

# Data curation and preservation

José María García



UNIVERSITÀ  
DEL SALENTO

International Summer School Data FAIRness. Lecce, Italy. July 2019

# Who am I?



- José María GARCÍA
- Associate Professor at **University of Seville**, Spain
- Research within the Applied Software Engineering (**ISA**) research group on semantic technologies, knowledge modelling, interoperability, cloud service engineering and blockchain
- LifeWatch Metadata & Semantics WG, **LW Spain** JRU
- Contact:



[josemgarcia@us.es](mailto:josemgarcia@us.es)



[@josemgarcia\\_us](https://twitter.com/josemgarcia_us)

# Outline

- Data curation in context
- Data curation activities
- Relevant metadata
- Data preservation
- Achieving digital preservation

# Outline

- Data curation in context
- Data curation activities
- Relevant metadata
- Data preservation
- Achieving digital preservation

# FAIR principles

F<sub>indable</sub>

A<sub>ccessible</sub>

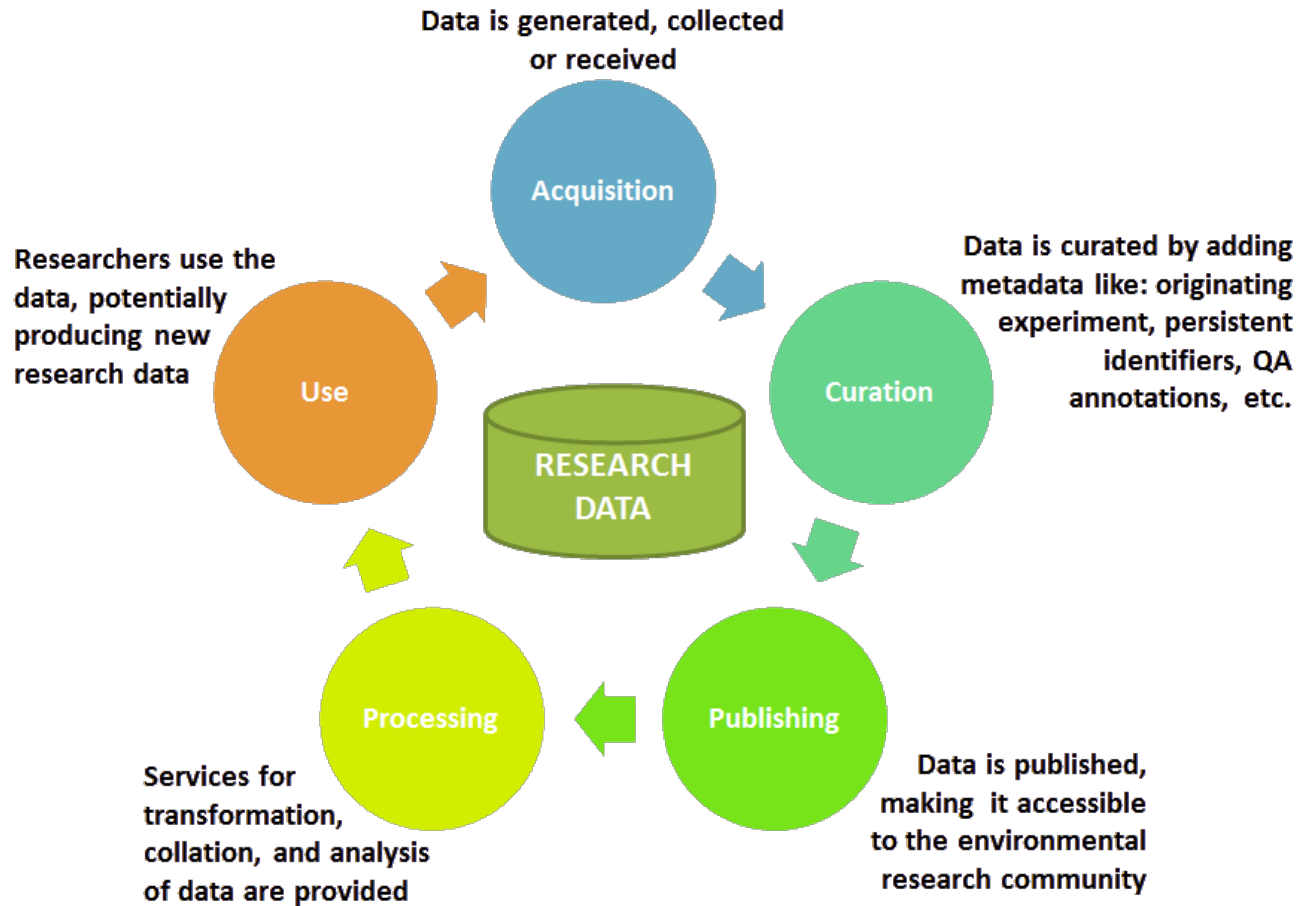
I<sub>nteroperable</sub>

R<sub>eusable</sub>

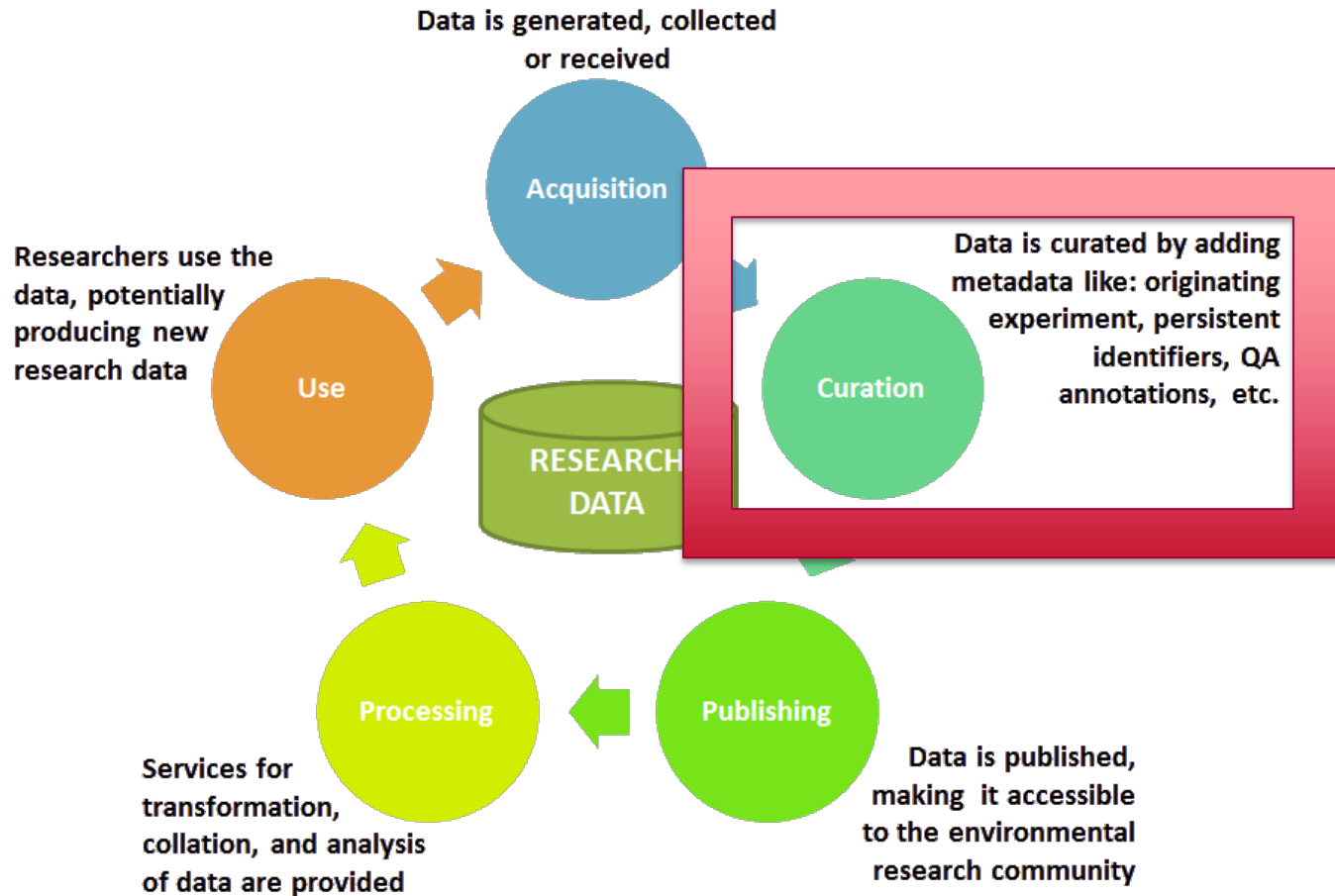


- To be FINDABLE
