data.gv.at/dataset/af8e02b6-1e03-4464-a69b-8533d8703ffe

@type	Dataset
@id	https://www.data.gv.at/dataset/af8e02b6-1e03-4464-a69b-8533d8703ffe
dateModified	2019-03-01T10:20:42
datePublished	2018-12-10T14:46:53
description	Radabstellplätze in Wien
keywords	fahrräder
keywords	verkehr
keywords	fahrrad
keywords	räder
license	https://creativecommons.org/licenses/by/4.0/deed.de
name	Radabstellplätze Wien
url	https://www.data.gv.at/katalog/dataset/radabstellplatze-wien
distribution	
@type	DataDownload
@id	https://www.data.gv.at/dataset/af8e02b6-1e03-4464-a69b-8533d8703ffe/resource/e5dea2b6-7246-4617-9a28-4460dfc22e16
encodingFormat	CSV
name	Radabstellplätze 2016
url	https://www.wien.gv.at/gogv/l9radabstellplaetze2016

Synergy between DOIs and schema.org

DataCite provides metadata on DOIs

- RDFs compliant with schema.org

Example

- https://data.datacite.org/application/vnd.schemaorg.ld+json/ **<DOI>**
- **https://data.datacite.org/application/vnd.schemaorg.ld+json/10.1371/journal.pcbi.1006750**

JSON-LD returned by DataCite can be directly embedded on web pages

- Performance issues!

https://data.datacite.org

https://data.datacite.org/application/vnd.schemaorg.ld+json/10.5284/1015

https://data.datacite.org

Repository registries

Directory of Open Access Repositories – DOAR

- Based on registrations
- http://www.opendoar.org/

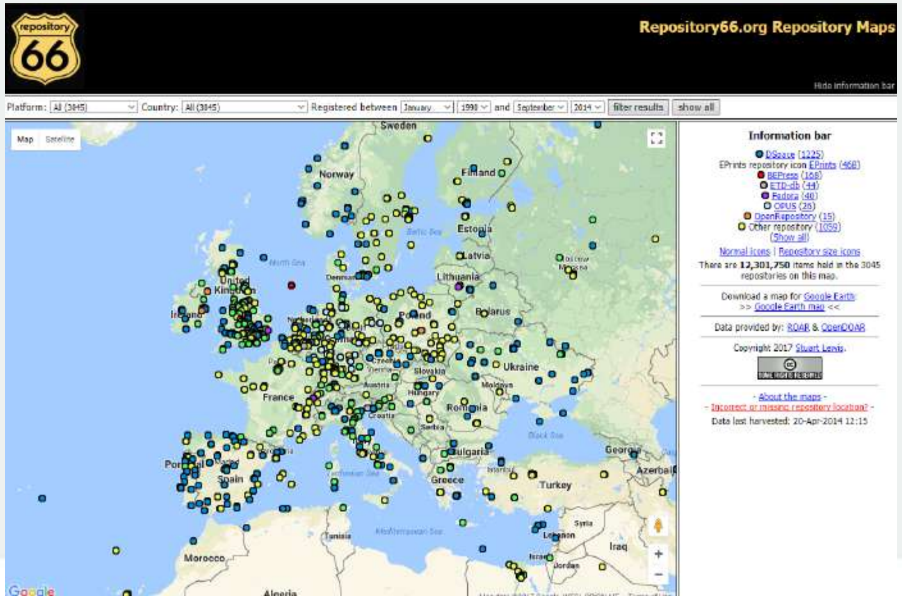
Registry of Open Access Repositories – ROAR

- Automatically harvested list based on OAI-PMH
- http://roar.eprints.org/

Projection of DOAR and ROAR on maps

- http://maps.repository66.org

- re3data.org
- FAIRsharing.org



Repository registries – re3data

Repository details

Phaidra Universität Wien

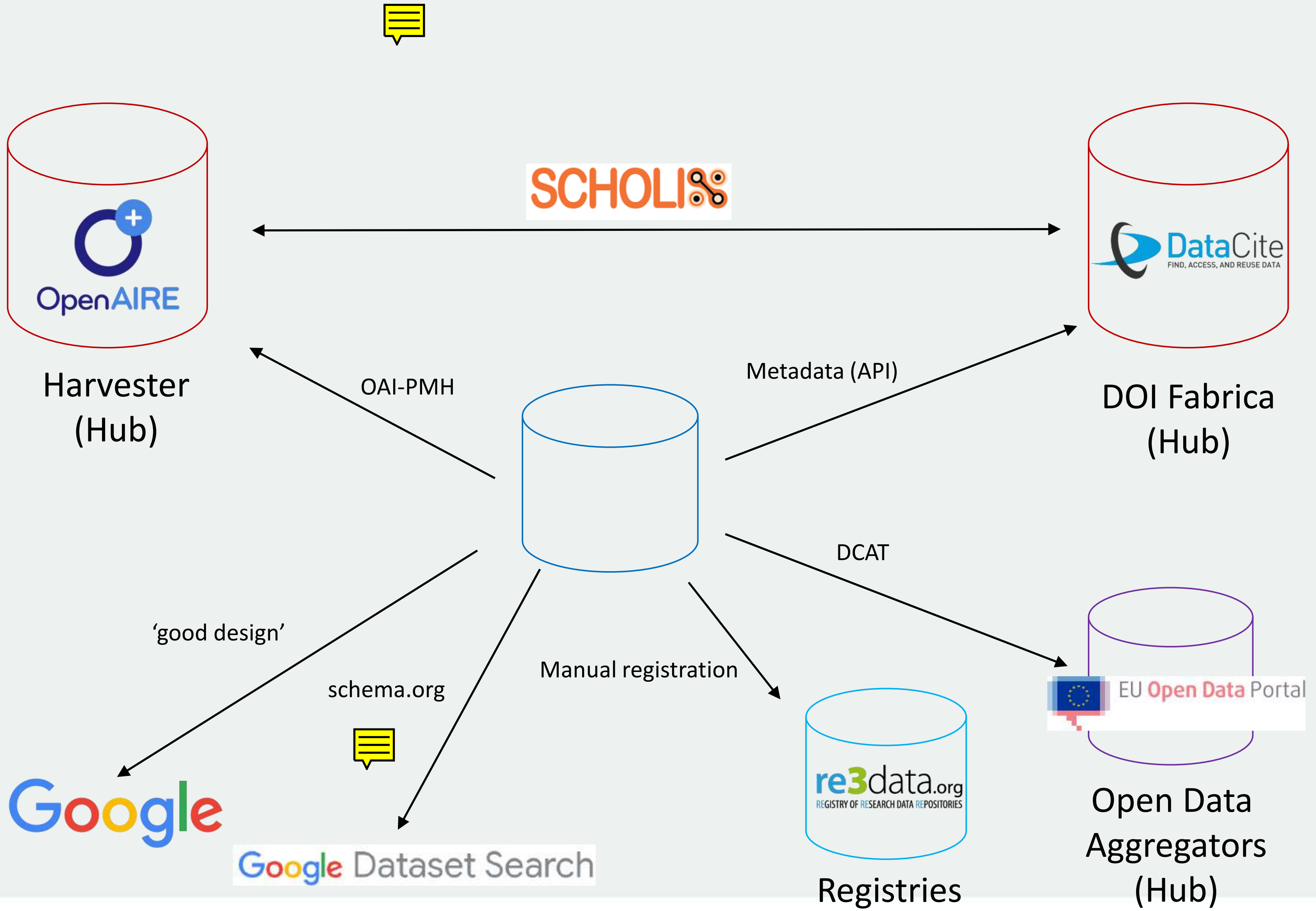
General | Institutions | Terms | Standards

Name of repository	**Phaidra Universität Wien**
Additional name(s)	Permanent Hosting, Archiving and Indexing of Digital Resources and Assets
Repository URL	https://phaidra.univie.ac.at/
Subject(s)	Humanities and Social Sciences, Life Sciences, Natural Sciences, Engineering Sciences
Description	Phaidra Universität Wien, is the innovative whole-university digital asset management system with long-term archiving functions, offers the possibility to archive valuable data university-wide with permanent security and systematic input, offering multilingual access using metadata (data about data), thus providing worldwide availability around the clock. As a constant data pool for administration, research and teaching, resources can be used flexibly, where continual citability allows the exact location and retrieval of prepared digital objects.
Contact	support.phaidra@univie.ac.at phaidra@univie.ac.at
Content type(s)	Images, Audiovisual data, Scientific and statistical data formats, Networkbased data, Plain text, other
Keyword(s)	hosting, long-term-archiving, multidisciplinary, digital objects, research
Repository type(s)	institutional other
Mission statement for designated community	https://datamanagement.univie.ac.at/en/about-phaidra/policy-of-phaidra/
Research data repository language(s)	eng deu ita
Data and/or service provider	dataProvider

https://www.re3data.org/repository/r3d100010472

SUMMARY

External visibility – your main options



You should know

How to make repository contents visible?

- what options there are and how to choose the best one for you setting?
- how to describe data using discussed standards
  - DCAT
  - DataCite

How do repositories support FAIRness?

- Which (sub-) principles specifically and how?

Next lecture

Developing Research Data Management Services

- Data lifecycle model
- Policies
- Costs and Business Models
- Data Stewards
- Repository Certification

Developing Research Data Management Services and Repository Certification

Tomasz Miksa

Agenda

Research Lifecycle models

Developing Research Data Management (RDM) Services

- Policies
- Costs
- Data Management
- Infrastructure

Repository Certification

Introduction

Perspective change

- ‘This lecture
  - You need to design a solution at an institution supporting data management and preservation
    - Taking into account technical, organisational, cultural, political problems…
  - **What do you do?!**

The size and governance structure of institutions has an impact

Large, hierarchical institutions

- Move slowly
- Require a lot of advocacy
- Have more resource
- Economies of scale

Smaller institutions

- More agile
- Simpler communication
- More focussed vision
- Less resource

Establish a long term governance group

Good mix of representatives from operational units

Senior management leadership

Avoid relying on one or two key individuals

Keep the group active

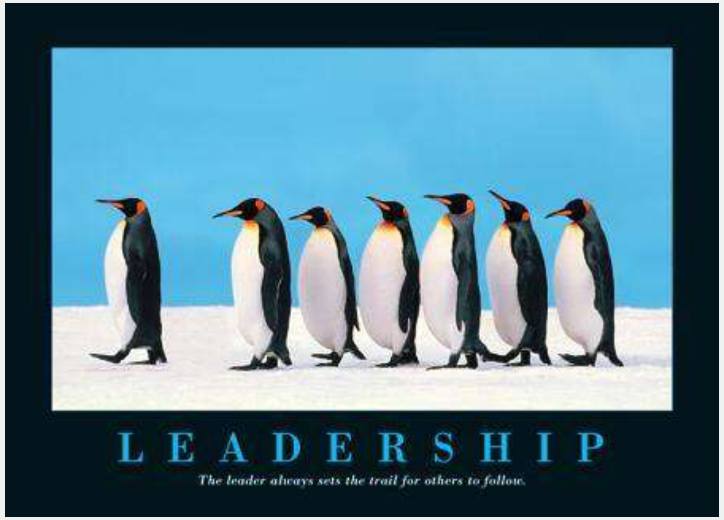


Image credit: http://www.executive-coaching-services.co.uk/executive-coaching/leaders.jpg

Be aware of existing infrastructure

In many HEIs services are beginning to be embedded but aren't joined up effectively

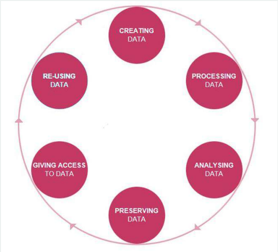
Research Organisation Description		
RDM profile component	**Record HEI Link - insert your URLs in the space provided**	**Guidance**
Means of raising staff awareness of funders' research data requirements		Provide a link to an information page on funders' policies. This could be internal or external (E.g., DCC's policy overview table)
Research data policy		Provide a link to research data policy or aspirational statement
Strategy or implementation plan for research data services		Provide a link to research data strategy page or roadmap
RDM advice and support services		Provide a link to page describing data management planning guidance and/or support services at this organisation
Active data storage		Provide link(s) to active research data storage information page(s). There may be multiple options at Research Group/School/College/Central levels.
Data register or catalogue		Provide a link to your internal research data registration homepage. This may be provided via the data repository and/or CRIS.
Persistent identification for datasets		Provide a link to any page(s) detailing schemes used to identify digital data items (e.g., DataCite).
Data access procedures		Provide a link to any information provided about research data access.
Secure data access		Provide a link to any information provided about secure data access and governance.
Institutional publications repository (if it includes research data or metadata)		Provide a link to your institutional repository homepage
Data repository for longer term access and preservation		Provide a link to your research data repository homepage. This may be an extension of your publications repository, a separate data repository or a pointer to an external data repository service (E.g., Zenodo).

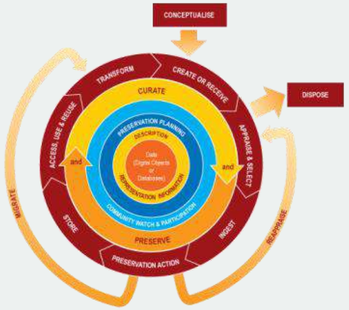
http://www.dcc.ac.uk/projects/opd-for-rdm

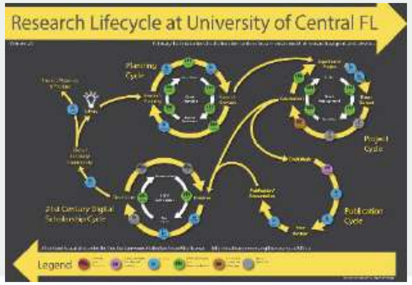
Understand processes

Research Data Lifecycle Models

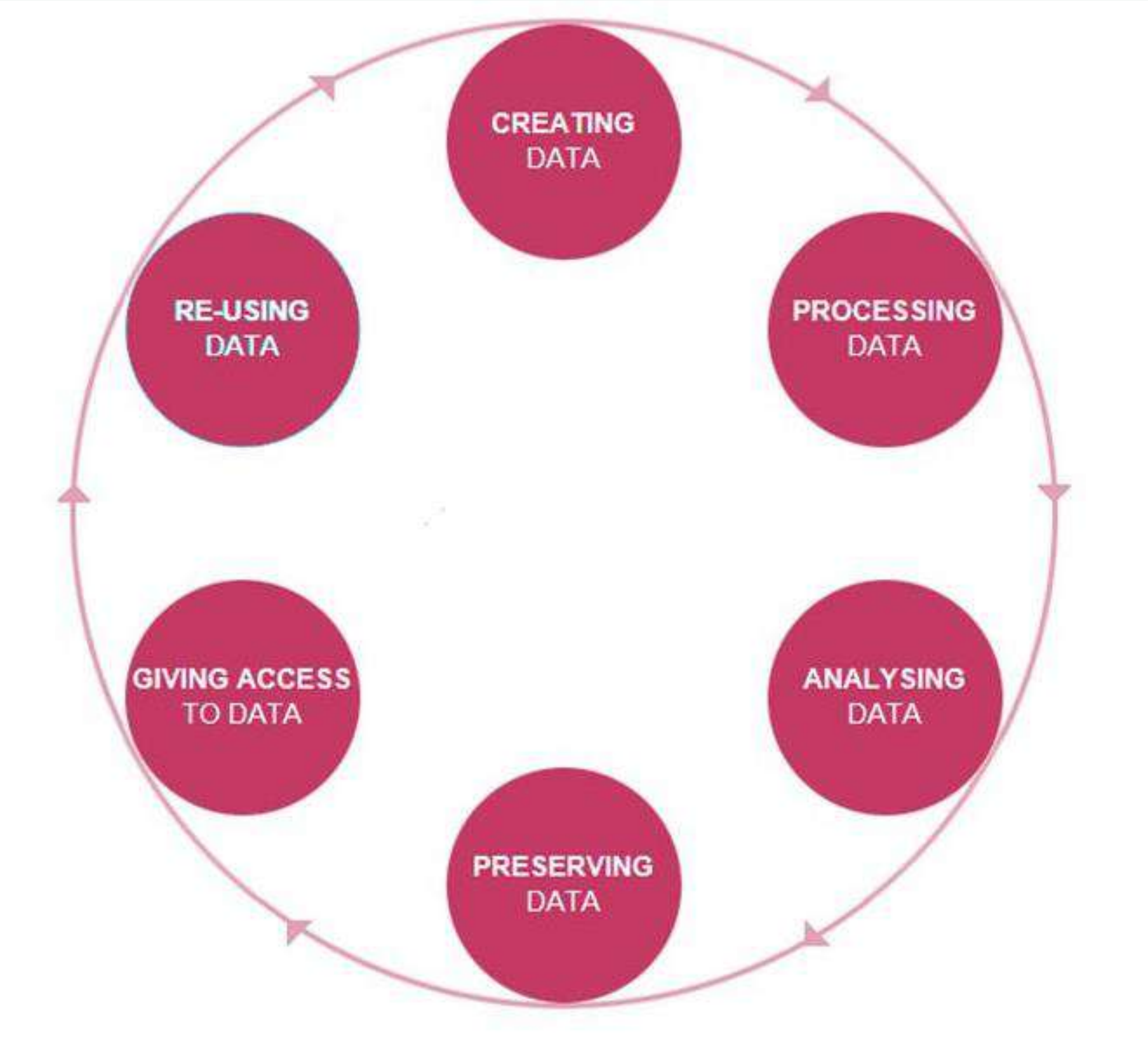
- Describe roles of stakeholders
- Help in
  - tailoring services
  - identifying responsibilities
  - defining infrastructure
- NOT to be used by researchers
- Examples
  - UK Data Archive
  - Digital Curation Centre
  - University of Central Florida



