FAIRSFAIR

Fostering Fair Data Practices in Europe

Clearing some of the highest FAIR hurdles: PIDs, Metadata, and Semantic Interoperability for Researchers

FAIRsFAIR "Fostering FAIR Data Practices in Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 Grant agreement 831558

Program

- Introduction - Leah Riungu-Kalliosaari
- Using persistent identifiers - Jessica Parland-von Essen
- Semantic interoperability and Metadata - Rob Hooft
- Q&A

The FAIRsFAIR project in a nutshell

Call: H2020-INFRAEOSC-5c

Budget: 10 million euro

Length: 36 months
Starting date: March 1 2019

22 partners from 8 MS
6 core partners

Data Archiving and Networked Services
DANS

eua EUROPEAN UNIVERSITY ASSOCIATION

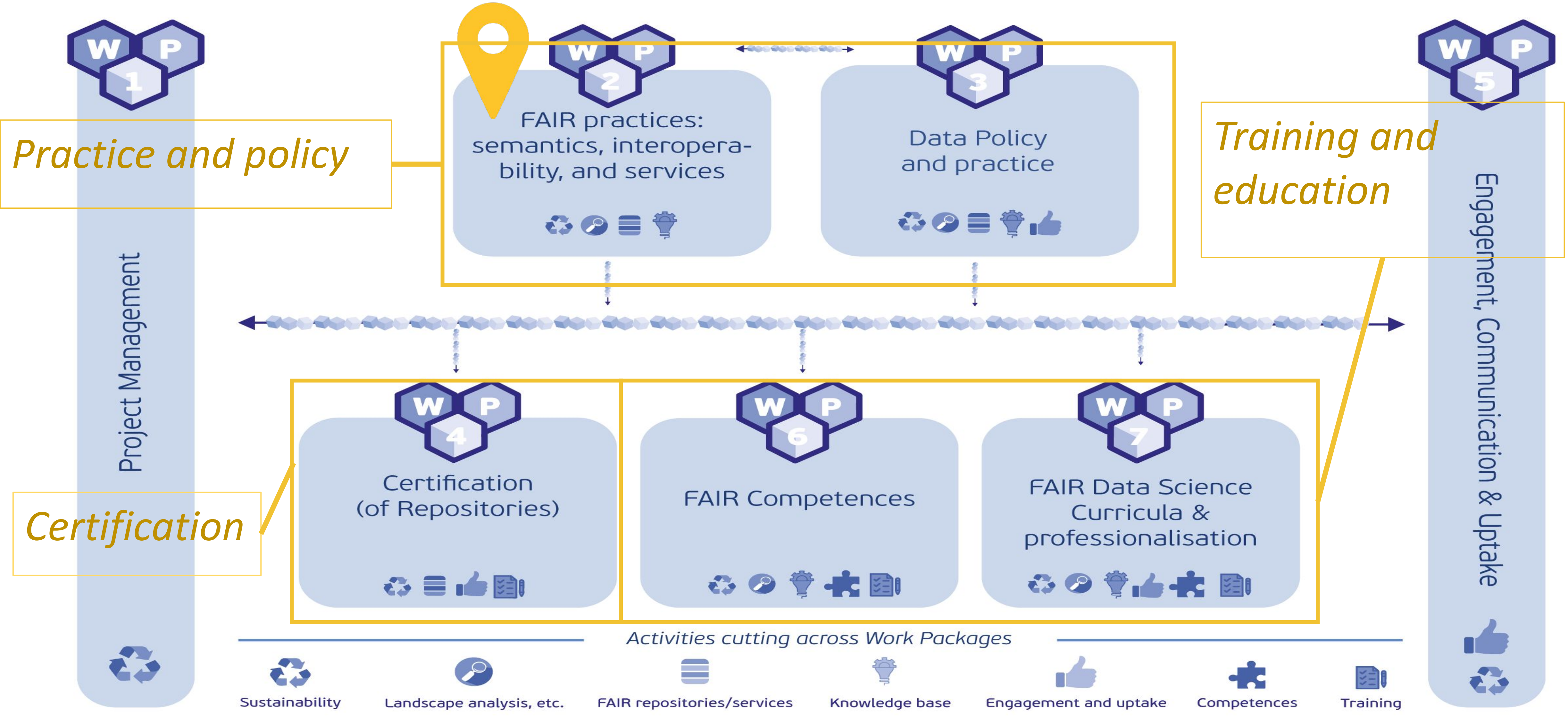
DCC

Trust-IT Services
Communicating ICT to markets

Science & Technology Facilities Council

CSC

FAIRSFAIR
Fostering Fair Data Practices in Europe



WP 1
Project Management

Practice and policy

WP 2
FAIR practices: semantics, interoperability, and services

WP 3
Data Policy and practice

Training and education

WP 5
Engagement, Communication & Uptake

Certification

WP 4
Certification (of Repositories)