  - From raw data to information
  - Data needs context
  - Validation and quality assurance
  - Long term preservation

# ENVRI Reference Model



# ENVRI Reference Model



# Data curation

*“all the processes needed for principled and controlled data creation, maintenance, and management, together with the capacity to add **value** to data”*

*Renée J. Miller: “Big Data Curation” (COMAD, 2014)*

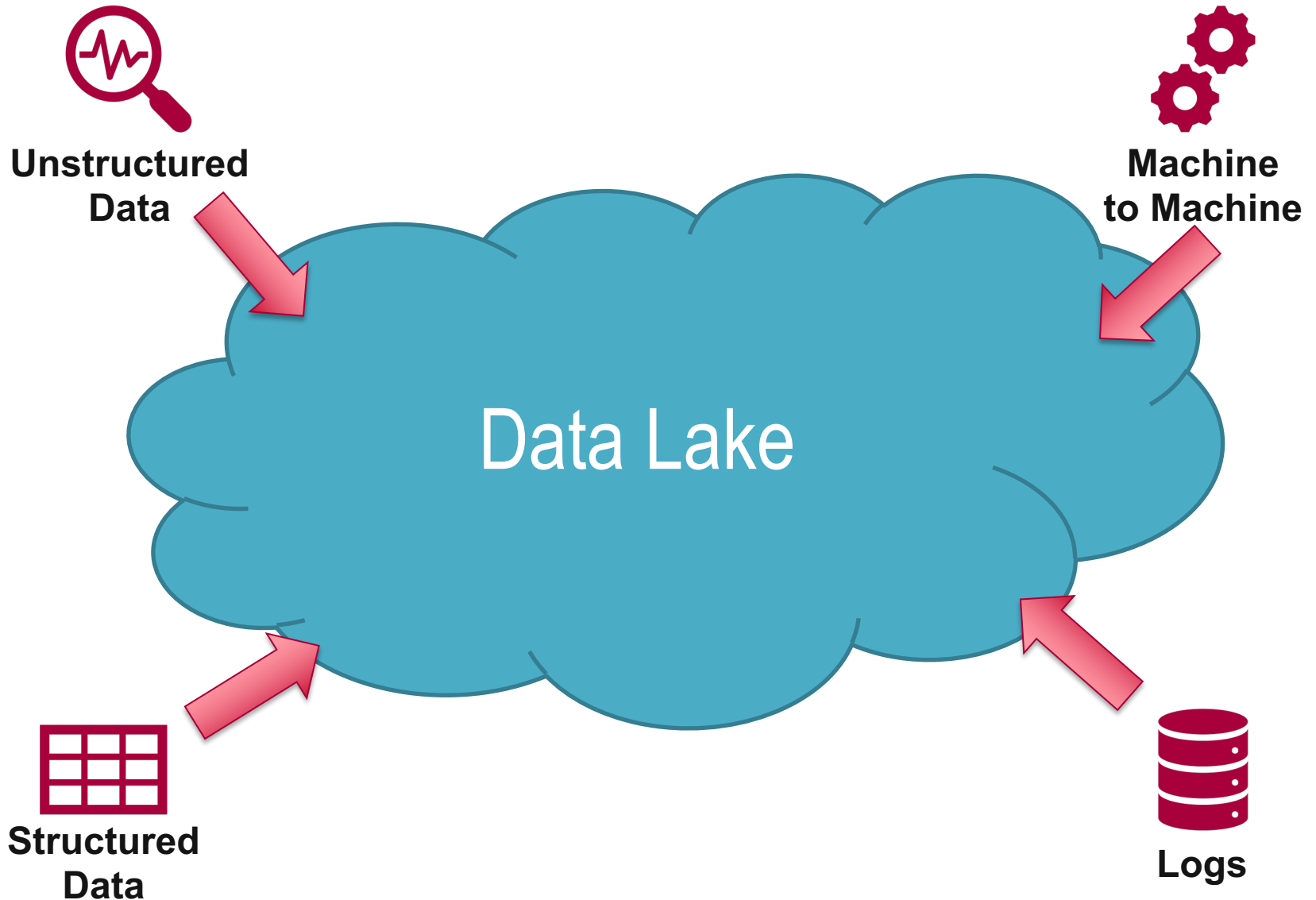
*“Digital curation involves maintaining, preserving and adding **value** to digital research data throughout its **lifecycle**”*

*Digital Curation Centre, UK*

*“Data curation is the active and on-going management of data through its lifecycle and interest and **usefulness** to scholarship, science, and education; curation activities enable **data discovery** and retrieval, maintain quality, add **value**, and provide for re-use over time.”*

*University of Illinois’ Graduate School of Library and Information Science*

# Why data curation is so important?



**Data lakes may become...**



## Avoiding data swamps

Data lakes easily become data dumps (swamps)

