ENVRI Community International Winter School on DATA FAIRness

https://www.lifewatch.eu/envri-iws-data-fairness-2020

Semantics

José María García, SCORE Lab, University of Seville, LifeWatch Spain JRU

AJ Sáenz-Albanés, LifeWatch ERIC

11-22 January 2021

ENVRI-FAIR has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824068

Outline

- Semantics 101
- Linked Data
- Machine-actionable FAIRness through semantic technologies
- Applying Semantics to the research data lifecycle
- Security and additional aspects
- Hands-on session

Semantics 101

The meaning of "semantics"

- The study of meaning, reference, or truth (Wikipedia)
- the branch of **linguistics** and **logic** concerned with **meaning**. The two main areas are *logical semantics*, concerned with matters such as sense and reference and presupposition and implication, and *lexical semantics*, concerned with the analysis of word meanings and relations between them. (Oxford)
- provides the rules for **interpreting the syntax** which do not provide the meaning directly but **constrains the possible interpretations** of what is declared. (J. Euzenat)

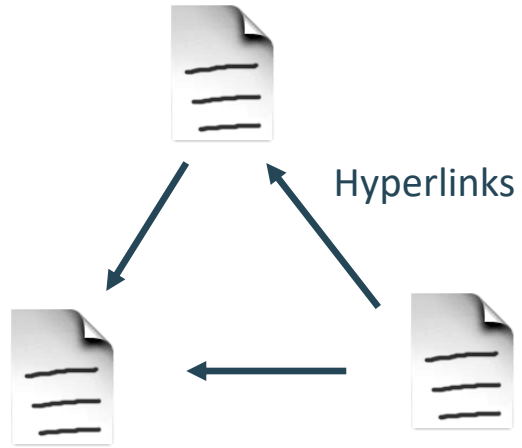
Semantic Technology

- **Semantic technology** uses formal **semantics** to help AI systems understand language and process information the way humans do.

- **Semantic technology** defines and links data on the Web (or within an enterprise) by developing languages to express rich, **self-describing interrelations** of data in a form that **machines can process**.

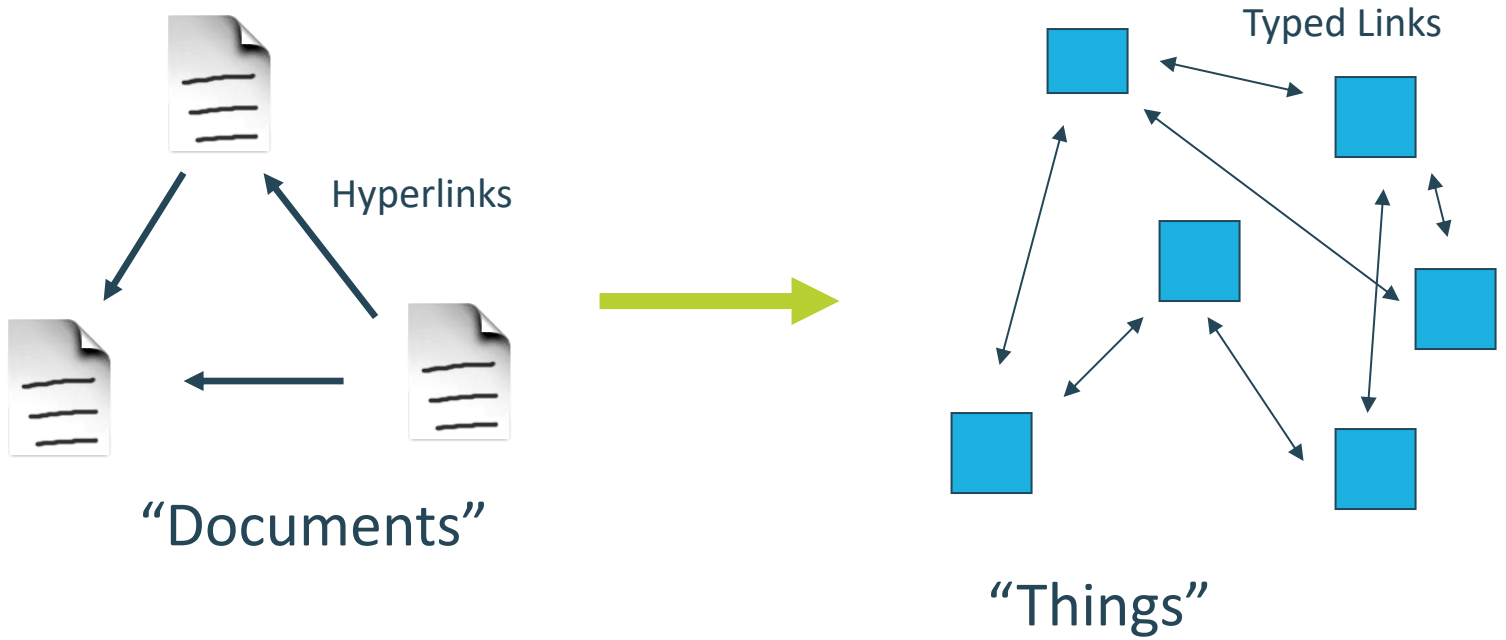
The "plain old" Web of Documents

- Fundamental elements
  - Names (URIs)
  - Documents (Resources) described by HTML, XML, etc.
  - Interactions via HTTP
  - (Hyper)Links between documents or anchors in these documents
- Shortcomings
  - Untyped links
  - Web search engines fail on complex queries



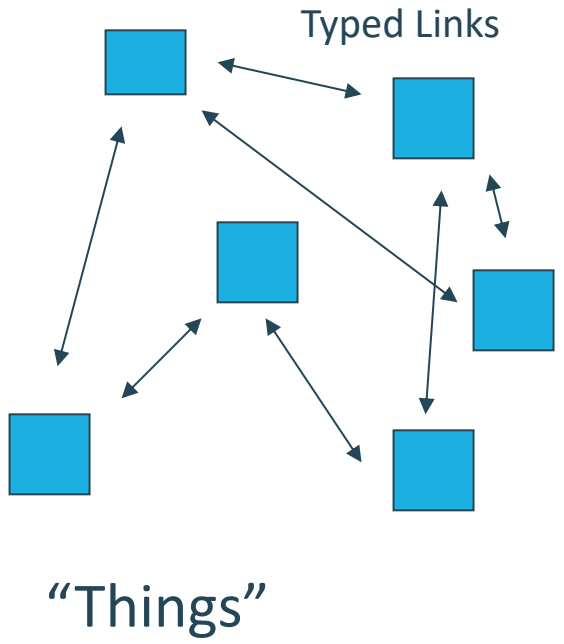
"Documents"