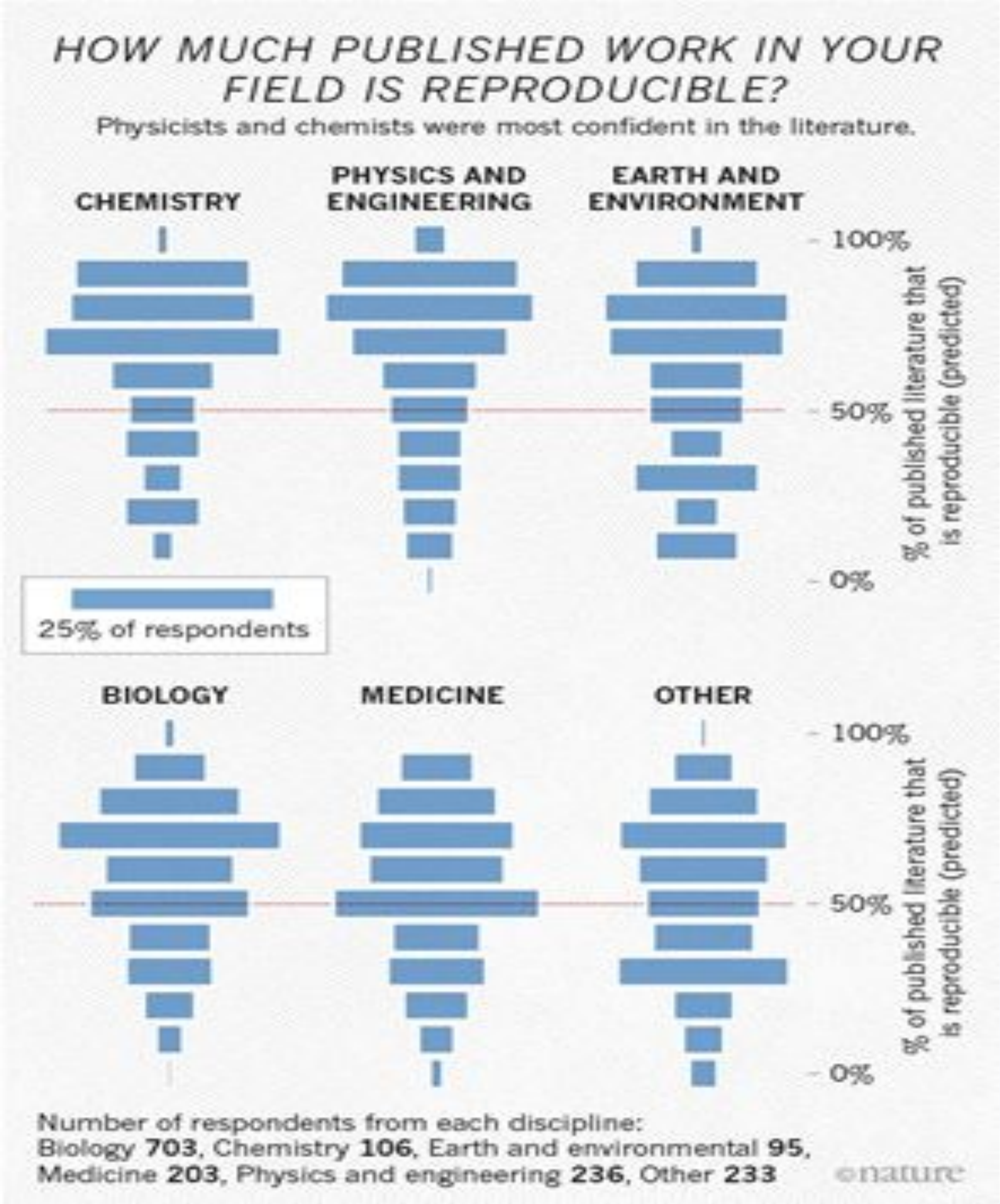
WP 6
FAIR Competences

WP 7
FAIR Data Science Curricula & professionalisation

Activities cutting across Work Packages

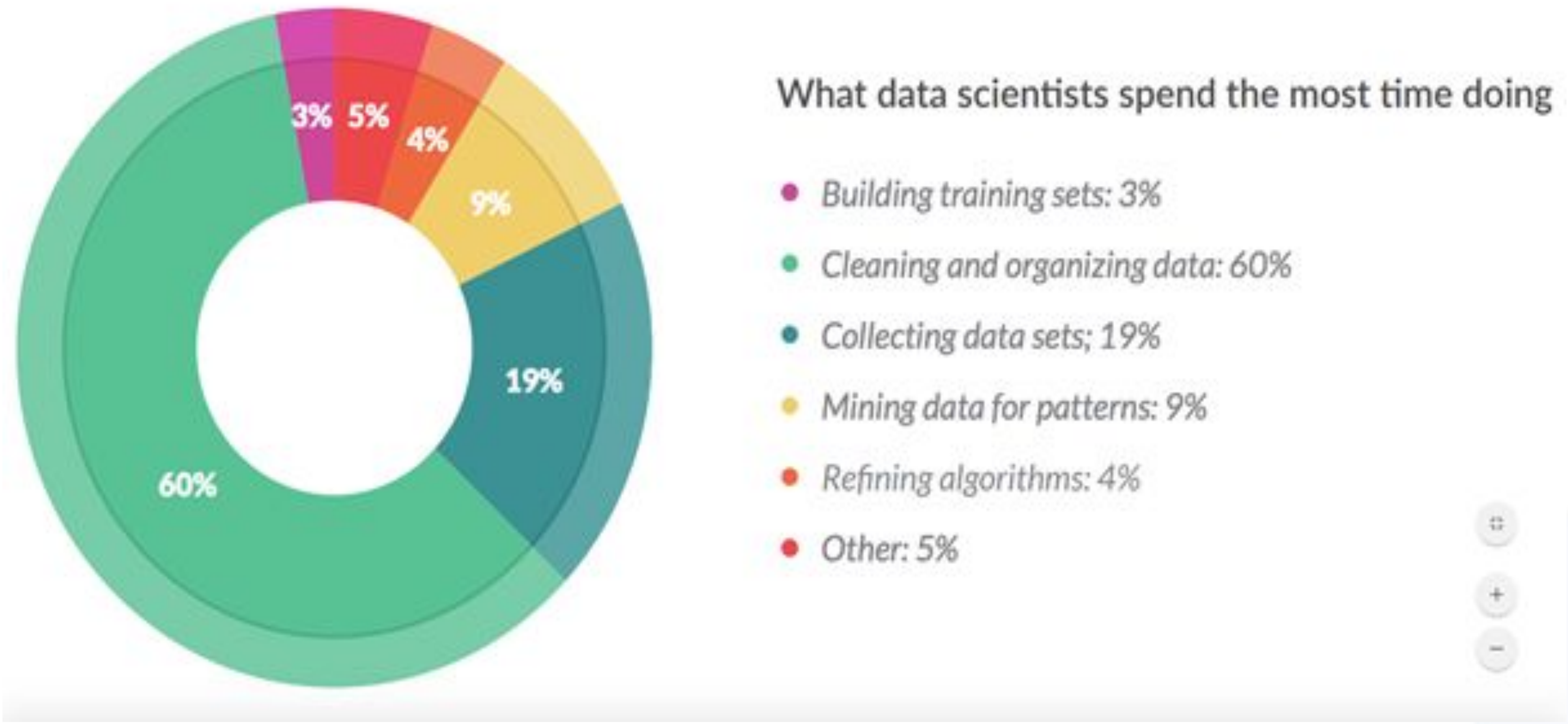
Sustainability
Landscape analysis, etc.
FAIR repositories/services
Knowledge base
Engagement and uptake
Competences
Training

FAIR Principles



Monya Baker: 1,500 scientists lift the lid on reproducibility. Survey sheds light on the ‘crisis’ rocking research. Nature 533, 2016. doi:10.1038/533452a

Working with data requires a lot of effort

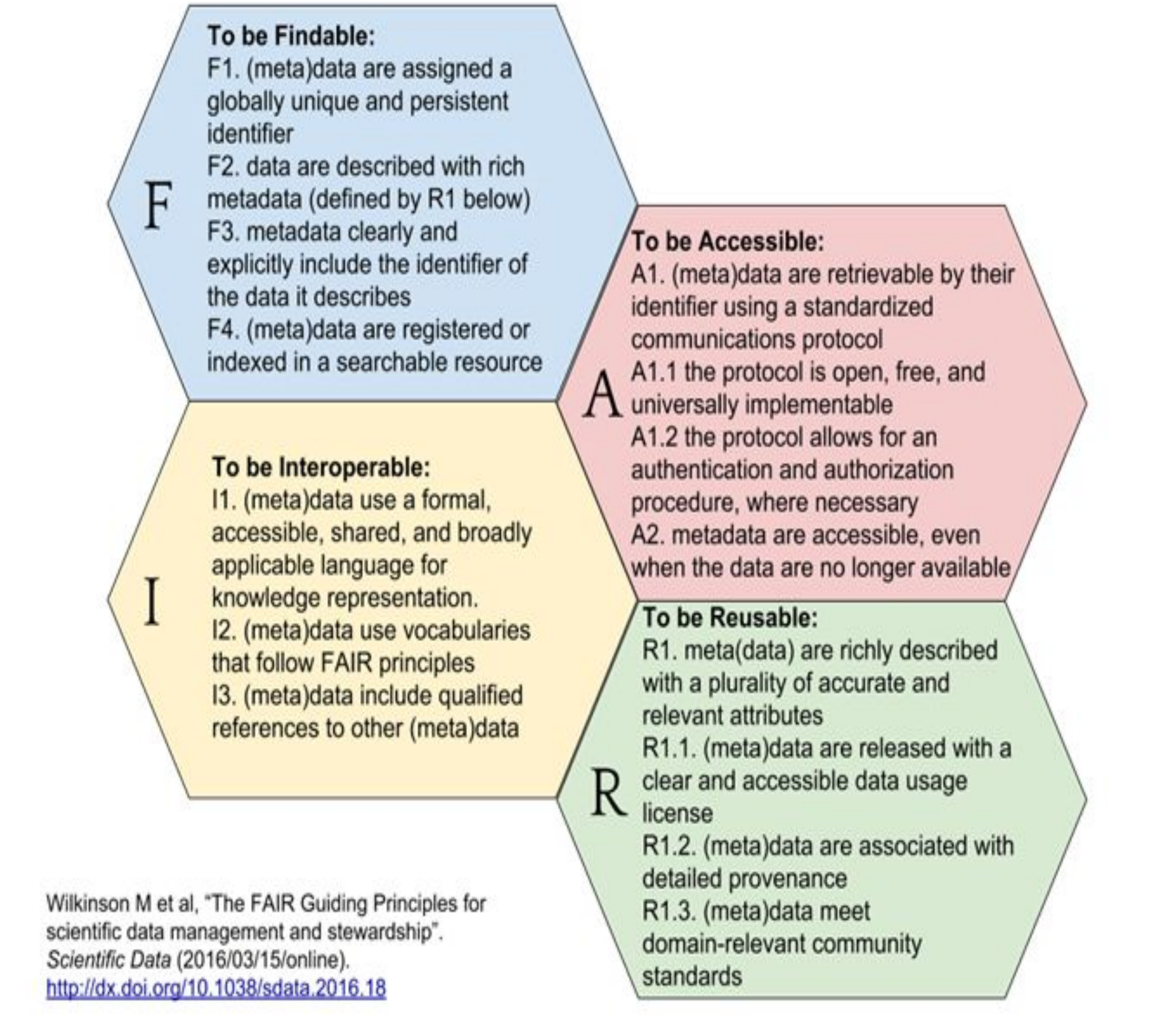


What data scientists spend the most time doing

- ● *Building training sets: 3%*
- ● *Cleaning and organizing data: 60%*
- ● *Collecting data sets; 19%*
- ● *Mining data for patterns: 9%*
- ● *Refining algorithms: 4%*
- ● *Other: 5%*

Data Science Report 2016
http://visit.crowdflower.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf

The FAIR Principles



To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be 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

To be 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

Wilkinson M et al, "The FAIR Guiding Principles for scientific data management and stewardship". *Scientific Data* (2016/03/15/online). http://dx.doi.org/10.1038/sdata.2016.18

F

- Data is described in a relevant **catalog** with sufficient information
- The dataset has a **landing page** and a unique and global **identifier**

A

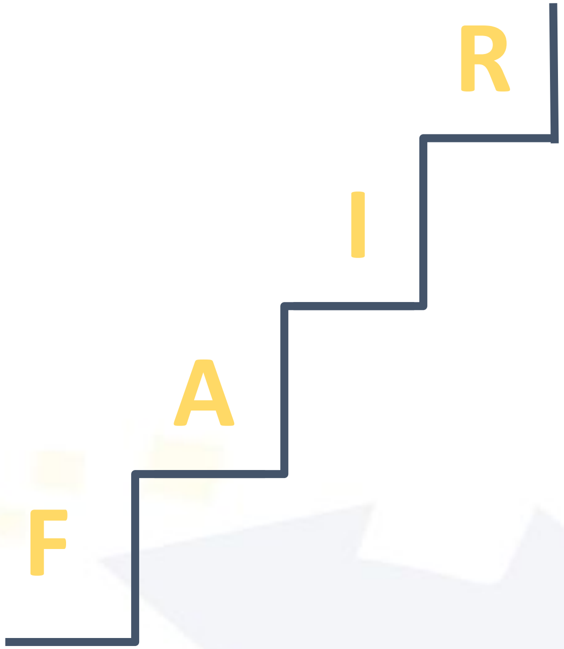
- The data can be **retrieved** over the internet
- **Versioning** and lifecycle are documented
- A tombstone page is available even if the data is deleted

I

- Use common or at least well **documented** and preferably **open formats**

R

- Rights and possible **licenses** are clearly stated
- Data is well documented and **understandable**

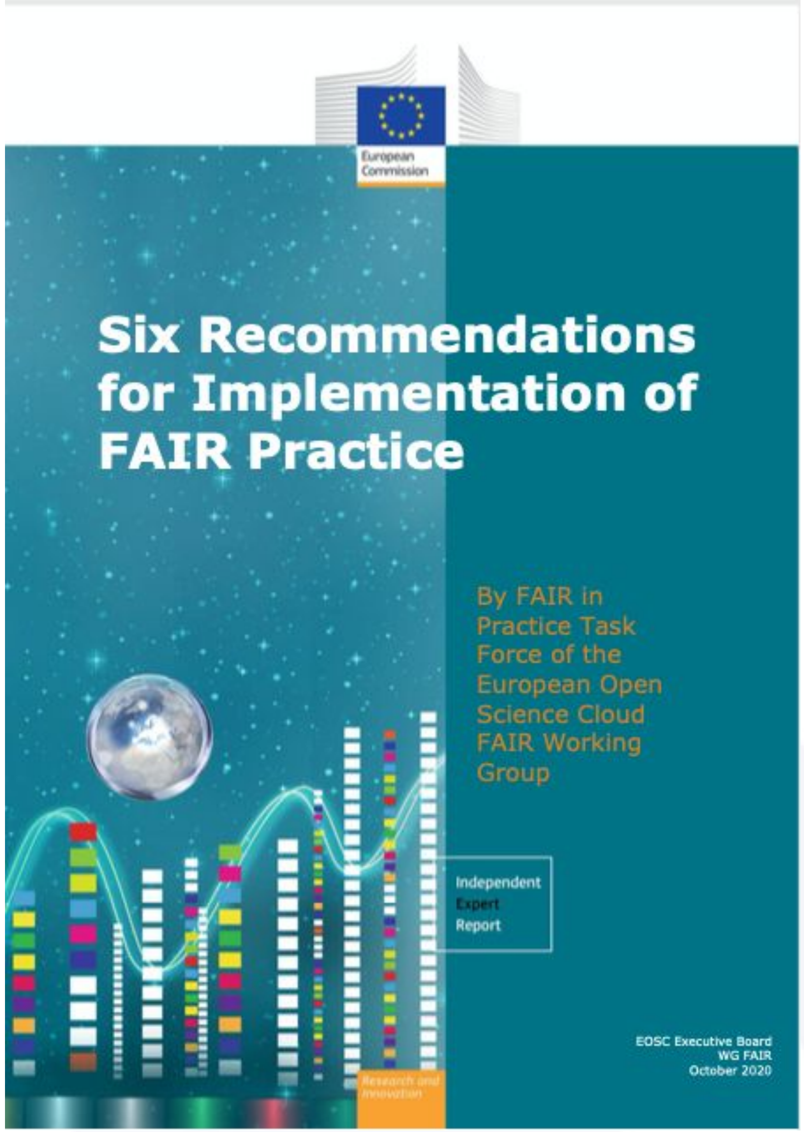


How can a researcher create FAIR data?