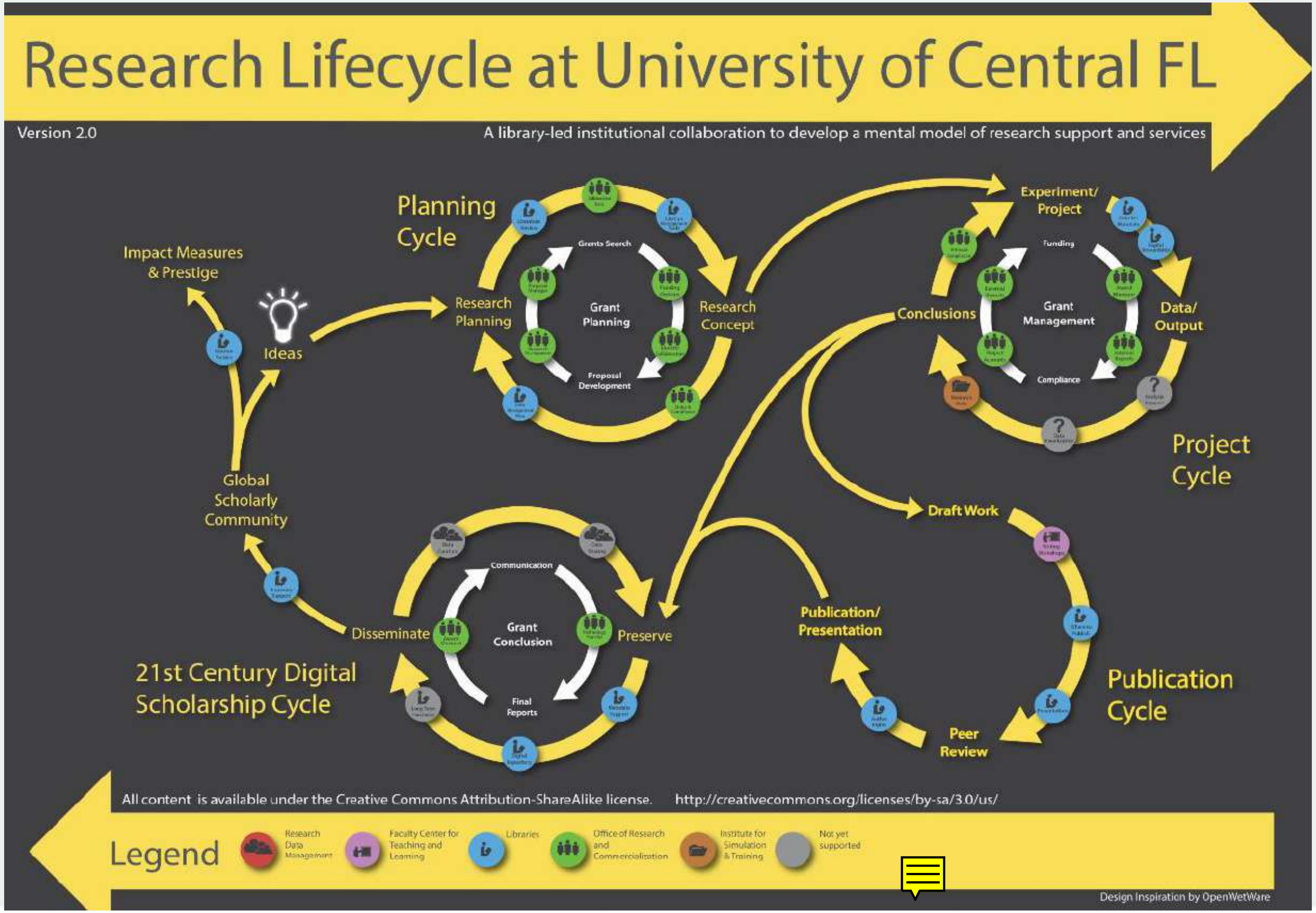




UK Data Archive Lifecycle model



University of Central Florida Lifecycle Model

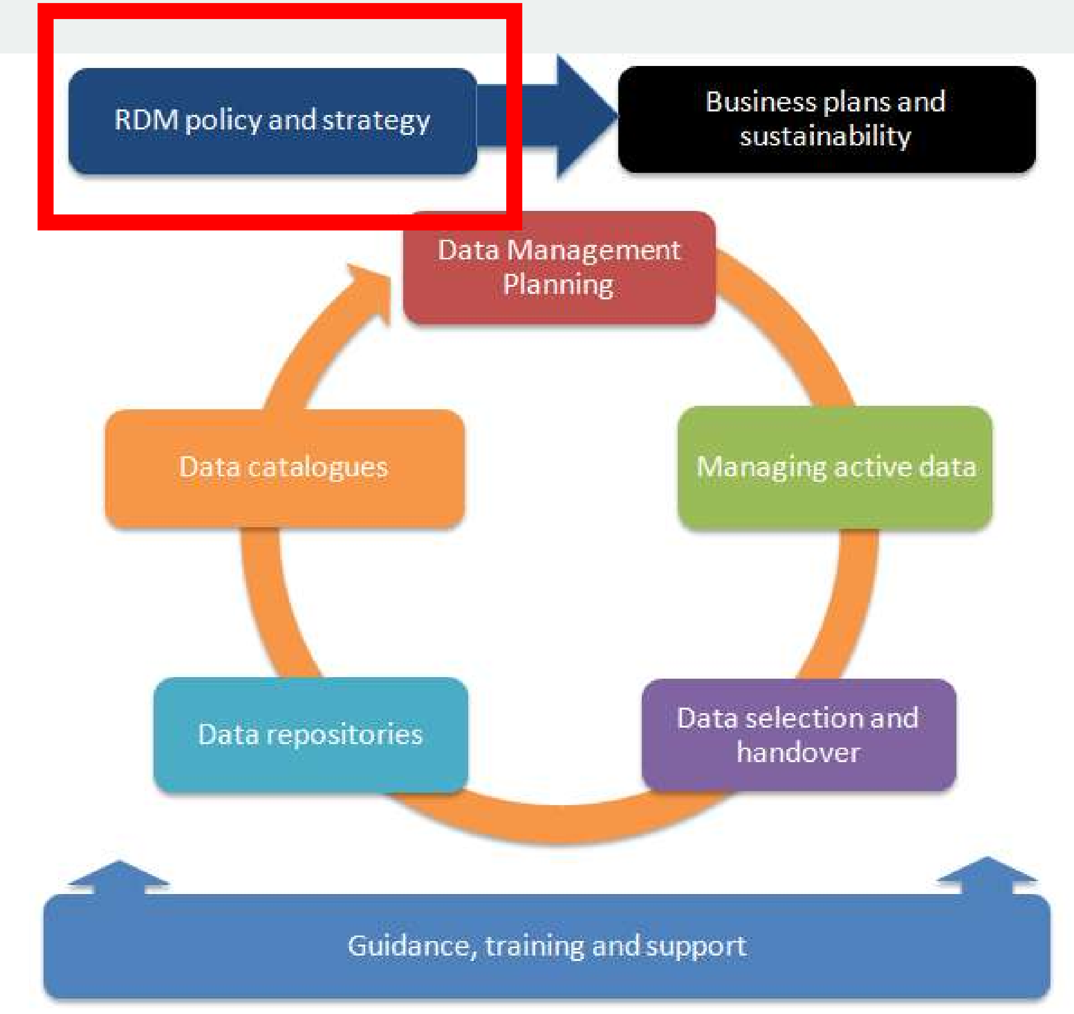


http://guides.ucf.edu/ScholarlyCommunication/ResearchLifecycle

Components of an RDM service infrastructure

- RDM policy and strategy
- Business plans and sustainability
- Guidance, training and support
- Data management planning
- Managing active data
- Data selection and handover
- Data repositories
- Data catalogues

Developing Research Data Management (RDM) Services



https://www.dcc.ac.uk/guidance/how-guides/how-develop-rdm-services

Understanding policies

1. Starting with some ***Taboos***
2. Going over to related ***Principles***
3. Going over to the creation of a ***Policy***
4. Going over to ***Rules***, ***Legislations*** and ***Regulations*** (canons, norms, guidelines)

Taboo

A **taboo** is something, which is **forbidden** or **disapproved of**, or placed under a social prohibition.

"Thou shalt not delete scientific data"
"Thou shalt not destroy infrastructures"

Usually a **negative** assertion.
In society and academic environment taboos are accepted only if they are just **a few.**

Principle

A principle is a fundamental truth or proposition that serves as the foundation for a system of belief

*Research **data** are to be preserved*

*Research **data** are to be kept **FAIR** - Findable, Accessible, Interoperable, Reusable.*

*Research data **infrastructures** are to be kept accessible*

Format: positive assertion

Policies

A policy is…
- a **course of action** or **principles** adopted by an organization

“The Institution [name XY] will preserve its research data infrastructure always accessible and free to its members according to the FAIR principles”

General assumptions concerning policies:

- A single Policy
  - the policy is a single entity, it should not be in competition with other policies
- Creators of Policy do not want to modify it
- Policy is usually accepted after a while
- Policy offers the frame for the generation of Rules

Rules, Regulations

Rules are prescribing conducts or actions
They are generated by the founder of “orders”
Characteristics of rules are:

- There may exist “**lots of rules**”
- Rules are not always **clear**
(they often need interpretation according to the situation)
- Rules are **usually accepted**, but **often imposed** procedures
- It is allowed to **modify Rules**
- The Law is an expression of rules

Example:

“Our University will maintain accessible our infrastructure each day from 9:00 a.m. to 12:00 a.m and offer support only on Friday from 7:00 a.m. to 8:00 a.m. The research data, that are publicly funded are to be kept free and accessible to all members of our University each Sunday, from 9:00 to 12:00 a.m.“

Taboos	Principles	Policies	Rules
Negative assertion few	Positive assertion more than „few“	A course or principle of action. Policy offers the frame for the generation of Rules, should not be in competition with other policies	Rules prescribe conducts or actions; define who what when and where should be done according to the Policy
“You shall not delete scientific data” *“Youl shall not destroy infrastrcutures”*	*“Research data are to be kept FAIR - Findable, Accessible, Interoperable, Reusable.”* *“Research data infrastructures are to be kept accessible”*	*“The Institution [name XY] will preserve its research data infrastructure always accessible and free to its members according to the FAIR principles”*	*“Our University will maintain accessible our infrastructure each day from 9:00 a.m. to 12:00 a.m and offer support only on Friday from 7:00 a.m. to 8:00 a.m. “*

Why these differentiations?

It is important to identify the different semantic levels

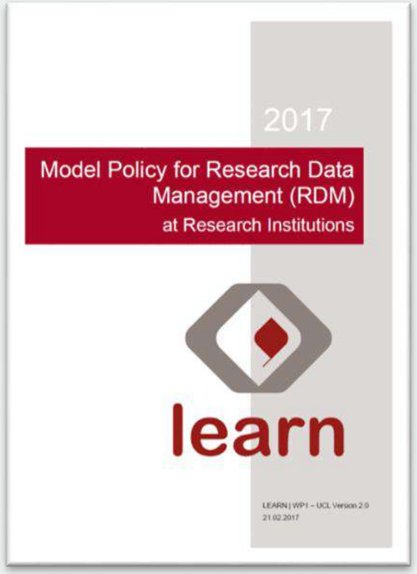
Understanding of the semantic hierarchy is useful in order to produce appropriate guidelines

Policy – LEARN project

Leader Activating Research Networks

Research Data Management Policy

- can be tailored by any University or Research Institution
- based on existing European policies



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654139.

Outreach example: Austria

Adaptation to needs of

- five Austrian art universities (started)
- three Medical Universities
- TU Wien

e-infrastructures
austria

Policy sections

1. Preamble
2. Jurisdiction
3. Intellectual Property Rights
4. Handling Research Data
5. Responsibilities, Rights, Duties
   5.1 Researchers are responsible for:...
   5.2 The [name of research institution] is responsible for:
6. Validity

http://learn-rdm.eu/wp-content/uploads/RDMToolkit.pdf?pdf=RDMToolkit

Policy section examples

5.1 Researchers are responsible for:

- Collecting, documenting, archiving, providing access to and storing or ensuring the proper destruction of research data and records. (…) Such information should be included in a Data Management Plan (DMP) that explicitly defines the approach to matters of data collection, administration, integrity, confidentiality, storage, use and publication.

5.2 The Institution is responsible for:

- Designing and deploying mechanisms and services for the storage, safekeeping, registration and deposition of research data to support current and future access to research data during and after the completion of research activities;

Tricky points:

- How do you define research data? (publications, source code, raw data..)
- Who are researchers? (students, employed formally, etc.)

Homework – read TUW policy

Research > RTI support > Research data > Research data management > Policy

Policy for Research Data Management (RDM) at the TU Wien

The TU Wien sees itself as playing an important role in the expansion of knowledge and technology transfer of research results and thus encouraging innovation and ultimately benefitting economy and society. A key to supporting academic research activities lies in the institution's ability to systematically manage, preserve and make available scientific output from different disciplines for reuse. In its Policy for Research Data Management, the TU Wien affirms the value of research data for research and teaching and the potential of their reuse by society.