Quality of data and analysis compromised

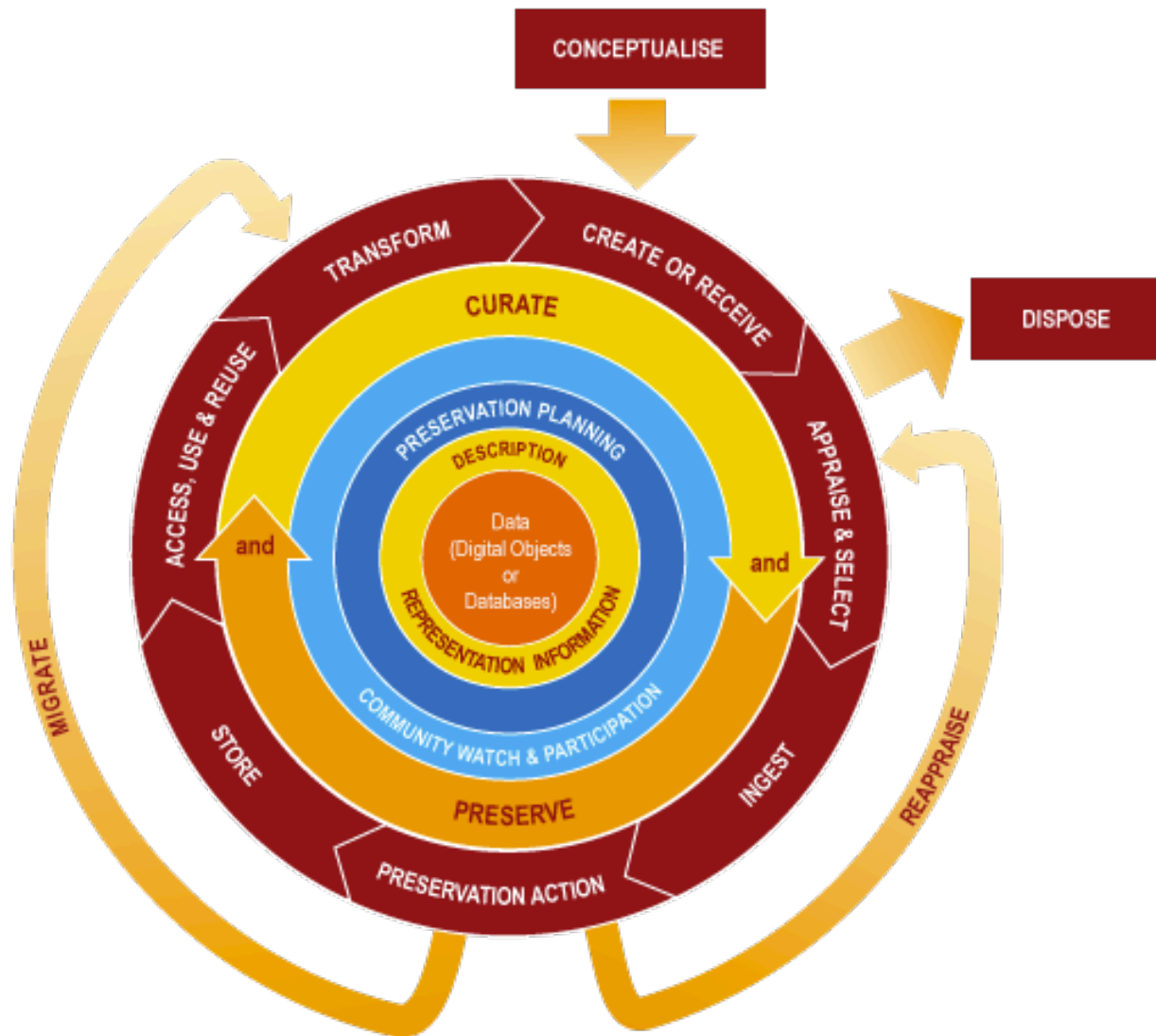
Complexity of raw (big) data needs curation

Add context to access meaningful subsets

# Outline

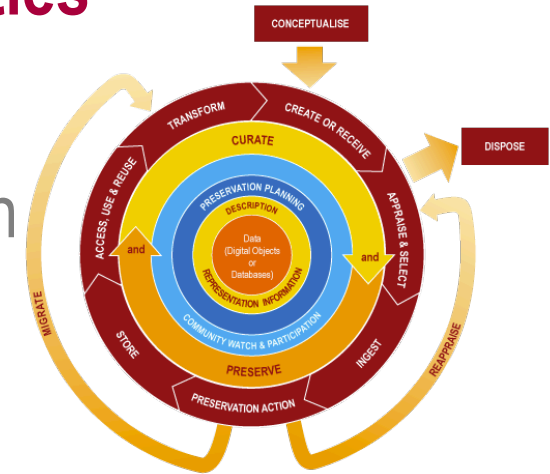
- Data curation in context
- **Data curation activities**
- Relevant metadata
- Data preservation
- Achieving digital preservation

# Data curation lifecycle



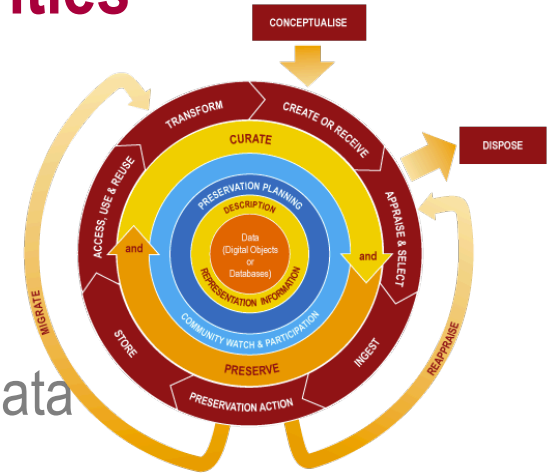
# Data curation generic activities

- Description and representation information
  - Assign metadata
  - Use standards
- Preservation planning
  - DMPs and planning of the curation activities
- Community watch and participation
- Curate and preserve
  - Management of the lifecycle



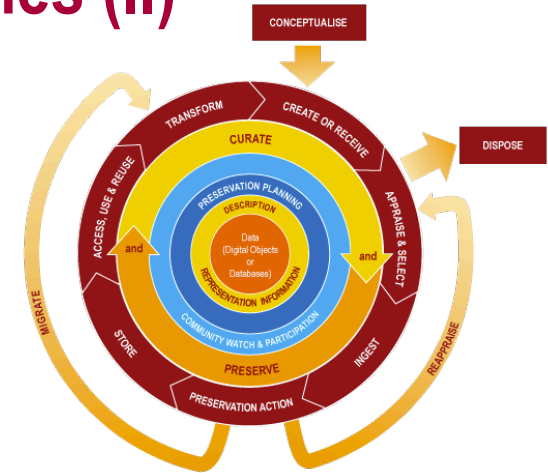
# Data curation sequential activities

- Conceptualize, data acquisition
- Create or Receive
  - Create datasets from the acquired/received data
  - Assign metadata if necessary
- Appraise and Select
  - Data evaluation, quality aspects, adherence to policies
  - Select data for long-term preservation
- Ingest
  - Transfer data to the archive, repository, data centre.
  - According to policies



# Data curation sequential activities (II)

- Preservation action
  - Ensuring authenticity, reliability, usability
  - Cleaning, validation, preservation metadata
- Store in a secure manner
- Access, use and re-use (as in FAIR)
- Transform into new data
- Occasionally:
  - Dispose
  - Reappraise
  - Migrate



# Outline

- Data curation in context
- Data curation activities
- **Relevant metadata**
- Data preservation
- Achieving digital preservation

# Metadata

Data that provides information about other data

Semantic annotations of data (and datasets) to  
provide context

## **5Ws (+1H) metadata can answer**

**Who** created the data? **Who** maintains it?

**When** were the data collected? **When** were they published?

**Where** was it collected?

**What** is the content of the data? **What** is their structure?

**Why** were the data created?

**How** were they produce or analysed?

## Metadata principles (RDA)

Different from data in mode of use

Not just for data, but also for users, services,  
computing resources...