1. Write a **data management plan** and keep it up to date
2. Pay attention to the **FAIR principles** from the start
3. Use common **data formats**
4. Document your data, create **rich metadata**
5. Apply suitable metadata standards that use **vocabularies**
6. Put your data in a **data repository**
7. Use a repository that offers **persistent identifiers**
8. **License** your data
9. **Cite** data, yours and others!

What part of FAIR is hard?

- Both FAIRsFAIR and the EOSC see that implementation of FAIR can be hard
- Some things are community-specific
- Some things are generic
    - PIDs
    - (semantic) interoperability
    - metadata



European Commission

Six Recommendations for Implementation of FAIR Practice

By FAIR in Practice Task Force of the European Open Science Cloud FAIR Working Group

Independent Expert Report

Research and Innovation

EOSC Executive Board
WG FAIR
October 2020

FAIRsFAIR

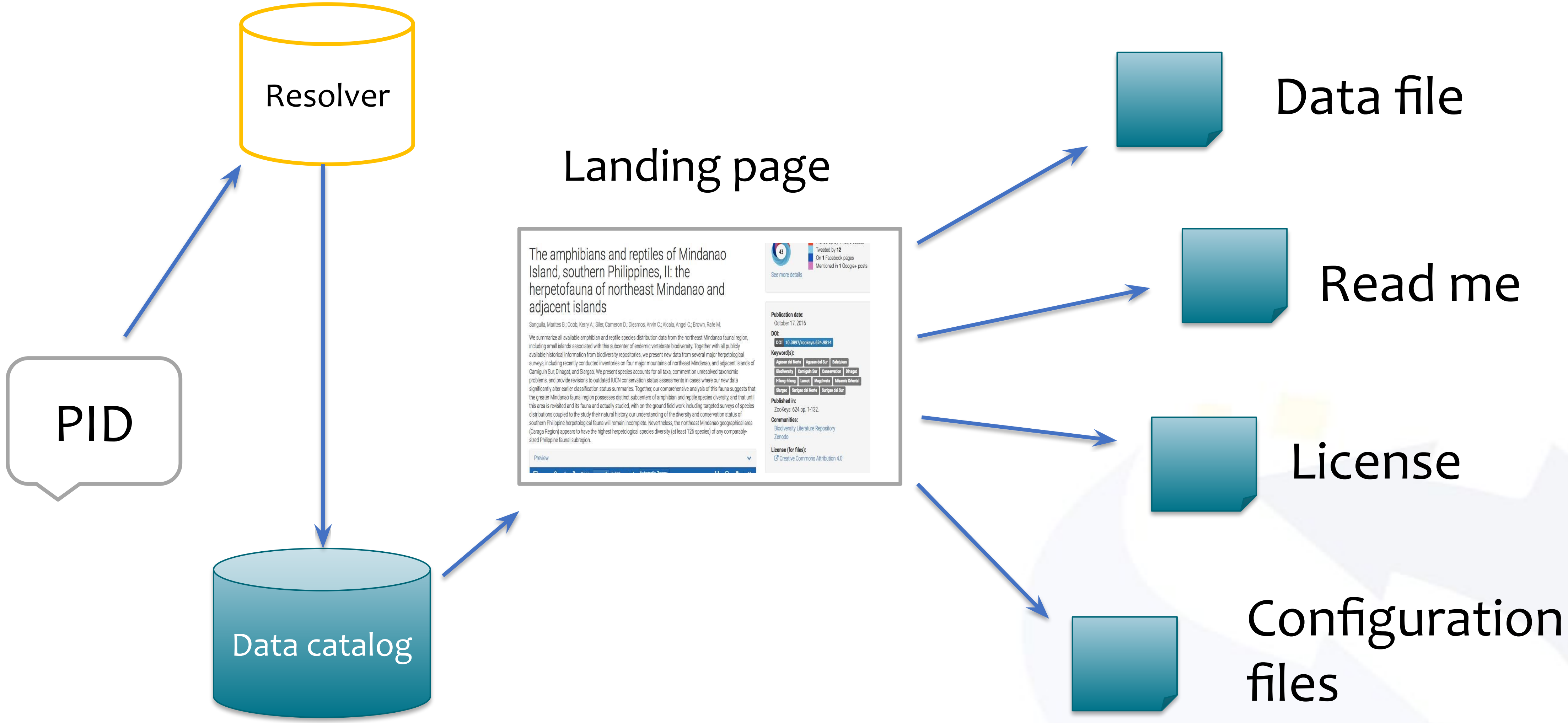
Fostering Fair Data Practices in Europe

Using Persistent Identifiers

FAIRsFAIR "Fostering FAIR Data Practices in Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 Grant agreement 831558

What is a persistent identifier?

- **Globally unique**, i.e. nobody else in the world should use the same string to refer to anything else
  - a controlled syntax and a governed namespace
  - be issued and managed by a clearly specified registration authority
- **Resolvable,** i.e. provide a way for both machines and humans to access the digital object itself, the state information and/or a landing page
- **Persistent**, i.e. remain unique and resolvable with a persistent syntax. The object it represents should also be persistent and protected against content drift
  - this requires metadata and curation



The amphibians and reptiles of Mindanao Island, southern Philippines, II: the herpetofauna of northeast Mindanao and adjacent islands

Sanguila, Marites B.; Cobb, Kerry A.; Siler, Cameron D.; Diesmos, Arvin C.; Alcala, Angel C.; Brown, Rafe M.

We summarize all available amphibian and reptile species distribution data from the northeast Mindanao faunal region, including small islands associated with this subcenter of endemic vertebrate biodiversity. Together with all publicly available historical information from biodiversity repositories, we present new data from several major herpetological surveys, including recently conducted inventories on four major mountains of northeast Mindanao, and adjacent islands of Camiguin Sur, Dinagat, and Siargao. We present species accounts for all taxa, comment on unresolved taxonomic problems, and provide revisions to outdated IUCN conservation status assessments in cases where our new data significantly alter earlier classification status summaries. Together, our comprehensive analysis of this fauna suggests that the greater Mindanao faunal region possesses distinct subcenters of amphibian and reptile species diversity, and that until this area is revisited and its fauna and actually studied, with on-the-ground field work including targeted surveys of species distributions coupled to the study their natural history, our understanding of the diversity and conservation status of southern Philippine herpetological fauna will remain incomplete. Nevertheless, the northeast Mindanao geographical area (Caraga Region) appears to have the highest herpetological species diversity (at least 126 species) of any comparably-sized Philippine faunal subregion.

Preview

43

Tweeted by **12**
On **1** Facebook pages
Mentioned in **1** Google+ posts

See more details

Publication date:
October 17, 2016

DOI:
DOI 10.3897/zookeys.624.9814

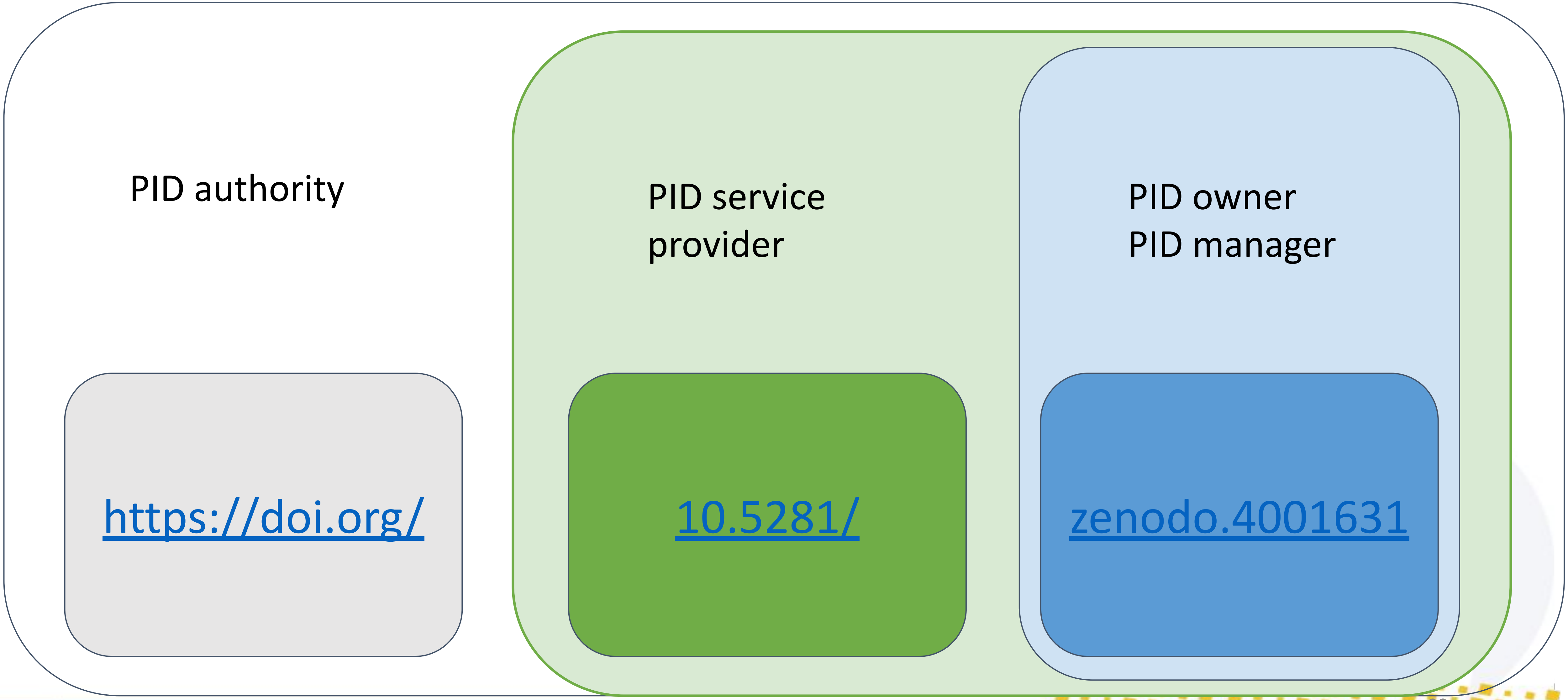
Keyword(s):
Agusan del Norte | Agusan del Sur | Balatukan | Biodiversity | Camiguin Sur | Conservation | Dinagat | Hilong-hilong | Lumot | Magdiwata | Misamis Oriental | Siargao | Surigao del Norte | Surigao del Sur

Published in:
ZooKeys: 624 pp. 1-132.