Towards a Web of Data



Characteristics of the Web of Data

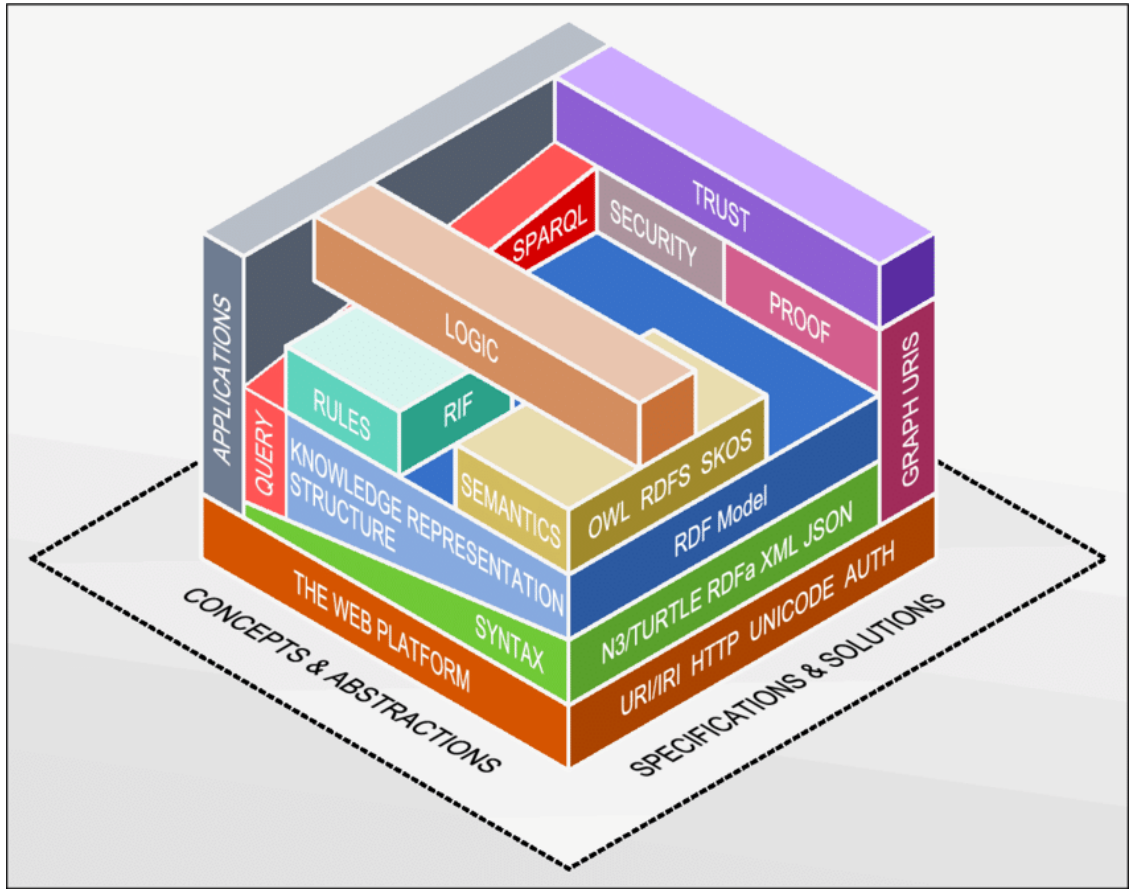
- Links between arbitrary things (e.g., persons, locations, events, buildings)
- Structure of data on Web pages is made explicit
- Things described on Web pages are named and get URIs
- Links between things are made explicit and are typed



The Semantic Web

- W3C vision of the Web of Data
- Its goal is to make the Web machine-readable
- Additional benefits
  - Automation of human tasks on data
  - Enables knowledge sharing and reuse
  - Separates data from its representation and tools
- Set of standards
- The original vision from Tim Berners-Lee did not succeed
  - Heavyweight standards
  - Little uptake from industry

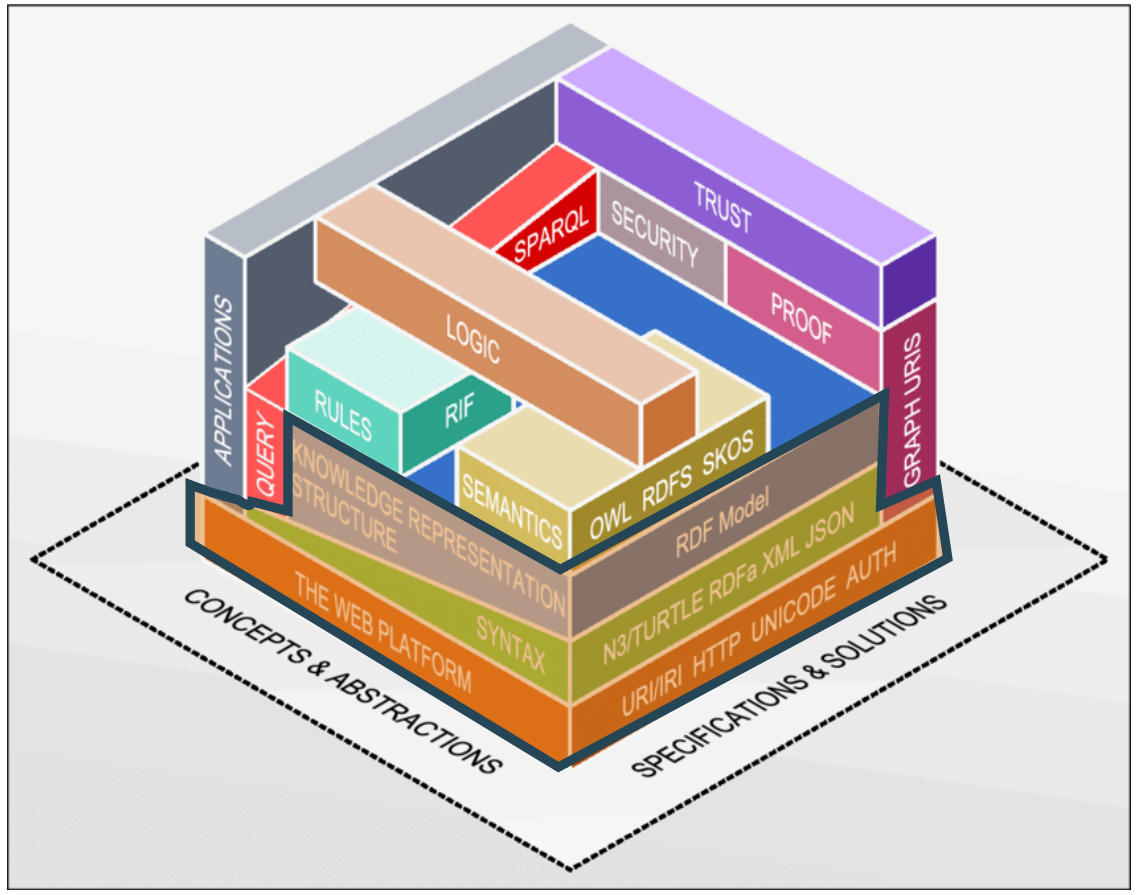
Semantic technology stack



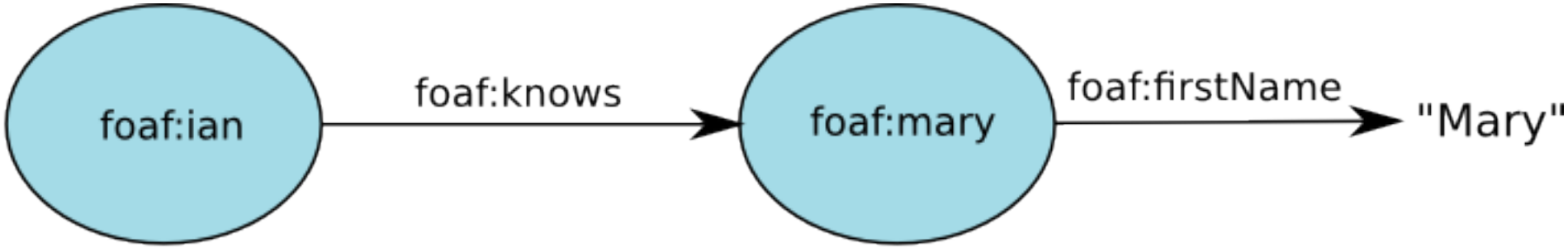
(B. Nowack)

