Data curation and preservation

José María García

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 @josemgarcia_us

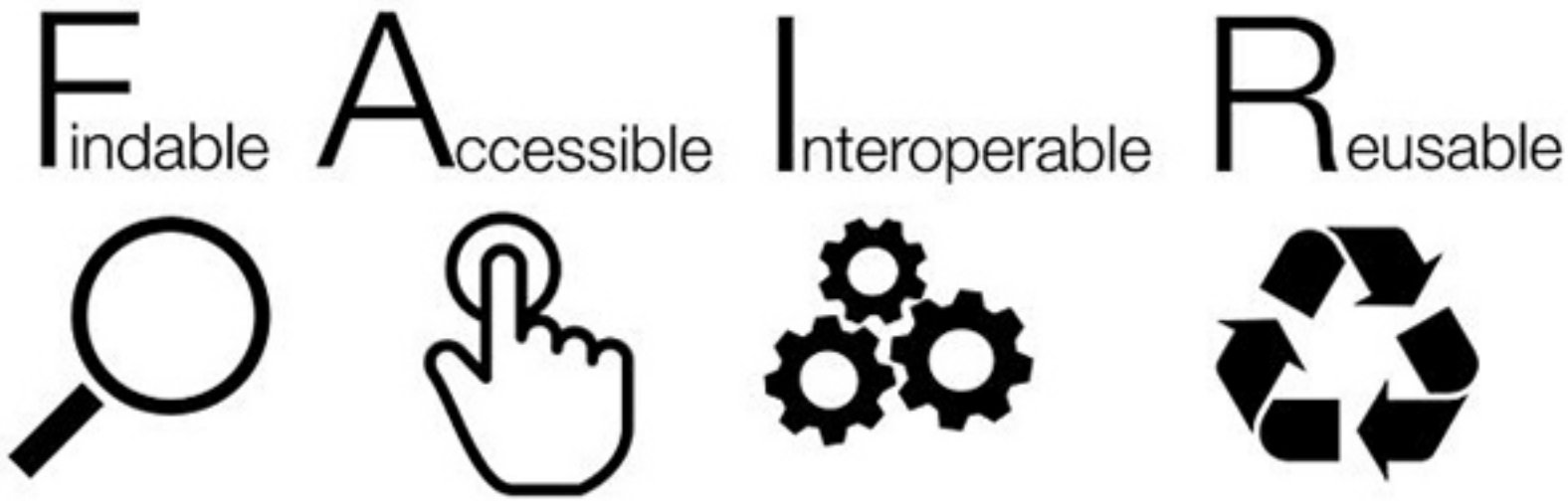
Outline

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

Outline

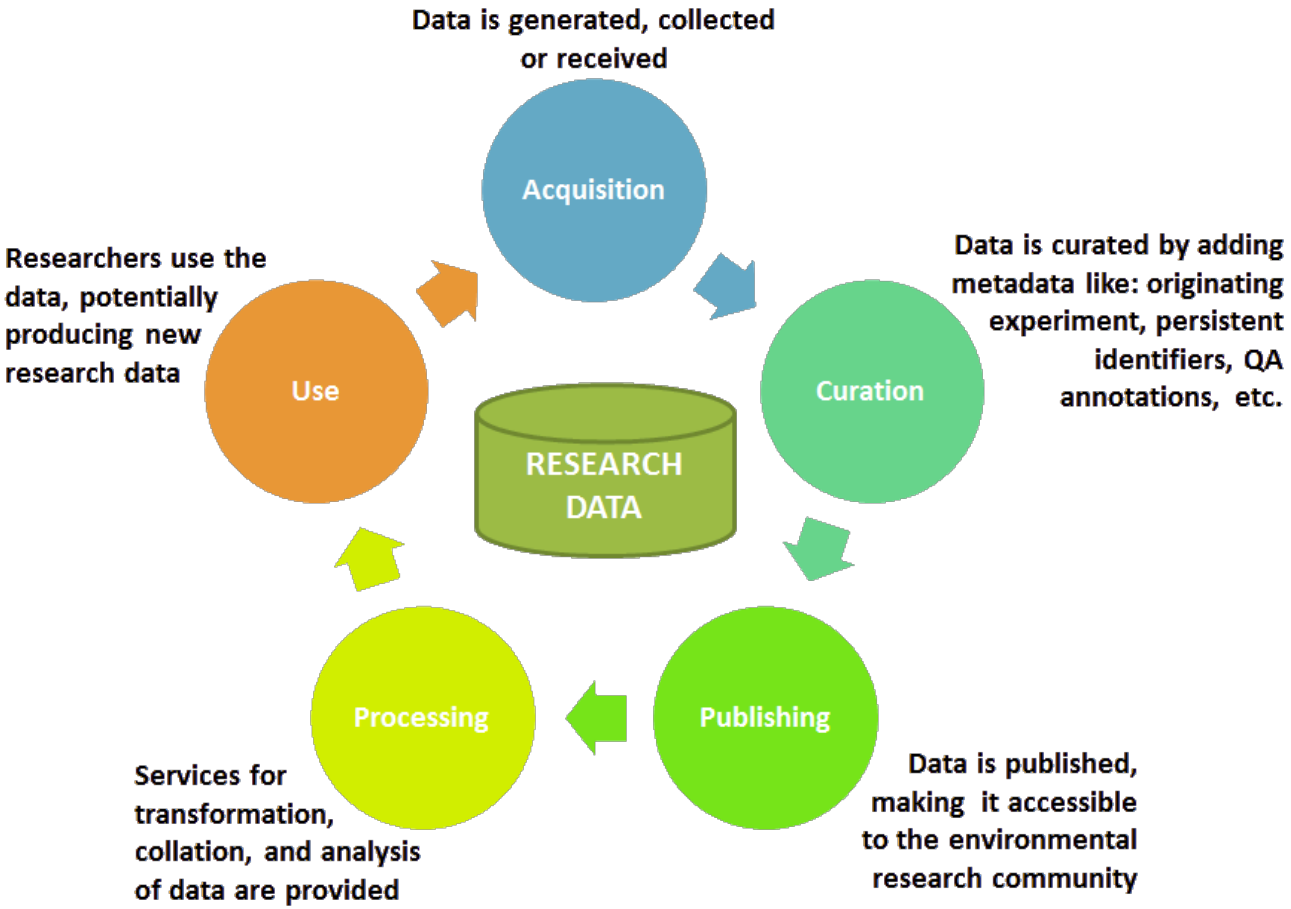
- Data curation in context
- Data curation activities
- Relevant metadata
- Data preservation
- Achieving digital preservation

