FAIRSFAIR

Fostering Fair Data Practices in Europe

# Clearing some of the highest FAIR hurdles: PIDs, Metadata, and Semantic Interoperability for Data Stewards and Service Providers

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: The FAIRsFAIR project - Jessica Parland-von Essen
- Implementing persistent identifiers - Jessica Parland-von Essen
- Semantic interoperability and Metadata - Rob Hooft
- Q&A

FAIRSFAIR

Fostering Fair Data Practices in Europe

# The FAIRsFAIR Project

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 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

FAIRSFAIR
Fostering Fair Data Practices in Europe

# FAIRsFAIR partners

Data Archiving and Networked Services
DANS

Trust-IT Services
ting ICT to markets

Universiteit van Amsterdam

DCC

Helsingin yliopisto
Helsingfors universitet
University of Helsinki

COMMITTEE ON DATA
CODATA
INTERNATIONAL SCIENCE COUNCIL

eua EUROPEAN UNIVERSITY ASSOCIATION

uc3m | Universidad Carlos III de Madrid

CSC

Inria

Universidade do Minho

SURF SARA

SPARC Europe

University of Essex

DataCite
FIND, ACCESS, AND REUSE DATA

Science & Technology Facilities Council | UK Research and Innovation

esciencedata FACTORY

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

INES
Centre Informatique National de l'Enseignement Supérieur

RDA
RESEARCH DATA ALLIANCE

DTL

Universität Bremen

FAIRSFAIR
Fostering Fair Data Practices in Europe

# Our objective

Help **survey the landscape of FAIR activities** in relation to EOSC and identify where dialogue and collaboration can be encouraged.

Create a **basis for harmonisation** efforts to bring together the various actors working in the **FAIR ecosystem** and build a **functioning EOSC** and **active community** around EOSC.

FAIRSFAIR

Fostering Fair Data Practices in Europe

# MAIN OUTPUTS

March 2019 - February 2022

## DATA PRACTICES

- **Reports**
  - FAIR requirements for persistence and interoperability
  - Guidelines for ontology design and vocabulary interoperability
  - Basic framework for services enabling FAIR (including software)
- **Solutions for interoperability and machine accessibility for FAIR-aligned repositories**
- **Prototype for interoperability of repositories**
- **Workshops and hackathons: Recommendations for FAIR Semantics and Semantics in FAIR**

## DATA POLICY

- **Reports**
  - Recommendations on data policy and analysis of practice
  - Integration of meta-data catalogues
  - White paper on alignment and synchronisation around FAIR, Open Science and EOSC
- **Support programme for repositories to reach FAIR compliance**

## CERTIFICATION

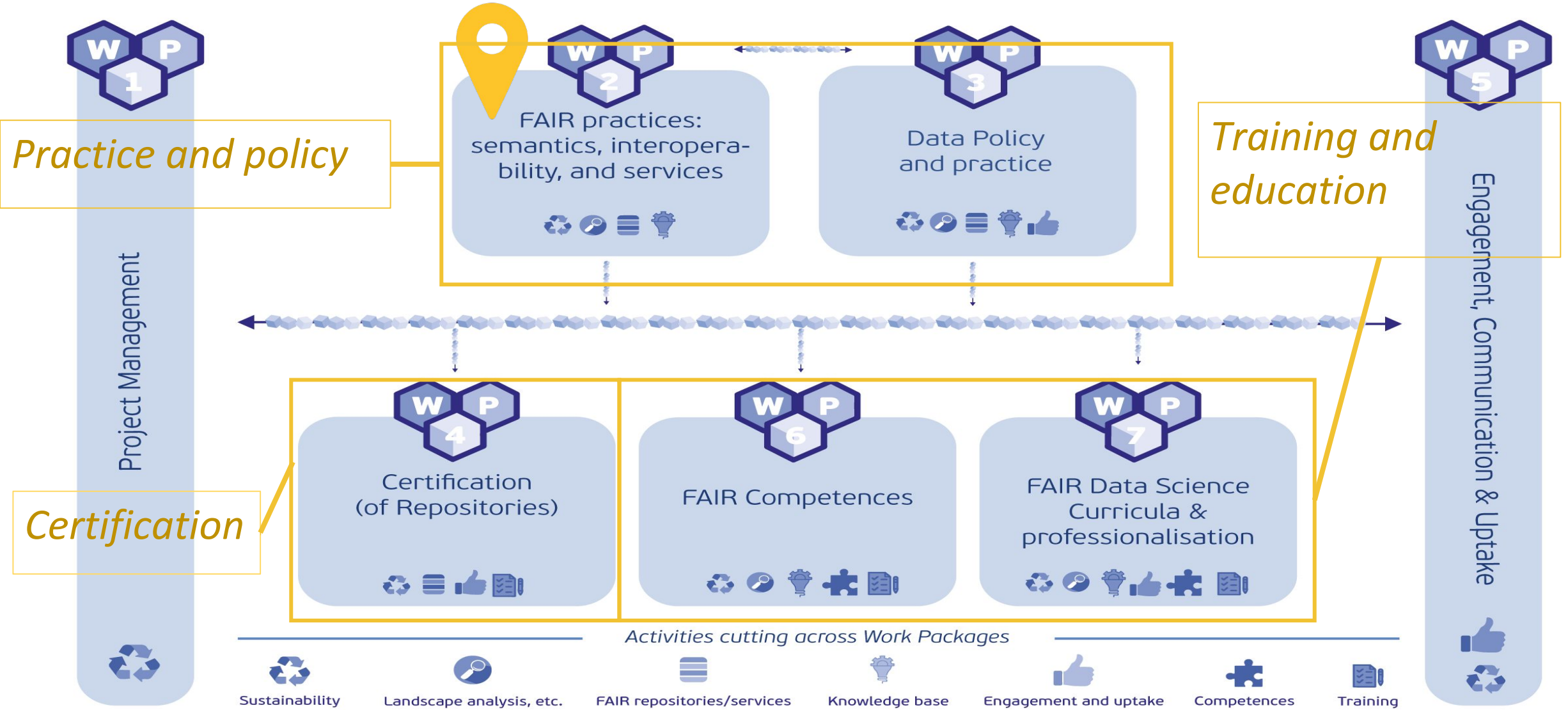
- **European network of trustworthy repositories enabling FAIR data**
- **Support and guidance for certification of data repositories**
- **Tool to identify relevant trustworthy certified repositories**
- **Pilots to support the assessment of FAIR data in trustworthy repositories**

## TRAINING, EDUCATION AND SUPPORT

- **Reports**
  - FAIR data in European higher education
  - Training for researchers in FAIR data science and its impact
- **FAIR competence centres tailored to different communities**
  - Three annual schools in core data skills for researchers
  - Five instructor training (train-the-trainer) events
- **FAIR competence framework for higher education**
  - Three annual FAIR data education stakeholder workshops
- **FAIR competences adoption handbook for universities**
  - Three workshops on integrating FAIR data competences
  - Case studies on good practices in FAIR competences education