Policy_for_Research_Data_Management.pdf PDF 548 KB

1. PREAMBLE

2. SCOPE

3. RIGHTS OF USE

4. RESEARCH DATA AND RESEARCH DATA MANAGEMENT

5. HANDLING RESEARCH DATA

6. RESPONSIBILITIES, RIGHTS AND DUTIES

7. VALIDITY

Definitions

Recommendations

https://www.tuwien.at/en/research/rti-support/research-data/research-data-management/policy/

RDM policy and strategy

Develop a strategy

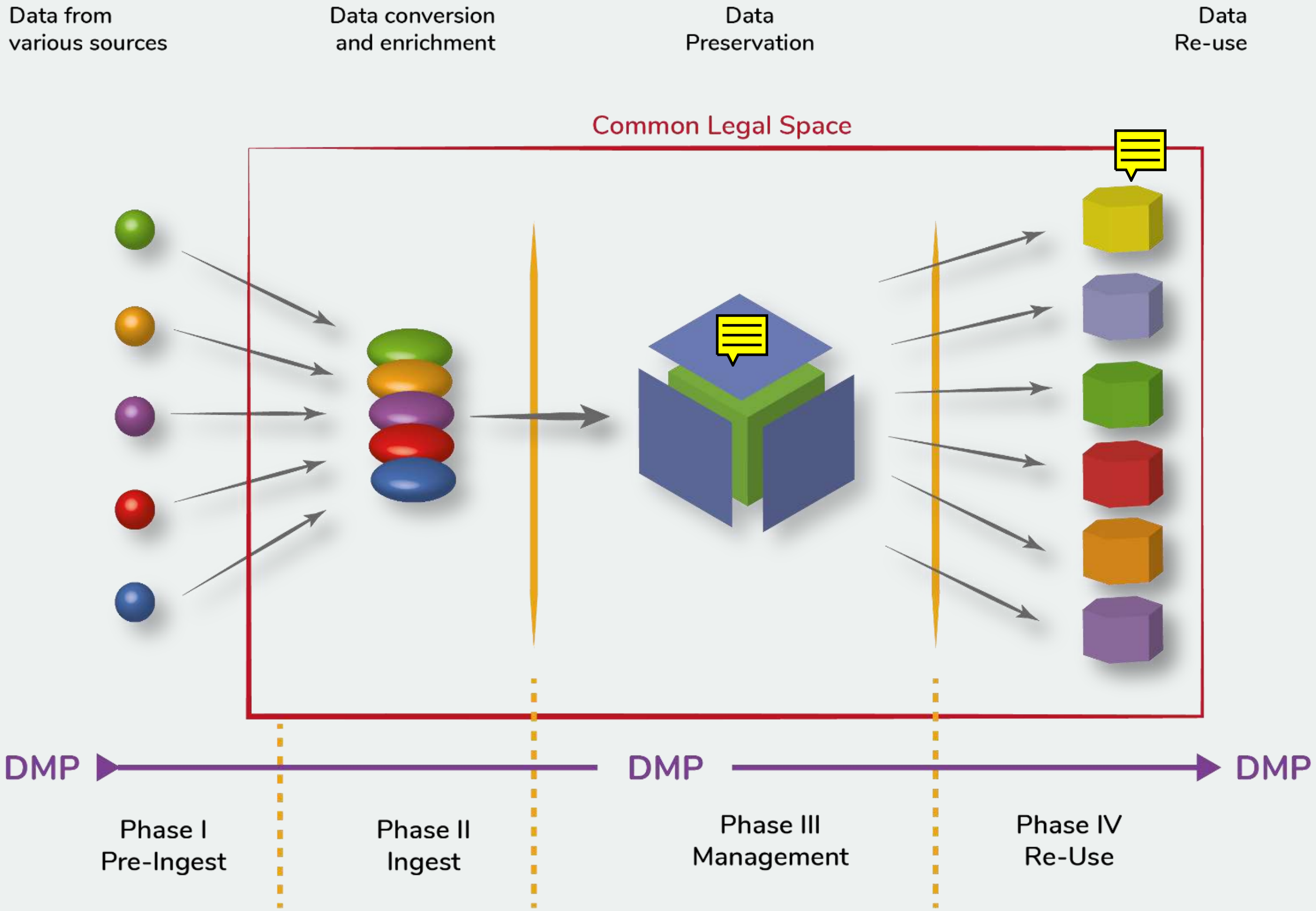
- Understand your current position and where you want to be to define your strategy
- Map out a programme of activity to deliver infrastructure and services

Develop a policy

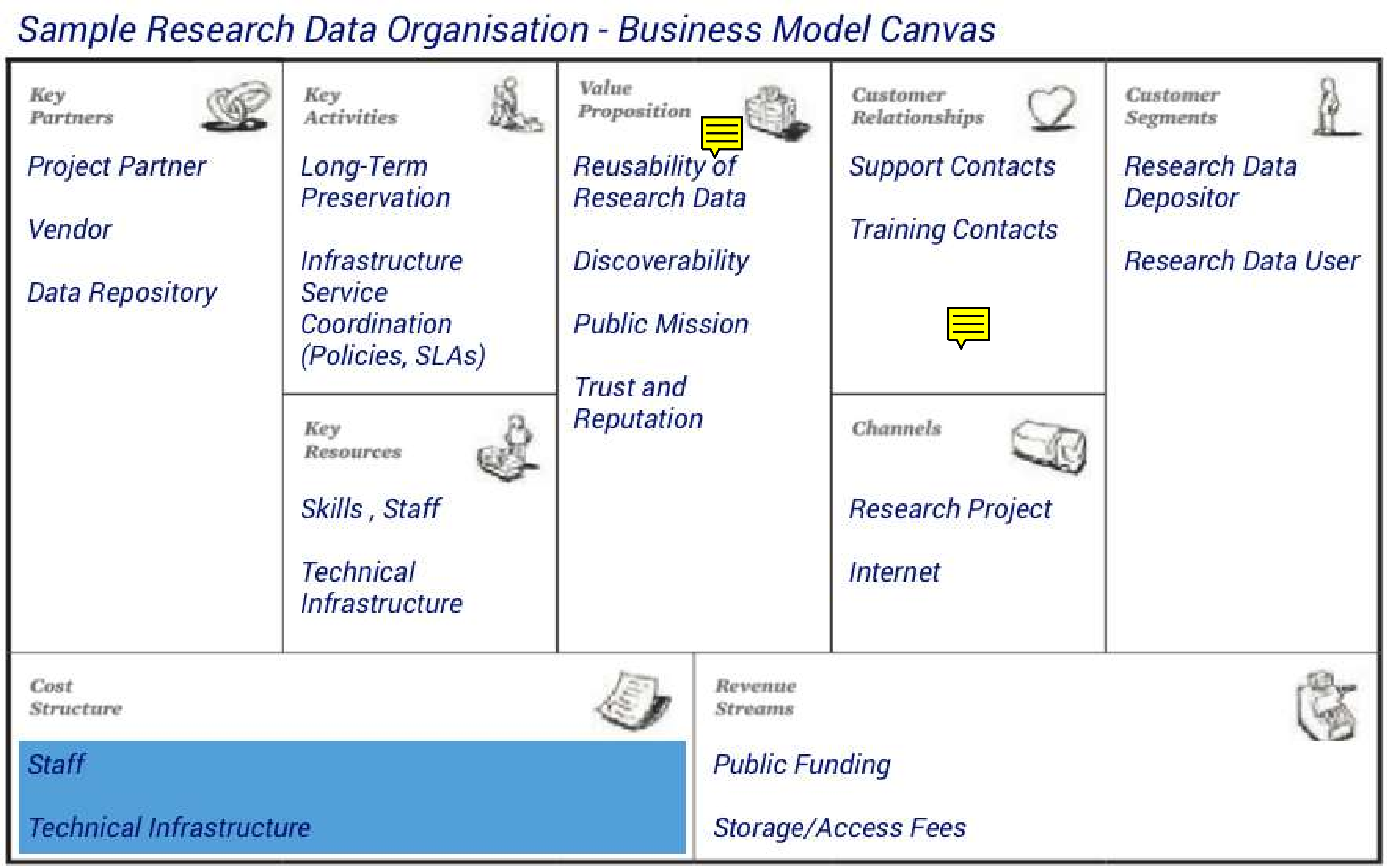
- Draft a policy based on external drivers and local context to establish your core RDM principles
- Ratify the policy then undertake advocacy work and pilot studies to aid implementation
- Consult broadly to gain consensus and secure support

BUSINESS PLANS AND SUSTAINABILITY

Who pays for what?



Identify the cost structure



Business plans and sustainability

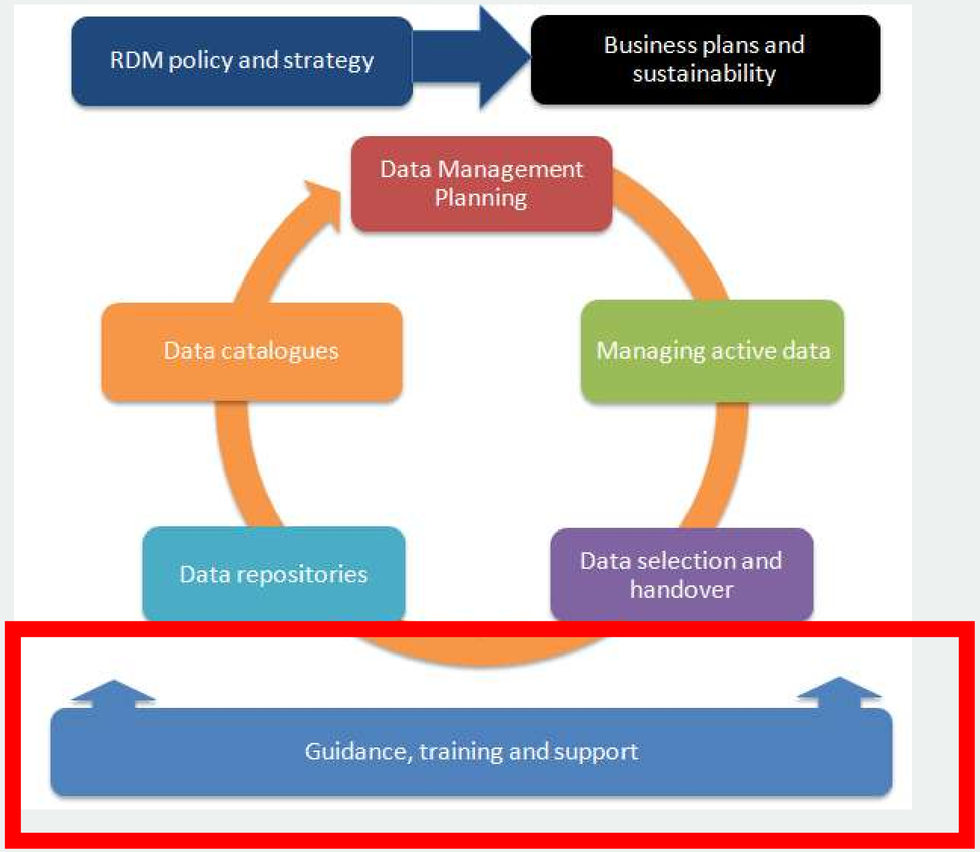
Based on your strategy, develop a phased **business plan** covering projections for 3, 5 and 10 years

Identify predicted **costs** and planned expenditures, together with an indication of the financial year in which expenditure will occur

Consider whether any costs can be recouped, for example by **charging for services**

Undertake a **cost/benefits analysis** to help make the case for investment

Address **sustainability** issues and the associated **long-term costs**



GUIDANCE, TRAINING AND SUPPORT

Guidance, training and support

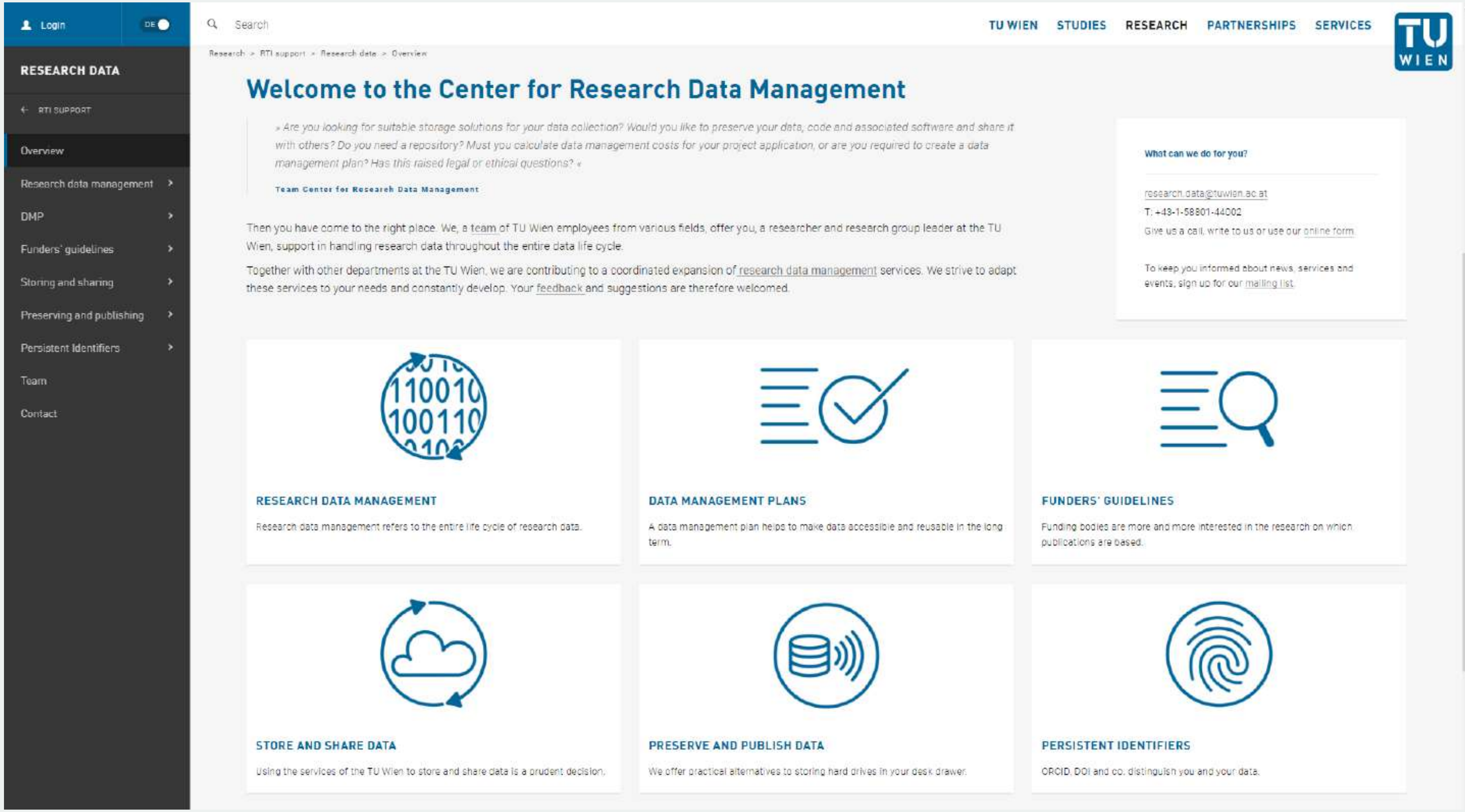
Single Point of Information

- Website
  - Collate details of existing institutional support to provide a single, joined-up RDM guidance website
- Helpdesk
  - Coordinate the provision of support, either through a central helpdesk or well-signposted contacts

Trainings and Consultations

- Provide RDM training to support researchers
and reskill support staff
- Offer more in-depth consultancy services

Singe point of information - example



https://www.tuwien.at/en/research/rti-support/research-data/overview/

Guidance, training and support

Who can do this?

What is the profile of the best candidate?

‘We need 500.000 respected data stewards to operate the European Open Science Cloud’
http://e-irg.eu/news-blog/-/blogs/we-need-500-000-respected-data-stewards-to-operate-the-european-open-science-cloud

Data stewards

Data Stewardship

TU Delft in NL as an example

- Subject-specific Data Steward at every Faculty
- Half-time positions
- PhD holders in a given area proffered
- Strategic coordination from the Library

What do data stewards do?