FAIR principles

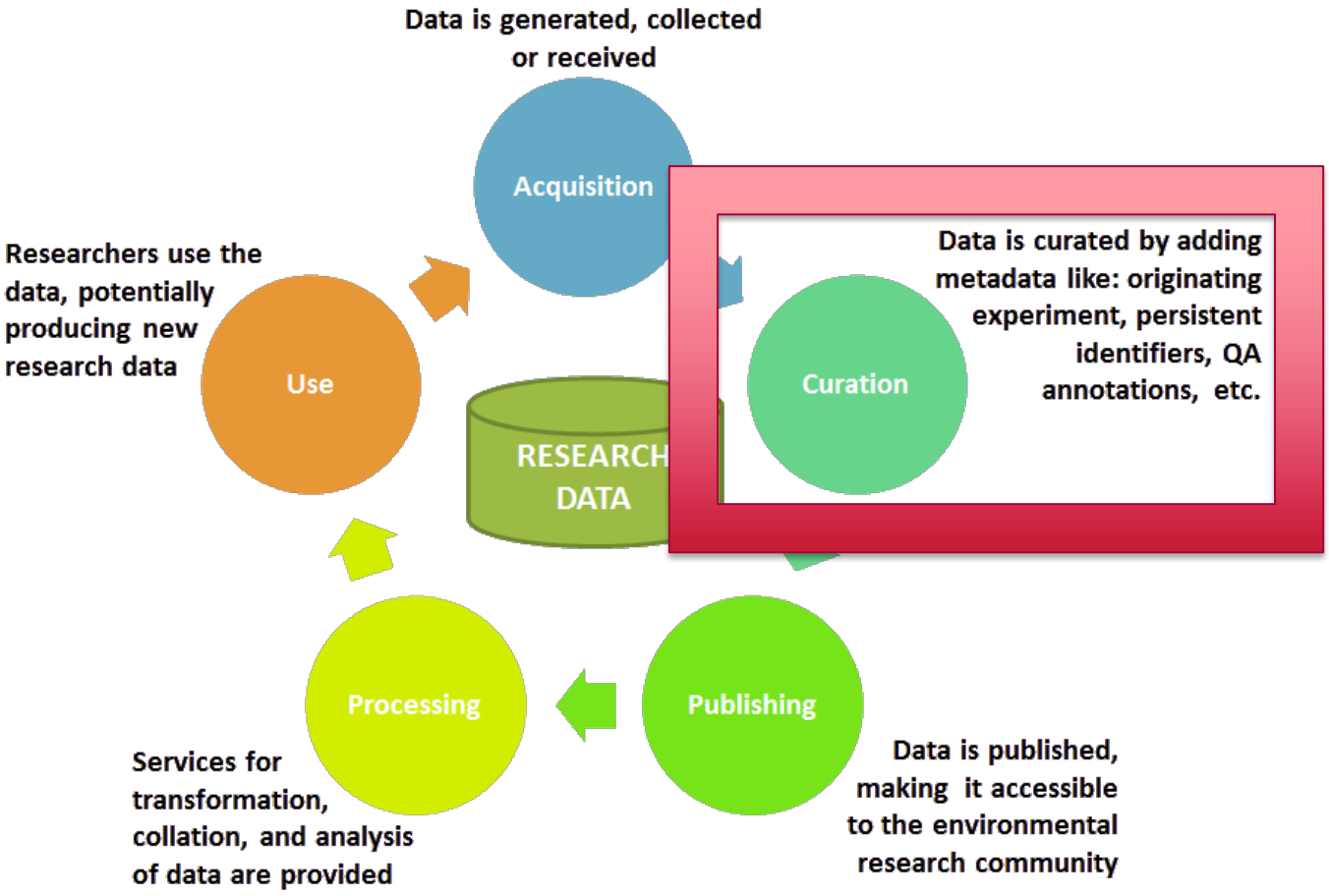


- To be FINDABLE
  - From raw data to information
  - Data needs context
  - Validation and quality assurance
  - Long term preservation

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)

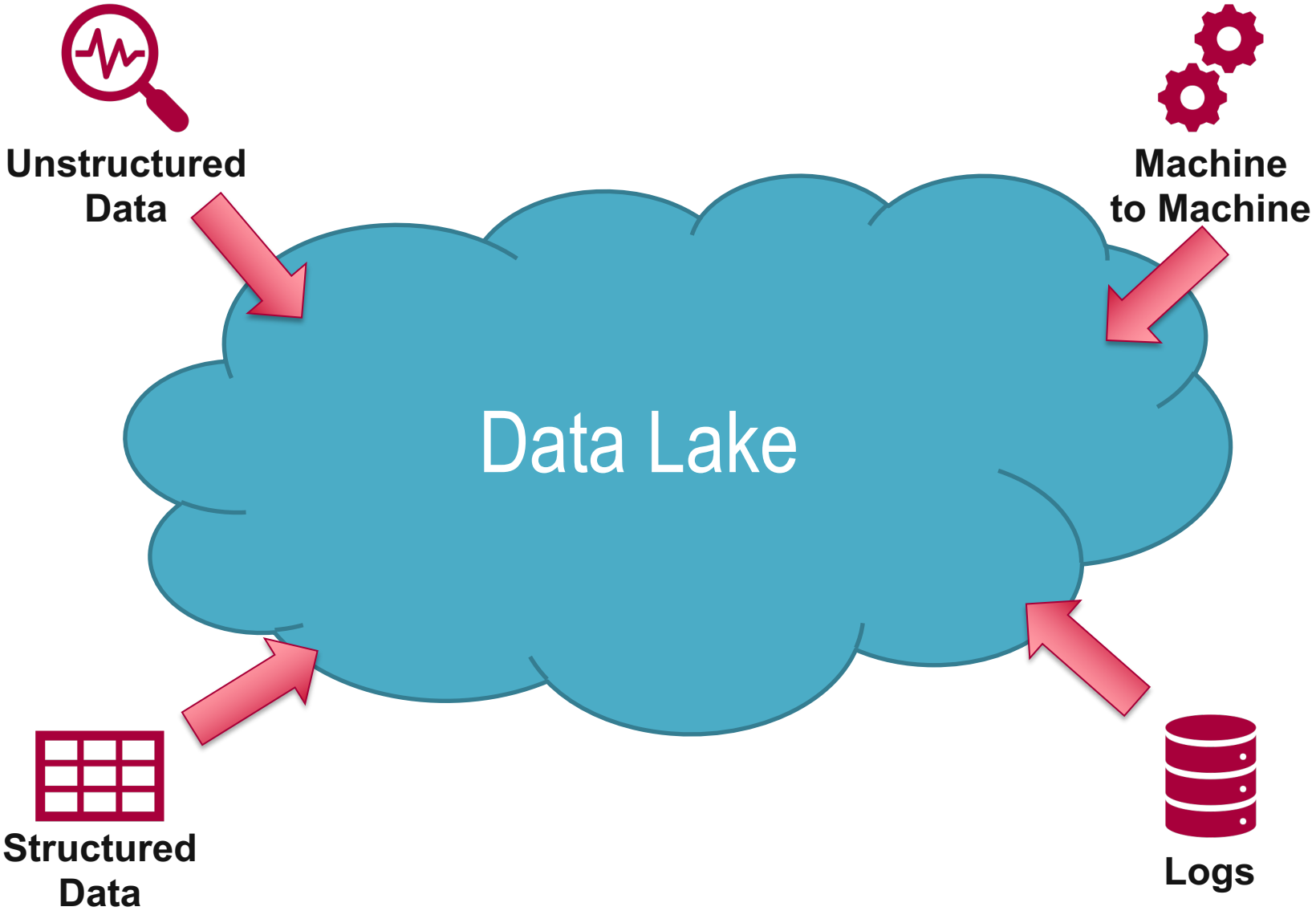
*"Digital curation involves maintaining, preserving and adding **value** to digital research data throughout its **lifecycle**"*