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
Fostering Fair Data Practices in Europe



# WP2 FAIR Practices: Semantics,Interoperability and Services

- **November 2019:** D2.1 Report on FAIR requirements for persistence and interoperability 2019 (https://doi.org/10.5281/zenodo.3557381)
- **February 2020**: Assessment report on FAIRness of services (https://doi.org/10.5281/zenodo.3688762)
- **March 2020**: FAIR Semantics: First recommendations (https://doi.org/10.5281/zenodo.3707985)
- **August 2020**: 2nd Report on FAIR requirements for persistence and interoperability (https://doi.org/10.5281/zenodo.4001631)
- **October 2020**: M2.15 Assessment report on 'FAIRness of software' (https://doi.org/10.5281/zenodo.4095092)

Contact: Jessica Parland-von Essen jessica.parland-vonessen@csc.fi

# T2.1: Review of the technical implementation of FAIR principles

- The 1st Report was mainly a **landscaping exercise** which reviewed and documented the **state of FAIR** in the **European scientific data ecosystem.**

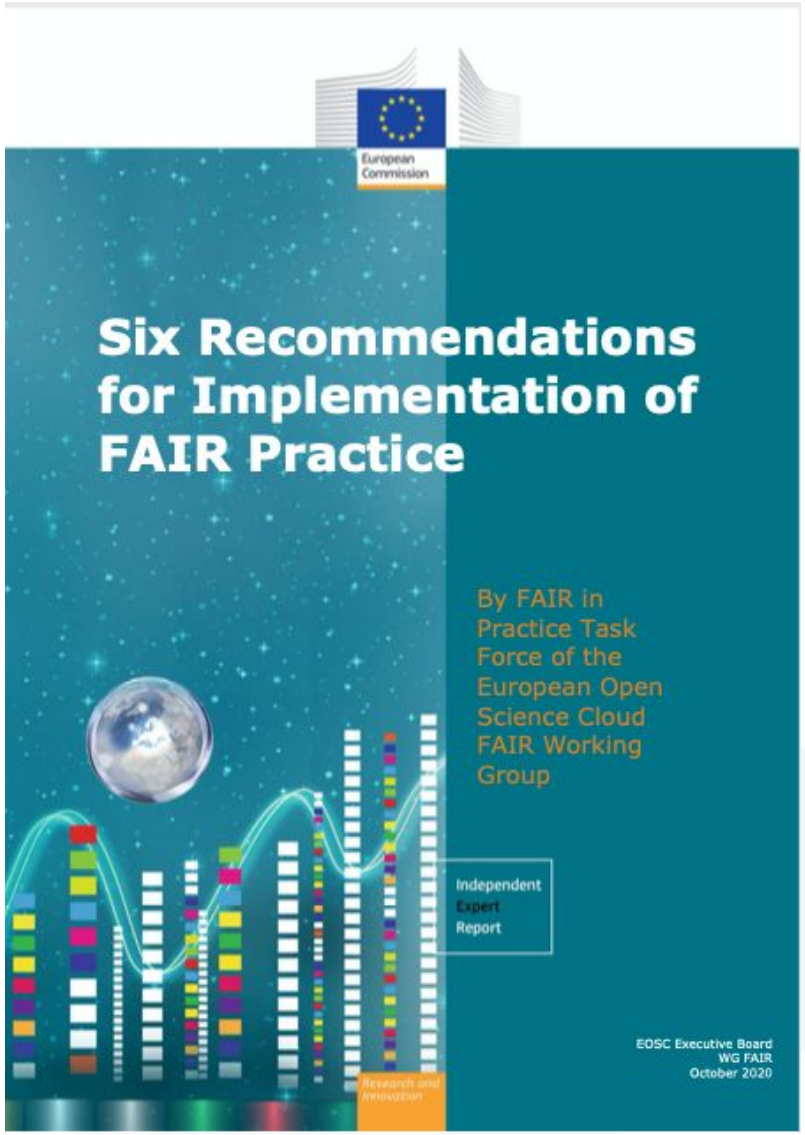
- The report identified commonalities and possible gaps in **semantic interoperability** and the use of **metadata and persistent identifiers** across infrastructures.

- The **2nd Report on FAIR requirements for persistence and interoperability** builds on the first report.

- It's been written specifically for **researchers, data stewards, and service providers,** and is a guide to the use of PIDs, metadata, and semantic interoperability.