Not just for description and discovery, but also  
for contextualization and interlinking

Must be machine and human understandable

Management (meta)data is also relevant

# Persistent Identifiers

Fundamental for data curation

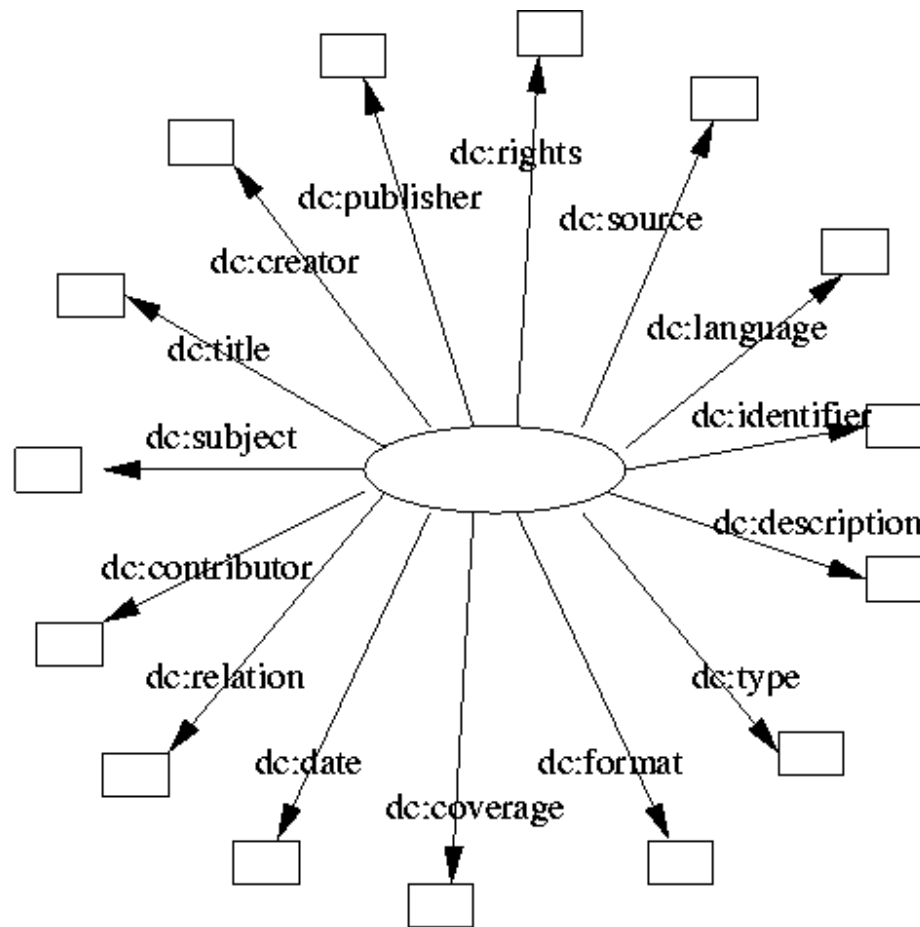
Uniqueness

Reusable between datasets

Helps long-term preservation

URIs, DOIs, ORCID

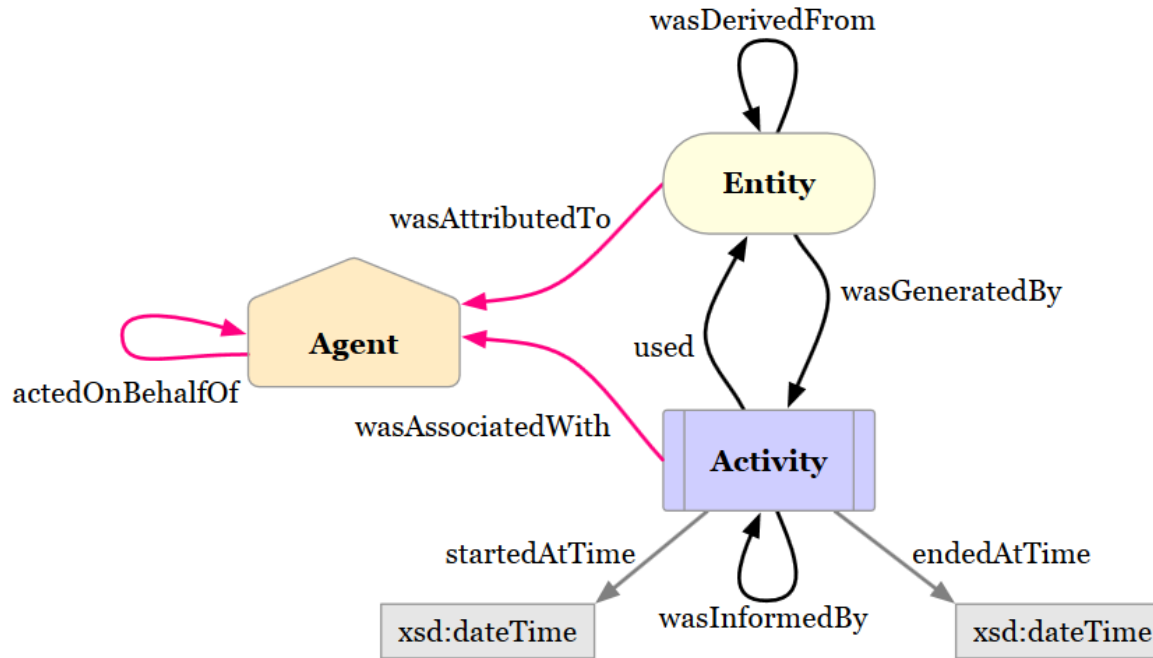
# Generic vocabularies for metadata



Dublin Core

Friend-of-a-friend

# Provenance

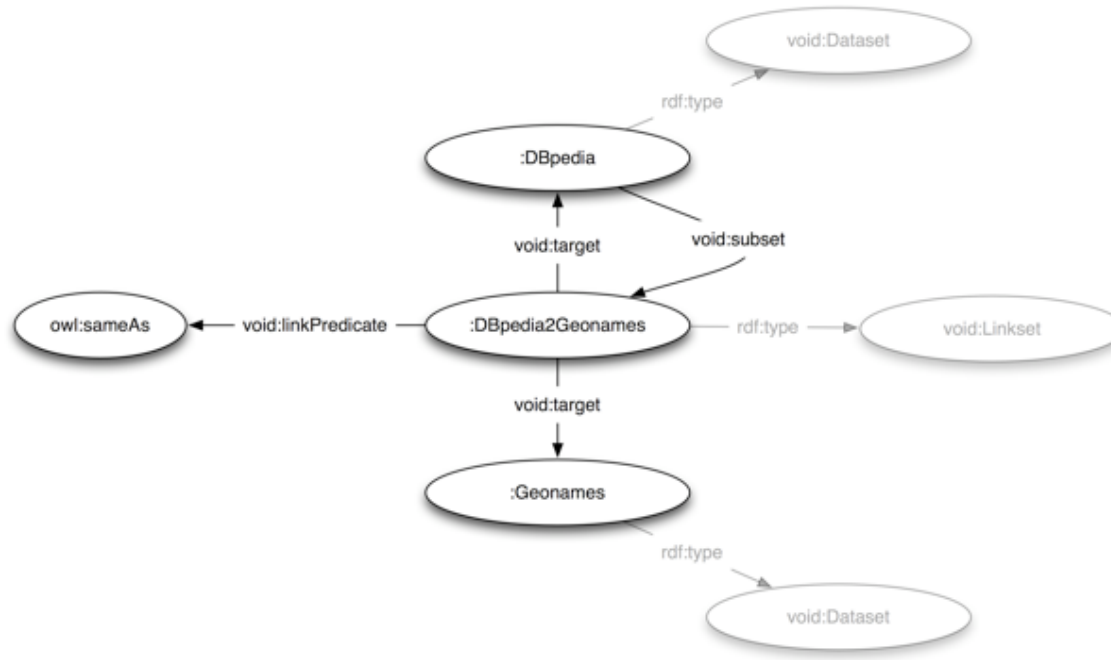


Information about entities, activities and agents involved in producing something