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

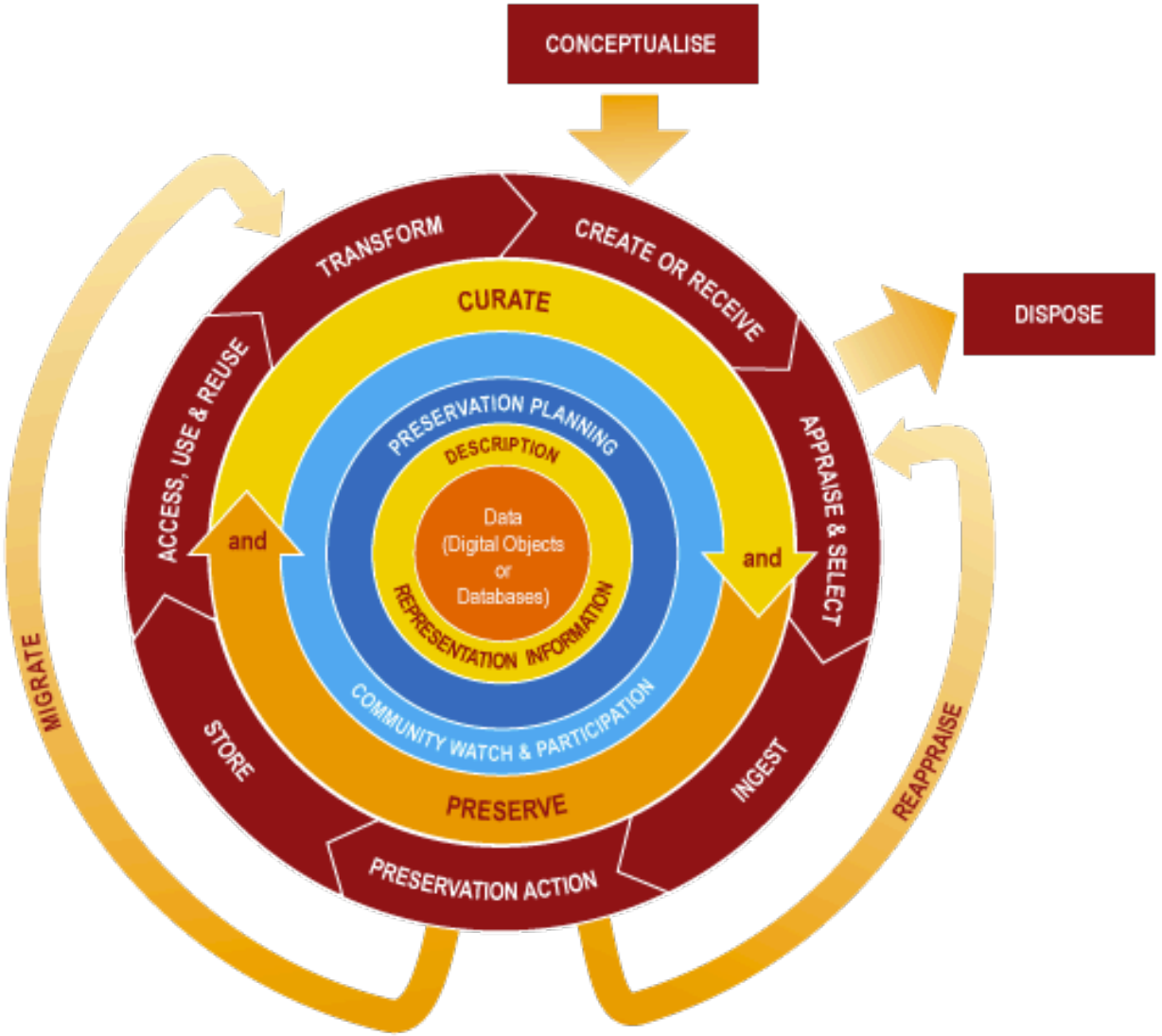
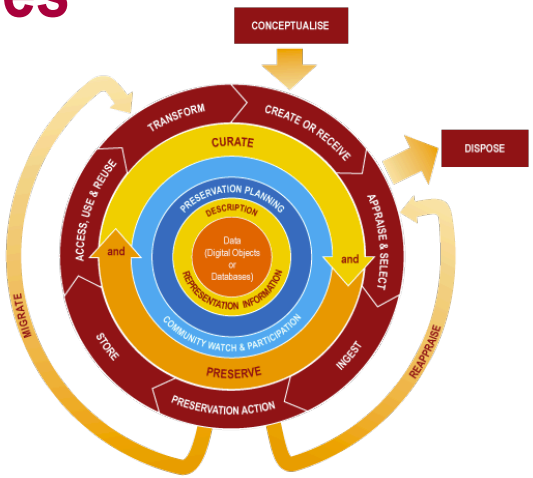


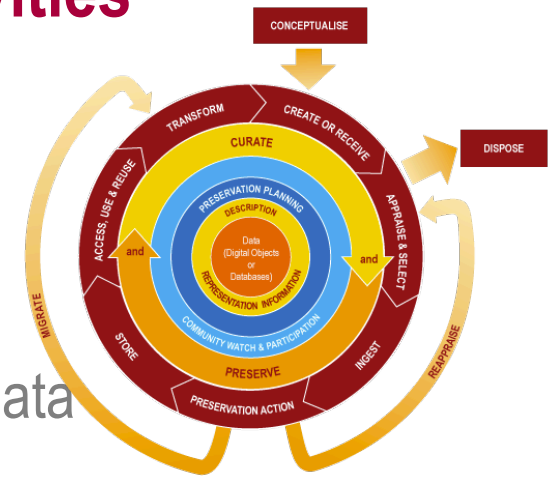
Image from the Digital Curation Centre, UK

Data curation generic activities



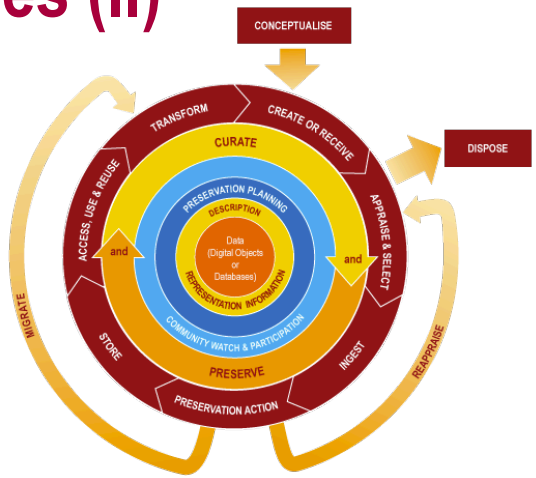
- Description and representation information
  - Assign metadata
  - Use standards
- Preservation planning
  - DMPs and planning of the curation activities
- Community watch and participation
  - Development of standards, tools, techniques…
- Curate and preserve
  - Management of the lifecycle