Representing information

- Resource Data Framework (RDF)
  - Foundations of all Semantic Web technologies
- RDF is based on triples
  - Labelled, unidirectional connection between two resources
  - Each entity represented by IRIs
  - Objects can be literal values
  - Namespaces



Resource Description Framework Model

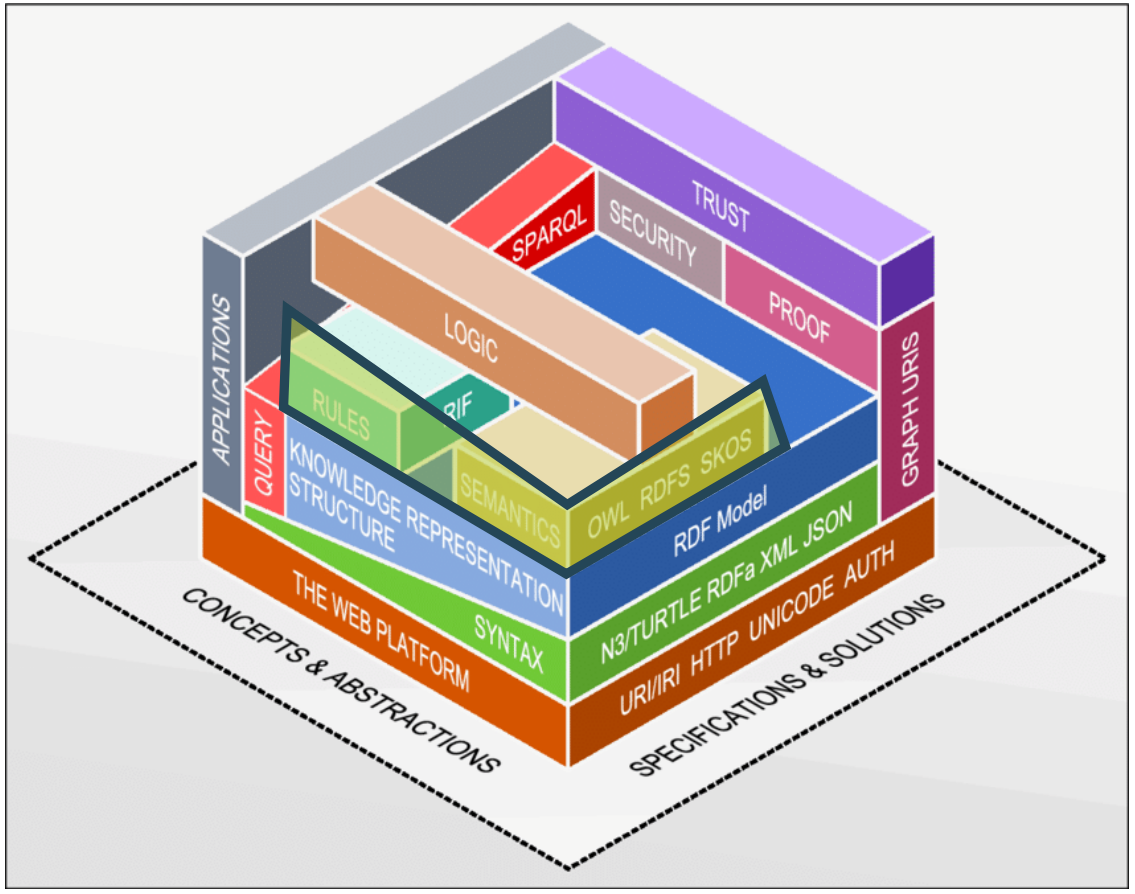



```
foaf:ian foaf:knows foaf:mary.
foaf:mary foaf:firstName "Mary".
```

- Serializations in many formats (RDF/XML, Turtle, N3, JSON-LD…)

Interpreting and reasoning

- RDF is only a data model
- We need additional layers on top to add semantics to the data
- RDF Schema (RDFS) provides a data-modelling vocabulary for RDF data
- Simple Knowledge Organization System (SKOS) can represent taxonomies, glossaries, thesauri…
- The Web Ontology Language (OWL) allows specification of RDF ontologies



RDF Schema

- https://www.w3.org/TR/rdf-schema
- Define the terms that can be used in an RDF graph
  - Classes (concepts)
  - Properties
  - Restrictions on RDF resources
- The main construct are the **classes** of resource
  - Everything in RDF is a resource, even RDFS classes
  - Resources can be instances (individuals) of a class (*rdf:type*)
  - Resources may belong to several classes at once
  - Class relationships can be inherited via subclassing (*rdfs:subClassOf*)

RDF Schema

- Properties between RDF resources can be constrained
  - Domains (*rdfs:domain*)
  - Ranges (*rdfs:range*)
  - Sub-properties (*rdfs:subPropertyOf*)
- Properties domain and ranges may be classes or data types, in case the objects are literals
  - Supports XML Schema data types
  - Language tags can be added to strings to support internationalized labels and descriptions of RDF resources
- Human readable annotations to describe resources (*rdfs:label*, *rdfs:comment*)

RDF Schema

- RDFS provides basic inference over RDF data
  - Entailment rules that extend the RDF graph making explicits the relationships
  - E.g. if *A rdfs:subClassOf B*, and *a rdf:type A*, then a triple *a rdf:type B* would be added by entailment rules.
- RDFS is useful to define vocabularies
  - Collections of semantically-related classes and properties
  - Relationships between those classes and properties and external classes and properties from other vocabularies (hierarchies of vocabularies)
  - Lightweight definitions and reasoning

SKOS

- https://www.w3.org/TR/skos-reference/
- Aimed at representing taxonomies, thesauri, etc.
- Adds additional classes and properties to RDFS to
  - Identify concepts (*skos:Concept*)
  - Annotate concepts with alternative and preferred labels (*skos:altLabel* and *skos:prefLabel* properties)
  - Specify the relative scope of related concepts (*skos:broader*, *skos:narrower* properties)
  - Define collections (*skos:Collection*)
  - …