Communities:
Biodiversity Literature Repository
Zenodo

License (for files):
Creative Commons Attribution 4.0

Structure of PIDs and example responsibilities



PID authority

https://doi.org/

PID service provider

10.5281/

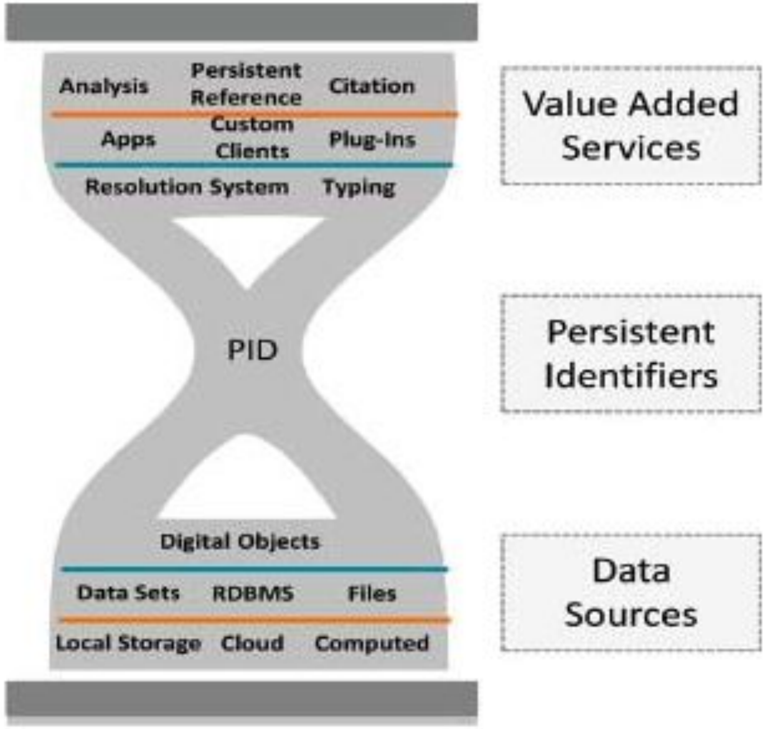
PID owner
PID manager

zenodo.4001631

THE PROMISE

THE COST

PID records and metadata



https://doi.org/10.1162/dint_a_00004

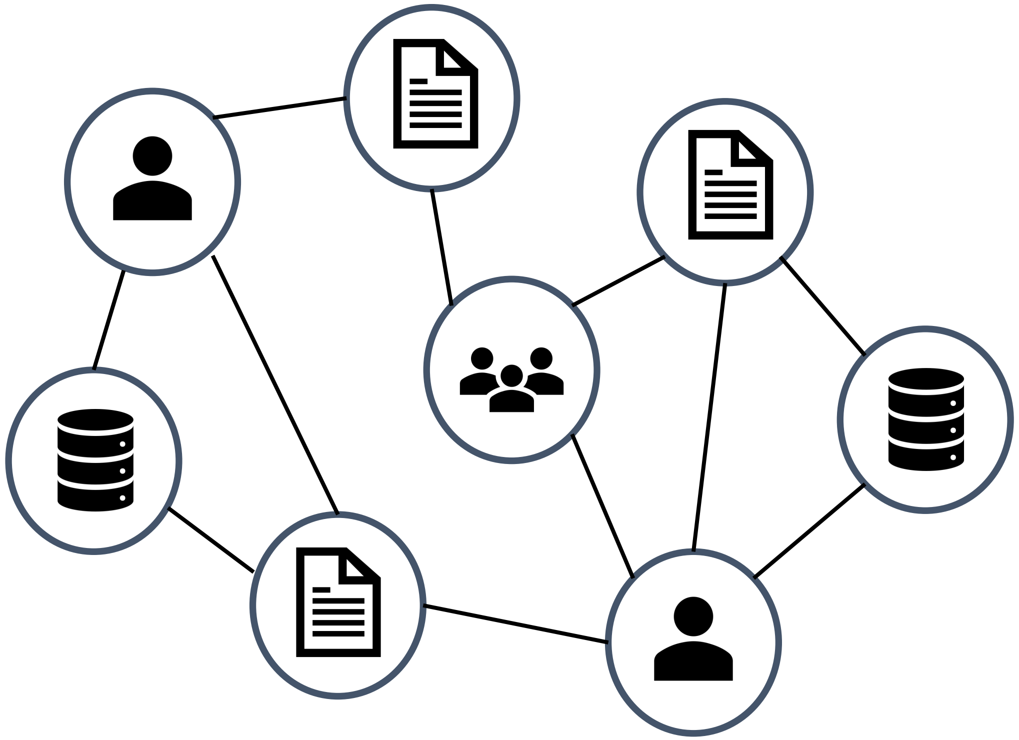
- The PID itself as a **string** contains information
  - the PID should be recognisable as a PID to the user (human and/or machine)
  - all other semantics poses a risk and should be carefully managed
- The PID contains **kernel metadata** that should be as minimal as possible
  - The PID record may be a non-authoritative source for arbitrary metadata and stored directly at the resolving service
- The **master metadata** is provided by the PID owner and manager

Resolving

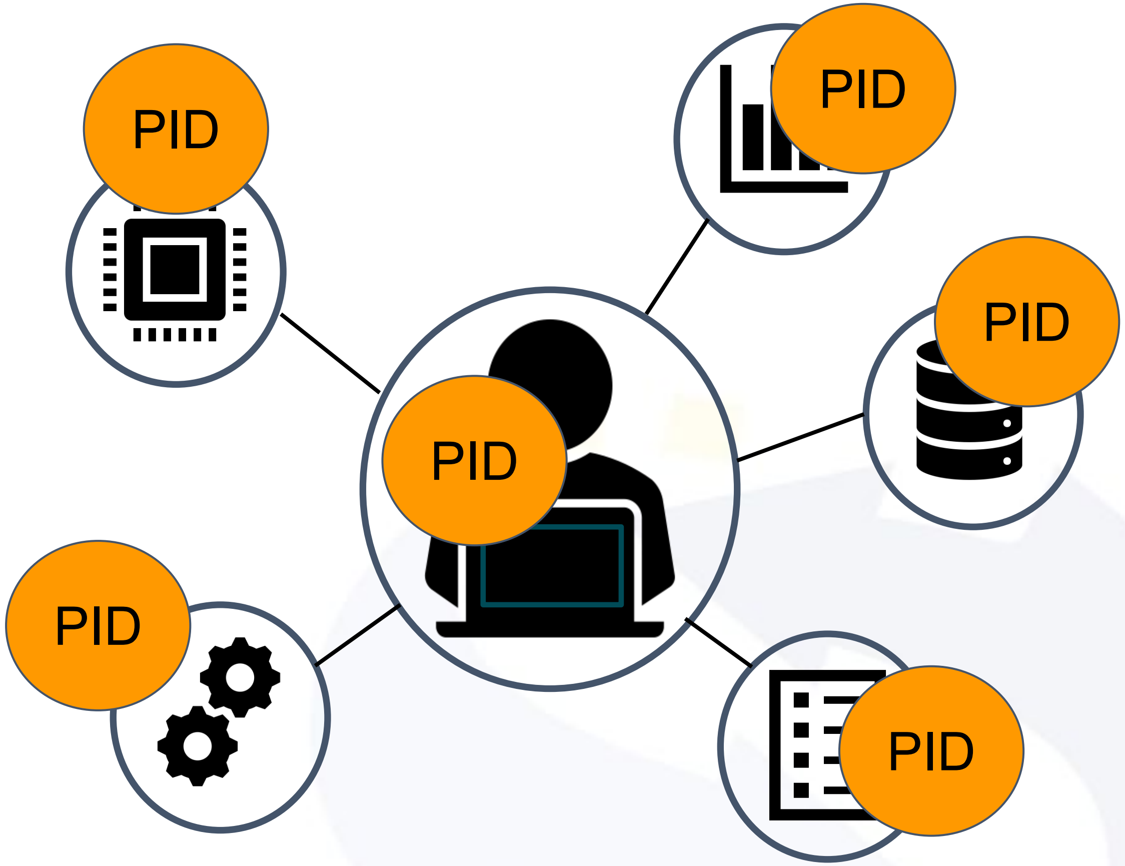
- **Domain Name Service** (DNS) resolver: Resolves a hostname to an IP address.
- **Local resolver**, e.g. load balancer, API gateway or web server: Redirects to a different host and/or path.
- **Full resolver**, e.g. handle system: Redirects to a URL either following a regular expression pattern, or a specific URL stored in the service.
- **Meta-resolver**, e.g. identifiers.org or n2t.net: Redirects to a URL following a regular expression pattern.
- **Single-service resolver**: some PIDs resolve to a single central resource, e.g. ORCID.