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

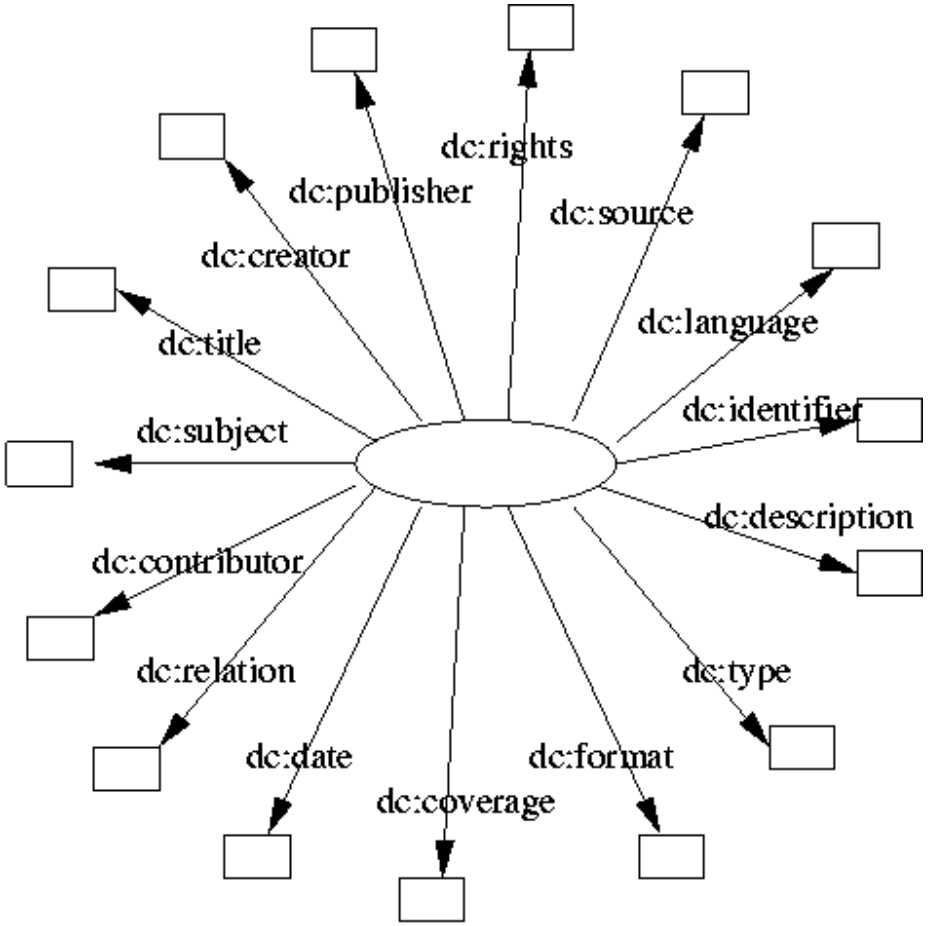
Uniqueness

Reusable between datasets

Helps long-term preservation

URIs, DOIs, ORCID

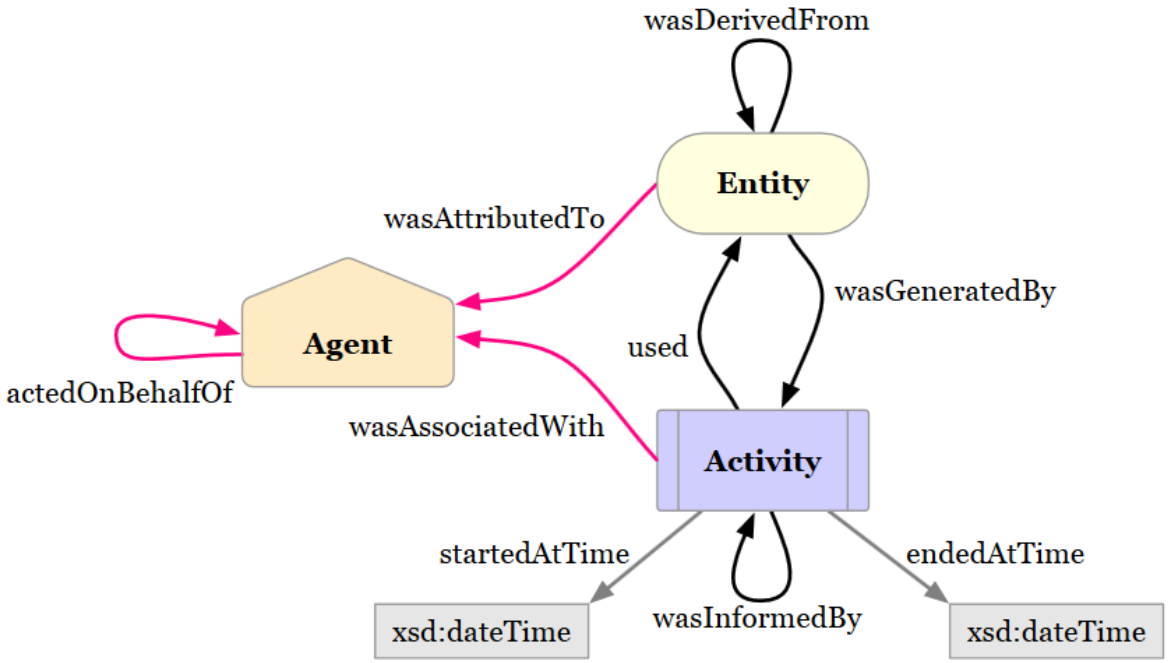
Generic vocabularies for metadata



Dublin Core

Friend-of-a-friend

Provenance

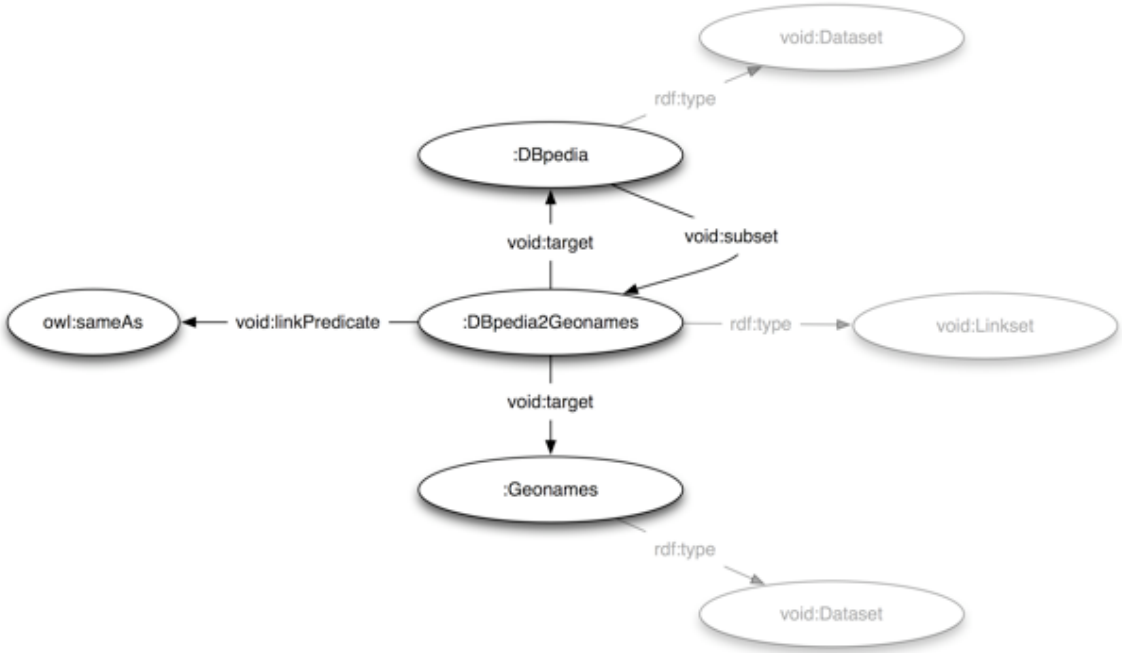


Information about entities, activities and agents involved in producing something

Chronology of ownership, custody, location, transformation

PROV framework from W3C

Dataset description



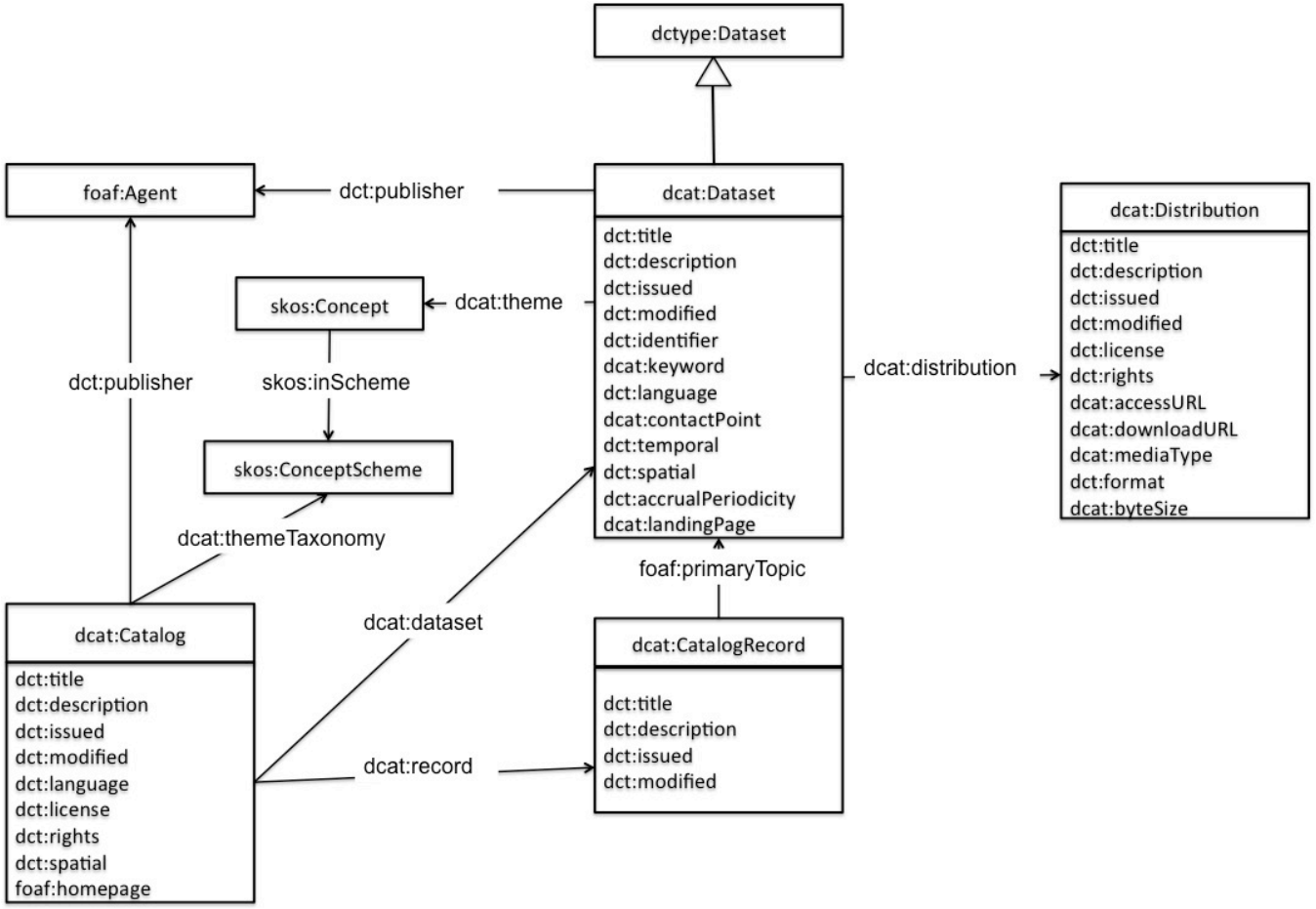
Vocabulary of Interlinked Datasets - VoID (W3C)