RDFS and SKOS limitations

- Only basic inference is provided (sub-classes, sub-properties, domains, ranges…)
- Simple constraints for properties
- SKOS only support concepts, but not individuals

Ontology Web Language (OWL)

- https://www.w3.org/TR/owl2-overview/
- *An ontology is a formal, explicit specification of a shared conceptualization* (Guarino et al.)
- OWL ontologies provide more reasoning facilities than RDFS and SKOS vocabularies
- Based on the 'open world' assumption
  - Absence does not imply negation
  - There are no unique names, so two entities may refer to the same real concept (*owl:sameAs*)
  - Leads to problems if constraints are missing from the ontology

OWL 2 facilities

- Algebraic properties (reflexive, symmetric, transitive properties…)
- Disjoint properties
- Qualified cardinality
- Restrictions on properties (useful for complex classifying)
- Property chains
- Union and intersections
- Disjunction and complement
- Equivalence
- Value restrictions on data properties

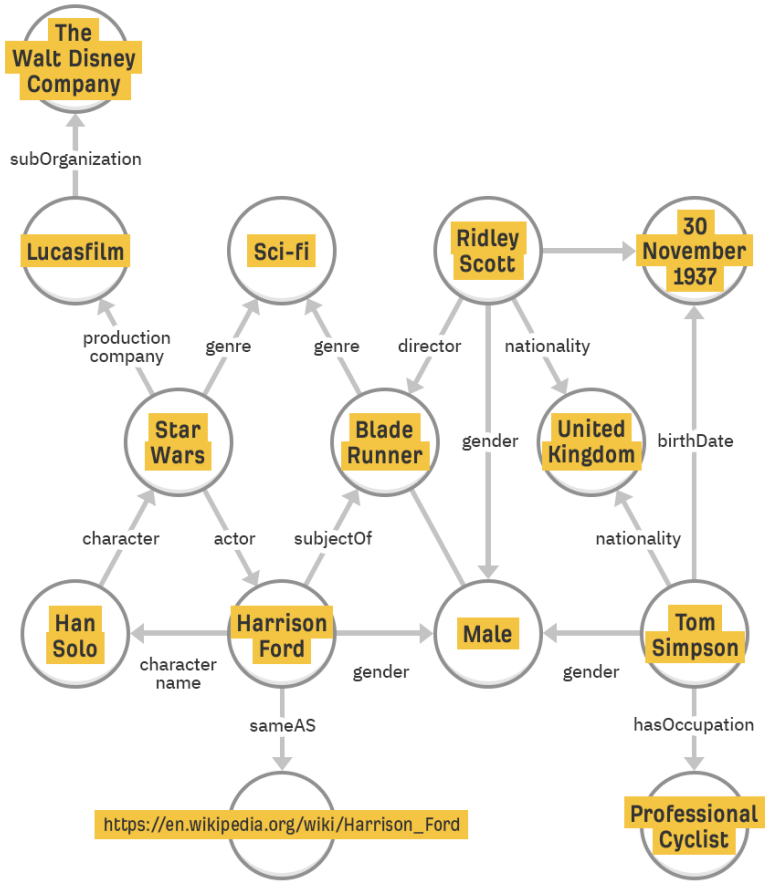
OWL limitations

- Full OWL inference may be too complex
  - Profiles limit reasoning capabilities for certain scenarios
- Heavyweight definitions
- Potential issues due to the ‘open world’ assumption
- Sometimes RDFS (and possibly SKOS) provide enough elements to describe *lightweight* ontologies

Linked Open Data

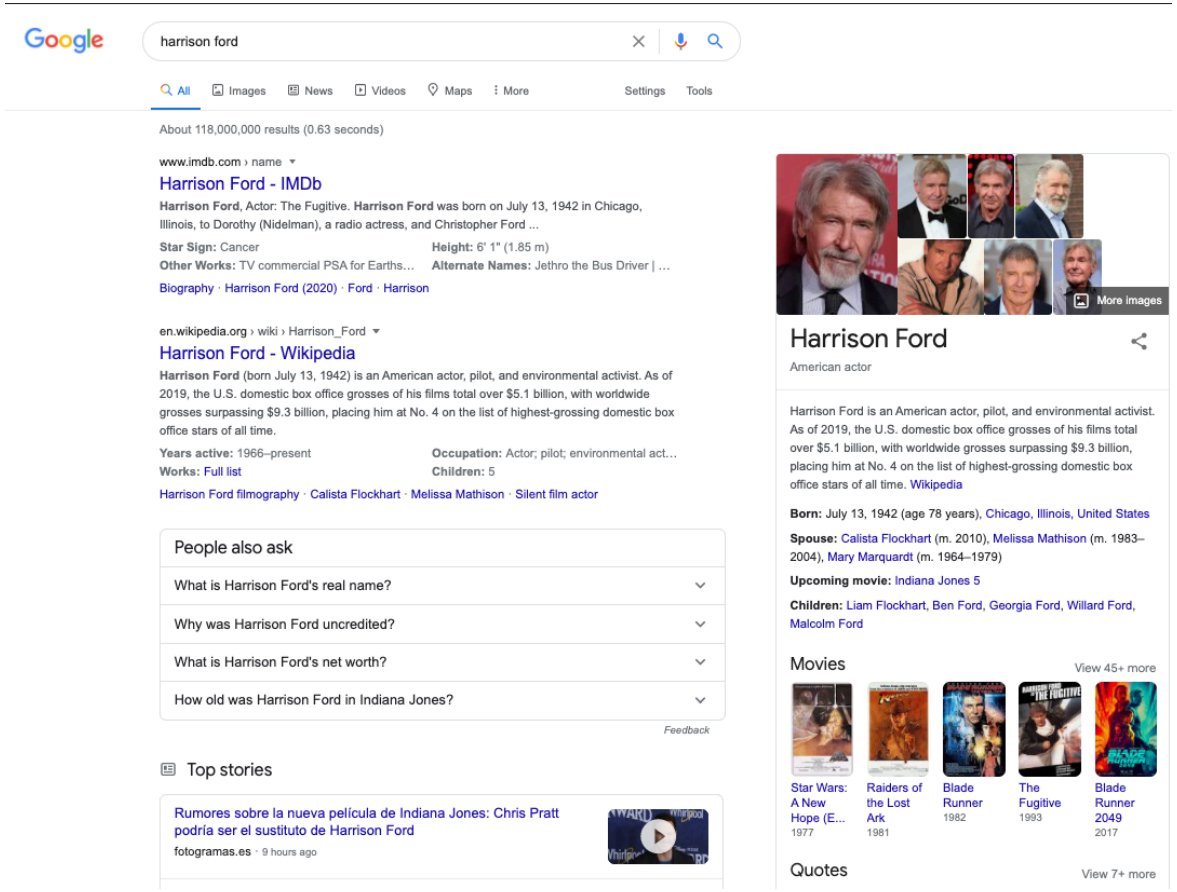
Knowledge Graphs

- A **knowledge graph** is a knowledge base that uses a graph-structured data model or topology to integrate data.
- Data comes from several sources and is interlinked to facilitate traversing
- There is no global schema, but a set of independent schemas
- New data and connections can be added without affecting the whole graph