Analyse data management needs

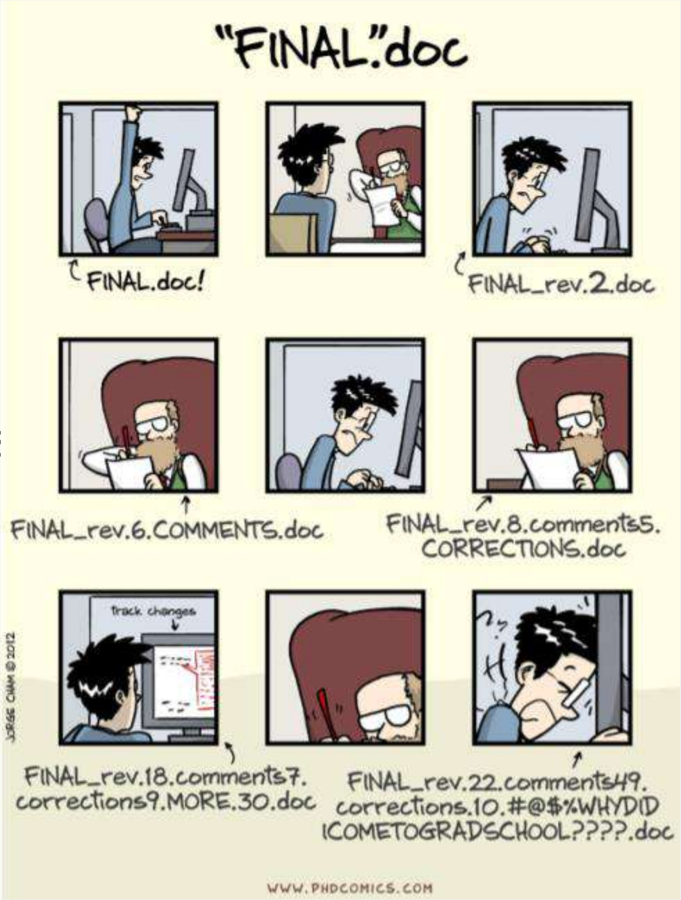
Provide advice to researchers

Train and inspire

Help comply with funders’ and journals’ policies

Develop faculty research data policies: roles and responsibilities

Trusted point of contact for data management questions



FEBRUARY 23, 2018

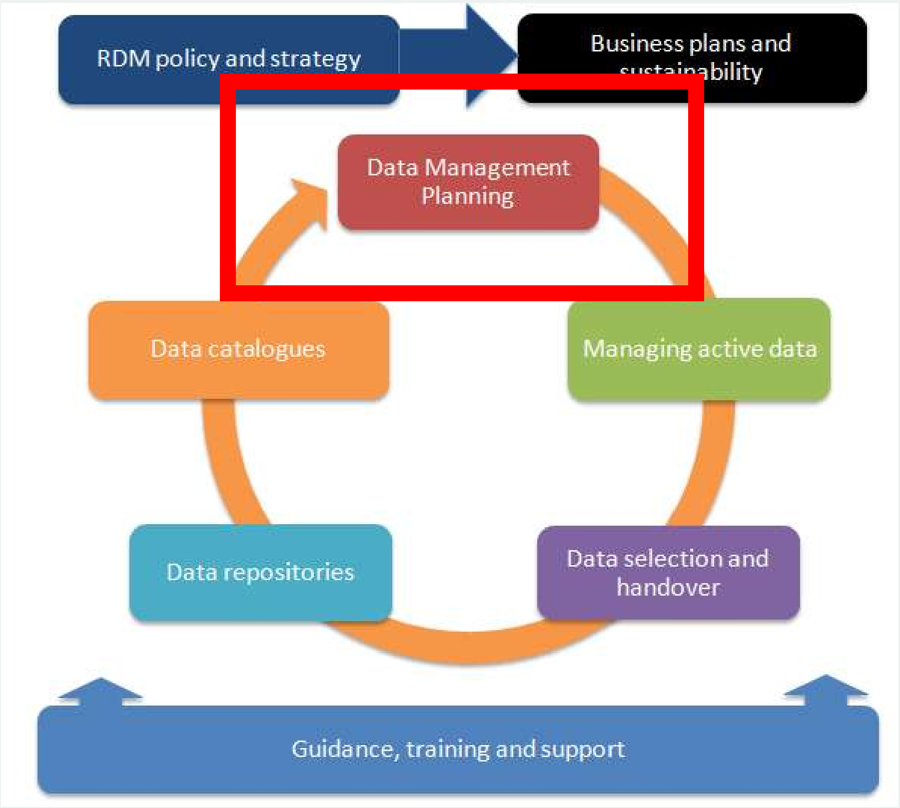
We are hiring (again!) – Data Steward position at TU Delft

We have an exciting job opening for a Data Steward at TU Delft at the Faculty of Architecture & Built Environment and the Faculty of Industrial Design (joint appointment): https://www.academictransfer.com/employer/TUD/vacancy/45483/lang/en/

- Closing date: 15 March 2018
- Salary: up to € 4084/month
- We are looking for individuals enthusiastic about data management and who have a PhD degree in the relevant subject area (or equivalent experience).

This is a great chance to join the dynamically growing team of Data Stewards at TU Delft and to contribute to a cultural change in research data management in a disciplinary manner. The job is really about inspiring the research community and improving day to day practices, and not about policy compliance.

All informal inquiries can be directed to me: M.Teperek@tudelft.nl



DATA MANAGEMENT PLANNING

Data management planning

Data Management Plan (DMP) describes:

- how the data will be created
- how it will be documented
- who will be able to access it
- where it will be stored
- who will back it up
- whether (and how) it will be shared & preserved

Do you really need your own DMP template?

1. Does your institution encourage/require that researchers write Data Management Plans, even if their funder does not require one? If so, what information are researchers asked to provide?

2. Do DMPs submitted to research funders supply the information your institution asks for (if any), or are there additional questions that you want to ask?

3. What guidance, examples and suggested answers can you provide to help researchers write DMPs?

[http://www.dcc.ac.uk/sites/default/files/documents/tools/dmpOnline/DMPonline-customisation-guidelines.pdf]

If you need one…

Identify who and why will need this information

Browse existing checklists for help

Provide a template

- questions
- guidance
- possible answers

Identify overlaps of your template (e.g. with Horizon 2020)

Test it and incorporate feedback

It is not only about DMPs…

Design the whole data management process

- do you need different DMP phases?
- who receives the DMPs?
- who reviews them?
- who provides assistance?
- etc.

Provide a reviewers' guide

- clear judgment criteria

Offer default data location

- if you require people to deposit data and describe in a DMP, provide a place for doing that

It is not only about DMPs…

Establish a policy

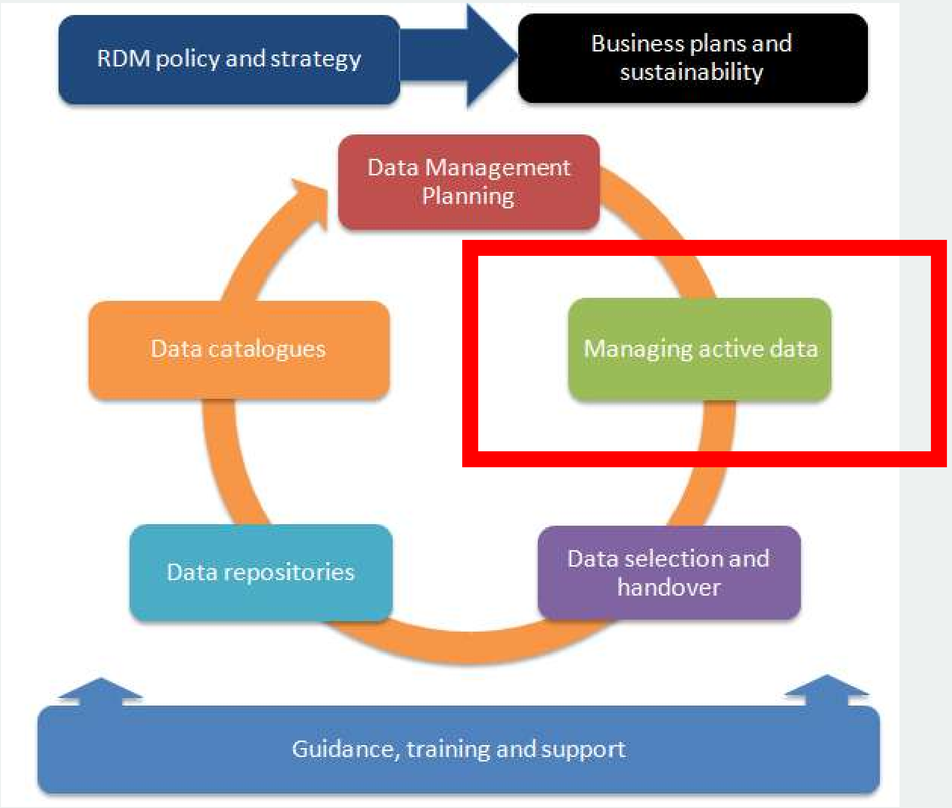
- make a use of DMPs obligatory
- consider incentives and penalties

Provide support

- tools
- examples
- trained staff for assistance in DMP creation
  - related problem: Where do you find people with expertise?

Awareness and understanding

- offer training material
- organise workshops
- related problem: How to teach correct data management practices at schools…



MANAGING ACTIVE DATA

Managing active data

Host on your own

Outsource?

- Sensitive data?
- Jurisdiction?

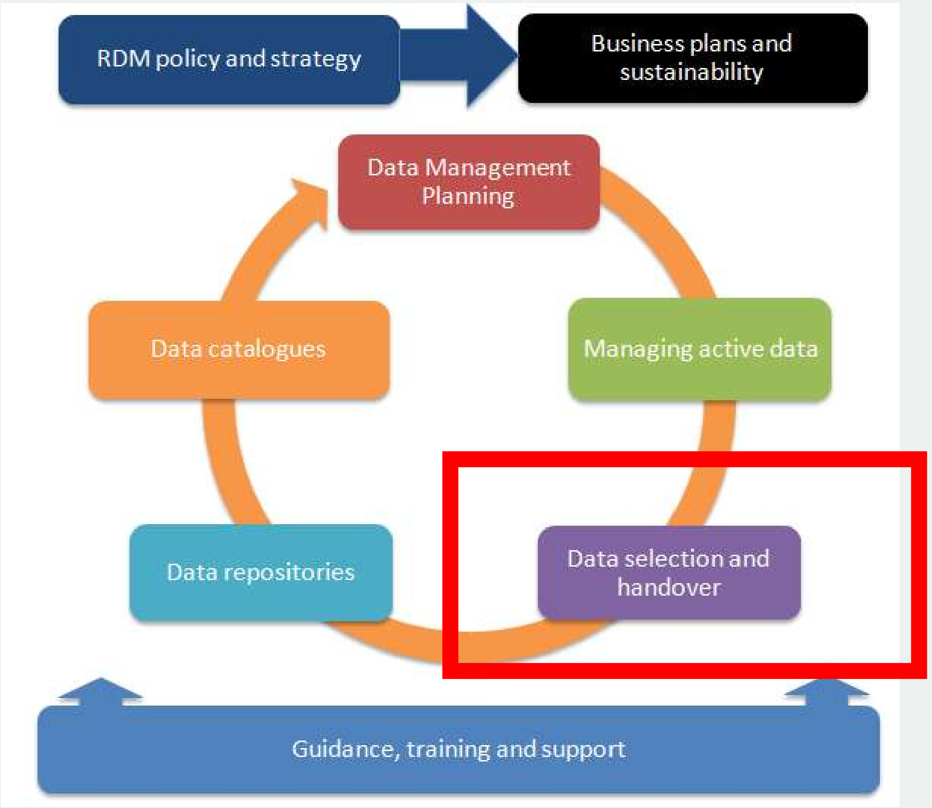
Managing active data

Review data holdings and RDM practices to see if the current infrastructure and systems are sufficient

Where appropriate, make a case for investment to provide additional research data storage

Develop procedures for the allocation and management of research data storage

Provide flexible RDM systems to support the creation, management and sharing of data that meet a diverse range of research contexts and needs



DATA SELECTION AND HANDOVER

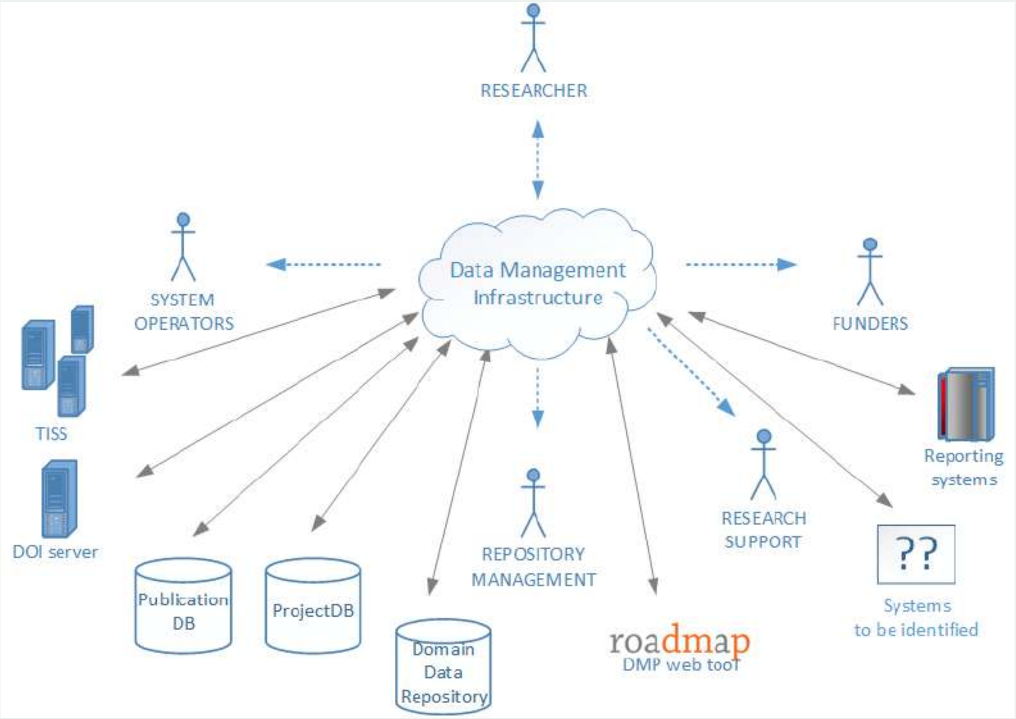
Data selection and handover

What needs to be kept?

- Do you encourage/require specific data to be deposited?
  - Into your own repository?