General dataset metadata (license, subject, features…)

Access (endpoints, distribution…)

Dataset structure (examples, vocabularies used, links)

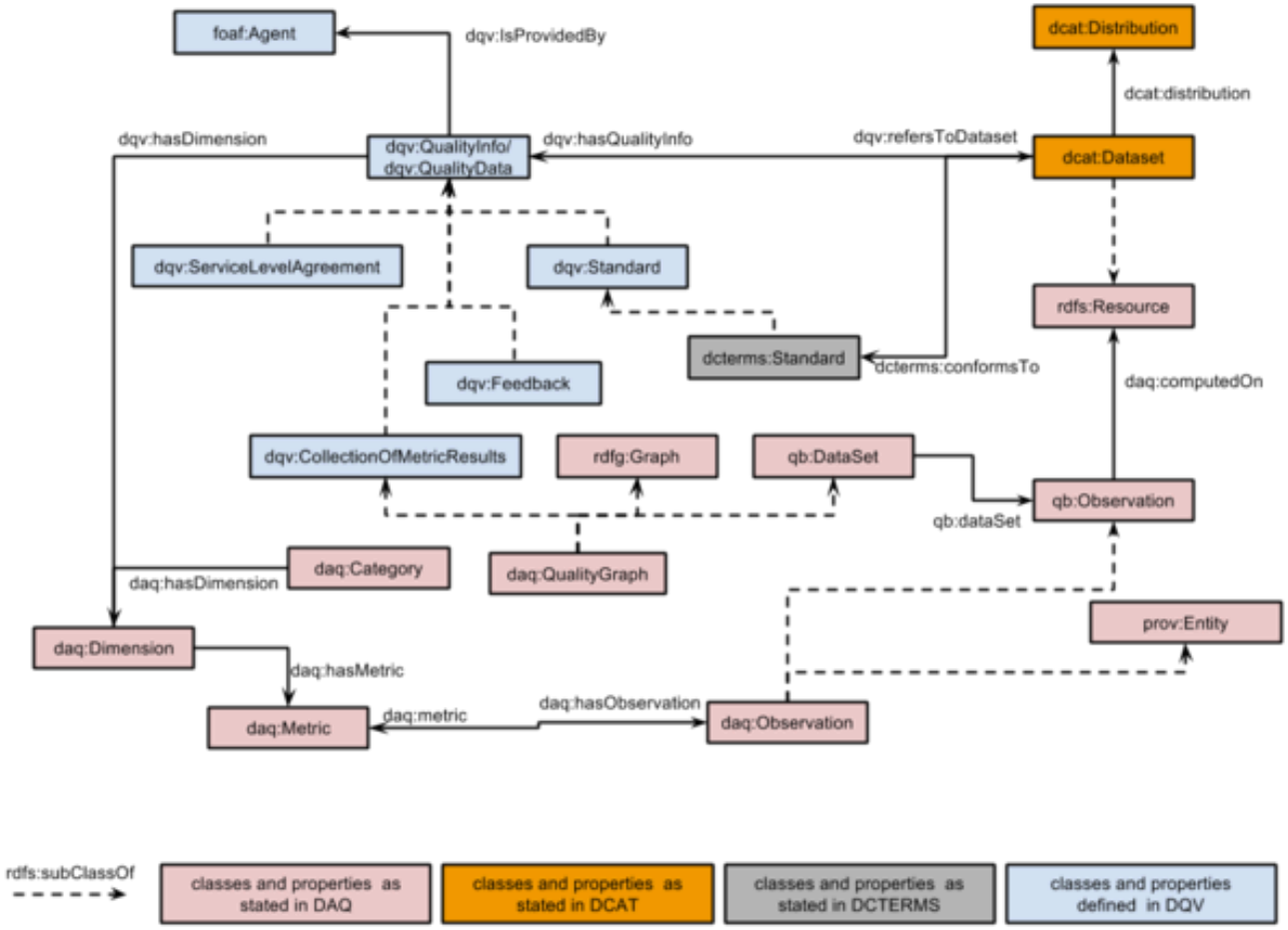
Data catalogs description



Data Catalog Vocabulary – DCAT (W3C)

Facilitate interoperability between data catalogs

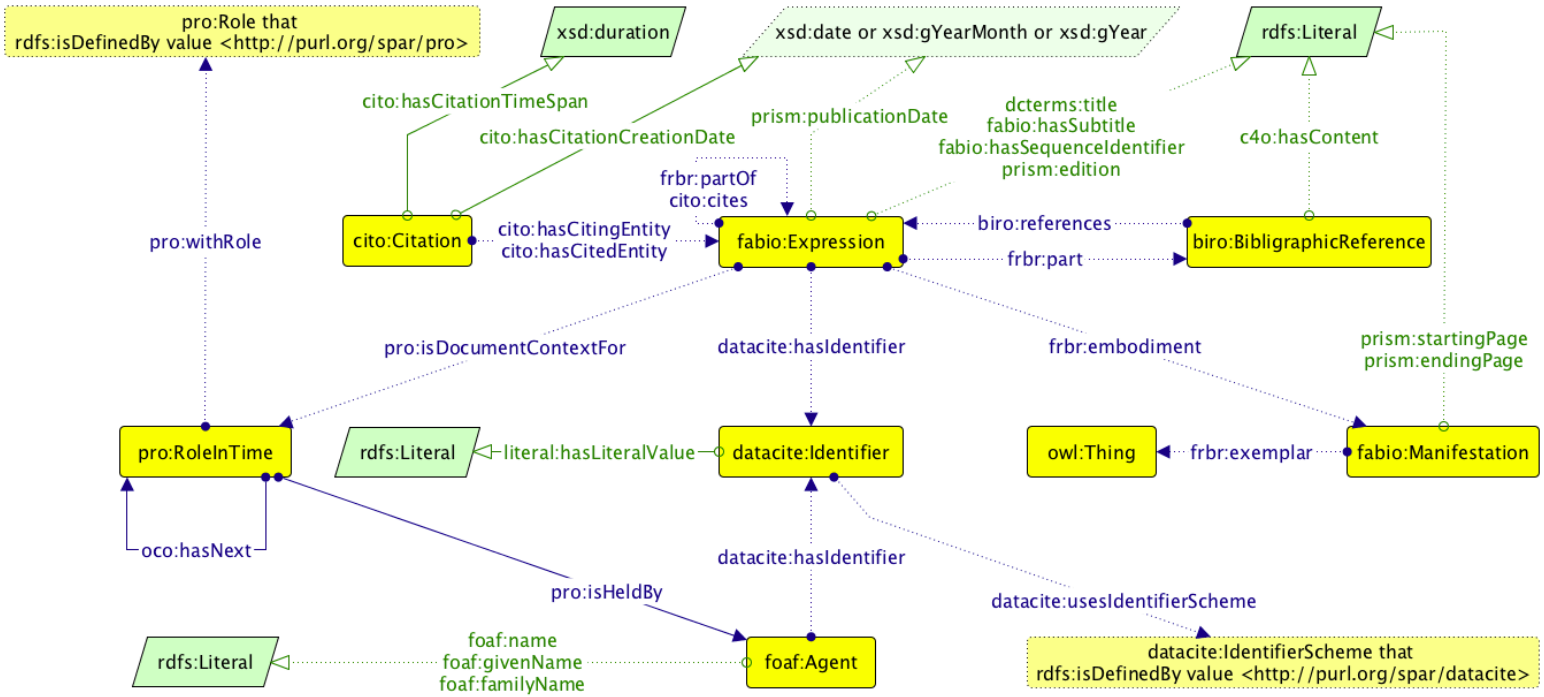
Quality assurance



Data Quality Vocabulary (W3C)