# 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



FAIRSFAIR

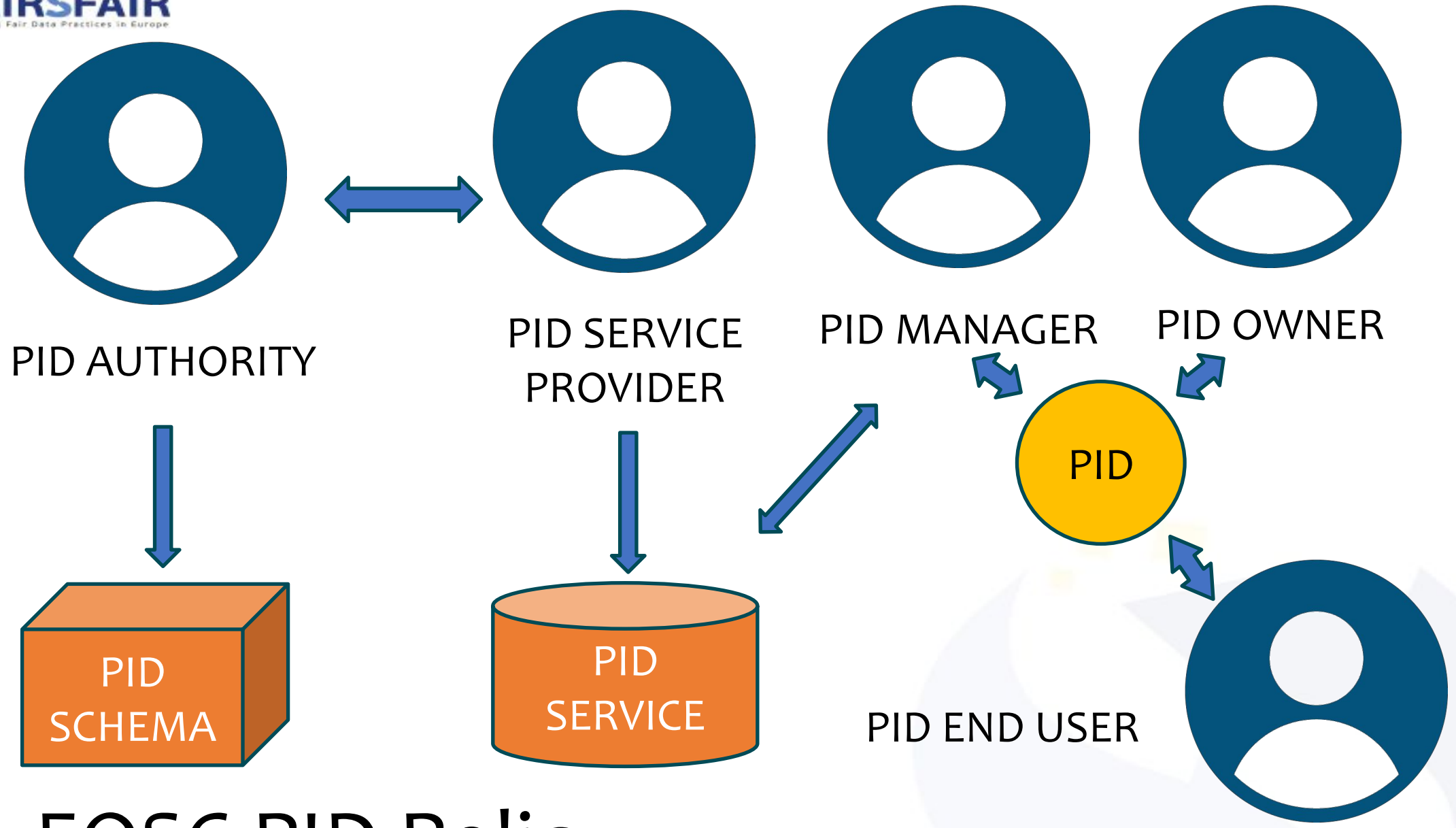
Fostering Fair Data Practices in Europe

# Implementing 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

# Persistent identifiers

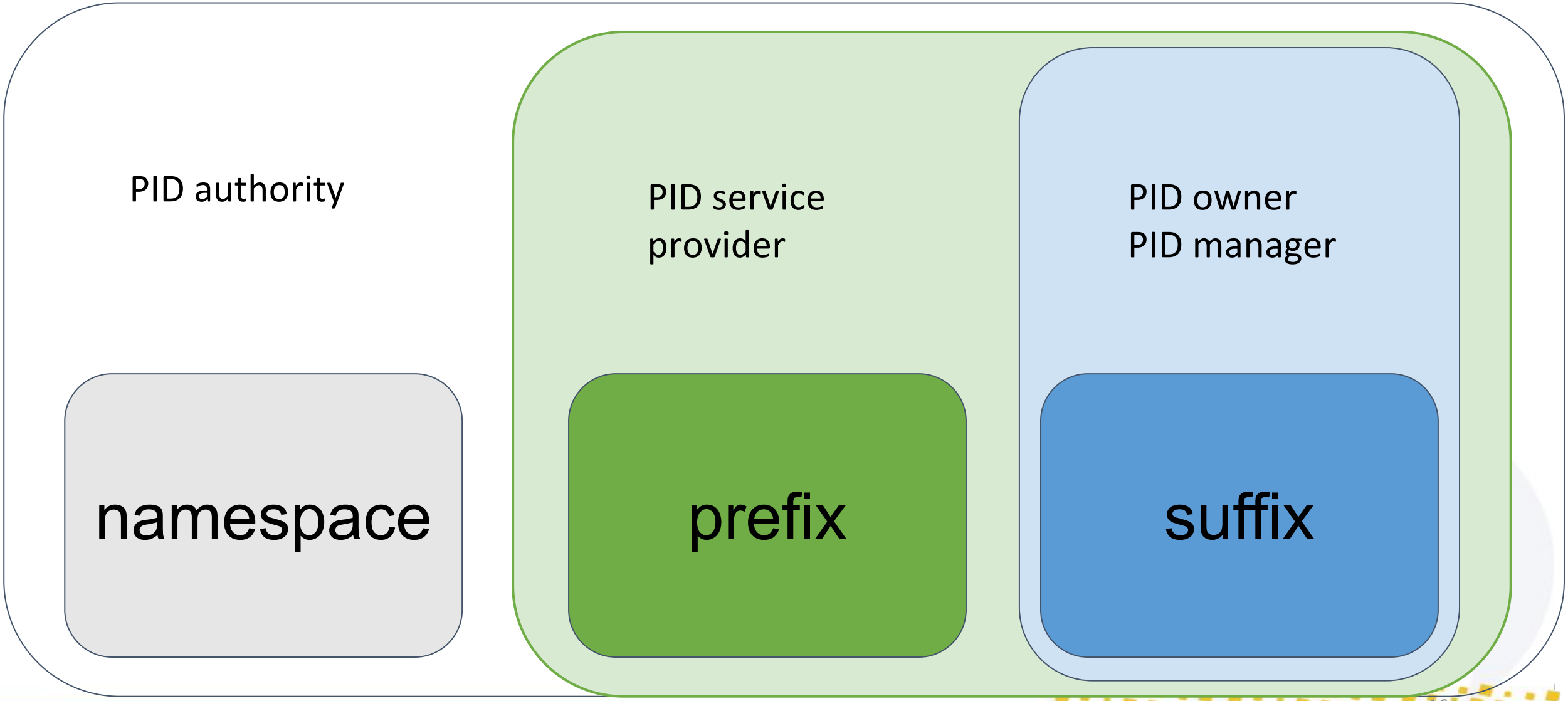
- **Globally unique**, i.e. nobody else in the world should use the same string to refer to anything else
  - a controlled syntax and a governed namespace
  - be issued and managed by a clearly specified registration authority
- **Resolvable,** i.e. provide a way for both machines and humans to access 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