Data Management Infrastructure

- Integrate existing systems
- Seamless data flow



Data selection and handover

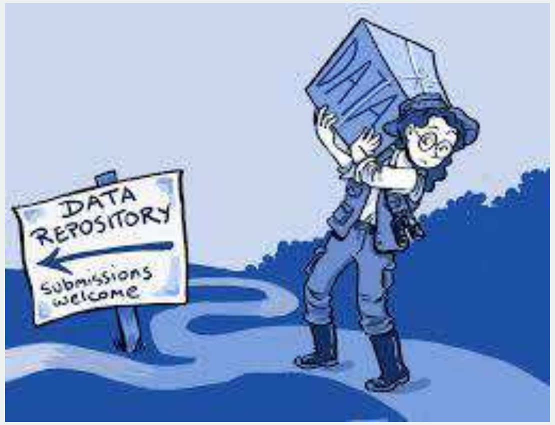
Identify which data fall under your remit and establish criteria to guide decisions on what to keep

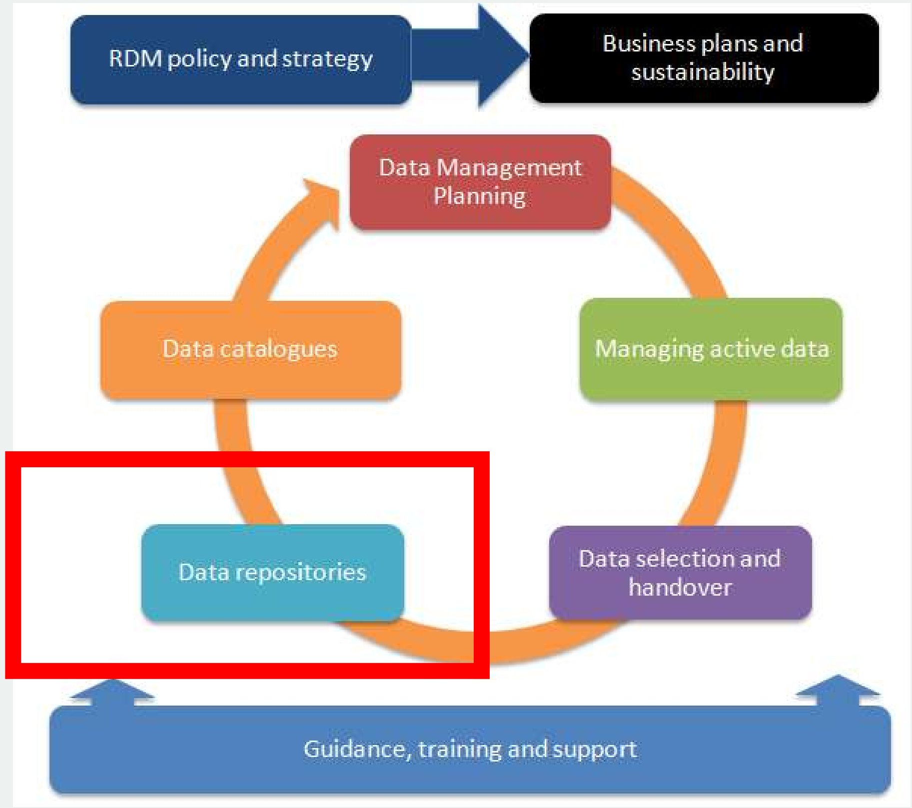
If establishing a repository, develop deposit agreements and high-level guidance

Develop or use existing deposit tools to ease the process of handover

Advocate the benefits of data deposit
to encourage uptake

Support research groups to develop
guidance and offer input to decisions





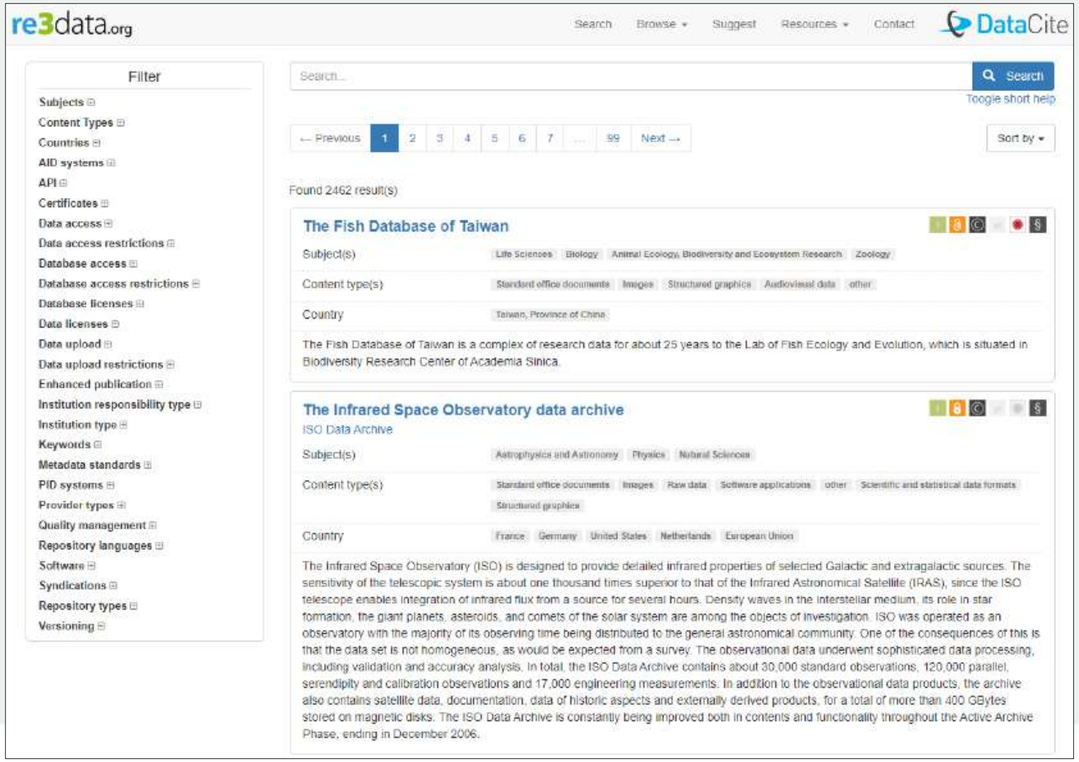
DATA REPOSITORIES

Data repositories

Which repository to use?

What are all these filters for?

Researcher



Data repositories

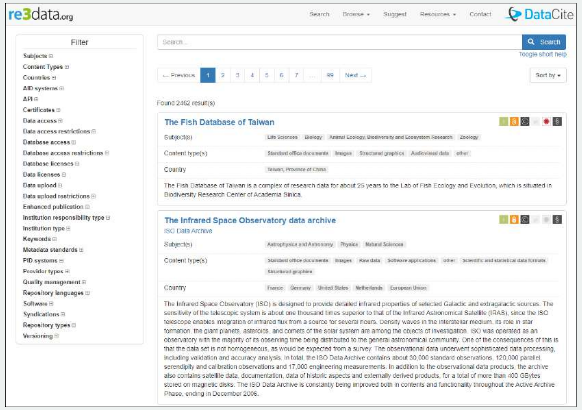
Research support

- Often has too little information on domain, data and needs

Not every researcher can get support

Researcher

Research Support



Data repositories: goal

Automate repository recommendation

- Set relevant filters automatically

Reduce effort

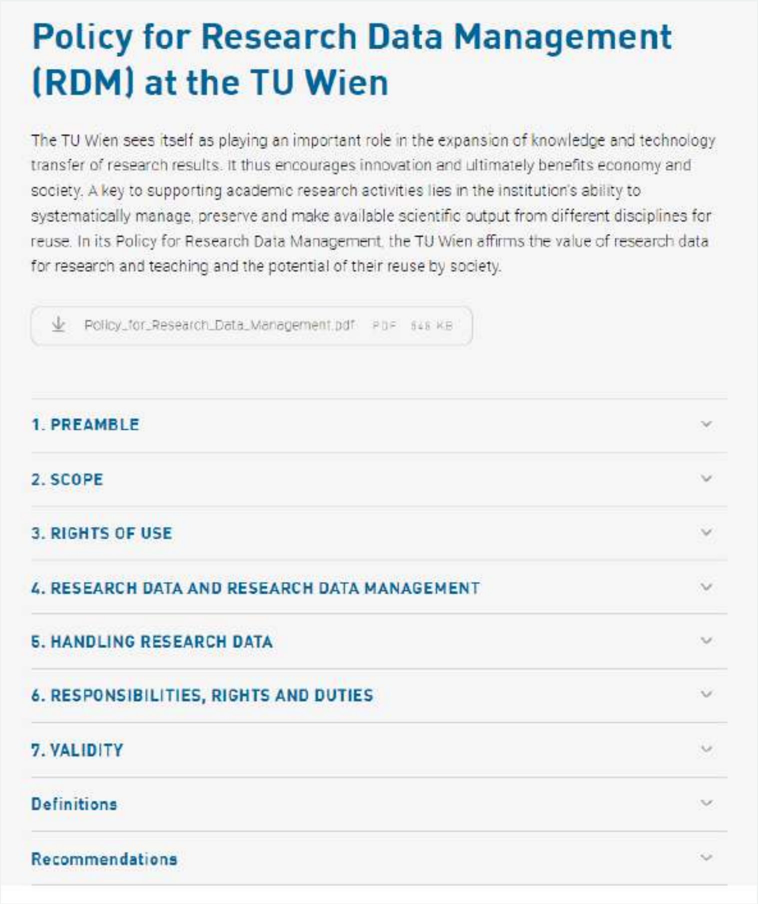
- Narrow down selection to relevant repositories

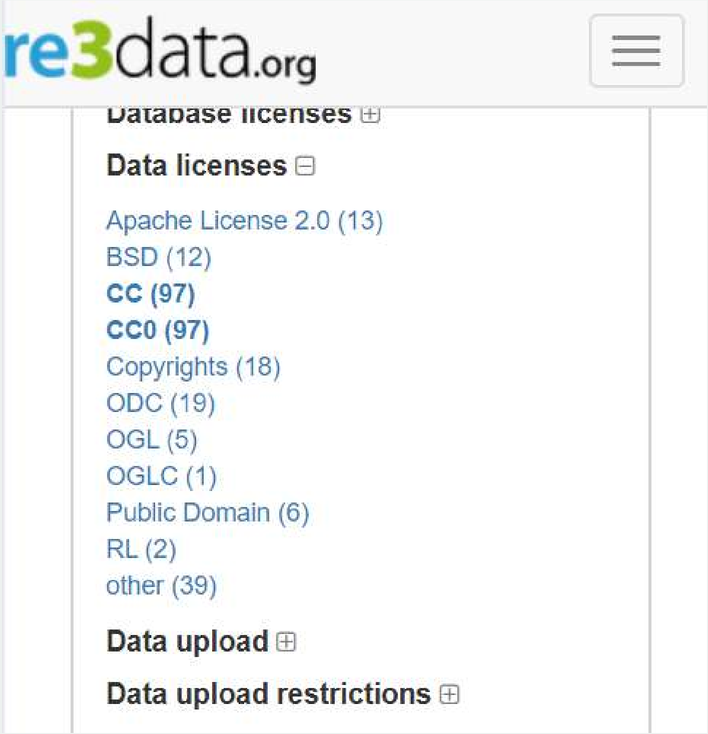
Lower the expertise needed

- Why should researchers deal with licenses if the funder/institutional policy regulates this anyway?

Automation example – RDM policy

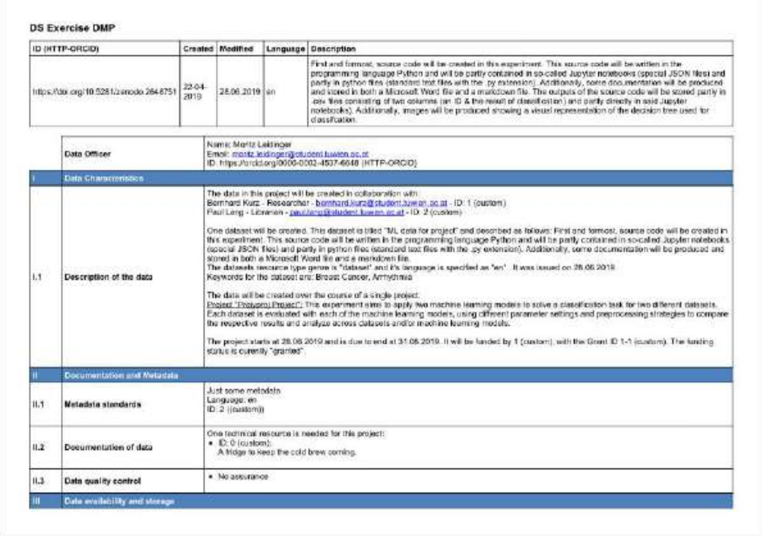
'… research data should be assigned an open use license'





Automation example - DMP

DMP: *‘Data is sensitive and ethical issues exist’*

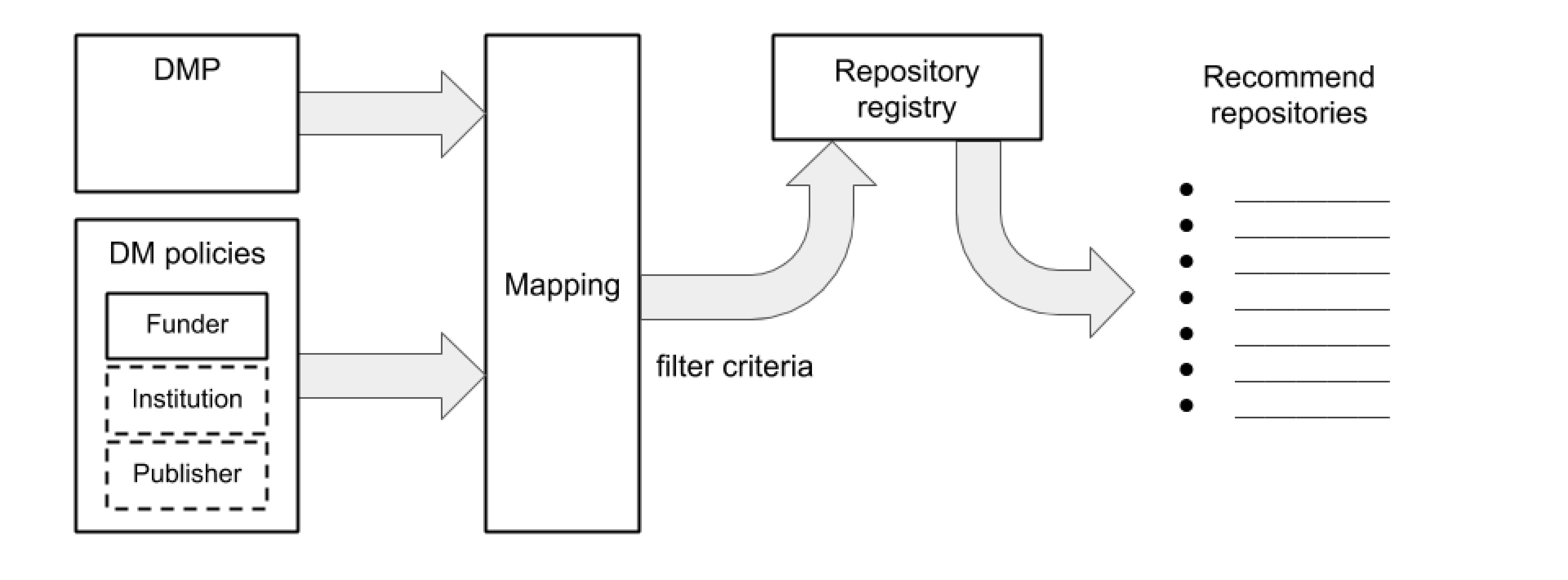


DS Exercise DMP

ID (HTTP-ORCID)	Created	Modified	Language	Description
https://doi.org/10.5281/zenodo.2648751	22-04-2019	28.06.2019	en	First and formost, source code will be created in this experiment. This source code will be written in the programming language Python and will be partly contained in so-called Jupyter notebooks (special JSON files) and partly in python files (standard text files with the .py extension). Additionally, some documentation will be produced and stored in both a Microsoft Word file and a markdown file. The outputs of the source code will be stored partly in .csv files consisting of two columns (an ID & the result of classification) and partly directly in said Jupyter notebooks). Additionally, images will be produced showing a visual representation of the decision tree used for classification.

	Data Officer	Name: Moritz Leidinger Email: moritz.leidinger@student.tuwien.ac.at ID: https://orcid.org/0000-0002-4537-6648 (HTTP-ORCID)
I	**Data Characteristics**	
I.1	**Description of the data**	The data in this project will be created in collaboration with: Bernhard Kurz - Researcher - bernhard.kurz@student.tuwien.ac.at - ID: 1 (custom) Paul Lang - Librarian - paul.lang@student.tuwien.ac.at - ID: 2 (custom) One dataset will be created. This dataset is titled "ML data for project" and described as follows: First and formost, source code will be created in this experiment. This source code will be written in the programming language Python and will be partly contained in so-called Jupyter notebooks (special JSON files) and partly in python files (standard text files with the .py extension). Additionally, some documentation will be produced and stored in both a Microsoft Word file and a markdown file. The datasets resource type genre is "dataset" and it's language is specified as "en". It was issued on 28.06.2019 Keywords for the dataset are: Breast Cancer, Arrhythmia The data will be created over the course of a single project: Project "Project": This experiment aims to apply two machine learning models to solve a classification task for two different datasets. Each dataset is evaluated with each of the machine learning models, using different parameter settings and preprocessing strategies to compare the respective results and analyze across datasets and/or machine learning models. The project starts at 28.06.2019 and is due to end at 31.08.2019. It will be funded by 1 (custom), with the Grant ID 1-1 (custom). The funding status is currently "granted".
II	**Documentation and Metadata**	
II.1	**Metadata standards**	Just some metadata Language: en ID: 2 ((custom))
II.2	**Documentation of data**	One technical resource is needed for this project: • ID: 0 (custom): A fridge to keep the cold brew coming.
II.3	**Data quality control**	• No assurance
III	**Data availability and storage**	

Possible solution



Simon Oblasser, Tomasz Miksa, & Asanobu Kitamoto. (2020, March). Finding a repository with the help of machine-actionable DMPs: opportunities and challenges. Zenodo. http://doi.org/10.5281/zenodo.3701564

DATA CATALOGUES

Data catalogues

Topic covered in the lecture on Repositories – external visibility

Define the metadata you need to record research datasets

Expose metadata for inclusion in national catalogues or other relevant services

Summary

It is not only about technical solutions!