Applied to the dataset

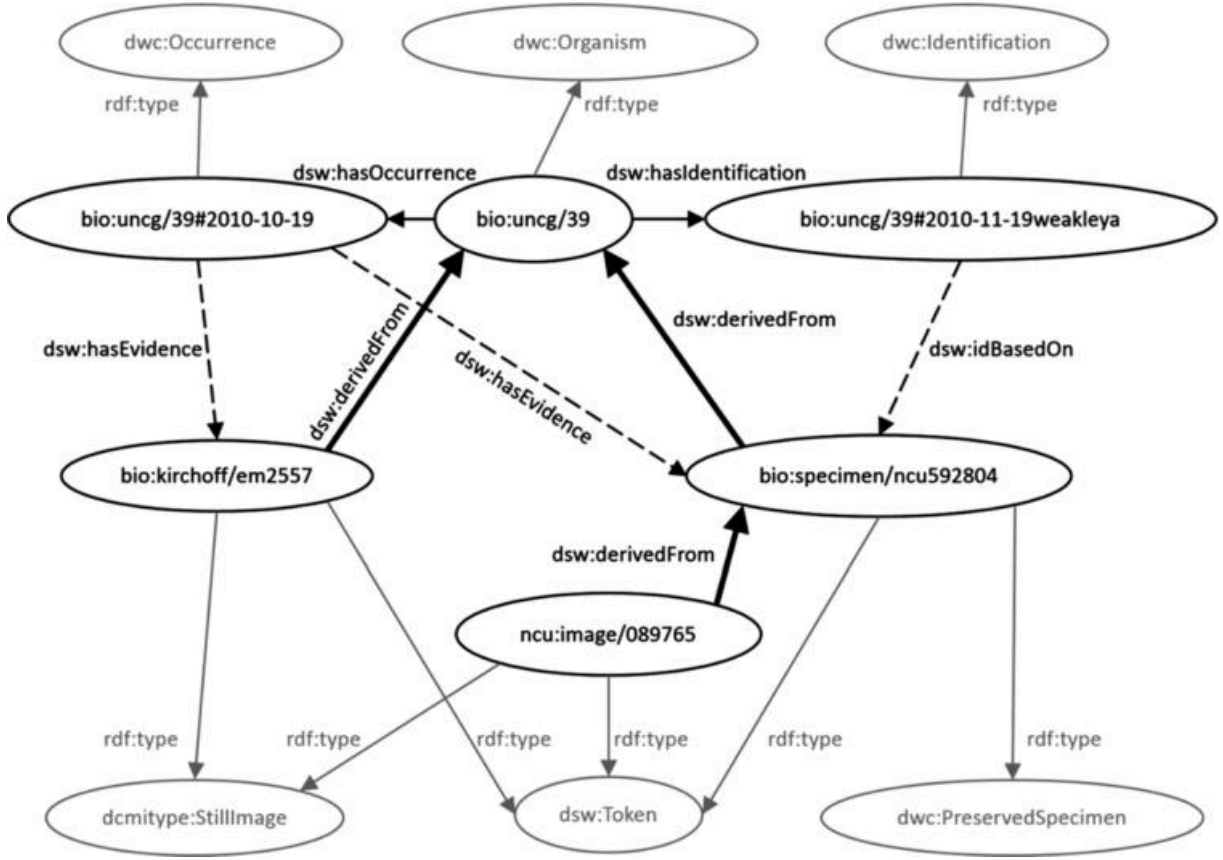
OpenCitations



Prefixes

biro:	http://purl.org/spar/biro/
cito:	http://purl.org/spar/cito/
c4o:	http://purl.org/spar/c4o/
datacite:	http://purl.org/spar/datacite/
dcterms:	http://purl.org/dc/terms/
fabio:	http://purl.org/spar/fabio/
foaf:	http://xmlns.com/foaf/0.1/
frbr:	http://purl.org/vocab/frbr/core#
literal:	http://www.essepuntato.it/2010/06/literalreification/
oco:	https://w3id.org/oc/ontology/
prism:	http://prismstandard.org/namespaces/basic/2.0/
pro:	http://purl.org/spar/pro/
owl:	http://www.w3.org/2002/07/owl#
rdfs:	http://www.w3.org/2000/01/rdf-schema#
xsd:	http://www.w3.org/2001/XMLSchema#

Metadata for environmental sciences and RI

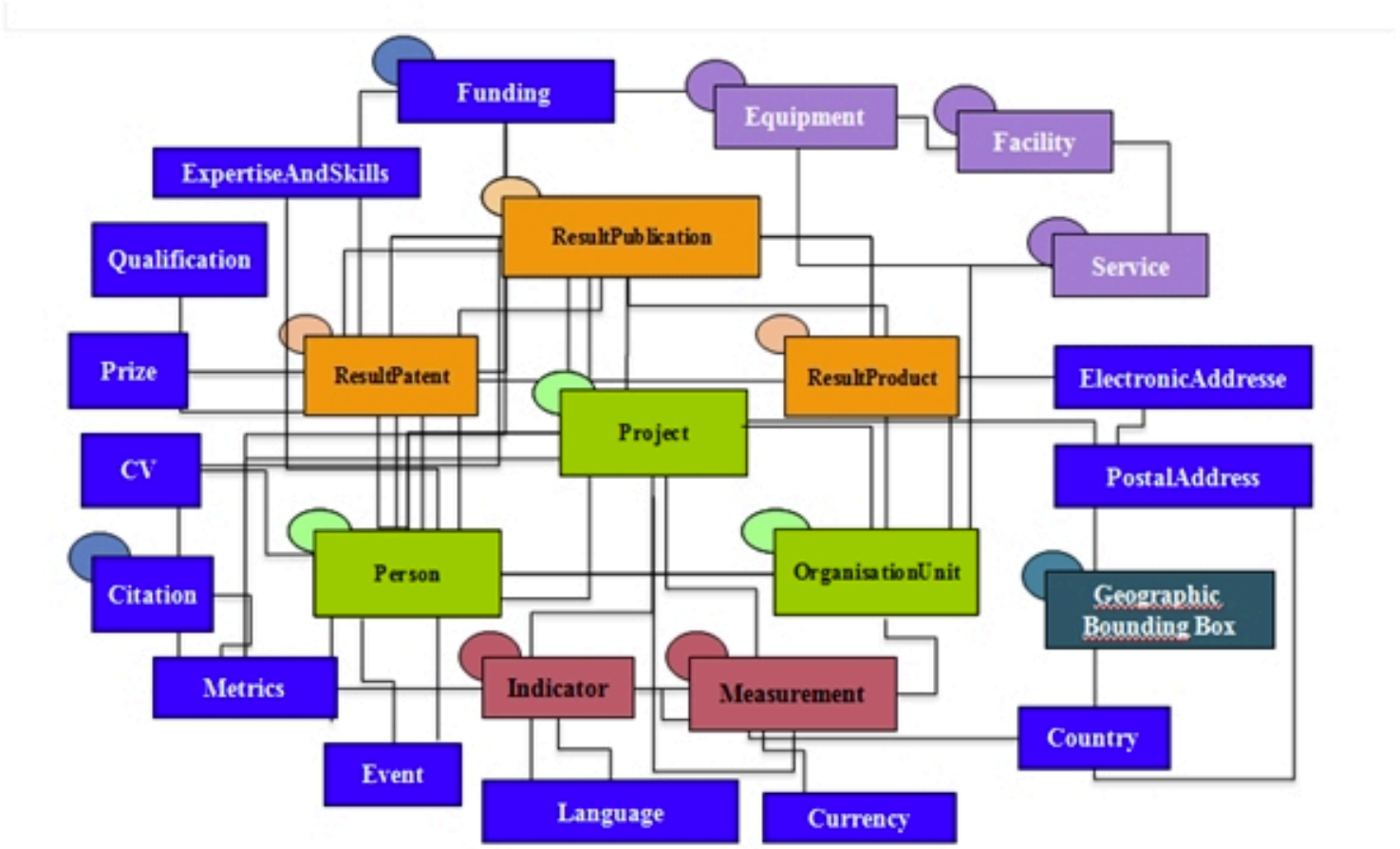


Darwin Core

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)