# EOSC PID Policy

https://doi.org/10.5281/zenodo.3574202

# Structure of PIDs and example responsibilities



# THE PROMISE

# THE COST

# Resolving

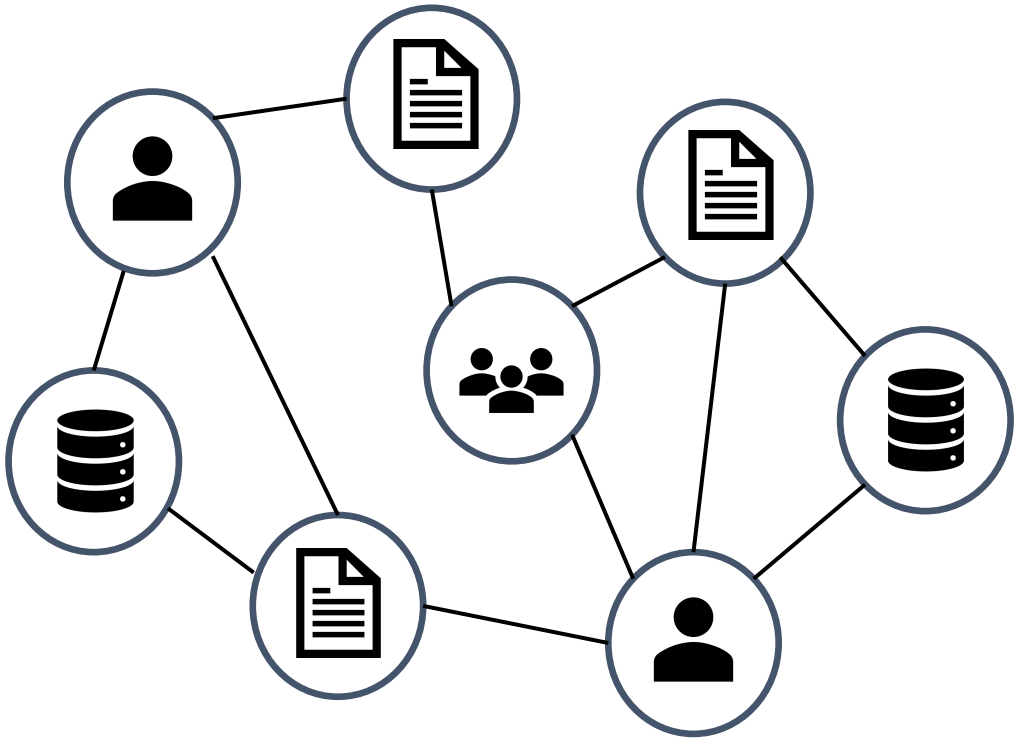
- **Domain Name Service** (DNS) resolver: Resolves a hostname to an IP address.
- **Local resolver**, e.g. load balancer, API gateway or web server: Redirects to a different host and/or path.
- **Full resolver**, e.g. handle system: Redirects to a URL either following a regular expression pattern, or 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

# PID records and metadata

- 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

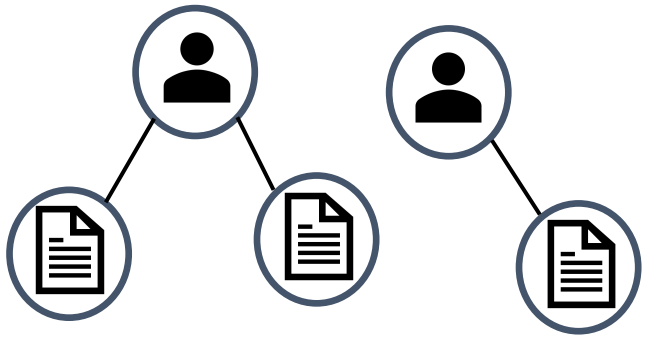
## Research information

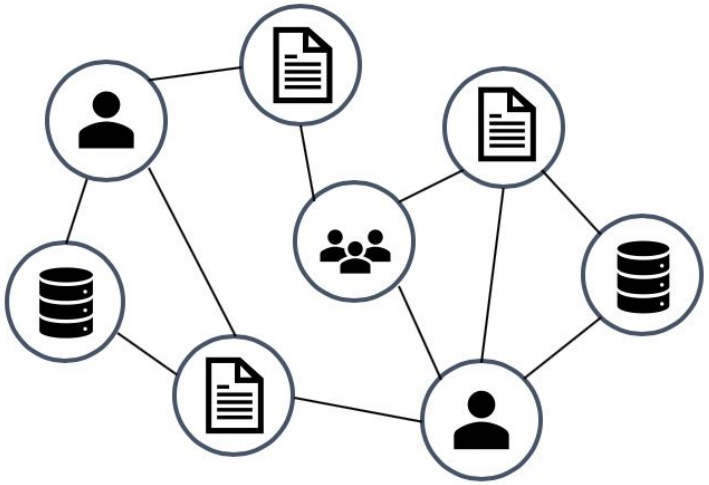


## Research data

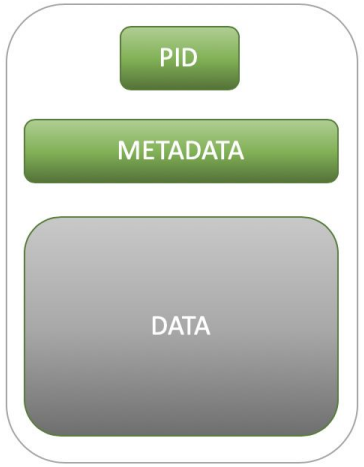
# Shallow FAIR and Deep FAIR

## Research Information

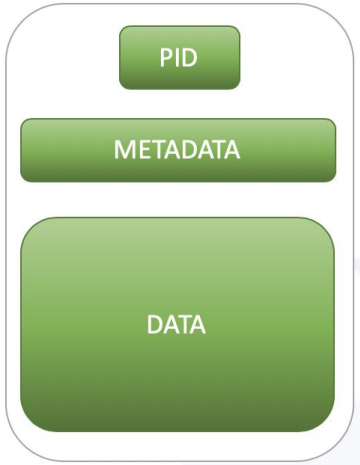




## Research Data




Necessary research information, PIDs, machine readable license

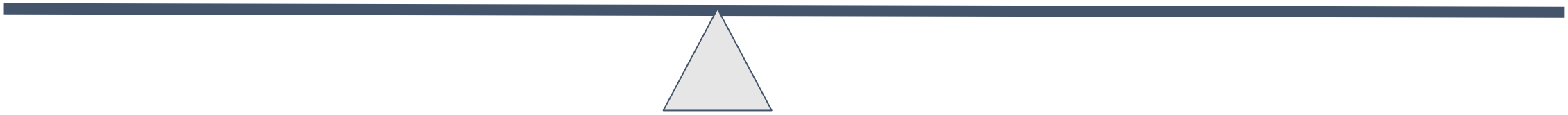



All data elements are machine accessible

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

# Providing PIDs

**designated community** **interoperability across domains**



Ontology

Life cycle

Use cases

# 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.

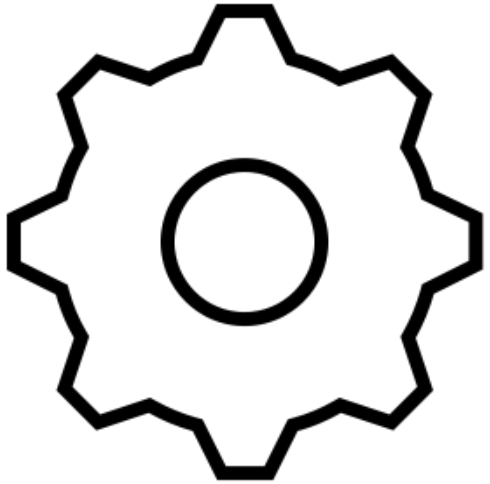
Rauber A, Asmi A, van Uytvanck D, Proell S. Data Citation of Evolving Data : Recommendations of the Working Group on Data Citation (WGDC). Published online October 20, 2015. https://doi.org/doi:10.15497/RDA00016