Chronology of ownership, custody, location, transformation

PROV framework from W3C

# Dataset description



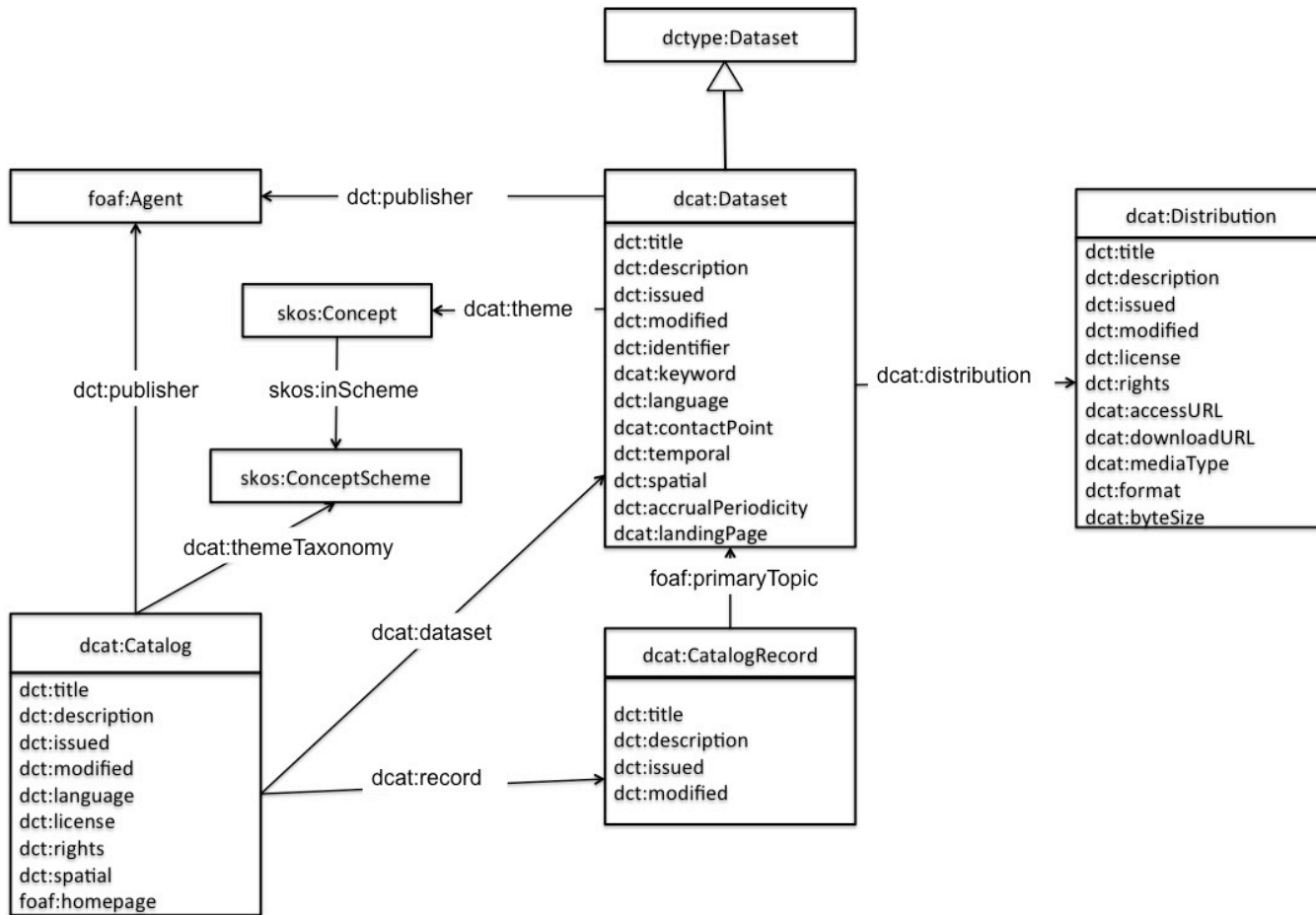
Vocabulary of Interlinked Datasets - VoID (W3C)

General dataset metadata (license, subject, features...)

Access (endpoints, distribution...)

Dataset structure (examples, vocabularies used, links)