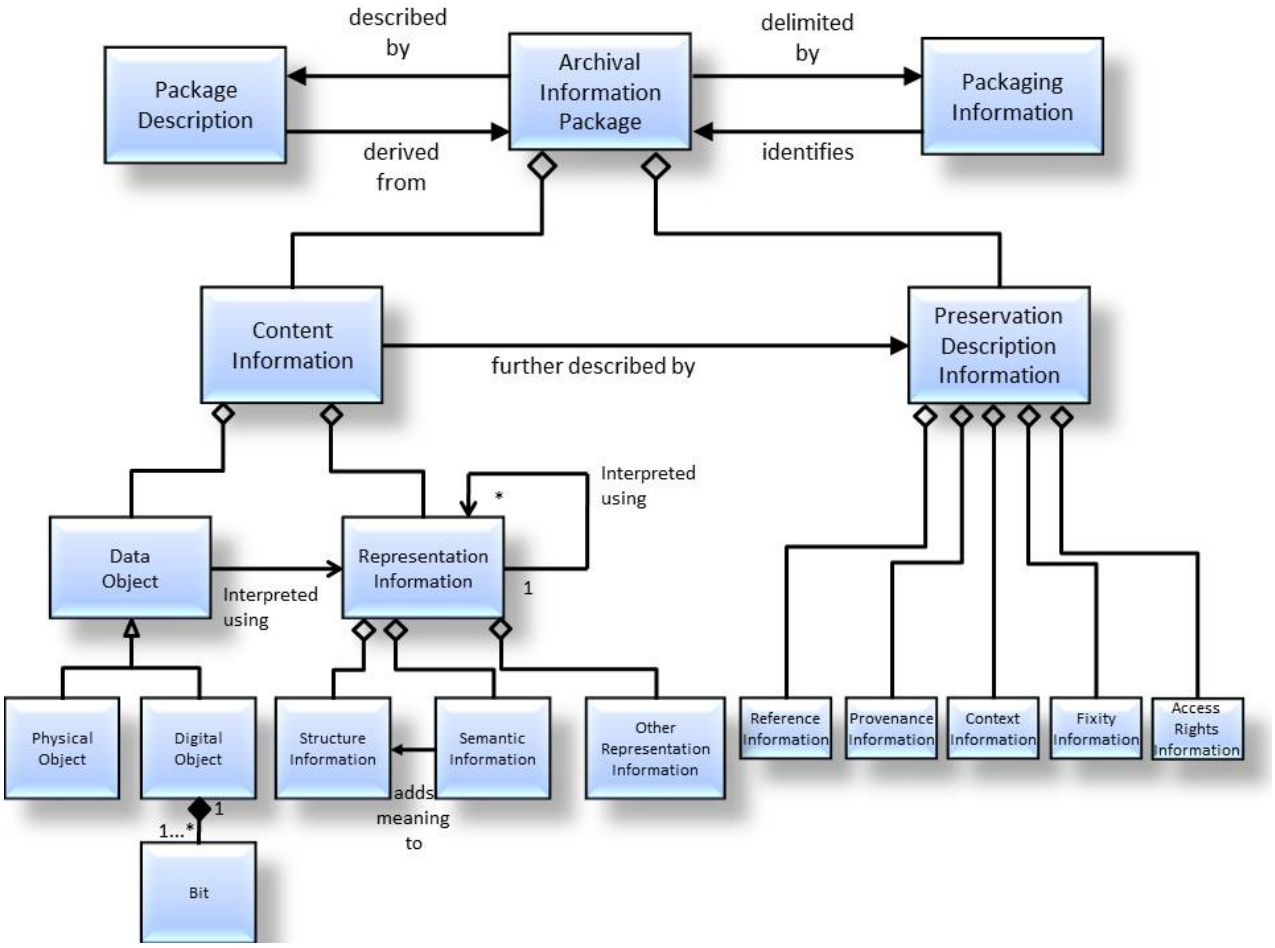
"The goal of digital preservation is, hence, the accurate rendering of authenticated content over time"