Source: Wimalaratne S, Fenner M D2.1 PID Resolution Services Best Practices. FREYA, 2018.
https://doi.org/10.5281/zenodo.1324300

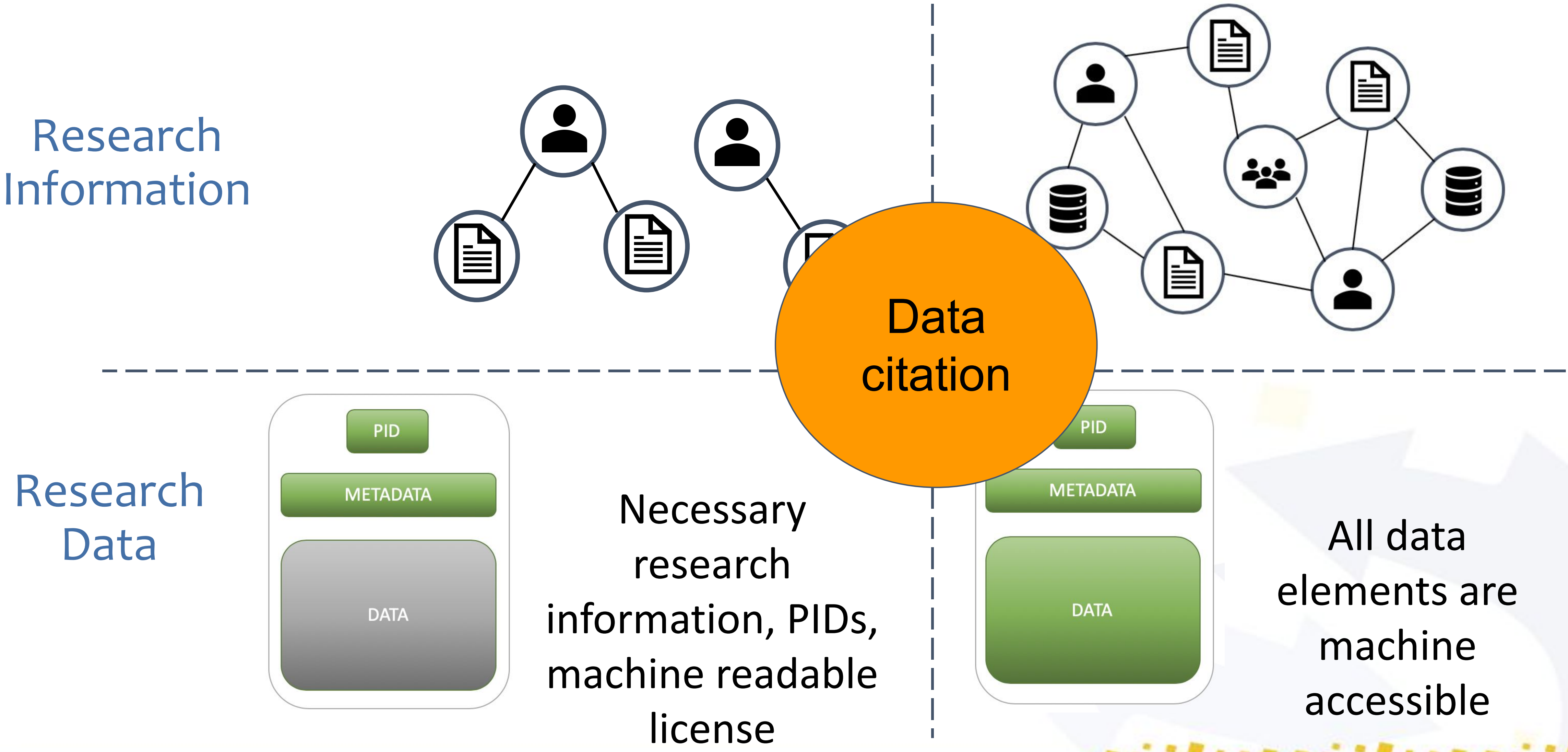
Research information



Research data



Shallow FAIR and Deep FAIR



Research Information

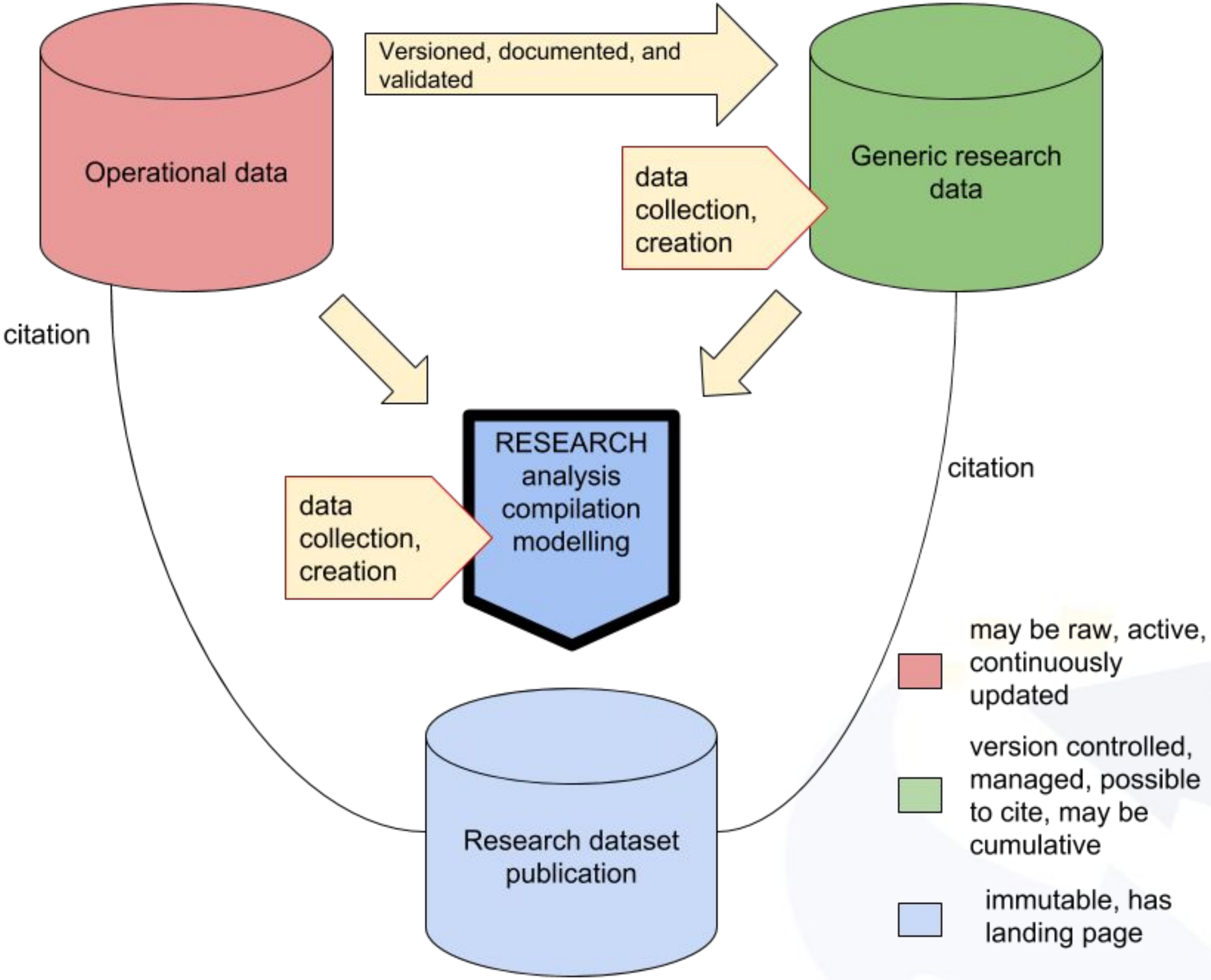
Research Data

Data citation

Necessary research information, PIDs, machine readable license

All data elements are machine accessible

https://www.swissuniversities.ch/fileadmin/swissuniversities/Dokumente/Organisation/SUK-P/SUK_P-2/2019_OpenScience_Poster12_RDM_UniBas_20190909_fundingFAIRcommunities.pdf (inspired by)



Operational data

Versioned, documented, and validated

Generic research data

data collection, creation

citation

RESEARCH analysis compilation modelling

data collection, creation

citation

Research dataset publication

🟥 may be raw, active, continuously updated

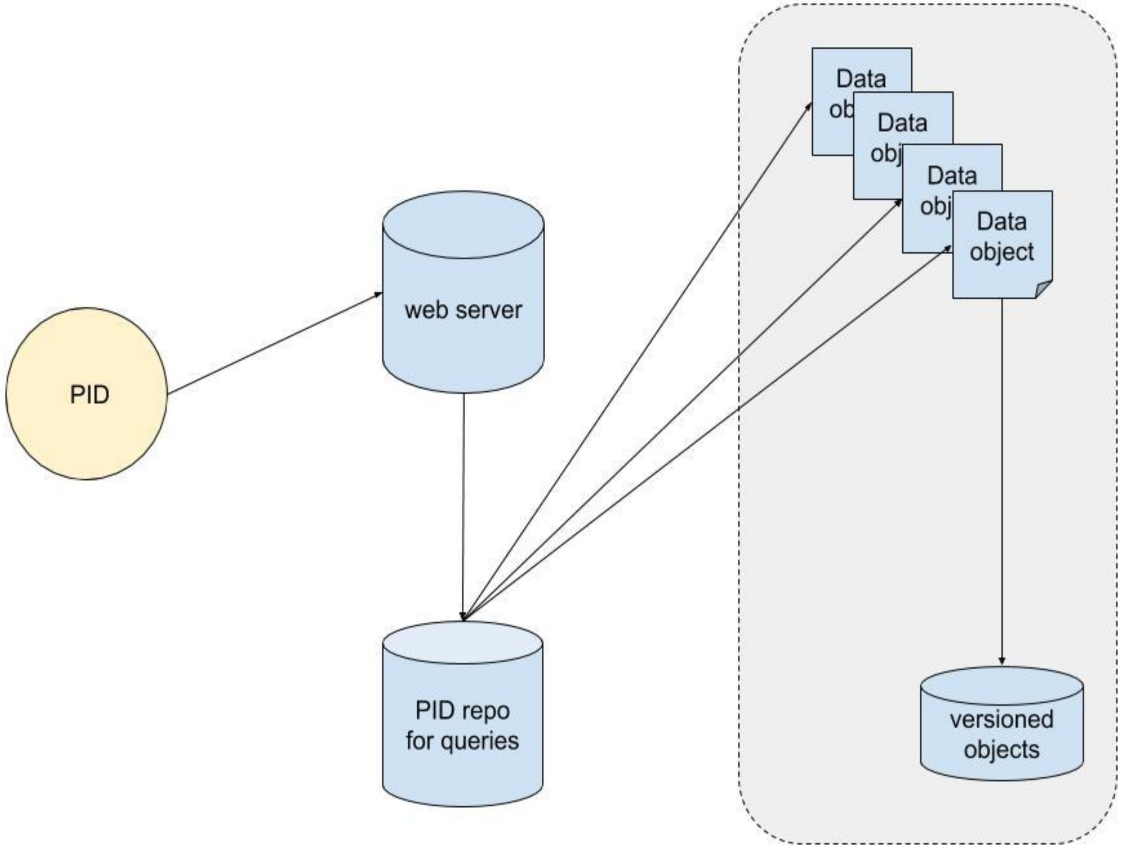
🟩 version controlled, managed, possible to cite, may be cumulative

🟦 immutable, has landing page

Evolving datasets and citation

- Data Versioning: For retrieving earlier states of datasets, the data needs to be versioned. Markers shall indicate inserts, updates and deletes of data in the database.
- Data Timestamping: Ensure that operations on data are timestamped, i.e. any additions, deletions are marked with a timestamp.
- Data Identification: The data used shall be identified via a PID pointing to a time-stamped query, resolving to a landing page.