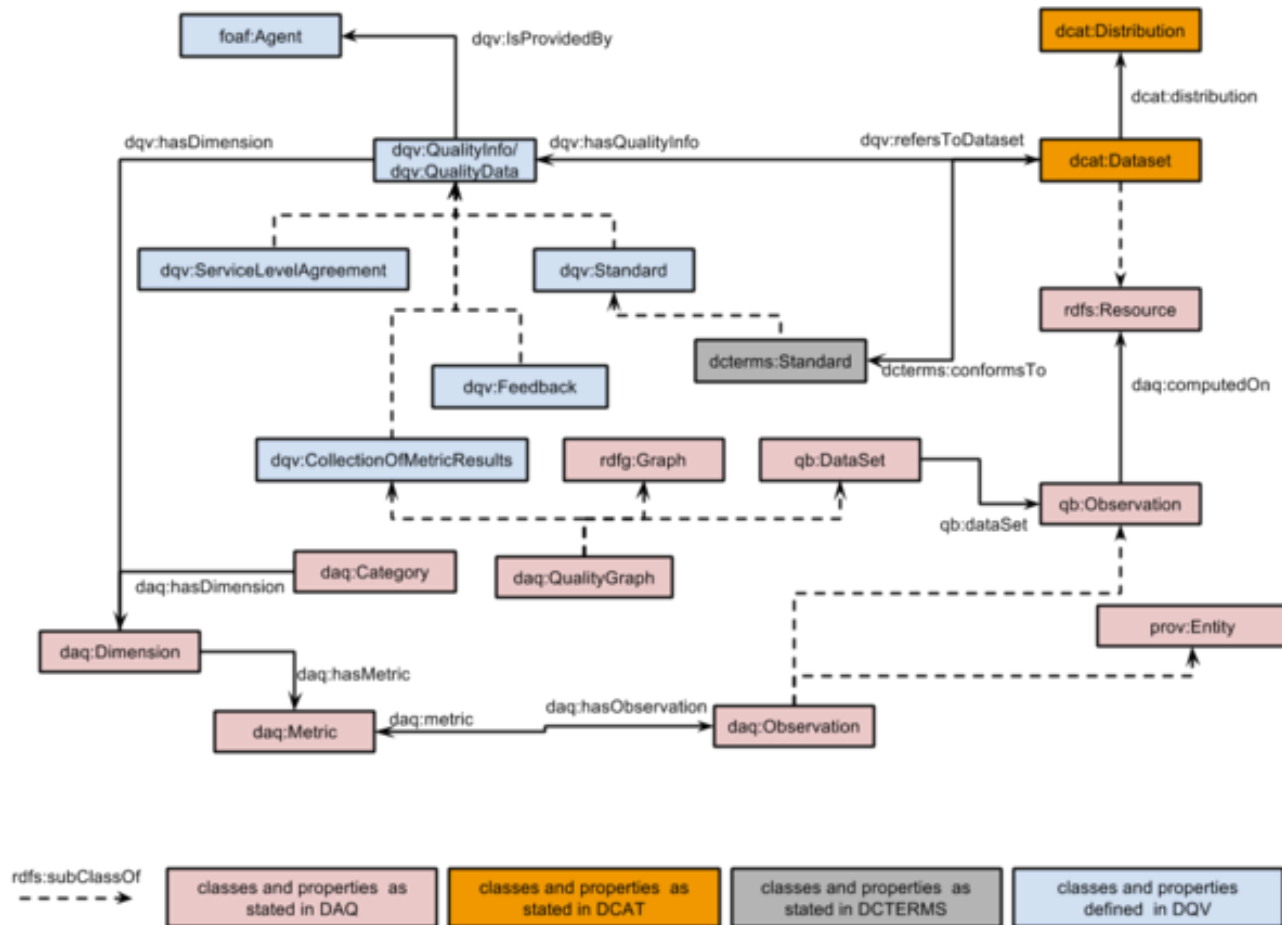
# Data catalogs description



Data Catalog Vocabulary – DCAT (W3C)

Facilitate interoperability between data catalogs

# Quality assurance

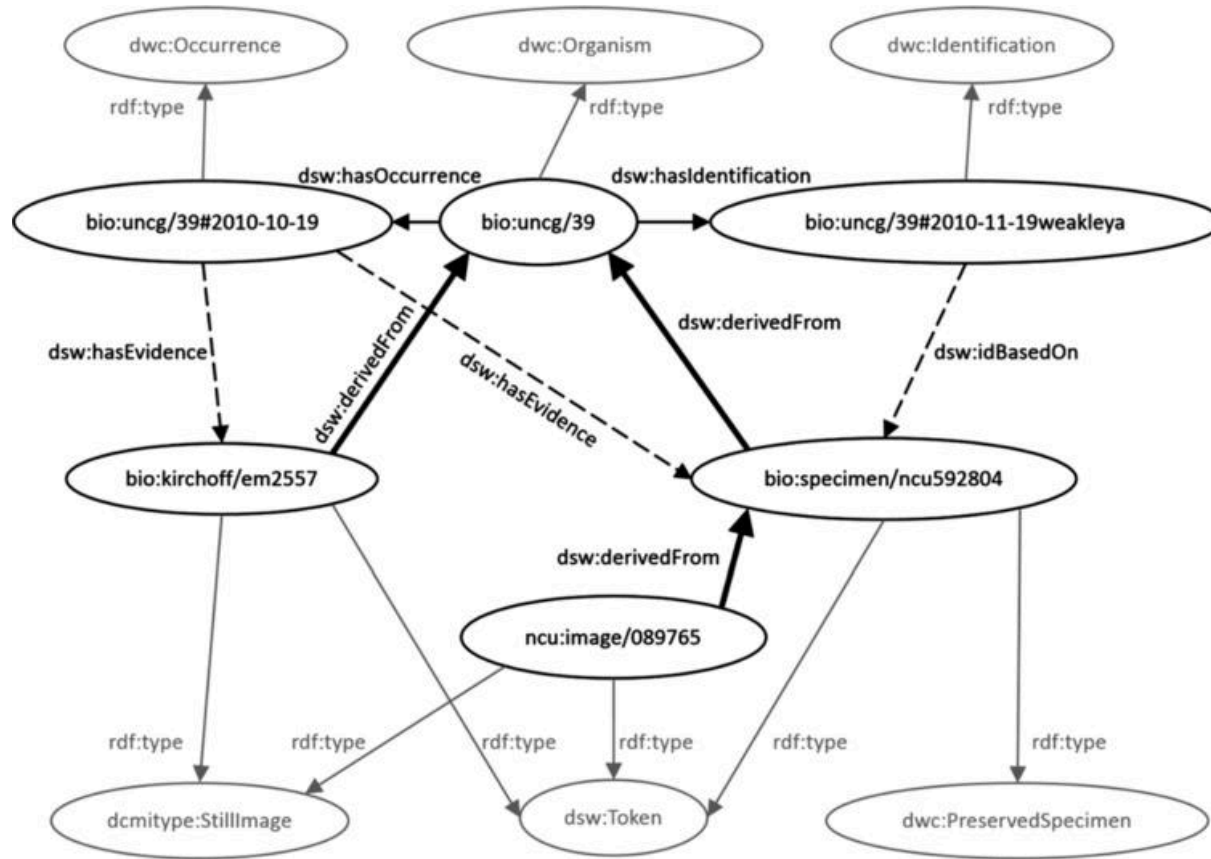


Data Quality Vocabulary (W3C)