FAIRSFAIR
Fostering Fair Data Practices in Europe

Open Science

1. Landscape analysis

- Define objects to identify
- Identify relevant practices, standards and regulations
- Map external information systems and services
- Analyse the current state and the target state of your organisation



2. Comparing alternatives

- Assess the different PID systems and services for each object type
- Compare costs
- Plan for the life cycle
- Assess the trustworthiness of the solutions

3. Deployment and maintenance

- Create policies, guidelines and documentation
- Implement and offer training

https://doi.org/10.5281/zenodo.4395767

# Service providers can support FAIR PIDs by

- providing and implementing a clear PID policy
- actively promoting and taking part in PID governance
- assigning and managing PIDs to master data and sharing these
- integrating external PIDs in their information architecture
- integrating external PIDs in the workflow of (meta)data creation
- automating the processes of metadata generation and linking as much as possible in user friendly, yet transparent ways

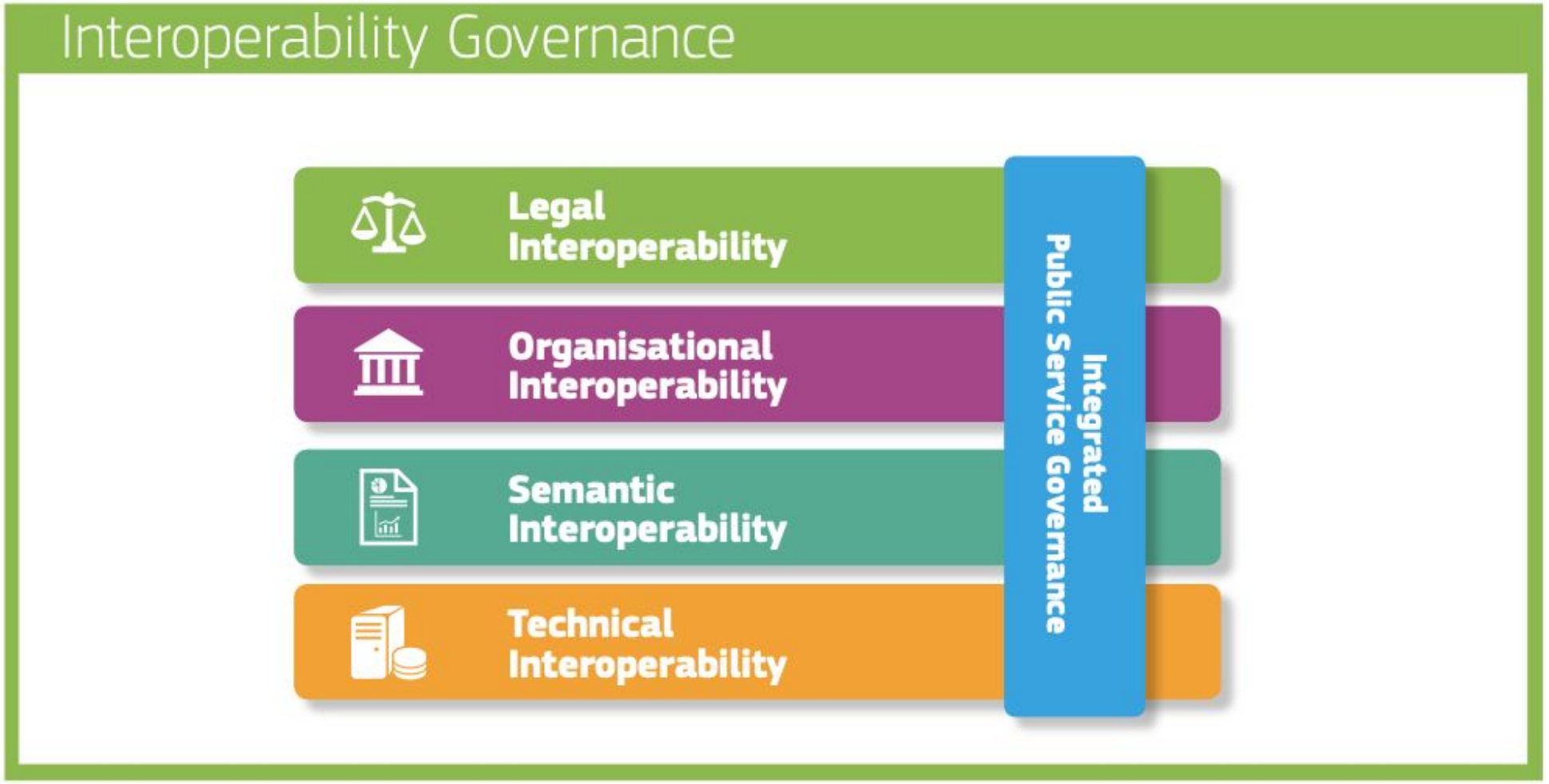
a PID
is a promise

FAIRSFAIR

Fostering Fair Data Practices in Europe

# Semantic Interoperability

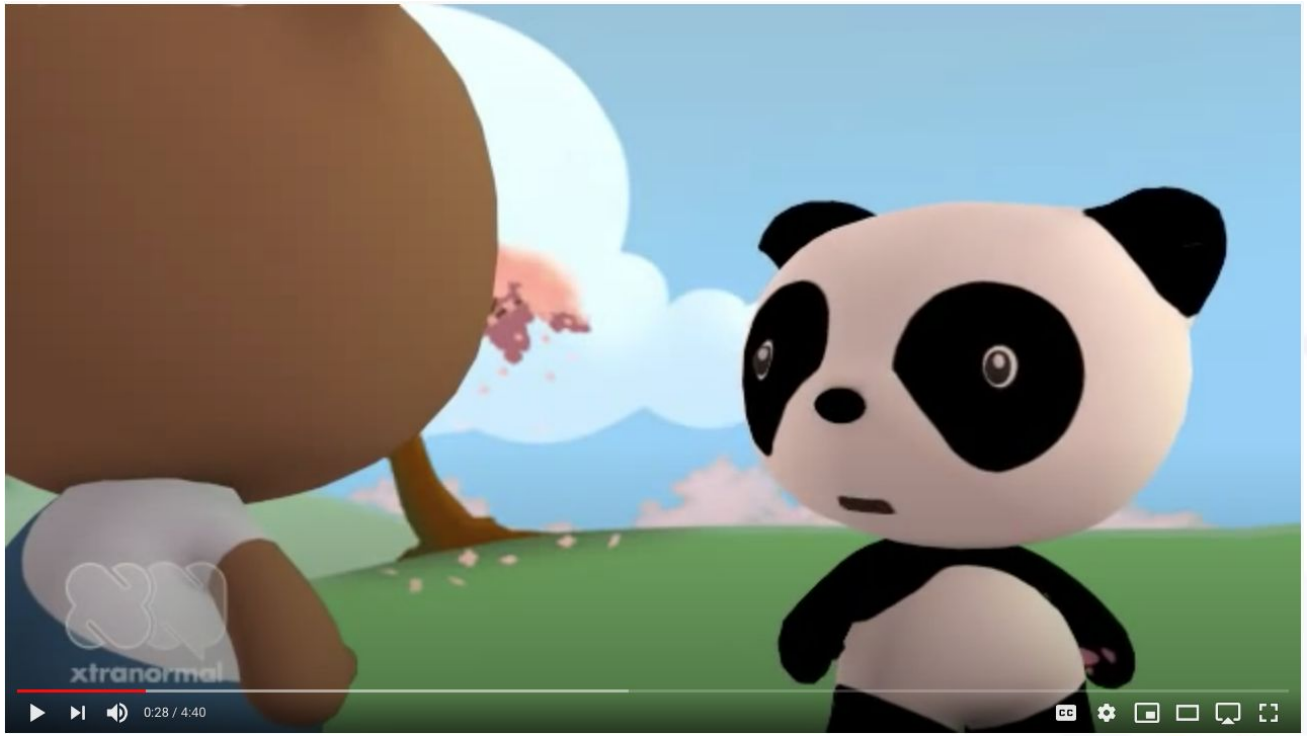
FAIRsFAIR "Fostering FAIR Data Practices in Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 Grant agreement 831558