Rauber A, Asmi A, van Uytvanck D, Proell S. Data Citation of Evolving Data : Recommendations of the Working Group on Data Citation (WGDC). Published online October 20, 2015. https://doi.org/doi:10.15497/RDA00016



Dynamic and growing datasets

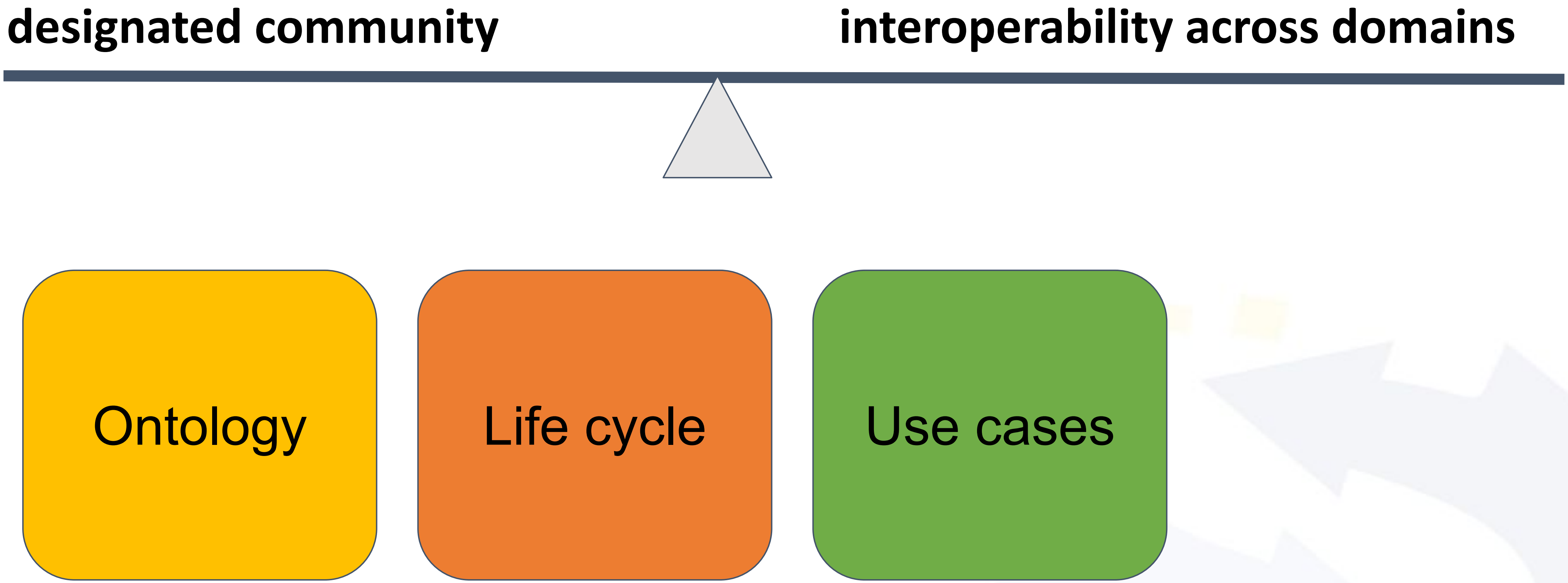
URN allows use of fragments

Avoid PID inflation

Consider costs and sustainability

Ad hoc creation rather than automatic minting and allocation?

Providing PIDs



designated community

interoperability across domains

Ontology

Life cycle

Use cases

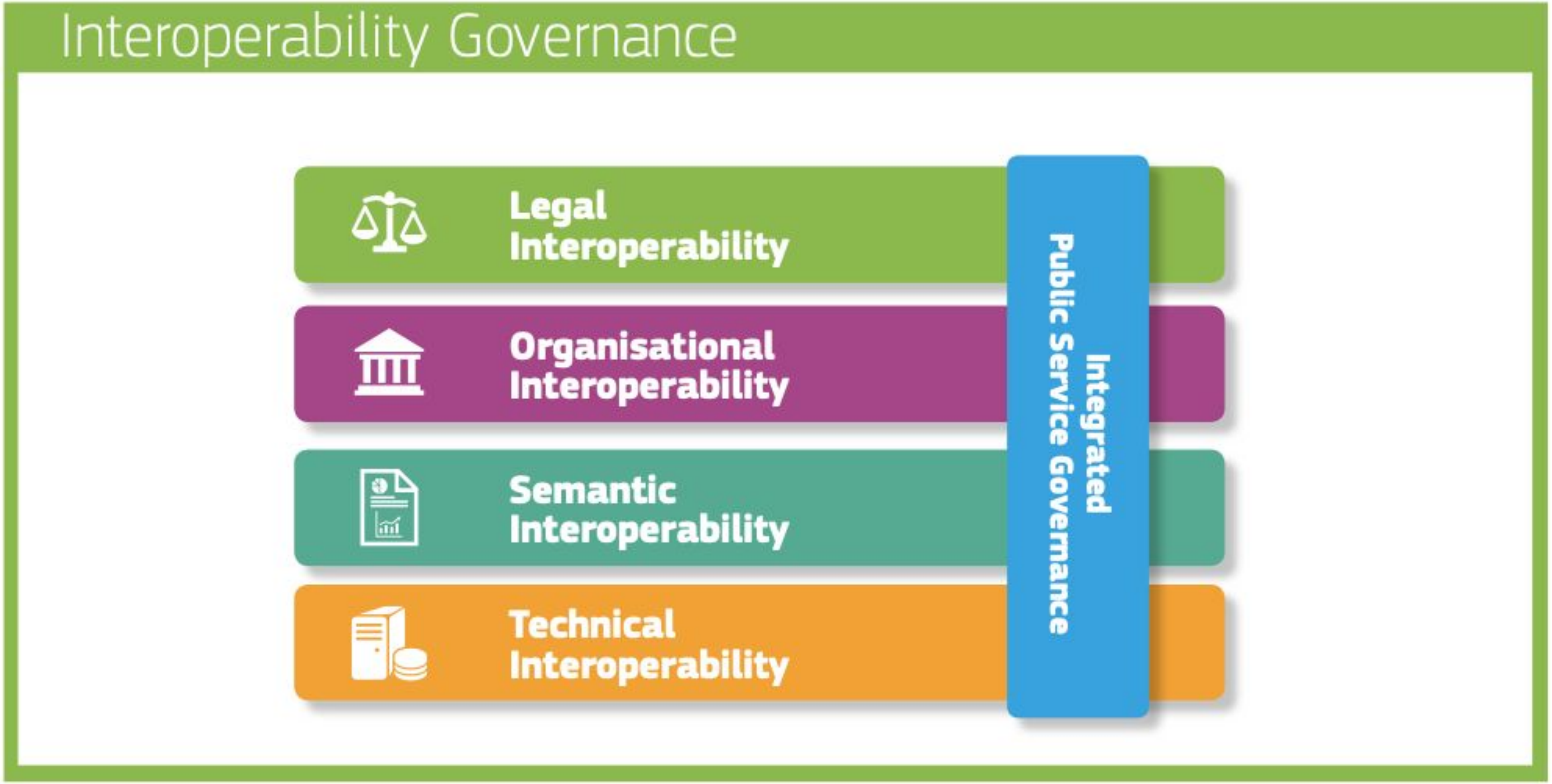
a PID
is a promise

FAIRsFAIR

Fostering Fair Data Practices in Europe

Semantic Interoperability

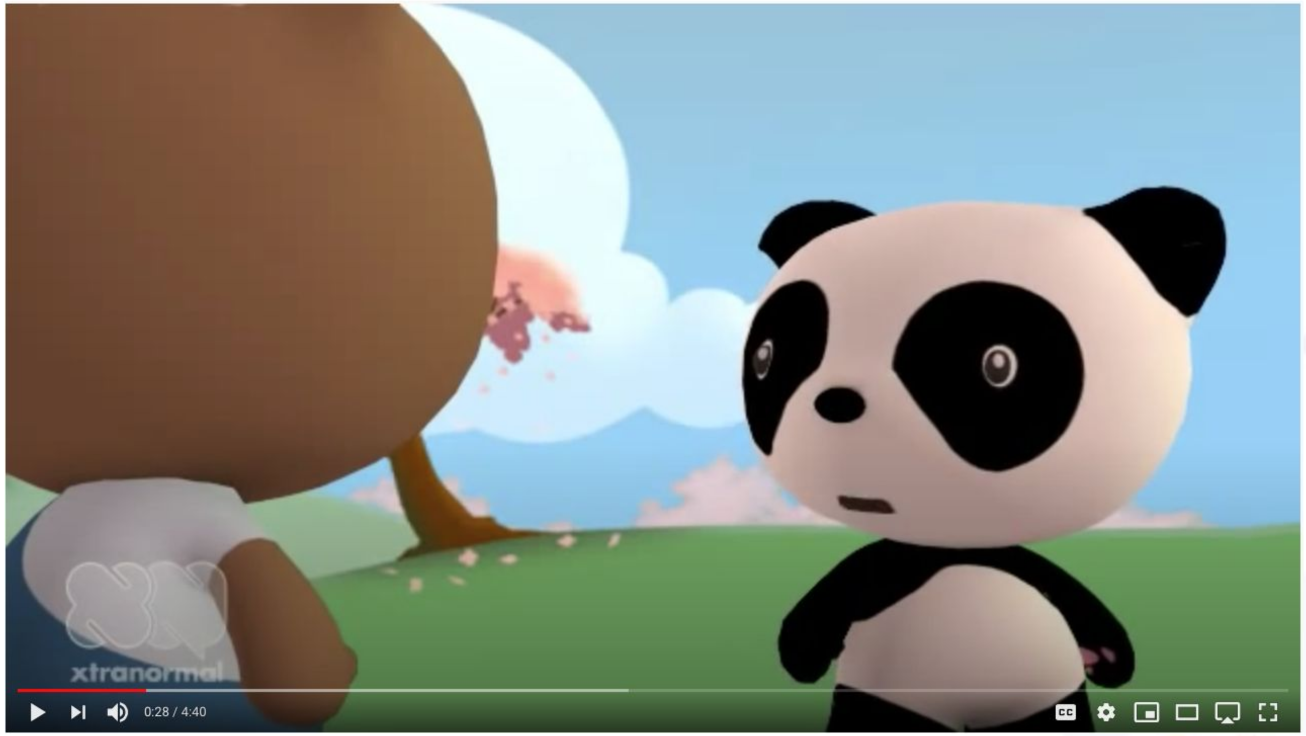
FAIRsFAIR "Fostering FAIR Data Practices in Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 Grant agreement 831558


Interoperability Governance
Legal Interoperability
Organisational Interoperability
Semantic Interoperability
Technical Interoperability
Integrated Public Service Governance

- Four layers in the New European Interoperability Framework
- https://ec.europa.eu/isa2/eif_en

Semantic Interoperability: Do I understand what you mean?