Applied to the dataset



# Metadata for environmental sciences and RI

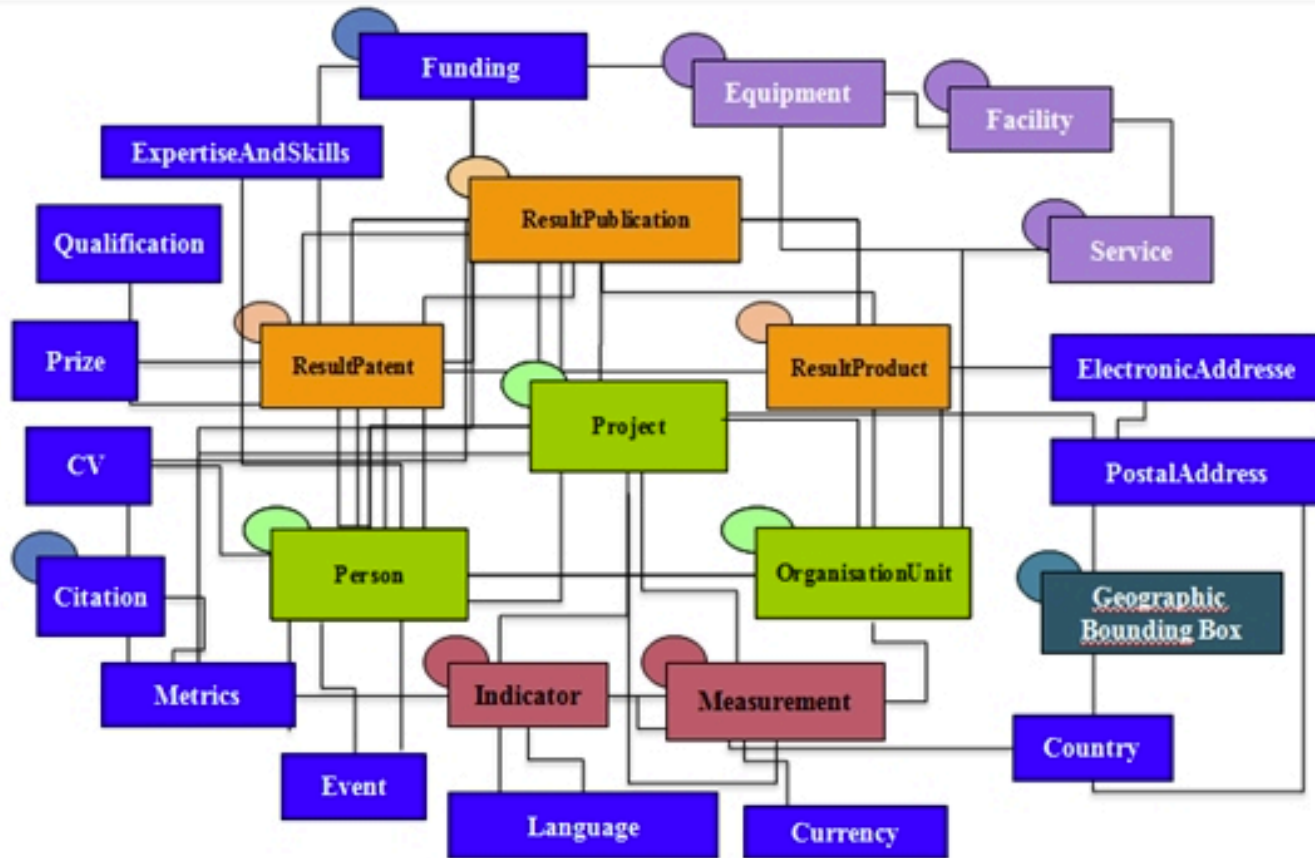


Darwin Core

Ecological Metadata Language (EML)

Environmental Monitoring Facility (EMF)

# Metadata for environmental sciences and RI



Common European Research Information Format (CERIF)

# Tooling support for data curation

Annotation tools

Assign PIDs to every object

Customize the metadata schema

Interlinking support

Versioning

Quality assessment

## Example tools

EDI metadata editor

Morpho data management software

metaphactory

# Outline

- Data curation in context
- Data curation activities
- Relevant metadata
- **Data preservation**
- Achieving digital preservation

# Data preservation

Essential activity in data curation lifecycle

Complementary to curation

Ensuring long-term accessibility and usability

Safety, authenticity, integrity of data and metadata

# Digital Preservation

*“Digital preservation refers to the series of managed activities necessary to ensure continued access to digital objects for as long as necessary”*

*Neil Beagrie and Maggie Jones: “Preservation management of digital materials: The Handbook”  
(Digital Preservation Coalition, 2008)*

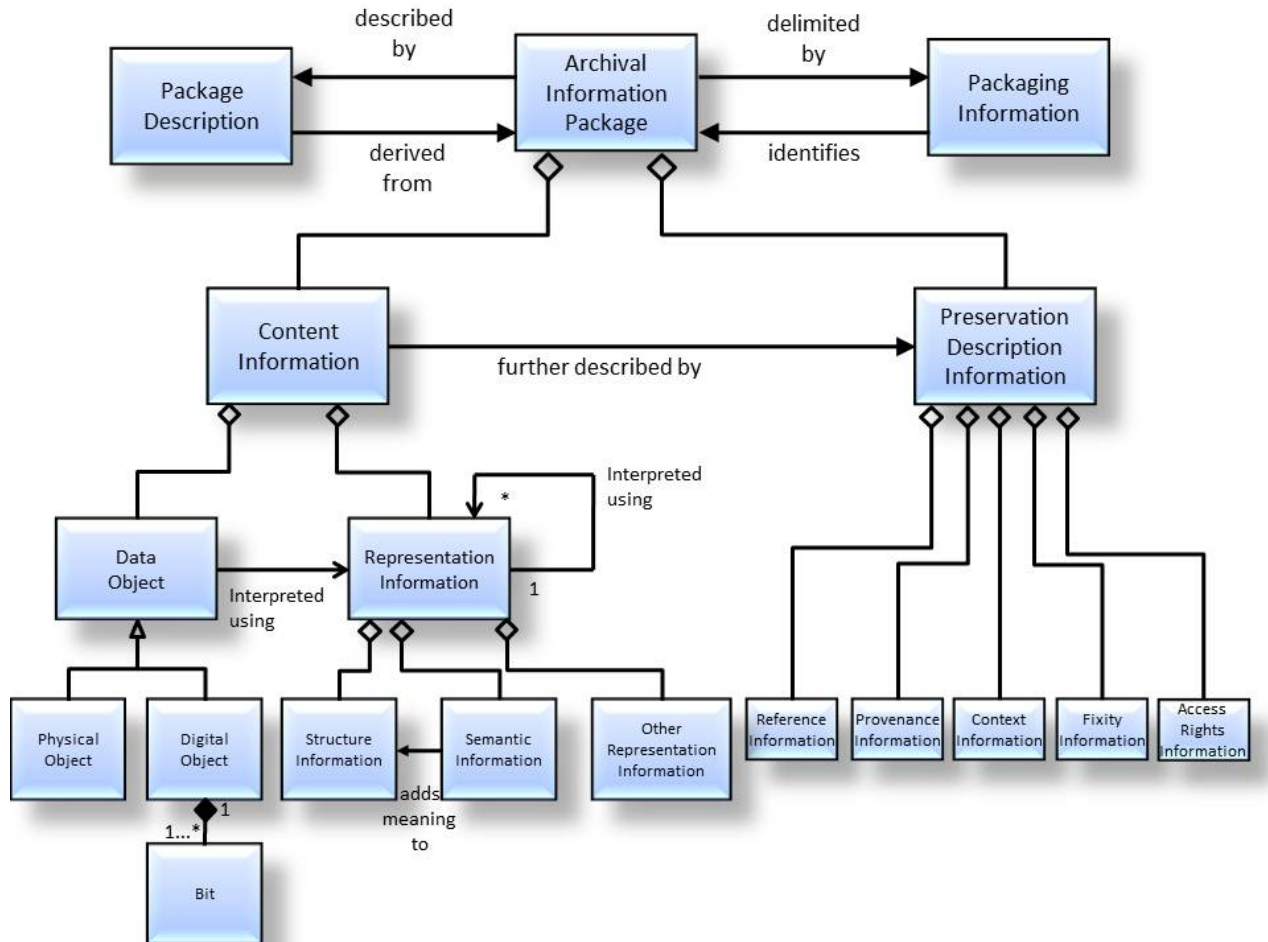
*“The goal of digital preservation is, hence, the accurate rendering of authenticated content over time”*

*Wikipedia*

*“The act of maintaining **information, Independently Understandable** by a **Designated Community**, and with evidence supporting its **Authenticity**, over the **Long Term**”*

*Open Archival Information System (OAIS)*

# OAIS Reference Model



## Usual solutions for digital preservation

Use file formats based on standards

Use services of digital archives to store documents for the long-term

Create and maintain high quality documentation

Use multiple storage facilities (the LOCKSS - Lots Of Copies Keeps Stuff Safe- method)

# Archiving – Internet Archive (archive.org)

The screenshot shows the top navigation bar of the Internet Archive website. It includes the Internet Archive logo, a menu with links for Web, Video, Texts, Audio, Software, About, Account, TVNews, and OpenLibrary, and a search bar. Below the navigation bar is a large banner for the WayBack Machine, featuring the text "INTERNET ARCHIVE WayBackMachine" and a search input field. The banner also displays the statistic "404 Billion web pages saved over time." and a "DONATE" button. Below the banner is a row of ten small thumbnail images representing various archived web pages. At the bottom of the page, there are three sections: "Tools" with a gear icon and links to "Wayback Machine Availability API", "WordPress Broken Link Checker", and "404 Handler for Webmasters"; "Subscription Service" with an Archive-It icon and text about capturing and searching digital content; and "Save Page Now" with a bookmark icon, a search input field, and a "SAVE PAGE" button. The footer contains links for "FAQ", "Contact Us", and "Terms of Use (10 Mar 2001)".