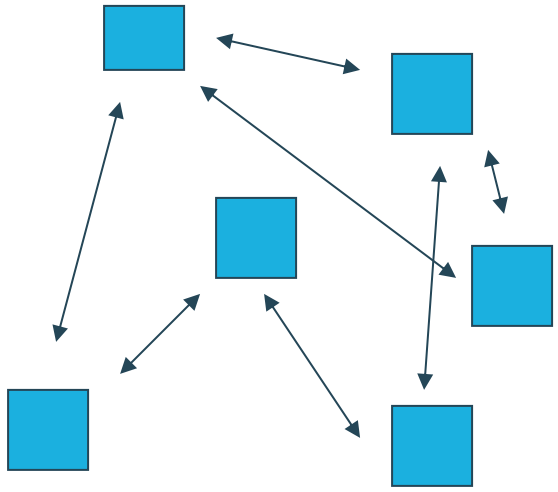
https://ahrefs.com/blog/google-knowledge-graph/

Google Knowledge Graph



Linked Data

- **Linked Data** is about the use of Semantic Web technologies to publish structured data on the Web and set links between data sources.
- Use URIs as names for things.
- Use HTTP URIs so that people can look up (dereference) those names.
- When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL)
- Include links to other URIs. so that they can discover more things.



5-star Linked Open Data

★ Available on the web (whatever format) but with an open licence, to be Open Data

★★ Available as machine-readable structured data (e.g. excel instead of image scan of a table)

★★★ as (2) plus non-proprietary format (e.g. CSV instead of excel)

★★★★ All the above plus, Use open standards from W3C (URIs, RDF and SPARQL) to identify things, so that people can point at your stuff

★★★★★ All the above, plus: Link your data to other people's data to provide context

★ Available on the web (+ open licence)

- Easy to publish web data
- Data can be easily accessed and stored locally
- Data can be entered manually into another system

Source [Bauer F., Kaltenböck, M.: Linked Open Data: The Essentials, 2012]

★★ Available as structured data

- All benefits from ★
- Data can be directly processed with proprietary software
- Easy to export it into another structured format

Source [Bauer F., Kaltenböck, M.: Linked Open Data: The Essentials, 2012]

★★★ Non-proprietary format is used

- All benefits from ★★
- No need to pay for a format controlled by a single organization

Source [Bauer F., Kaltenböck, M.: Linked Open Data: The Essentials, 2012]

★★★★ Use open standards to identify things

- All benefits from ★★★
- Link to data from anywhere, either on the web or locally
- It can be bookmarked and parts of the data can be reused
- Access to data items can be optimized (caching, load balancing, etc.)
- BUT the publisher needs to identify separable items, assign URIs to each one, and allow to access it independently

Source [Bauer F., Kaltenböck, M.: Linked Open Data: The Essentials, 2012]

★★★★★ Data is linked to provide context

- All benefits from ★★★★
- New data of interest can be discovered while consuming other
- Data schema can be obtained
- Added value to the data
- Linked datasets are discoverable
- BUT resources have to be invested to link datasets

Source [Bauer F., Kaltenböck, M.: Linked Open Data: The Essentials, 2012]

The Linked Open Data Cloud (May 2007)

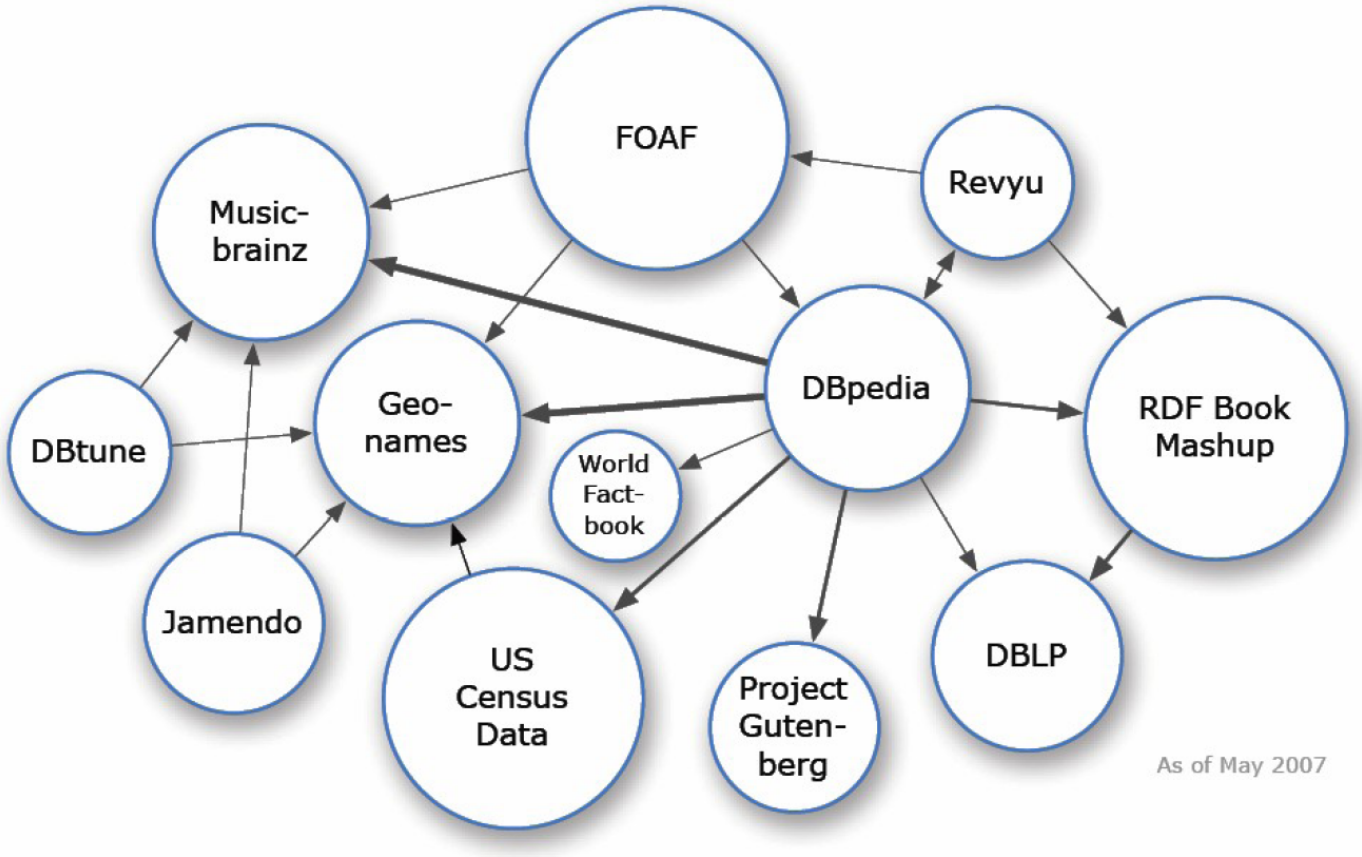
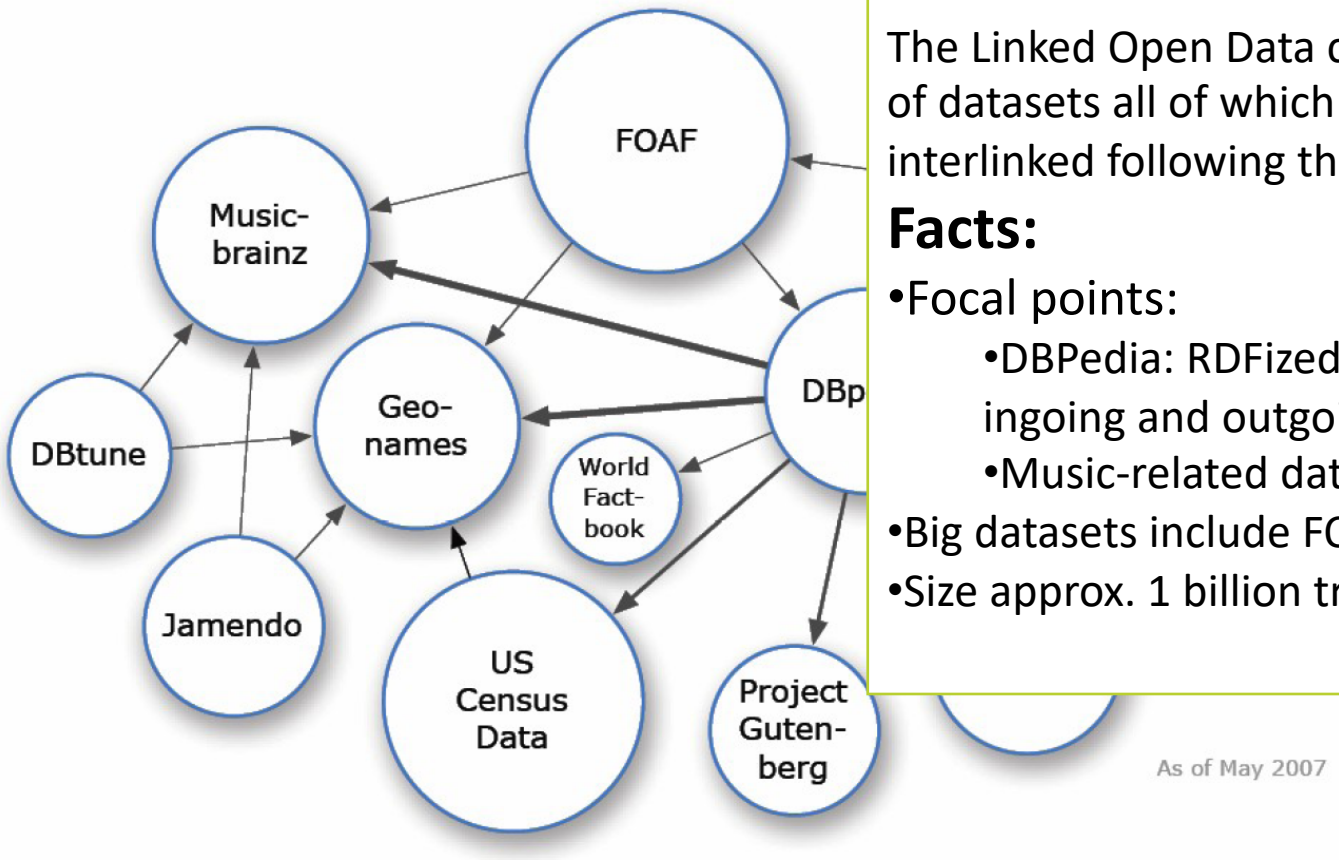