- 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?

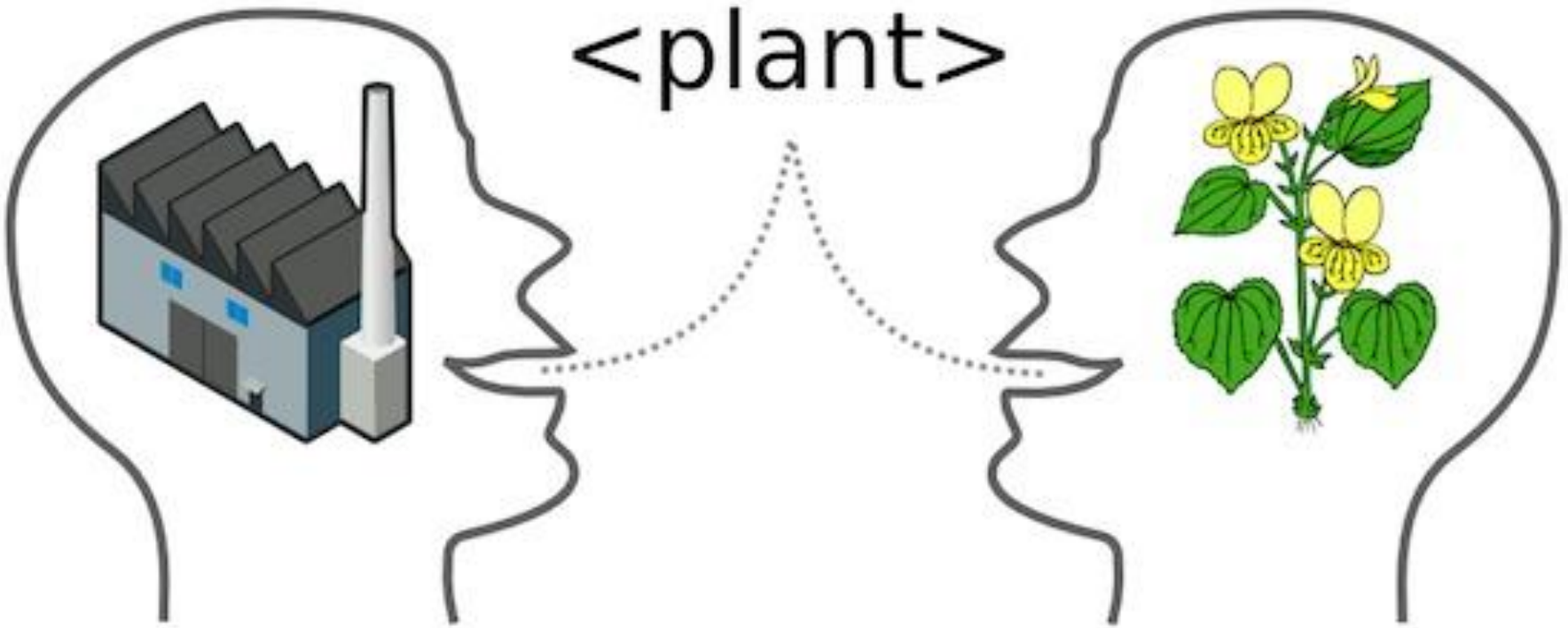
Does using “standard file formats” solve the interoperability problem?

- 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!

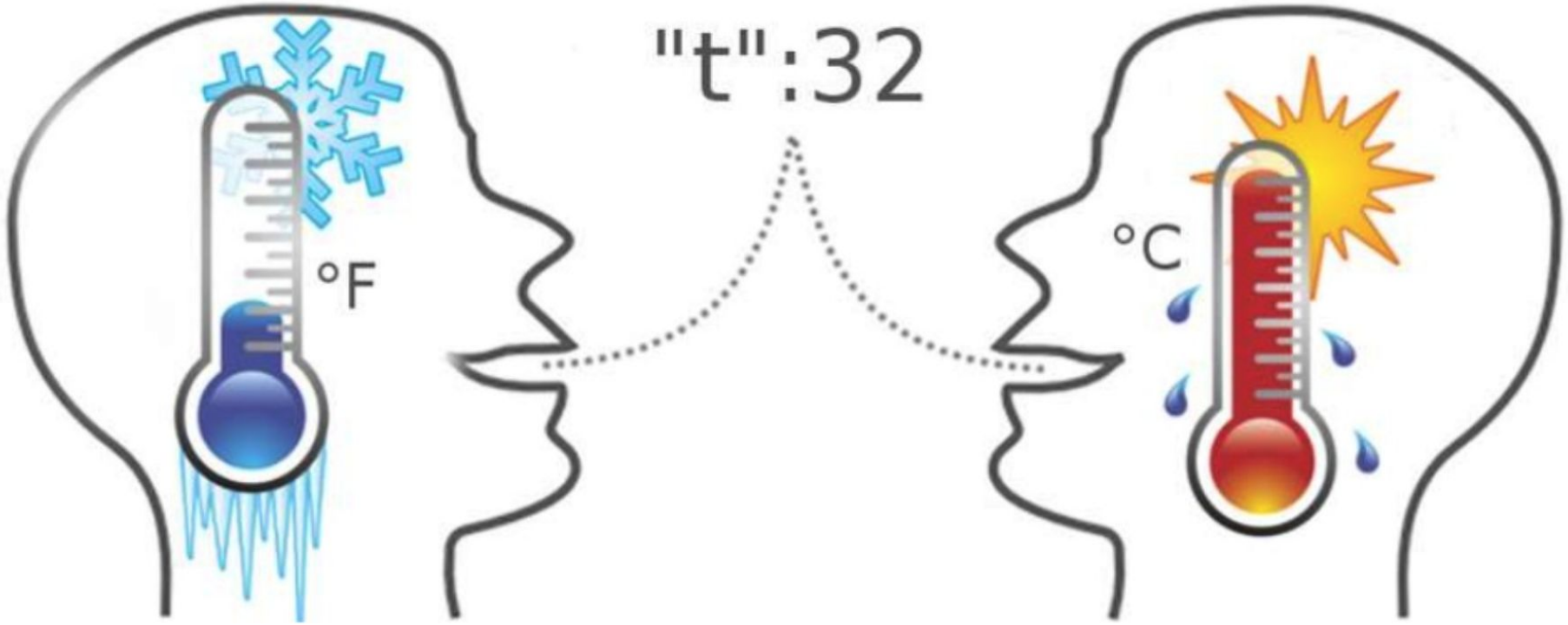
- All content needs to be unambiguous




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”