- If you don’t know it, worth watching https://www.youtube.com/watch?v=66oNv_DJuPc



Semantic Interoperability: File Format?

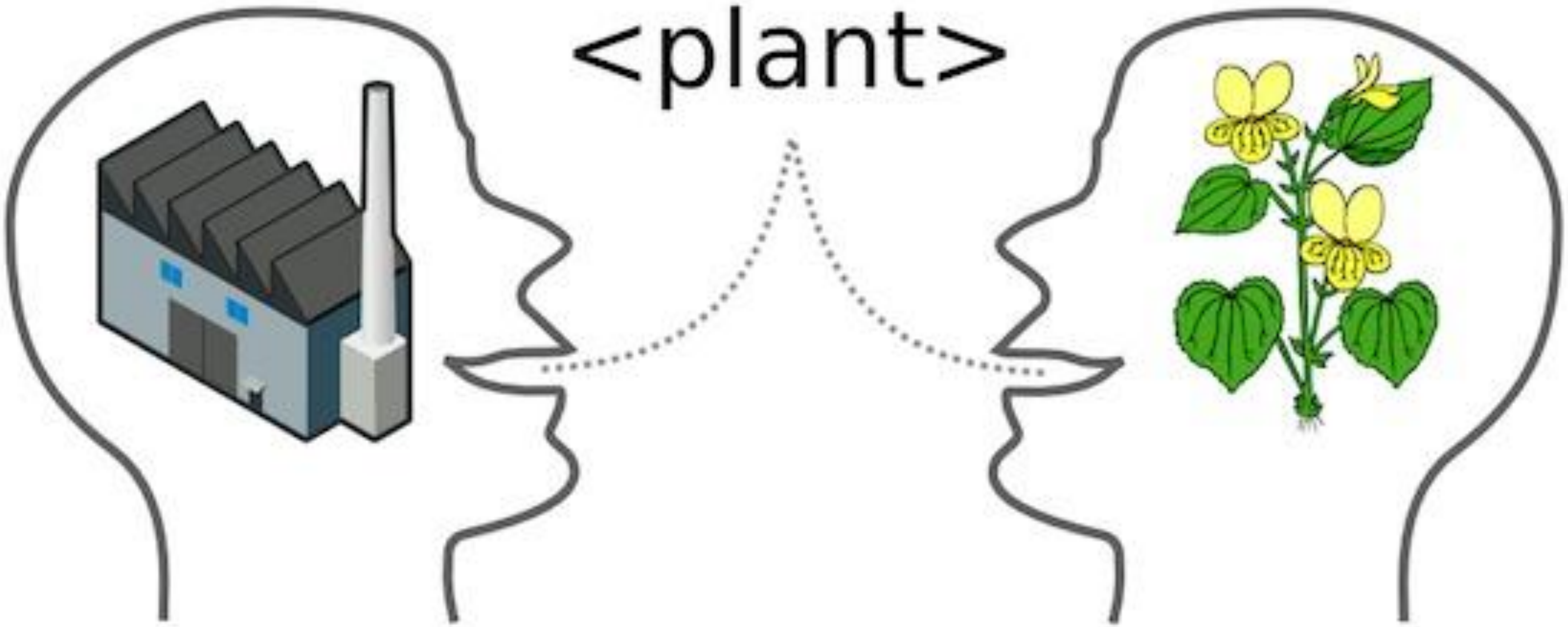
Does using “standard file formats” solve the interoperability problem?

- No!
- File format is “technical interoperability”
- PDF is standard, but no help for data reuse
- CSV doesn’t save the day
- SPSS doesn’t guarantee that data is understandable

So, what do we need in addition?

Semantic Interoperability: Contents!

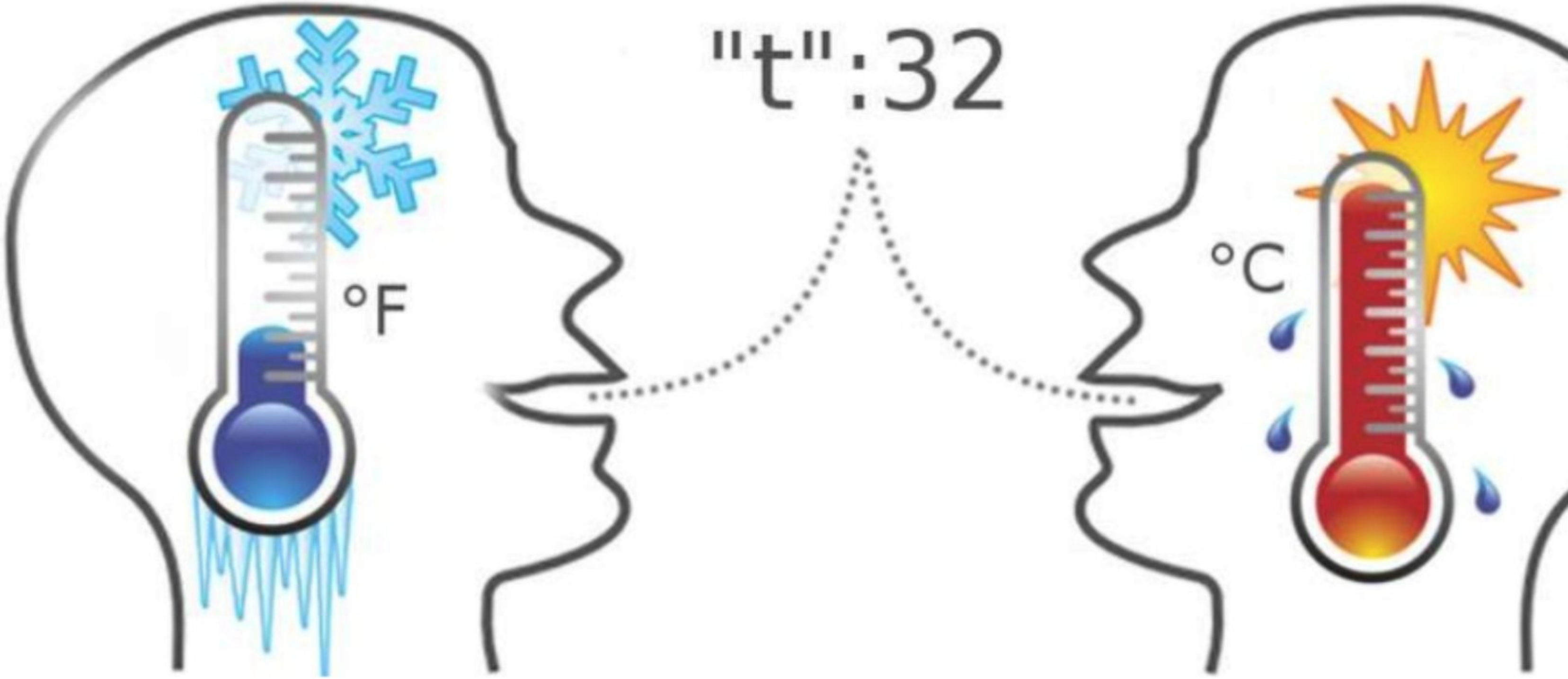
- All content needs to be unambiguous



https://www.peterkrantz.com/2010/semantic-interoperability/

Semantic Interoperability: Contents!

- All terms need to be unambiguous
- All numbers need to be unambiguous



https://www.slideshare.net/maximelefrancois86/reference-knowledge-models-for-smart-application

Semantic Interoperability: Contents!

- All terms need to be unambiguous
- All numbers need to be unambiguous
- What does it mean when data is missing/empty/”zero”/”-1”

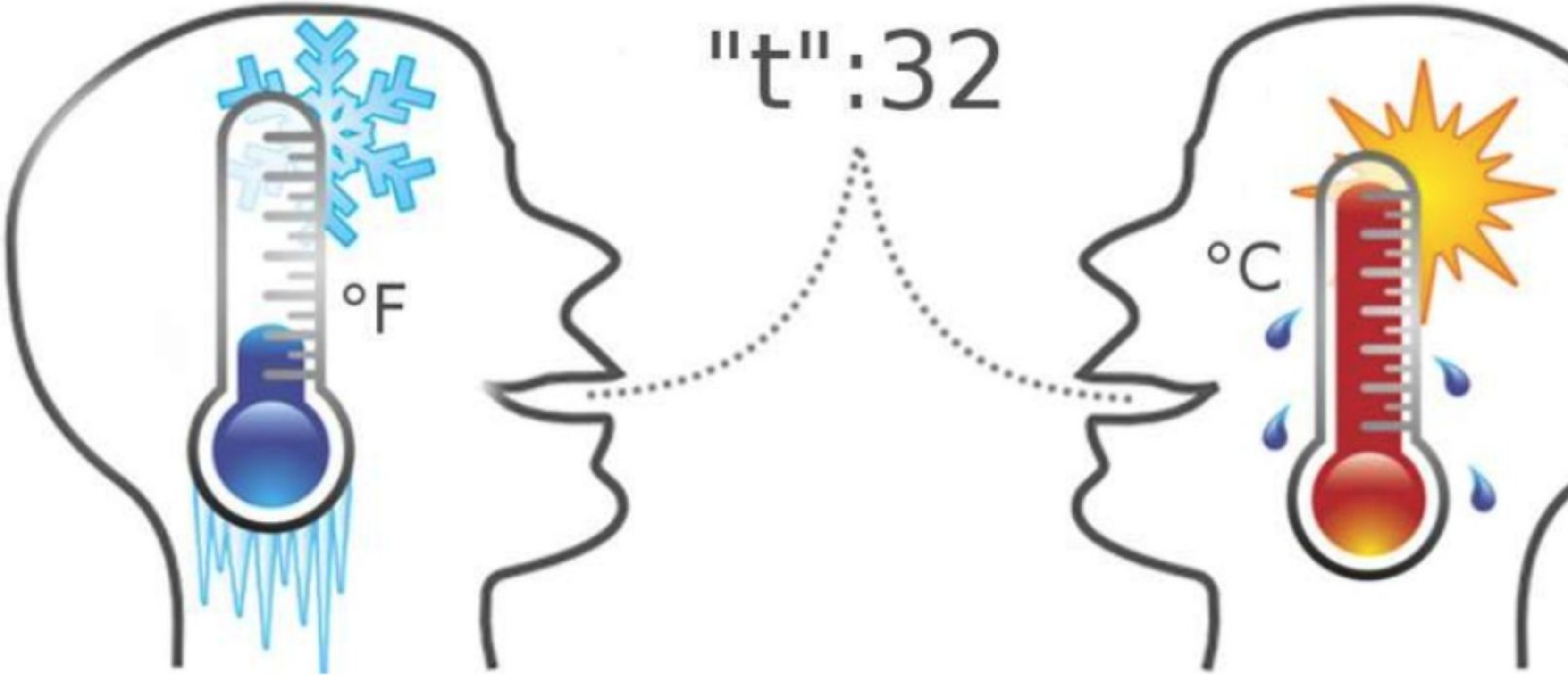
- It should be impossible to misunderstand

Misunderstanding?