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

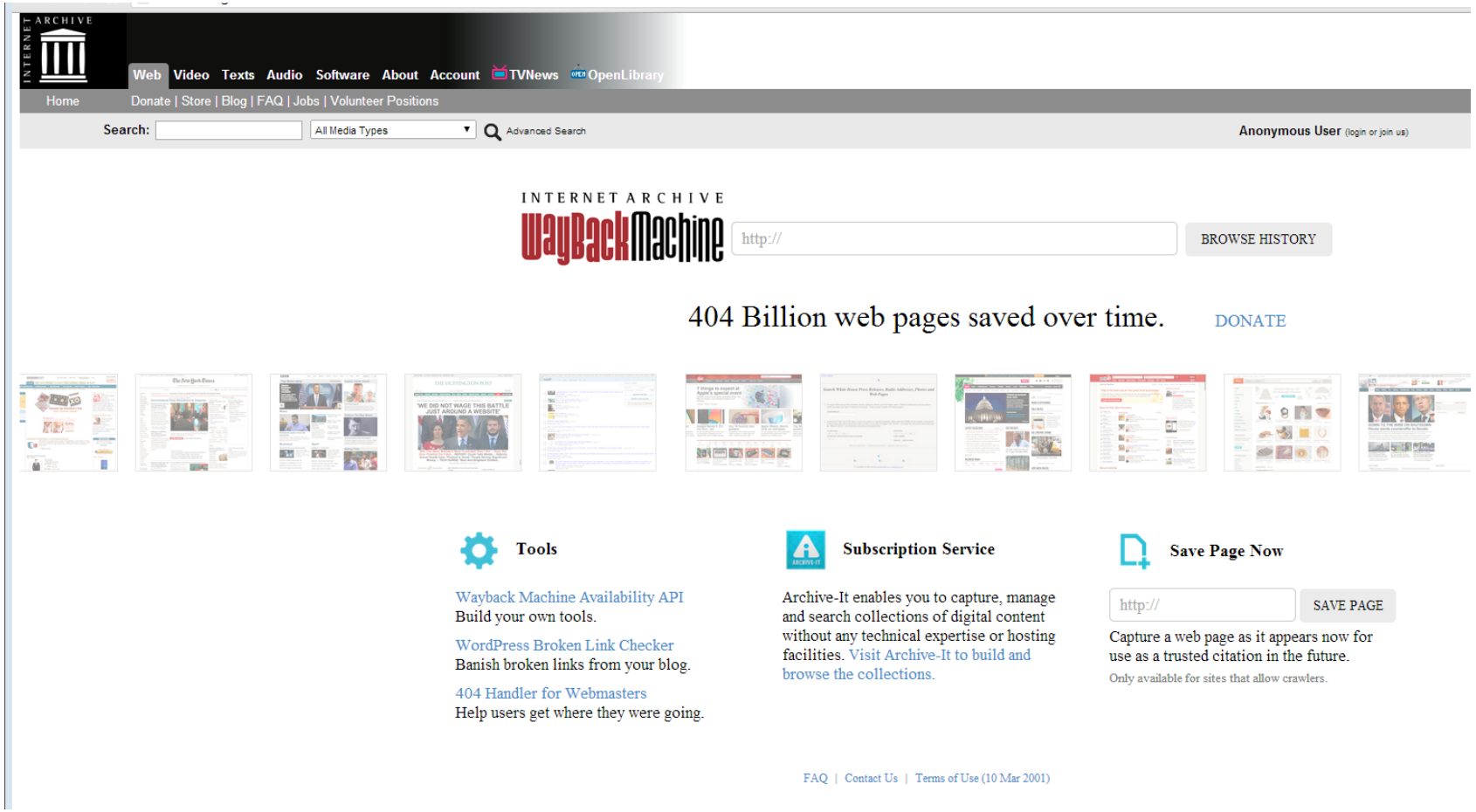
Use file formats based on standards

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

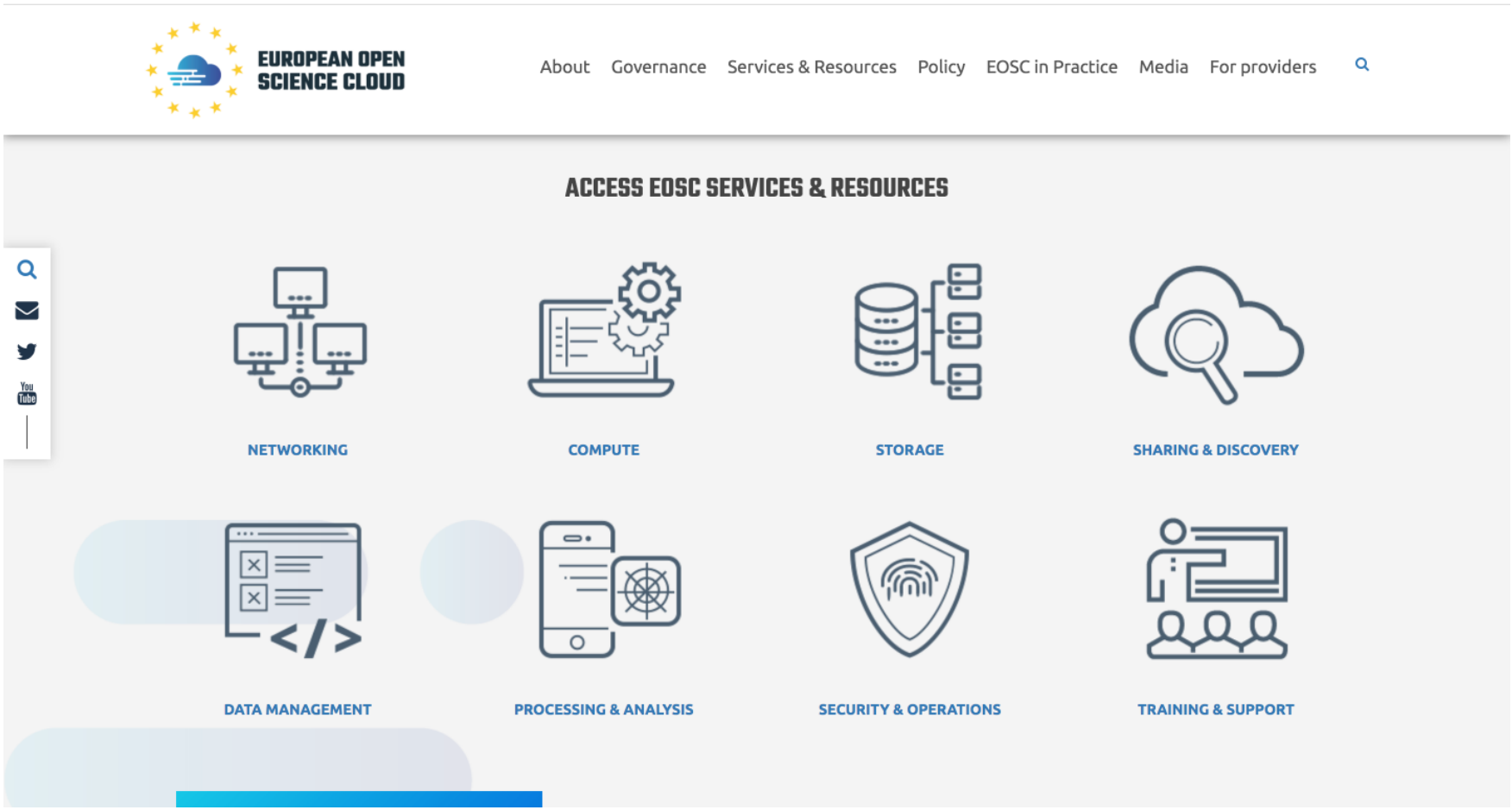
Create and maintain high quality documentation

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

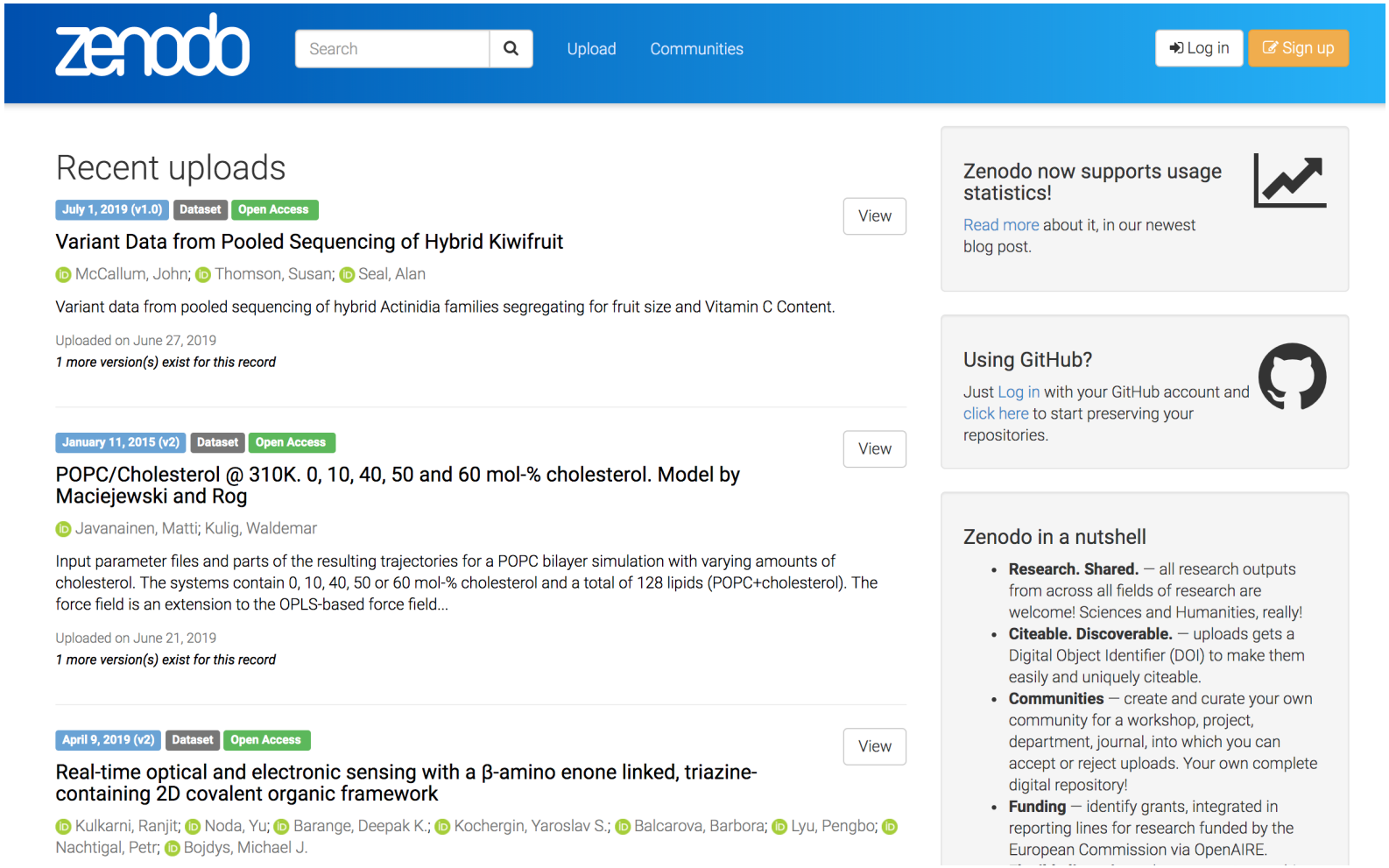
Archiving – Internet Archive (archive.org)



Services – Research Infrastructures, EOSC



Repositories – Zenodo



Zenodo policies for data preservation

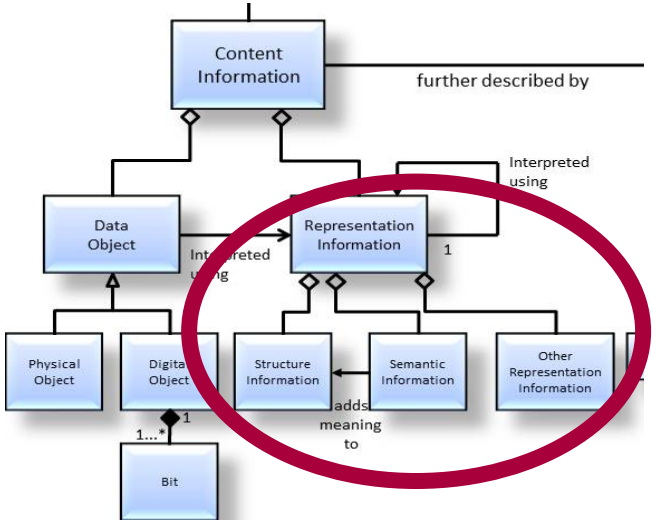
- Data files are **versioned** (not at records level) and preserved
- **Ingestion** as Submission Information Package
- **Replicas** in CERN data centres
- **Retention peri**od as long as CERN survives (+20 years)
- **Functionality** is **NOT** preserved
- **File preservation** with nightly backups
- **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