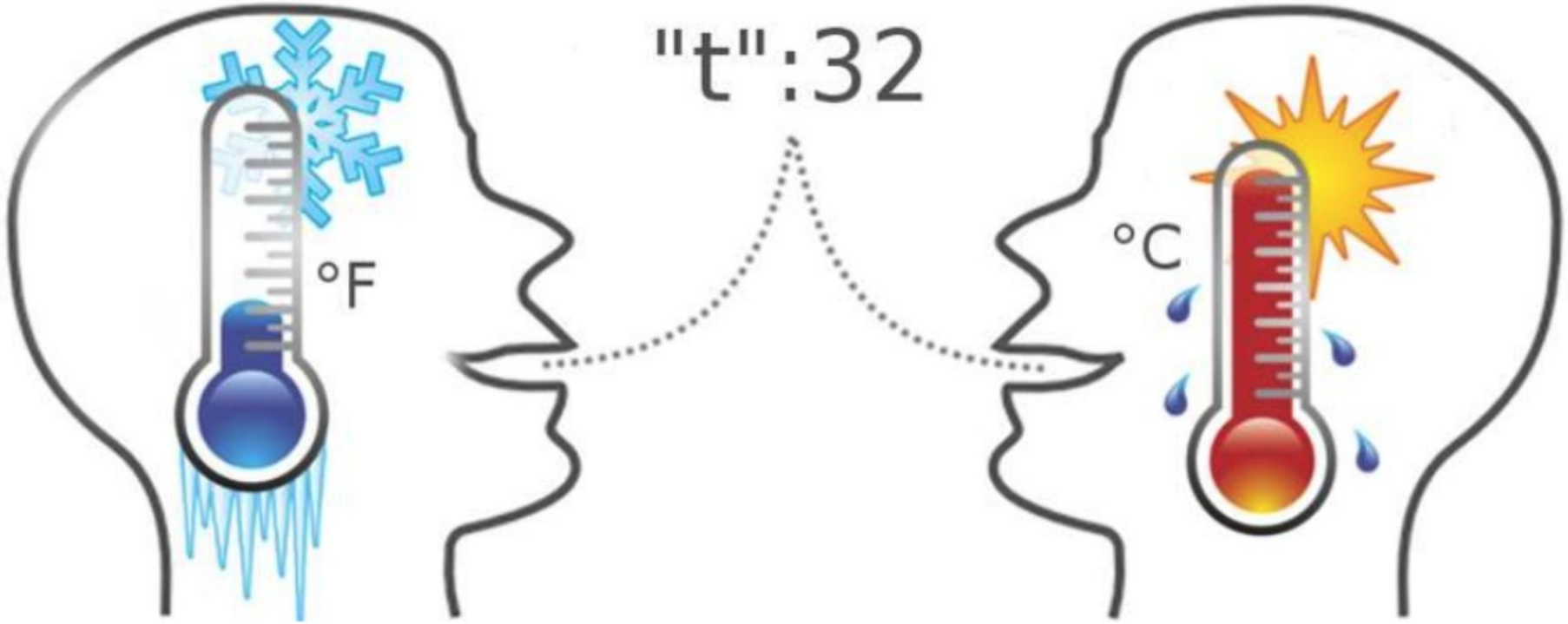
- It should be impossible to misunderstand

# 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?

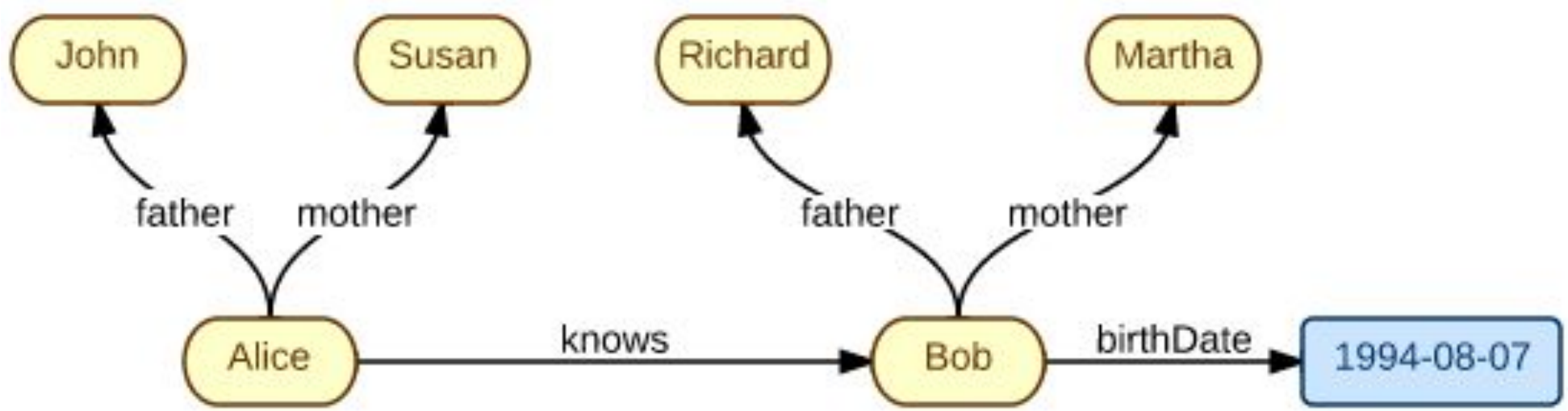



https://www.slideshare.net/maximelefrancois86/reference-knowledge-models-for-smart-application