- It should be impossible for a user of the data to misunderstand it

Who is that user?

- You, right now?
- You, in 5 years?
- A coworker?
- Someone from a similar lab elsewhere?
- Someone performing wide-scale data integration?

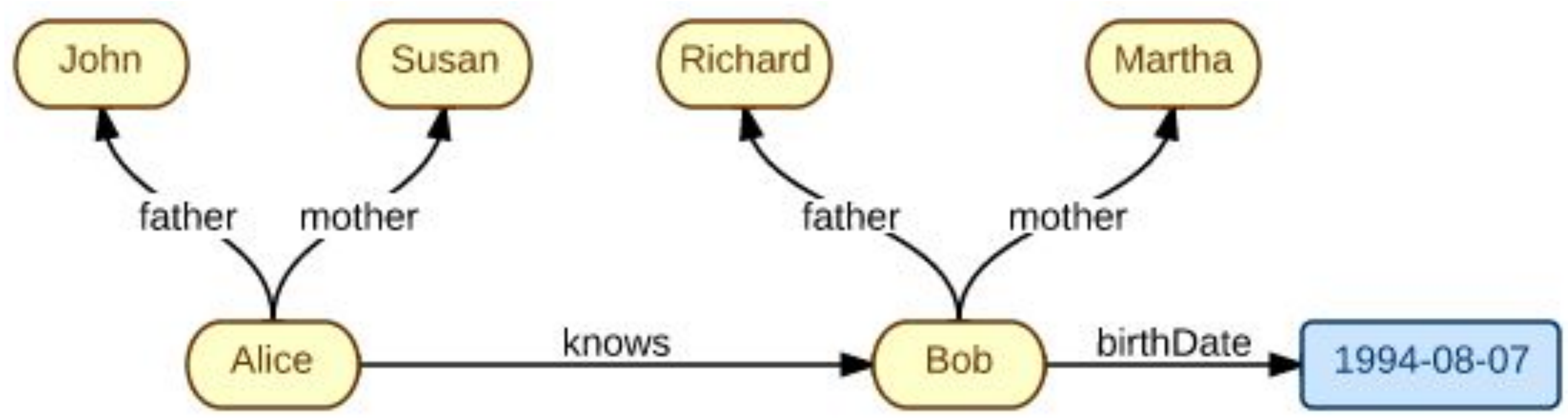


https://www.slideshare.net/maximelefrancois86/reference-knowledge-models-for-smart-application

Semantic interoperability: How?

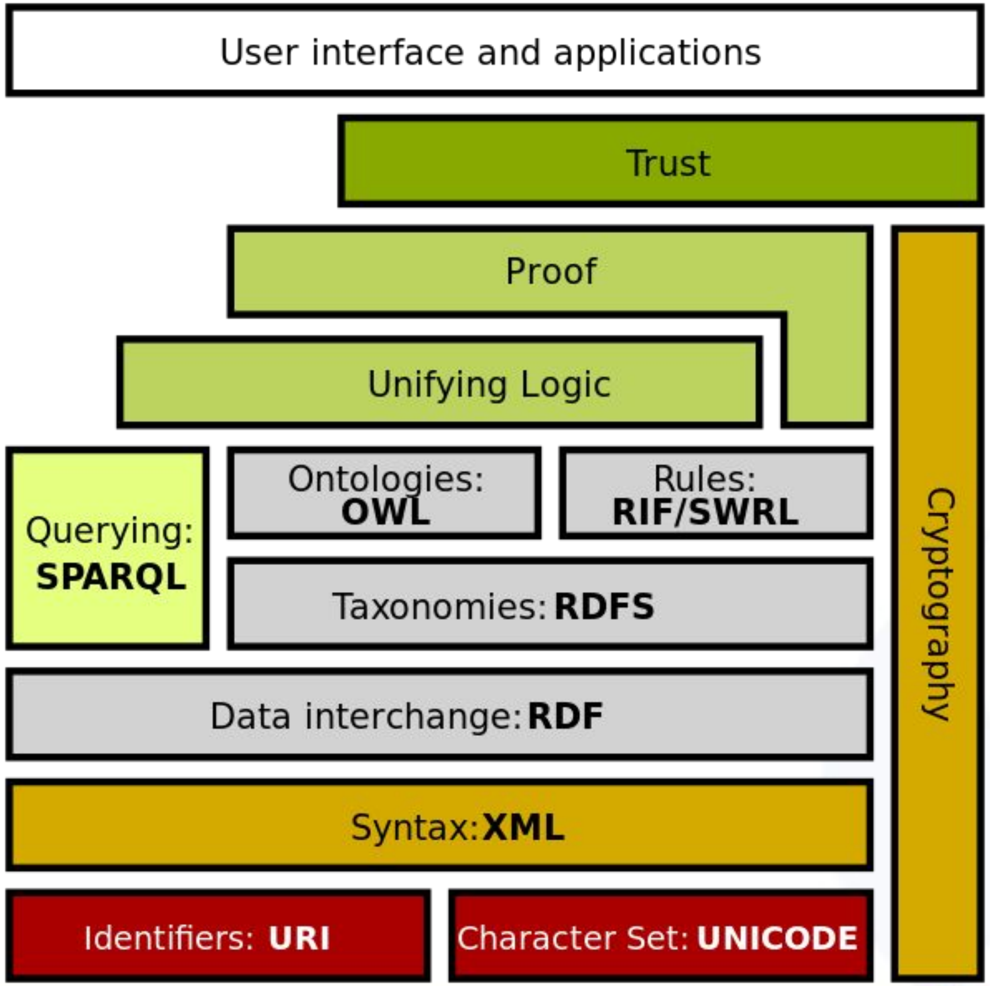
- FAIR principles paper does not make a choice
- “Linked Data” is a very good technical implementation
  - Describe relationships and properties
  - Instead of terms, use “URIs” (or…. PIDs)
  - Very generic, many tools have been developed to work with it

Relationships and properties



https://sites.google.com/site/restframework/introduction-to-rdf

“Semantic Web” stack



User interface and applications

Trust

Proof

Unifying Logic

Querying: **SPARQL**

Ontologies: **OWL**

Rules: **RIF/SWRL**

Taxonomies: **RDFS**

Data interchange: **RDF**

Syntax: **XML**

Identifiers: **URI**

Character Set: **UNICODE**

Cryptography

Data Structure

Unstructured data?

RDF?

Please do not convert all data to RDF!

But make sure it is unambiguously possible.

Semantic interoperability for machines

- Rather than using a term to describe something that is specific to a human language and may be context-dependent in practice, refer to concep represented by a unique identifier
- Each data value should be associated with a precise data type, documented to such precision that misunderstandings are avoided
- Semantic artefacts are tools used for this
- Sharing and curating these to be FAIR and trustworthy is the basis for sustainable semantic interoperability
- Follow the recommendations for FAIR semantics (FAIRsFAIR D2.5)

Researchers can support FAIR Interoperability by

- using open standards for data and metadata formats
- using existing shared and curated semantic artefacts (code lists, vocabularies, ontologies etc)
- supporting creation, curation and linking of FAIR semantic artefacts
- using shared protocols and registers for data types, instruments etc

FAIRsFAIR
Fostering Fair Data Practices in Europe

Metadata

FAIRsFAIR "Fostering FAIR Data Practices in Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 Grant agreement 831558

“Don’t give me books for Christmas, I already have a book”

--- Jean Harlow

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 it describes;

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

Accessible:

A1. (meta)data are retrievable by their identifier using a standardized communications protocol;

A1.1 the protocol is open, free, and universally implementable;

A1.2. the protocol allows for an authentication and authorization procedure, where necessary;

A2. metadata are accessible, even when the data are 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;

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;

DMP

“Which metadata standard(s) will you use”

With options:

- I will use Dublin Core
- I will use discipline specific standards

But: It should always be both!

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 it describes;

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

Accessible:

A1. (meta)data are retrievable by their identifier using a standardized communications protocol;

A1.1 the protocol is open, free, and universally implementable;

A1.2. the protocol allows for an authentication and authorization procedure, where necessary;

A2. metadata are accessible, even when the data are 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;

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;

What is in a metadata standard?

- Optional: Format
- Fields
  - Definition
  - Priority: Obligatory, Recommended, Optional
  - Ontology / Vocabulary

How to find a metadata standard

- FAIRsharing
  - http://fairsharing.org
- RDA metadata directory
  - http://rd-alliance.github.io/metadata-directory/
- CEDAR:
  - http://metadatacenter.org
- Component MetaData Infrastructure, CDMI
  - https://www.clarin.eu/content/component-metadata

Researchers can support metadata by

- choosing trustworthy services that enable FAIR data
- using open and shared ontologies and standards for metadata
- creating rich metadata (not only mandatory or minimal metadata)
- following guidelines and applications profiles

FAIRsFAIR

Fostering Fair Data Practices in Europe

Q&A

FAIRsFAIR "Fostering FAIR Data Practices in Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 Grant agreement 831558

FAIRsFAIR

FOLLOW US

www.fairsfair.eu

@FAIRsFAIR_eu

www.linkedin.com/company/fairsfair/

www.youtube.com/channel/UCE4wSBnNIBfu6SqlBalMfNg