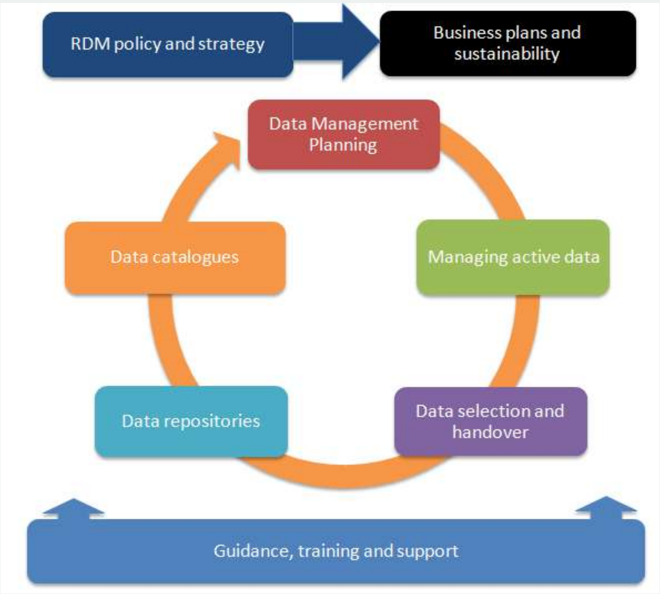
Integration is crucial

- systems
- stakeholders groups
- services

Develop a vision and a plan

Include all stakeholders

Make incremental development



REPOSITORY CERTIFICATION

Motivation

Trust that data remains useful and meaningful into the future

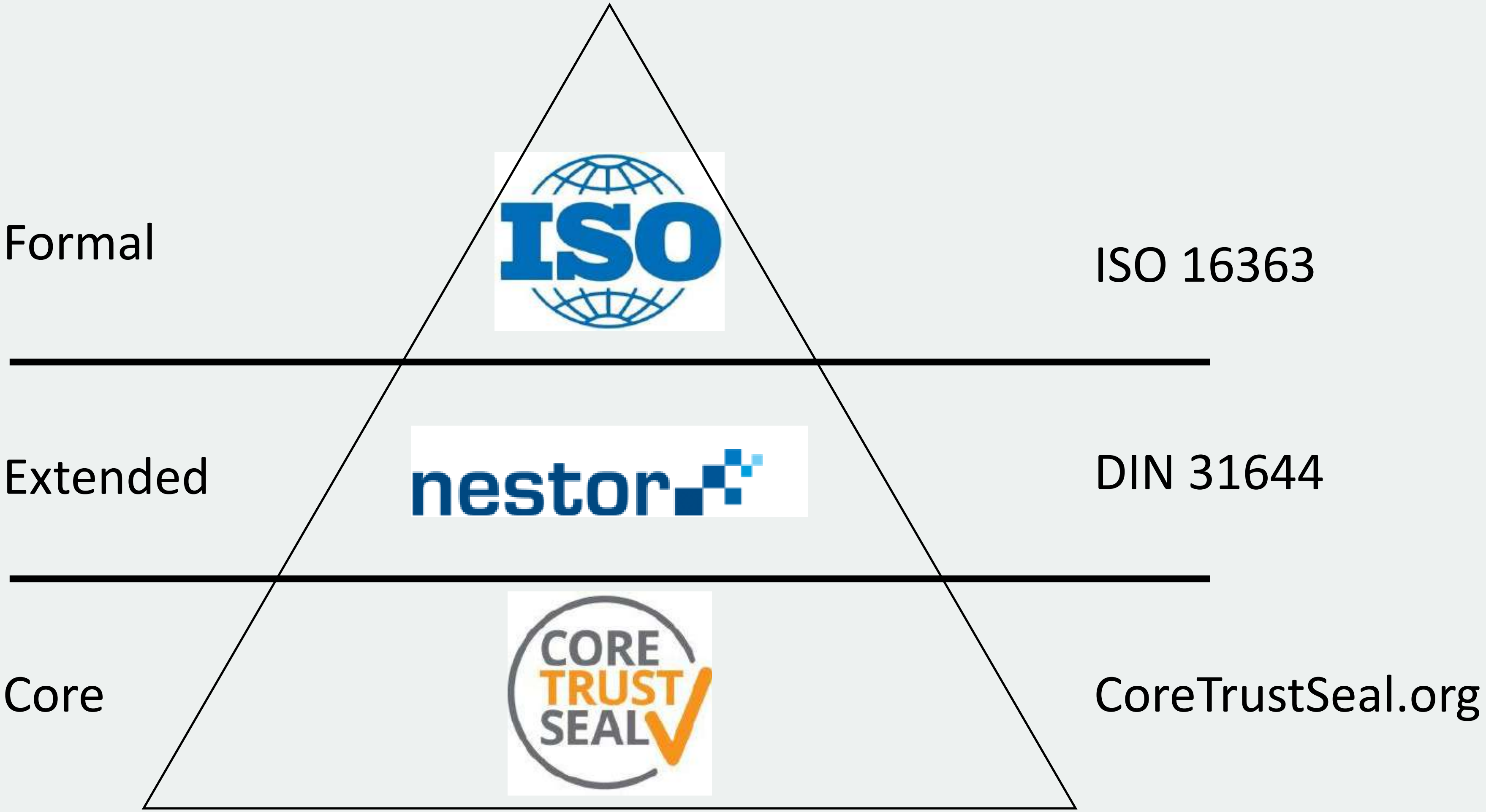
- Will the repository exist in the future?
- How is data stored?
- Is a preservation plan in place?
- etc.

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

Are all of them trusted repositories?

CERN

Certification Standards



Core Trust Seal (CTS) - procedure

Self assessment based on a checklist

Guidance

- online tools
- documents and webinars

Review of the self assessment by two reviewers

Assessments publically available

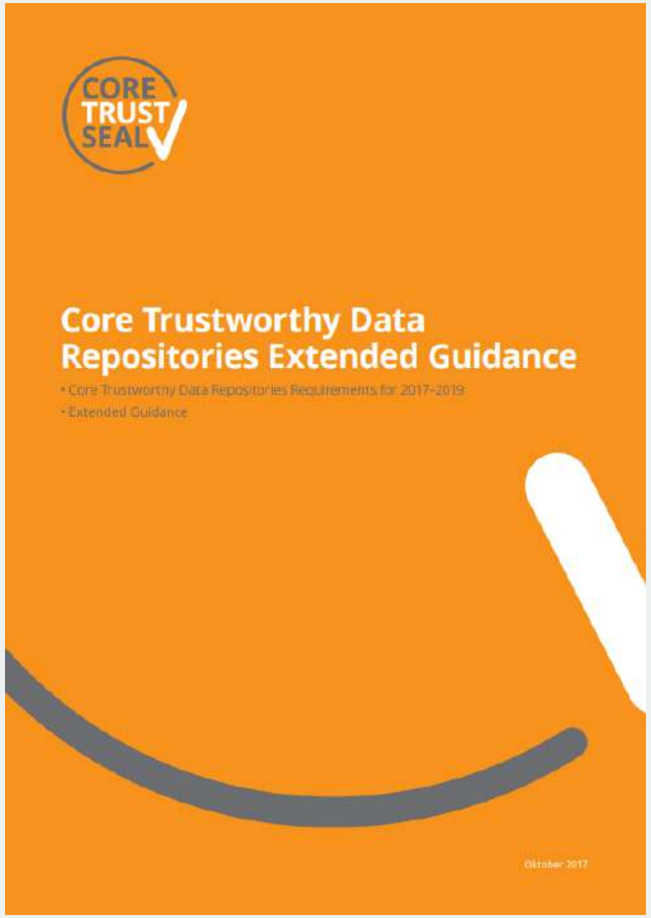
Renewal every three years

CTS Requirements

Organisational Infrastructure (6)

Digital Object Management (8)

Technology (2)



https://www.coretrustseal.org/wp-content/uploads/2017/01/20171026-CTS-Extended-Guidance-v1.0.pdf

Organisational Infrastructure (6)

R1. The repository has **an explicit mission** to provide access to and preserve data in its domain.

R3. The repository has a **continuity plan** to ensure ongoing access to and preservation of its holdings.

R5. The repository has **adequate funding** and sufficient numbers of **qualified staff** managed through a clear system of governance to effectively carry out the mission.

Digital Object Management (8)

R7. The repository guarantees **the integrity and authenticity** of the data.

R10. The repository assumes responsibility for **long-term** preservation and manages this function in a planned and documented way.

R12. Archiving takes place according to **defined workflows** from ingest to dissemination.

R13. The repository enables users to **discover the data** and **refer to them in a persistent way** through proper citation.

Technical (2)

R15. The repository functions on **well-supported operating system** and other core infrastructure software is using hardware and software technologies appropriate to service it provides to its Designated Community.

R16. The technical infrastructure of the repository provides for **protection** of the facility and its data, products, services, and users.

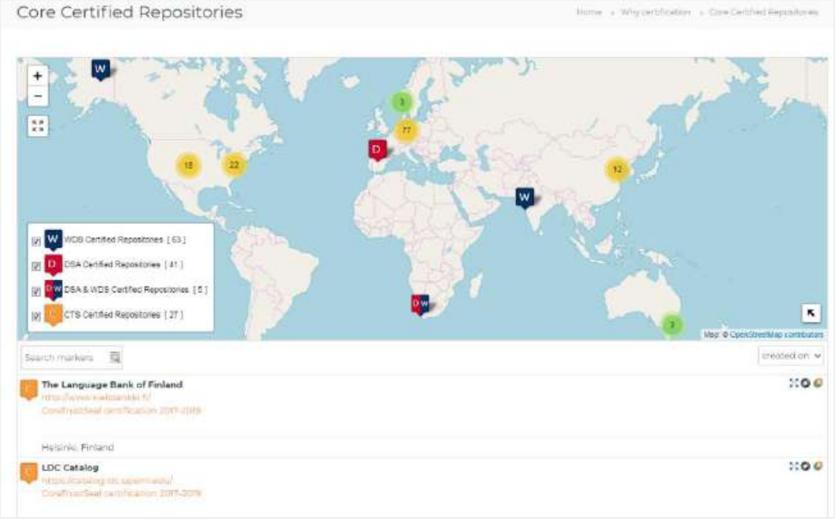
Self-assessment examples

Austrian example

- A Resource Centre for Humanities Related Research in Austria
  - https://arche.acdh.oeaw.ac.at
- Self assessment
  - https://www.coretrustseal.org/wp-content/uploads/2018/03/ARCHE.pdf

More examples

- https://www.coretrustseal.org/why-certification/certified-repositories/



X. Preservation plan

R10. The repository assumes responsibility for long-term preservation and manages this function in a planned and documented way.

Compliance Level:

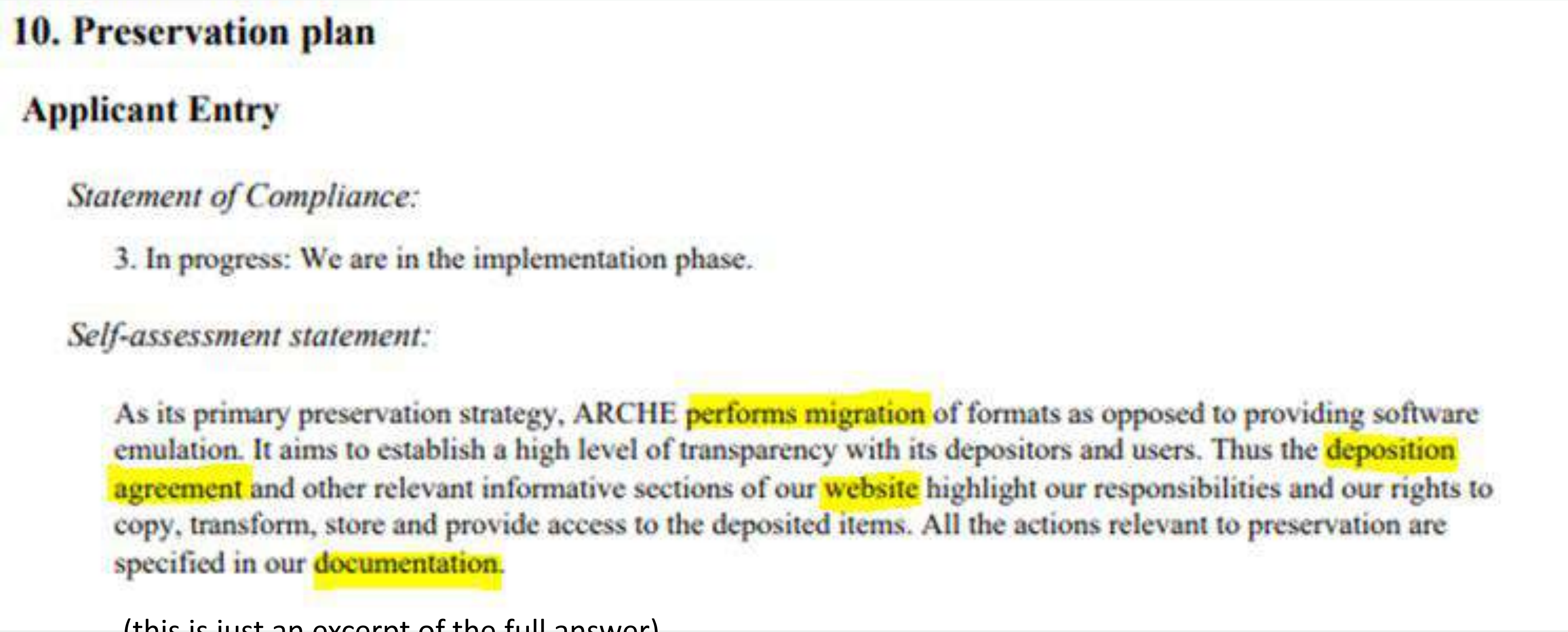
Response

Guidance:
The repository, data depositors, and Designated Community need to understand the level of responsibility undertaken for each deposited item in the repository. The repository must have the legal rights to undertake these responsibilities. Procedures must be documented and their completion assured.

For this Requirement, responses should include evidence related to the following questions:

- Is a preservation plan in place?
- Is the 'preservation level' for each item understood? How is this defined?
- Does the contract between depositor and repository provide for all actions necessary to meet the responsibilities?
- Is the transfer of custody and responsibility handover clear to the depositor and repository?
- Does the repository have the rights to copy, transform, and store the items, as well as provide access to them?
- Are actions relevant to preservation specified in documentation, including custody transfer, submission information standards, and archival information standards?
- Are there measures to ensure these actions are taken?

Self-assessment example



10. Preservation plan

Applicant Entry

Statement of Compliance:

3. In progress: We are in the implementation phase.

Self-assessment statement:

As its primary preservation strategy, ARCHE performs migration of formats as opposed to providing software emulation. It aims to establish a high level of transparency with its depositors and users. Thus the deposition agreement and other relevant informative sections of our website highlight our responsibilities and our rights to copy, transform, store and provide access to the deposited items. All the actions relevant to preservation are specified in our documentation.