# Semantic interoperability: How?

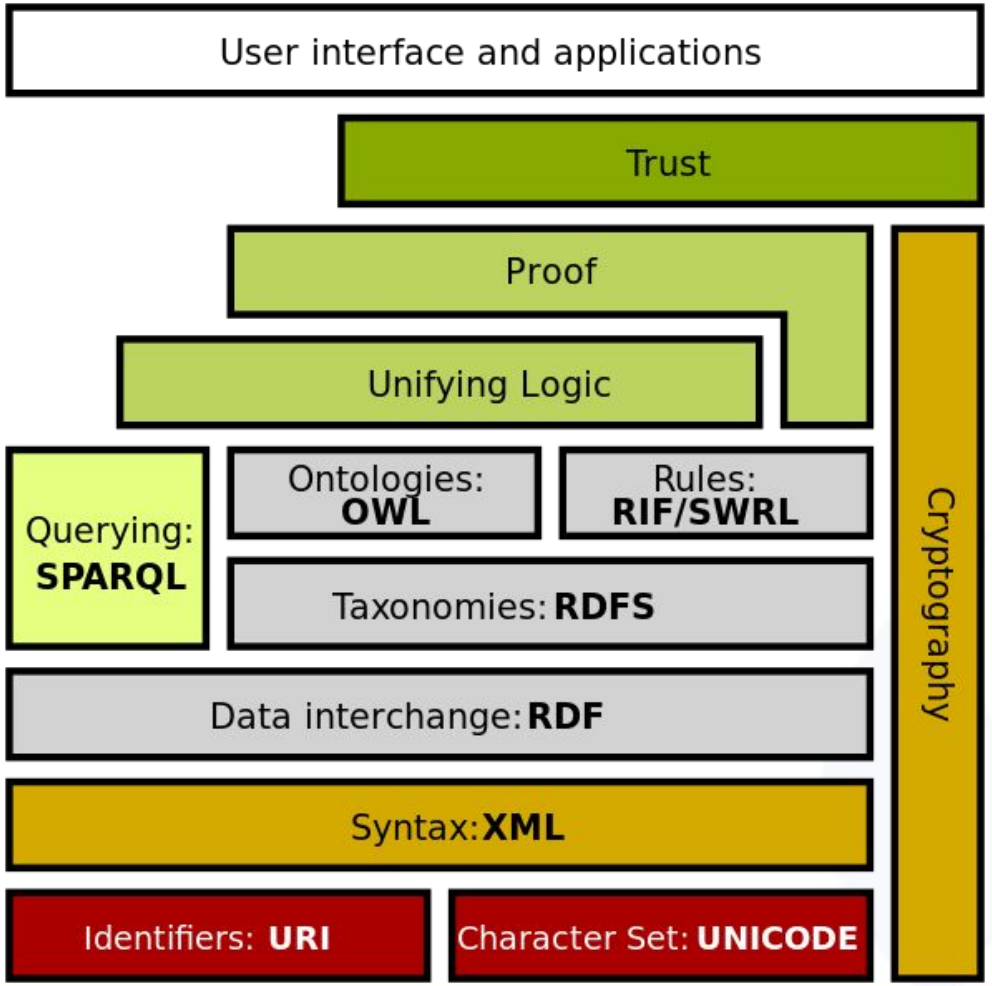
- FAIR principles paper does not make a choice
- “Linked Data” is a very good technical implementation
  - Describe relationships and properties
  - Instead of terms, use “URIs” (or…. PIDs)
  - Very generic, many tools have been developed to work with it

# Relationships and properties




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

# “Semantic Web” stack



# Data Structure

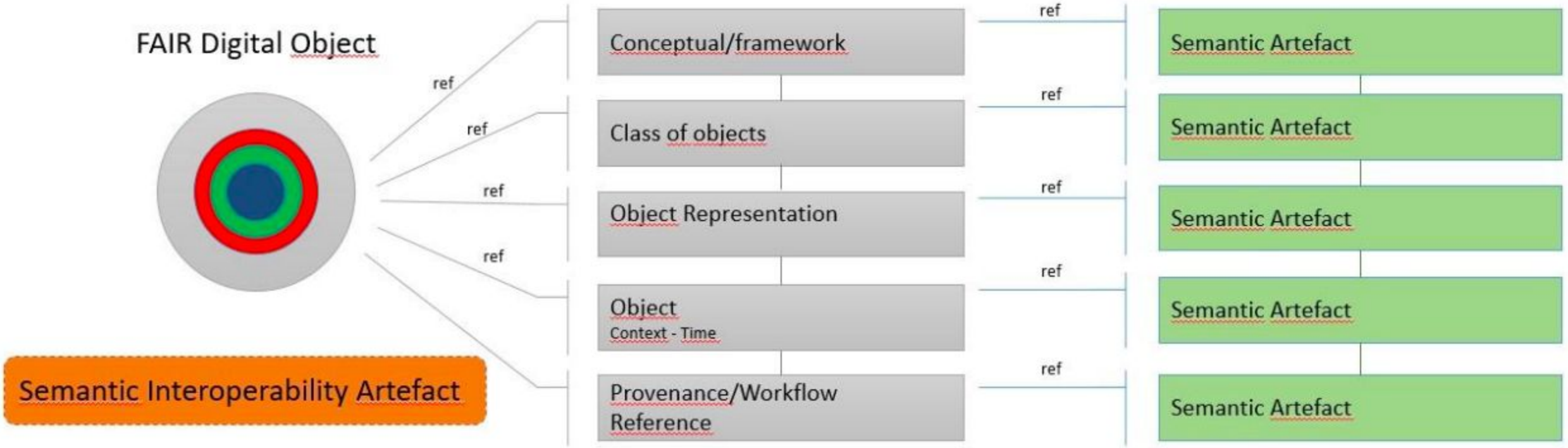
Unstructured data?

# RDF?

Please do not convert all data to RDF!

But make sure it is unambiguously possible.

# Rich semantics also codes relations




The EOSC Interoperability Framework
https://www.eoscsecretariat.eu/sites/default/files/eosc-interoperability-framework-v1.0.pdf

# 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 concepts 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)

# Service providers can support FAIR Interoperability by

- providing and implementing a clear data management policy
- actively promoting and taking part in developing common semantic artefacts and tools
- creating, curating and linking FAIR semantic artefacts and sharing these when possessing authoritative resources
- integrating external semantic artefacts in the information architecture
- integrating semantic artefacts in the workflow of (meta)data creation
- automating the processes of metadata generation using shared and controlled vocabularies

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!

FAIRSFAIR
Fostering Fair Data Practices in Europe

# 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

# Service providers can support metadata by

- providing and implementing a clear metadata policy
- actively promoting and taking part in developing common semantic artefacts and tools
- creating, curating and linking FAIR semantic artefacts and sharing these when possessing authoritative resources
- integrating semantic artefacts in the workflow of (meta)data creation
- automating the processes of metadata generation using shared and controlled vocabularies

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
Fostering Fair Data Practices in Europe

# FAIRsFAIR

FOLLOW US

www.fairsfair.eu

@FAIRsFAIR_eu

www.linkedin.com/company/fairsfair/

www.youtube.com/channel/UCE4wSBnNIBfu6SqIBaIMfNg