INTERNET ARCHIVE  
WayBackMachine

404 Billion web pages saved over time. [DONATE](#)

**Tools**  
[Wayback Machine Availability API](#)  
Build your own tools.  
[WordPress Broken Link Checker](#)  
Banish broken links from your blog.  
[404 Handler for Webmasters](#)  
Help users get where they were going.

**Subscription Service**  
Archive-It enables you to capture, manage and search collections of digital content without any technical expertise or hosting facilities. [Visit Archive-It to build and browse the collections.](#)

**Save Page Now**  
 [SAVE PAGE](#)  
Capture a web page as it appears now for use as a trusted citation in the future.  
Only available for sites that allow crawlers.

[FAQ](#) | [Contact Us](#) | [Terms of Use \(10 Mar 2001\)](#)

# Services – Research Infrastructures, EOSC



[About](#) [Governance](#) [Services & Resources](#) [Policy](#) [EOSC in Practice](#) [Media](#) [For providers](#) [Search](#)

## ACCESS EOSC SERVICES & RESOURCES



NETWORKING



COMPUTE



STORAGE



SHARING & DISCOVERY



DATA MANAGEMENT



PROCESSING & ANALYSIS



SECURITY & OPERATIONS



TRAINING & SUPPORT



# Repositories – Zenodo

zenodo   [Upload](#) [Communities](#) [Log in](#) [Sign up](#)

## Recent uploads

July 1, 2019 (v1.0) [Dataset](#) [Open Access](#)

[View](#)

### Variant Data from Pooled Sequencing of Hybrid Kiwifruit

[McCallum, John](#); [Thomson, Susan](#); [Seal, Alan](#)

Variant data from pooled sequencing of hybrid Actinidia families segregating for fruit size and Vitamin C Content.

Uploaded on June 27, 2019

*1 more version(s) exist for this record*

January 11, 2015 (v2) [Dataset](#) [Open Access](#)

[View](#)

### POPC/Cholesterol @ 310K. 0, 10, 40, 50 and 60 mol-% cholesterol. Model by Maciejewski and Rog

[Javanainen, Matti](#); [Kulig, Waldemar](#)

Input parameter files and parts of the resulting trajectories for a POPC bilayer simulation with varying amounts of cholesterol. The systems contain 0, 10, 40, 50 or 60 mol-% cholesterol and a total of 128 lipids (POPC+cholesterol). The force field is an extension to the OPLS-based force field...

Uploaded on June 21, 2019

*1 more version(s) exist for this record*

April 9, 2019 (v2) [Dataset](#) [Open Access](#)

[View](#)

### Real-time optical and electronic sensing with a $\beta$ -amino enone linked, triazine-containing 2D covalent organic framework

[Kulkarni, Ranjit](#); [Noda, Yu](#); [Barange, Deepak K.](#); [Kochergin, Yaroslav S.](#); [Balcarova, Barbora](#); [Lyu, Pengbo](#); [Nachtigal, Petr](#); [Bojds, Michael J.](#)

## Zenodo now supports usage statistics!



[Read more](#) about it, in our newest blog post.

## Using GitHub?



Just [Log in](#) with your GitHub account and [click here](#) to start preserving your repositories.

## Zenodo in a nutshell

- **Research. Shared.** — all research outputs from across all fields of research are welcome! Sciences and Humanities, really!
- **Citeable. Discoverable.** — uploads gets a Digital Object Identifier (DOI) to make them easily and uniquely citeable.
- **Communities** — create and curate your own community for a workshop, project, department, journal, into which you can accept or reject uploads. Your own complete digital repository!
- **Funding** — identify grants, integrated in reporting lines for research funded by the European Commission via OpenAIRE.

# Zenodo policies for data preservation

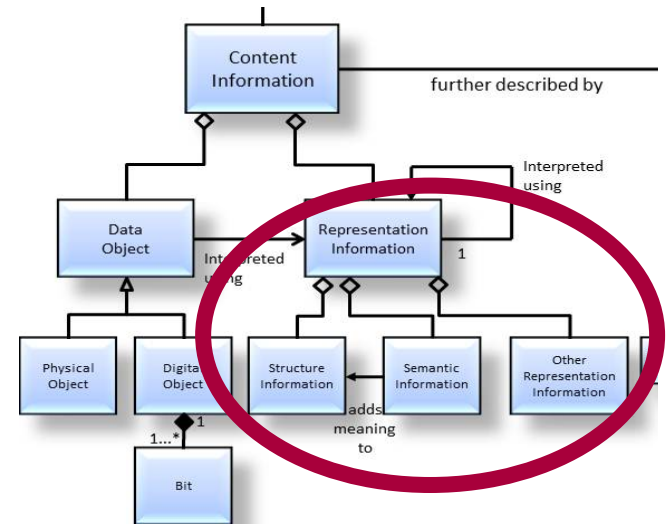
- Data files are **versioned** (not at records level) and preserved
- **Ingestion** as Submission Information Package
- **Replicas** in CERN data centres
- **Retention period** as long as CERN survives (+20 years)
- **Functionality** is **NOT** preserved
- **File preservation** with nightly backups
- **Fixity and authenticity** regularly checked against checksums
- **Succession plans** in case of closure of the repository

# Outline

- Data curation in context
- Data curation activities
- Relevant metadata
- Data preservation
- **Achieving digital preservation**

# Digital Preservation – topics relevant

- Object classification and validation
  - Rendered vs non-rendered
  - Complex vs simple
  - Dynamic vs static
  - Active vs passive
- Representation information (representation network)
  - The information model is key
  - Recursion ends at KNOWLEDGE BASE of the DESIGNATED COMMUNITY (this knowledge will change over time and region)
- Persistent identifiers
- Audit & Certification / Trustworthy Digital Repositories



# Choices for data preservation

- Particularities of dataset formats
  - CSVs can be treated as databases
  - What about Linked Data?
- Do we need to preserve functionality over data?
- Consider data evolution
  - Versioning
  - Other datasets directly or indirectly connected
- Linked datasets
  - Stakeholders rights
  - Ownership of interlinked datasets
  - Ownership of archived versions

# Achieving digital preservation of data

- Selection of data sources
  - Which data sources should be preserved?
  - When we stop crawling?
- Who is responsible for the preservation?
- Which formats can we distinguish?
- Database approach
- Access rights and licenses
- Ownership and authenticity

# Achieving digital preservation of data

- Storage
  - Multiple redundancy to reduce risks
  - Trust
  - Scalability
- Metadata and definitions
  - Self-descriptiveness requires preserving the ontologies too
  - Provenance and additional information

# Thanks for your attention

## Questions?



josemgarcia@us.es



@josemgarcia\_us



UNIVERSITÀ  
DEL SALENTO