(this is just an excerpt of the full answer)

https://www.coretrustseal.org/wp-content/uploads/2018/03/ARCHE.pdf

Advice from Uni Graz

one team lead

collaborative document to work jointly

put procedures and workflows in writing

enhance publicly available information

use the glossary and extended guidance

- ‘we’ – ‘the repository’ – ‘the system’

read successful applications on website

allow sufficient time for the process of internal coordination and data gathering

technical procedures do not pose problems/attract inquiries from reviewers, it is organizational issues that do

Based on presentation by Elisabeth Steiner at the Certification Workshop on fair-aligned repositories in Austria, 14.11.2019, Vienna

NESTOR

DIN 31644

- Deutsches Institut für Normung
- ‘Criteria for trustworthy digital archives’

Derives from Trustworthy Repositories Audit & Certification (TRAC)

- document describing the metrics of an OAIS-compliant digital repository
- TRAC is discontinued, replace by the ISO 16363
- translated to German

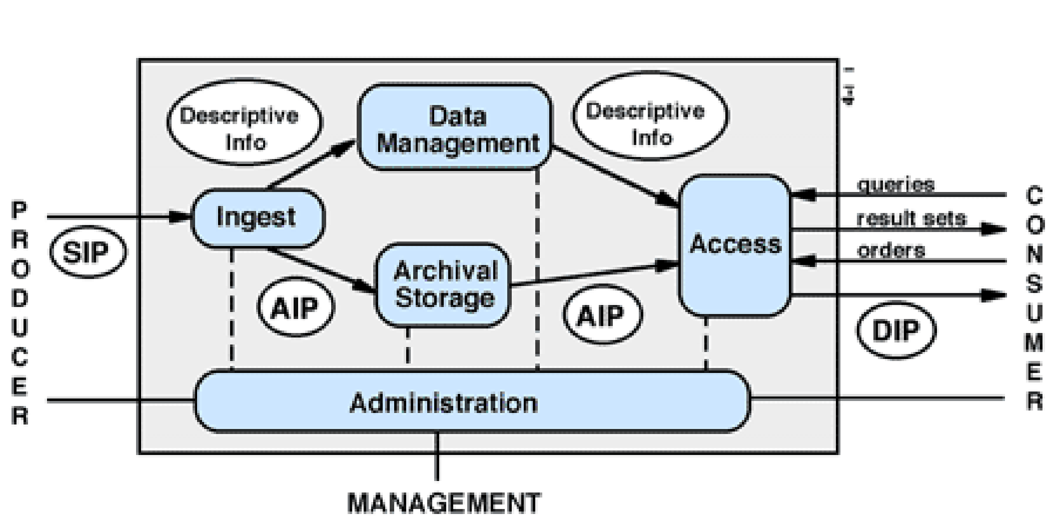
Uses OAIS terminology

Open Archival Information System (OAIS)

Discussed in a separate lecture

OAIS

- conceptual framework for an archival system dedicated to preserving and maintaining access to digital information over the long term
- defines concepts, terminology
- not to be instantiated!



NESTOR requirements overview

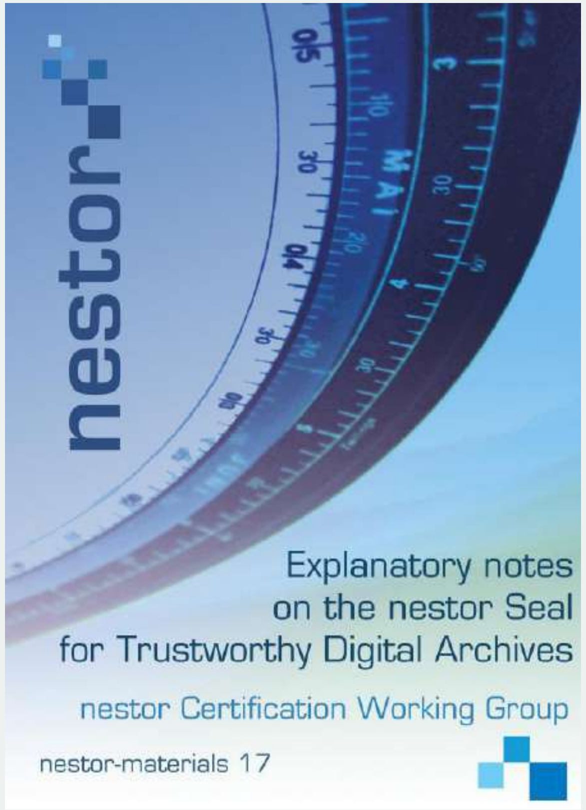
34 criteria

- 12 obligatory
- other can be excluded when justified

auxiliary questions

self-assessment

2 external reviewers



https://d-nb.info/1047613859/34

C27	Identification

A digital archive should use internal identifiers to manage the information objects and their representations and, where applicable, their parts and relationships (part/totality, different variants, versions etc.), especially to ensure unique assignment of the content data to the metadata.

The use of externally visible, standardised persistent identifiers ensures reliable tracing of the information objects and their representations, and consequently also access.

To what extent must the criterion be met?

An average of 7 points must be achieved in the assessment of the applicable criteria C13 - C34.

Explanation: The information objects, representations and their parts are permanently linked to each other. These links can only be preserved through the use of persistent identifiers. The identifiers should not change over the course of time (i.e. be permanent) and should be created using uniform specifications. They should be recognisable to external users, producers and others. By entering the identifier, external users should be able to find and use the required object. Possible specific requirements for identifiers are described e.g. in DIN 13646 "Requirements for the long-term handling of persistent identifiers".

Questions

- Which identifiers does the digital archive use?
- Which procedure has been used to give unique identifiers to all information objects, representations and their parts, and to all content and metadata?
- How is the identifier-based assignment conducted?
- How is the permanence of the identifiers ensured?
- How are the identifiers made available to external users?

Documents: Specification of the internal and external identifiers

ISO 16363

Uses OAIS terminology

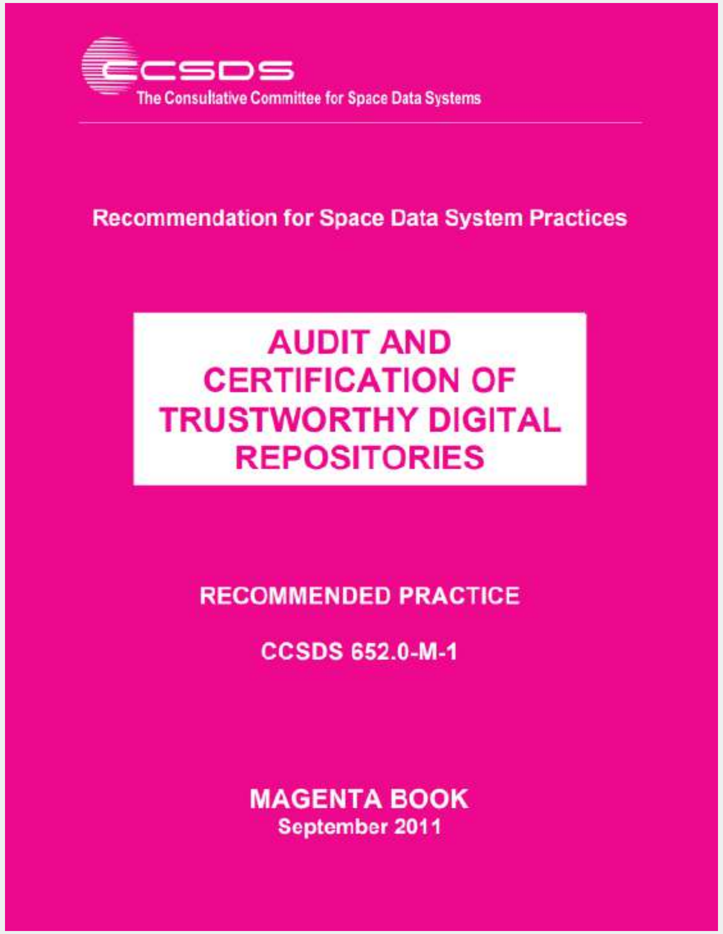
Based on TRAC

Over 100 metrics

Detailed instructions

Full external auditing process

ONE repository
certified since 2012!



CCSDS
The Consultative Committee for Space Data Systems

Recommendation for Space Data System Practices

AUDIT AND CERTIFICATION OF TRUSTWORTHY DIGITAL REPOSITORIES

RECOMMENDED PRACTICE

CCSDS 652.0-M-1

MAGENTA BOOK
September 2011

https://public.ccsds.org/Pubs/652x0m1.pdf

Structure

Organizational infrastructure

- Governance and organizational viability
- Financial sustainability
- …

Digital object management

- Ingest: acquisition of content
- AIP Preservation
- …

Infrastructure and security risk management

- Security risk management
- ….

Metrics

Metrics and their structure:

- Statement of requirement
- Supporting text
- Examples of Ways the Repository can Demonstrate it is Meeting this Requirement
- Discussion

4.3 PRESERVATION PLANNING

4.3.1 The repository shall have documented preservation strategies relevant to its holdings.

Supporting Text

This is necessary in order that it is clear how the repository plans to ensure the information will remain available and usable for future generations and to provide a means to check and validate the preservation work of the repository.

Examples of Ways the Repository Can Demonstrate It Is Meeting This Requirement

Documentation identifying each preservation risk identified and the strategy for dealing with that risk.

Discussion

These documented preservation strategies will describe how the repository will act upon identified risks, as part of the preservation strategic plan. These preservation strategies and the preservation strategic plan will typically address the degradation of storage media, the obsolescence of media drives, and the obsolescence or inadequacy of Representation Information (including formats) as the knowledge base of the Designated Community changes, and safeguards against accidental or intentional digital corruption. For example, if

Metrics – self assessment

4.3 PRESERVATION PLANNING					
	Metric	**Supporting Text**	**Examples of Documents the Repository can use to demonstrate it is Meeting this Requirement:**	**Brief description of evidence (add rows if necessary to list all relevant documents for a metric) Use short titles for documents. Provide detailed**	**Explanation of how the repository addresses this metric**
4.3.1	**THE REPOSITORY SHALL HAVE DOCUMENTED PRESERVATION STRATEGIES RELEVANT TO ITS HOLDINGS.**	This is necessary in order that it is clear how the repository plans to ensure the information will remain available and usable for future generations and to provide a means to check and validate the preservation work of the repository.	Documentation identifying each preservation risk identified and the strategy for dealing with that risk.		
4.3.2	**THE REPOSITORY SHALL HAVE MECHANISMS IN PLACE FOR MONITORING ITS PRESERVATION ENVIRONMENT.**	This is necessary so that the repository can react to changes and thereby ensure that the preserved information remains understandable and usable by the Designated Community.	Surveys of the Designated Community of the repository.		
4.3.2.1	*The repository shall have mechanisms in place for monitoring and notification when Representation Information is inadequate for the Designated Community to understand the data holdings.*	This is necessary in order to ensure that the preserved information remains understandable and usable by the Designated Community.	Subscription to a Representation Information registry service; subscription to a technology watch service, surveys amongst its Designated Community members, relevant working processes to deal with this information.		

http://www.iso16363.org/?smd_process_download=1&download_id=30

ISO Process for Audits

Preparatory work by repository

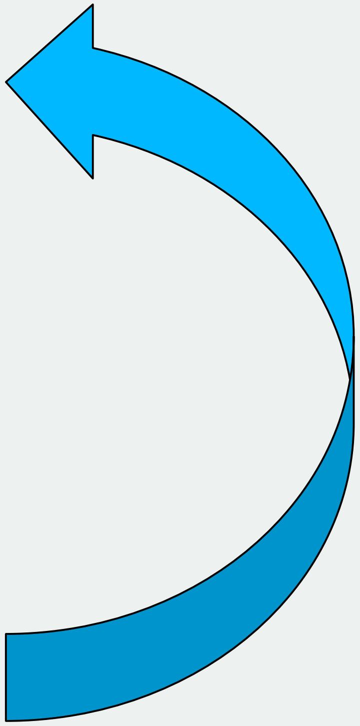
First audit and resulting certification

- Identifies improvements needed
- Repository prepares improvement plan

Repository implements improvement plan

Surveillance audit after a period

Re-certification



Why certify a repository (using any scheme)?

Builds stakeholder confidence in the repository

- researchers', funders', publishers', etc.

Improves communication within the repository

- roles and responsibilities (implicit knowledge becomes explicit)
- processes

Ensures transparency

- most assessments are public

Enables comparison of repositories

Attracts funding

Summary

You should know

What are the lifecycle models and how to use them?

What are the components of an RDM service infrastructure?

What is the scope of policies and how they drive DM activities and obligations?

How do develop support services and who are data stewards?

What to consider when implementing DMPs in an institution?

What standards exist and why certification matters?

What are the certification criteria?

Acknowledgments

Jones, S., Pryor, G. & Whyte, A. (2013). ‘How to Develop Research Data Management Services - a guide for HEIs’. DCC How-to Guides. Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/resources/how-guides

Andreas Rauber. (2017, November). Research Outputs Management: what services are involved? File (Version 1.00). Zenodo. http://doi.org/10.5281/zenodo.1063548LEARN

Teperek, Marta. (2017, November). Data Stewardship - addressing disciplinary data management needs at TU Delft. Zenodo. http://doi.org/10.5281/zenodo.1064794

Budroni, Paolo. (2017, May). The LEARN Project Using the LEARN RDM Policy & Guidance. Zenodo. http://doi.org/10.5281/zenodo.579993

data and legal aspects

Verena Dolovai – Compliance Officer Research

agenda

- types of data
- ownership of data
- inventor, author, creator
- fields of law
- rights in research data
- copyright in research data?
- rights holders in research data
- data protection law & research project
- examples

TECHNISCHE UNIVERSITÄT WIEN

types of data

- instrument measurements
- experimental observations
- still images, video and audio
- text documents, spreadsheets, databases
- quantitative data (e.g. survey data)
- survey results & interview transcripts
- simulation data, models & software
- slides, artefacts
- specimens, samples, questionnaires

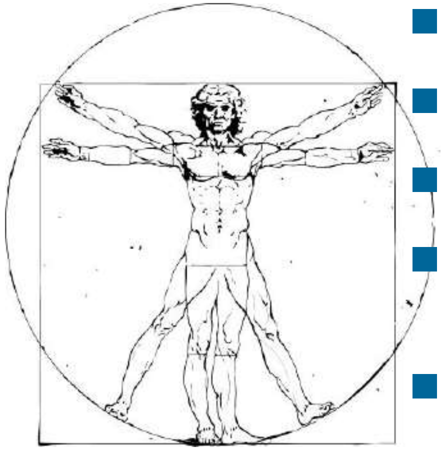


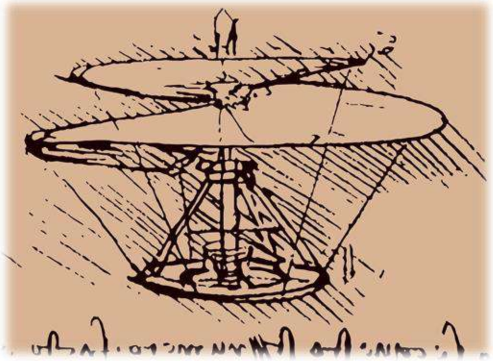

TECHNISCHE UNIVERSITÄT WIEN

- owner
  - a term of civil law
  - not exactly suitable for data ownership
  - however, „owner“ has become familiar in terms of data management

- rather, legally speaking:
  - author (copyright law)
  - creator (if not protected by copyright law)
  - inventor (patent law)
  - legal successor

inventor, author, creator





- **inventor → patent law**
- patent law: protects inventor, idea, technology
- -> exclusive right, registration, fees
- 20 years protection

- **author → copyright law**
- copyright law: protects author, work, intellectual creation (works of literature, science, art, databases, computer programmes)
- -> no registration, no fees
- 70 years protection

- **creator (person who creates data(sets) → licenses**
- licenses: protect creator

fields of law

- copyright law (databanks, images)
- data protection law
- data security law
- competition law (protection of know-how)
- contract law
- labour law (towards employees of an institution)

rights in research data

TECHNISCHE UNIVERSITÄT WIEN

data protection law in research	rights in research data
GDPR*: personal data & informed consent	general legal aspects of research data
*as well as national data protection laws: Austrian Data Protection Act (DSG) and Research Organsiation Act (FOG)	licenses, open source software, IP, storage, erasure, publication, user rights in data, user agreements, confidentiality**, liability etc)
	**data secrecy and protection of confidential information

rights in research data

data*

copyright and neighbouring rights
data protection law

software

copyright

copyright relevance: **qualitative research data**; data in social sciences and humanities

database

right in the database (work)

data medium

ownership

*as a rule, data in a literal sense are hardly subject to copyright; however, term „data“ must be defined in each indiviual case!

copyright in research data?

TECHNISCHE
UNIVERSITÄT
WIEN

essays
articles
papers
monographs
graphics
maps
tables with individual graphic design
databases
research software
photographs
video recordings

possible copyright:
result of individual
creation

case by case assessment!

measurement data
metadata
results of a software simulation

usually no copyright:
no result of individual
creation

rights holders in research data

TECHNISCHE
UNIVERSITÄT
WIEN

researchers, students

generated data subject to copyright

written software

ownership in personal data media

data protection rights

transfer of rights*?

university

ownership in computers & servers

third parties (contractual partners, clients, funders)

maybe joint copyright (joint data collection)

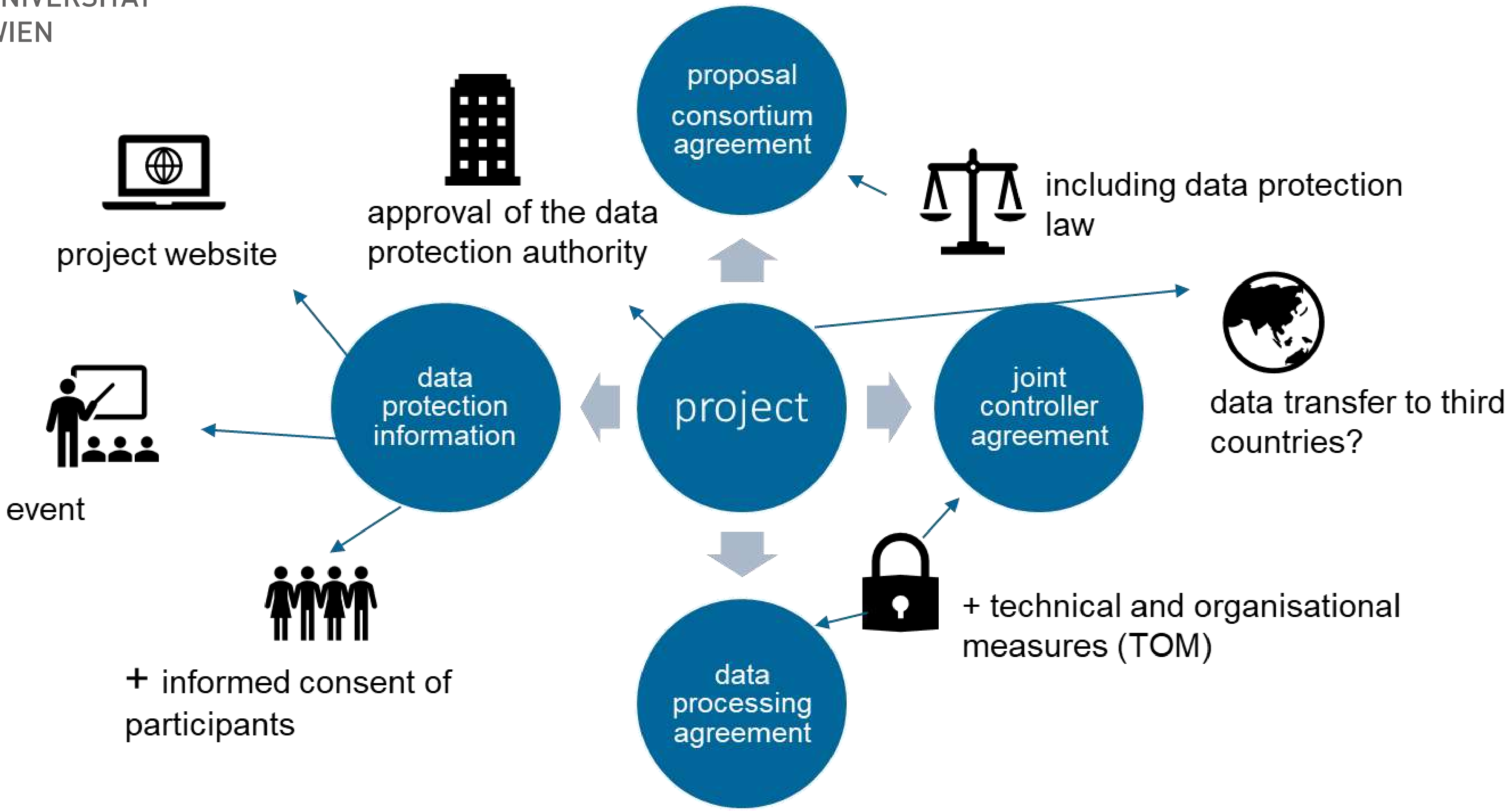
exploitation rights

data protection rights

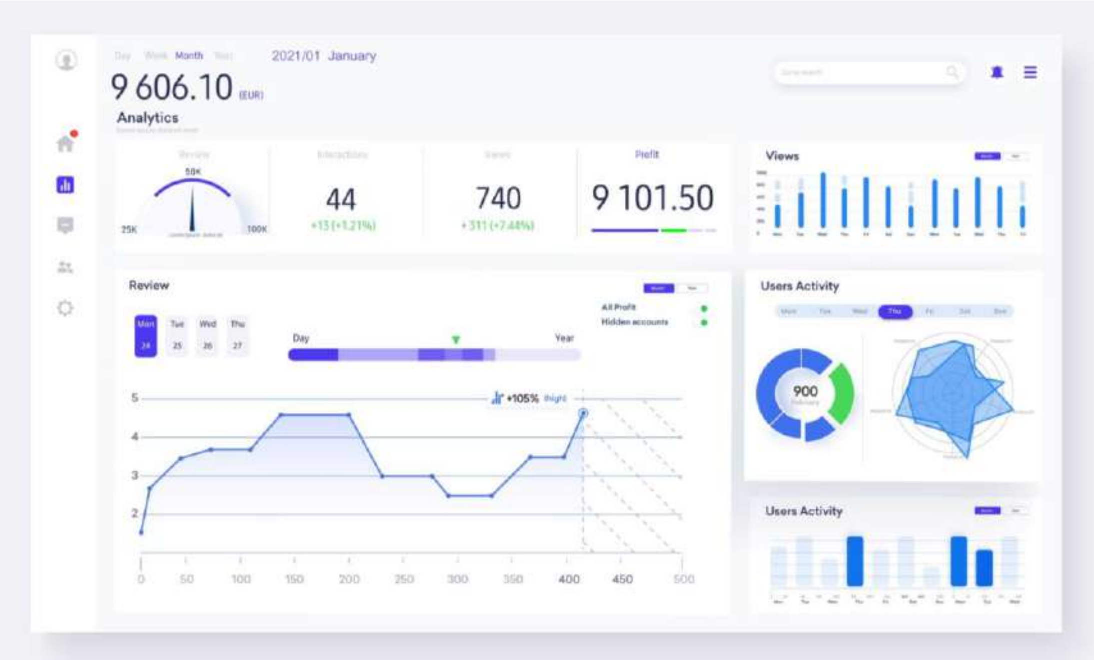
*employment contract, exploitation rights, and user rights

data protection law & research project

TECHNISCHE UNIVERSITÄT WIEN

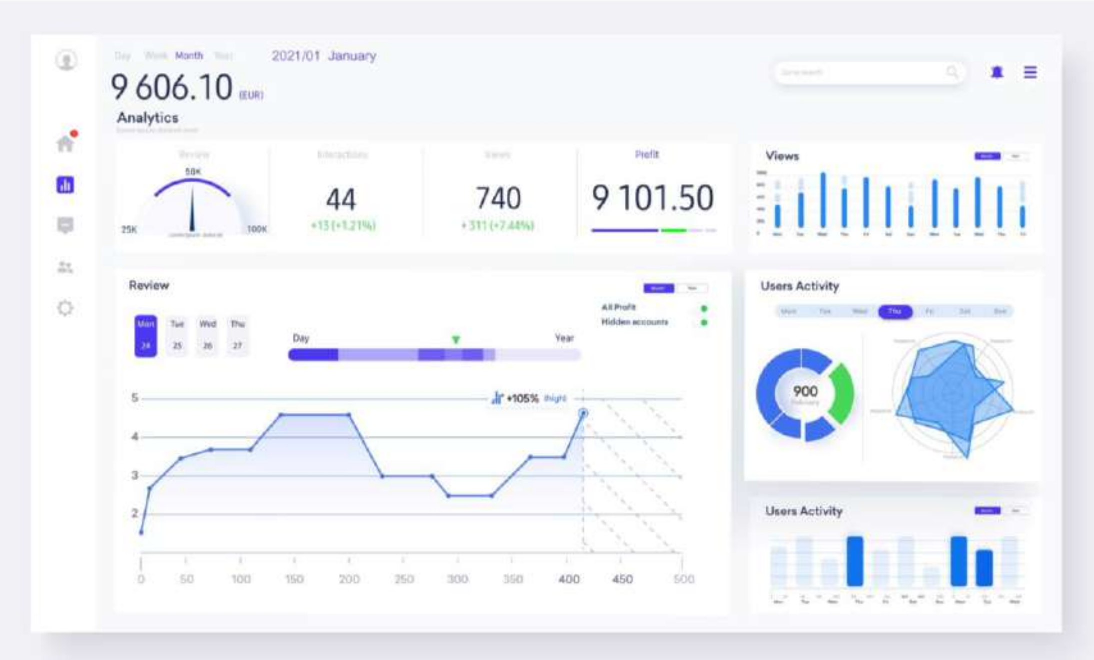


TECHNISCHE UNIVERSITÄT WIEN



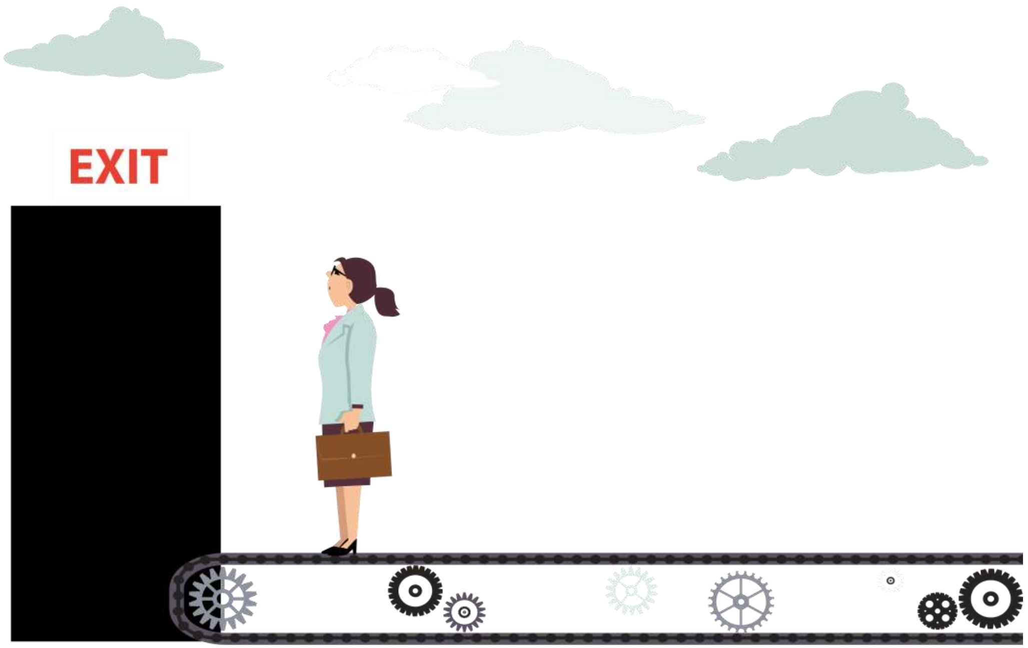
I want to use data from a website

TECHNISCHE UNIVERSITÄT WIEN



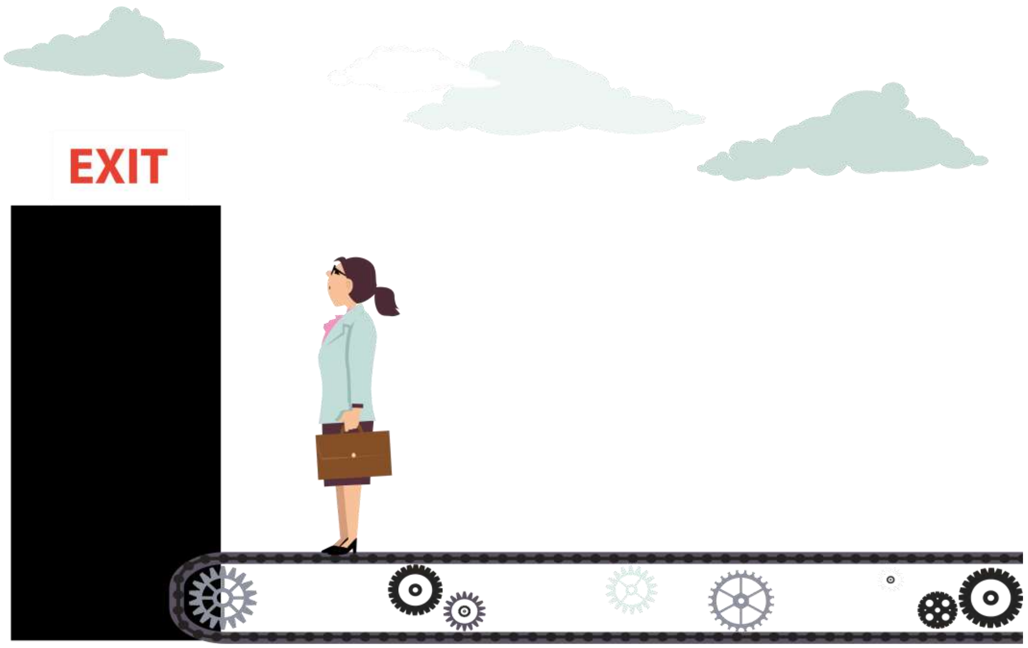
I want to use data from a website

- check for license or other legal requirements / rules
- not available?
- seek legal advice
- include reference



I‘m leaving my organisation
What to do with „my data“?

TECHNISCHE UNIVERSITÄT WIEN



I‘m leaving my organisation

What to do with „my data“?

- check employment contract
- check project contract / funding agreement
- seek legal advice & agreement with employer

TECHNISCHE UNIVERSITÄT WIEN

I am employee of xx and I want to license „my data“. Am I free to choose any license?

TECHNISCHE UNIVERSITÄT WIEN

I am employee of xx and I want to license „my data“. Am I free to choose any license?

- **check for internal guidelines (types of licenses, powers of attorney, regulations)**
- **check employment contract**
- **check project contract / funding agreement**

TECHNISCHE
UNIVERSITÄT
WIEN

Someone used my data illegally or without my permission… what should I do?

examples

Someone used my data illegally or without my permission… what should I do?

- seek legal advice

contact

Verena Dolovai
Compliance Officer Research
Technische Universität Wien
Favoritenstraße 16/DG
1040 Wien
Telefon: +43 1 58801 406635
verena.dolovai@tuwien.ac.at