Figure from http://lod-cloud.net/

The Linked Open Data Cloud (May 2007)



Basics:

The Linked Open Data cloud is an interconnected set of datasets all of which were published and interlinked following the Linked Data principles.

Facts:

- Focal points:
  - DBPedia: RDFized vesion of Wikipiedia; many ingoing and outgoing links
  - Music-related datasets
- Big datasets include FOAF, US Census data
- Size approx. 1 billion triples, 250k links

Figure from http://lod-cloud.net/

The Linked Open Data Cloud (March 2009)

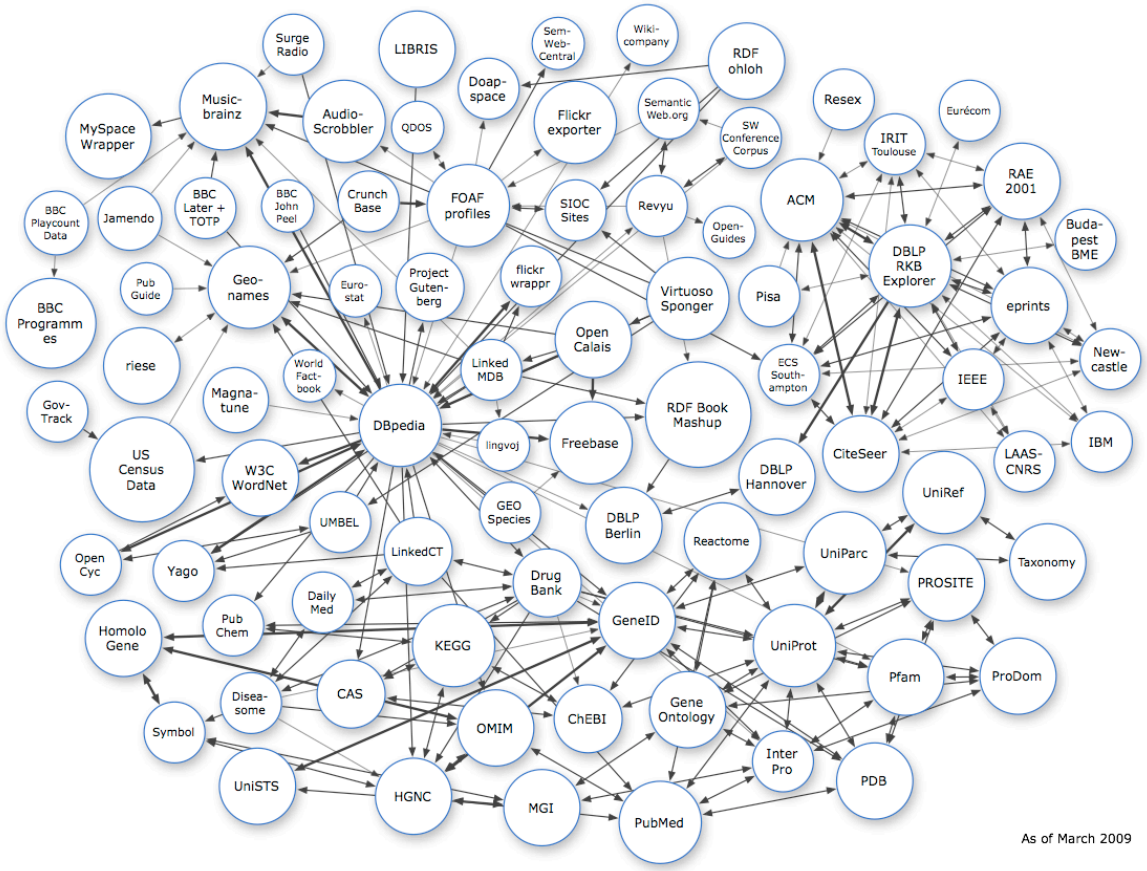
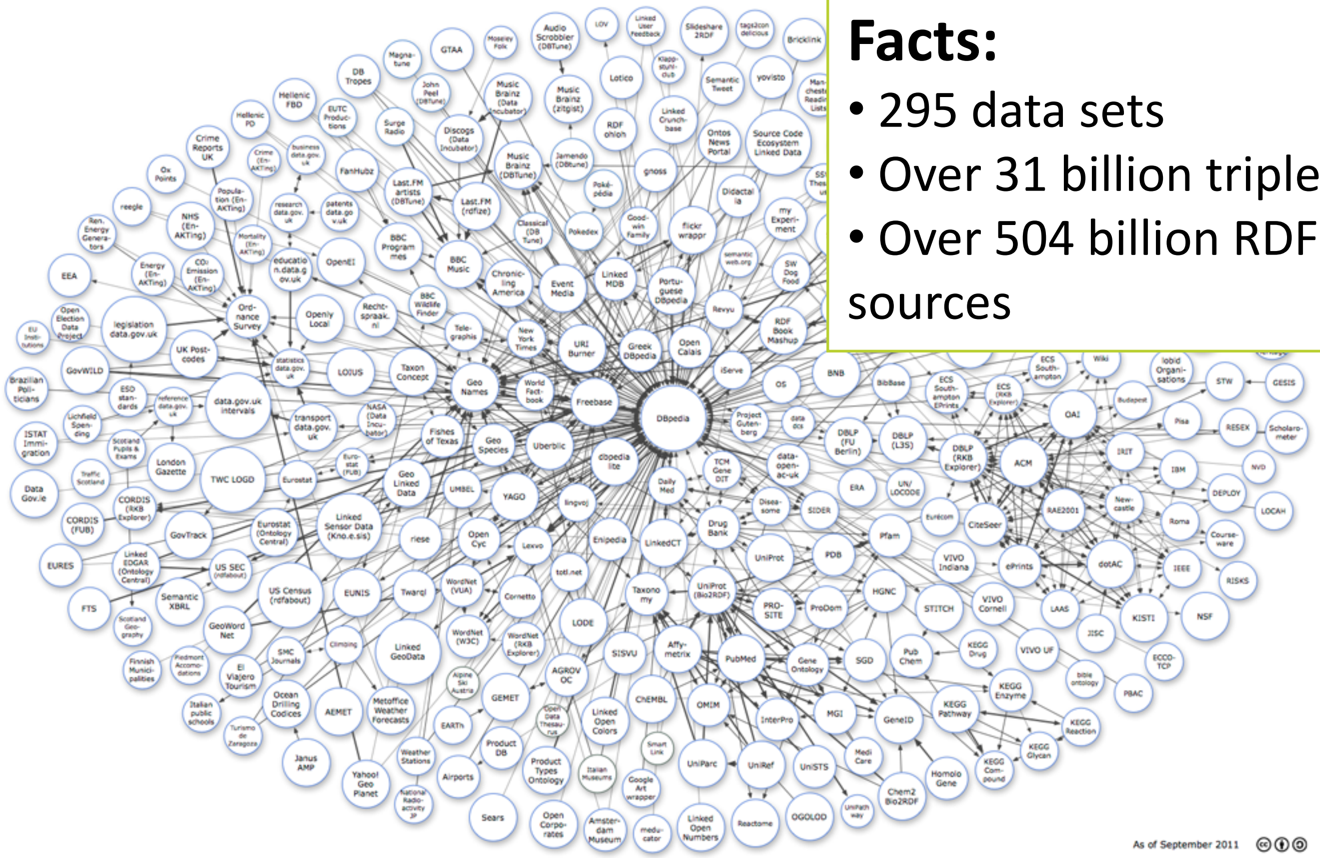


Figure from http://lod-cloud.net/

The Linked Open Data Cloud (Sept. 2011)

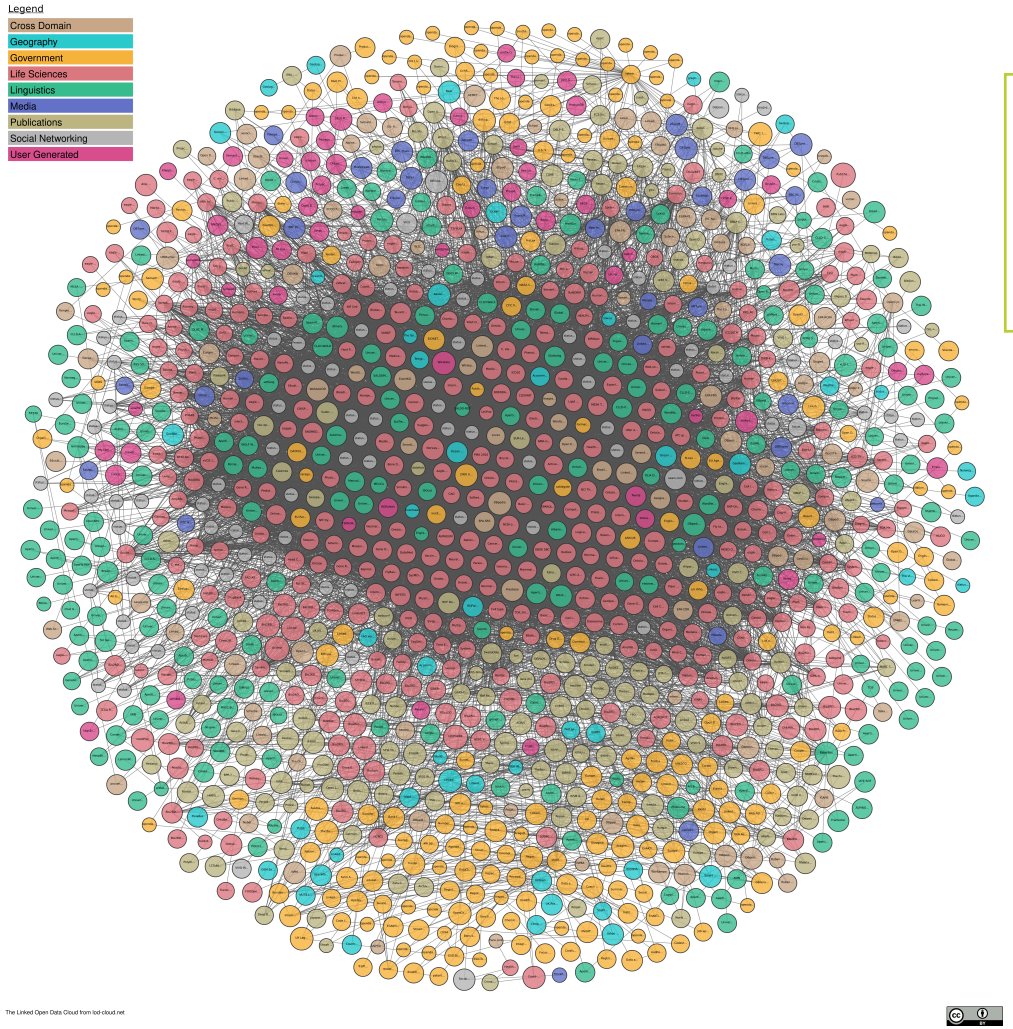


Facts:

- 295 data sets
- Over 31 billion triples
- Over 504 billion RDF links between data sources

Figure from http://lod-cloud.net/

The Linked Open Data Cloud (Nov. 2020)



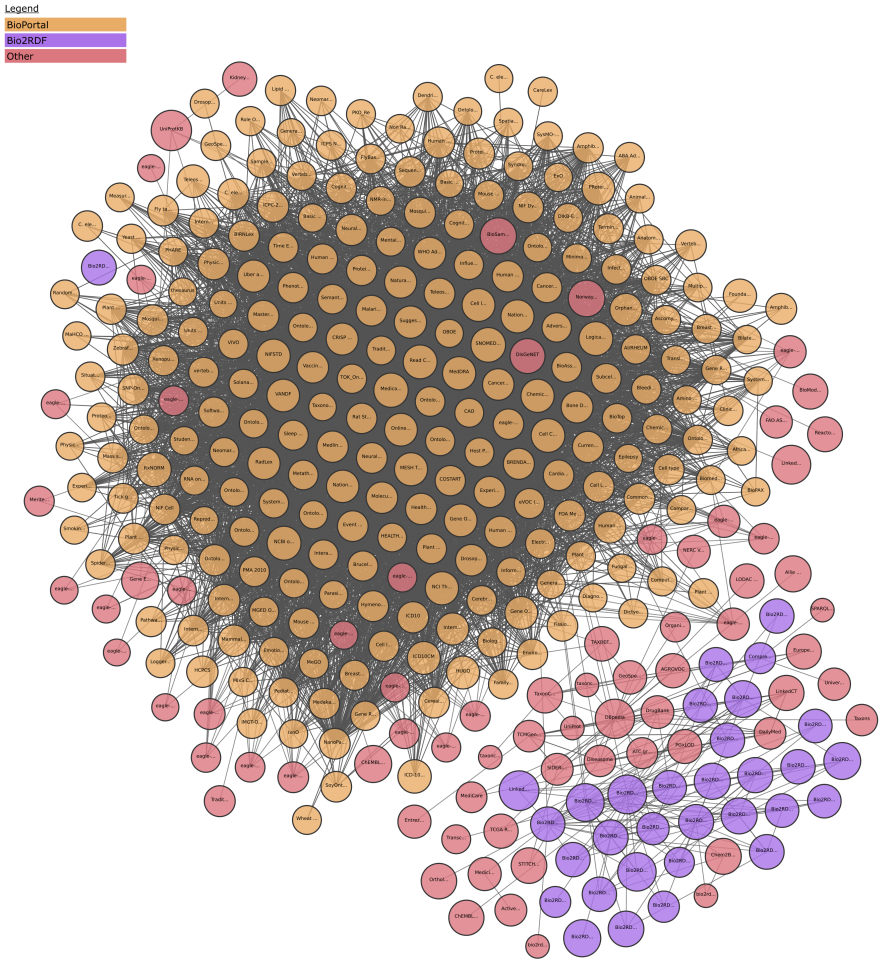
Facts:

- 1269 data sets
- Organized in 9 subdomains

Figure from http://lod-cloud.net/

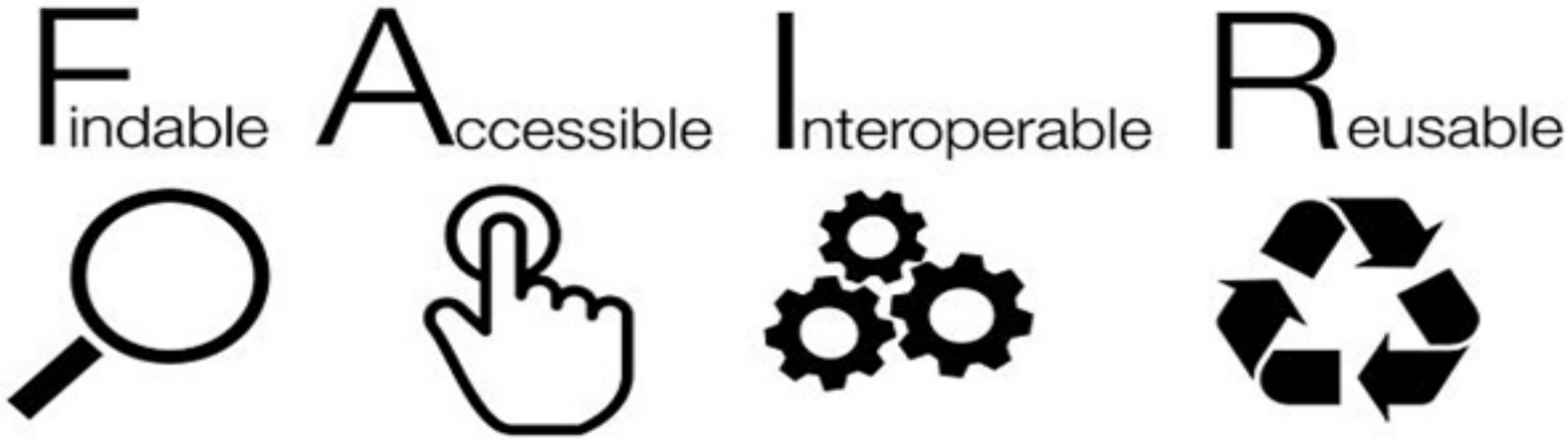
The Linked Open Data Cloud: Life Sciences

- The biggest subcloud in LOD Cloud
- Substantial contribution from BioPortal
- Similar efforts are being promoted for environmental science (LW EcoPortal)



Machine-actionable FAIRness through semantic technologies

FAIR principles



- Allow computational systems to find, access, interoperate, and reuse data
- As automatic as posible
- Semantic technologies play a fundamental role in achieving FAIRness of your data

To be Findable

- (Meta)data are assigned a globally unique and persistent identifier
  - URIs, PIDs
- Data are described with rich metadata
  - Use of LD vocabularies
- Metadata clearly and explicitly include the identifier of the data they describe
  - Link between data and metadata
- (Meta)data are registered or indexed in a searchable resource
  - Triple stores

To be Accessible

- (Meta)data are retrievable by their identifier using a standardised communications protocol
  - E.g. HTTP URIs
- The protocol is open, free, and universally implementable
- The protocol allows for an authentication and authorisation procedure, where necessary
- Metadata are accessible, even when the data are no longer available
  - Independent graphs

To be Interoperable

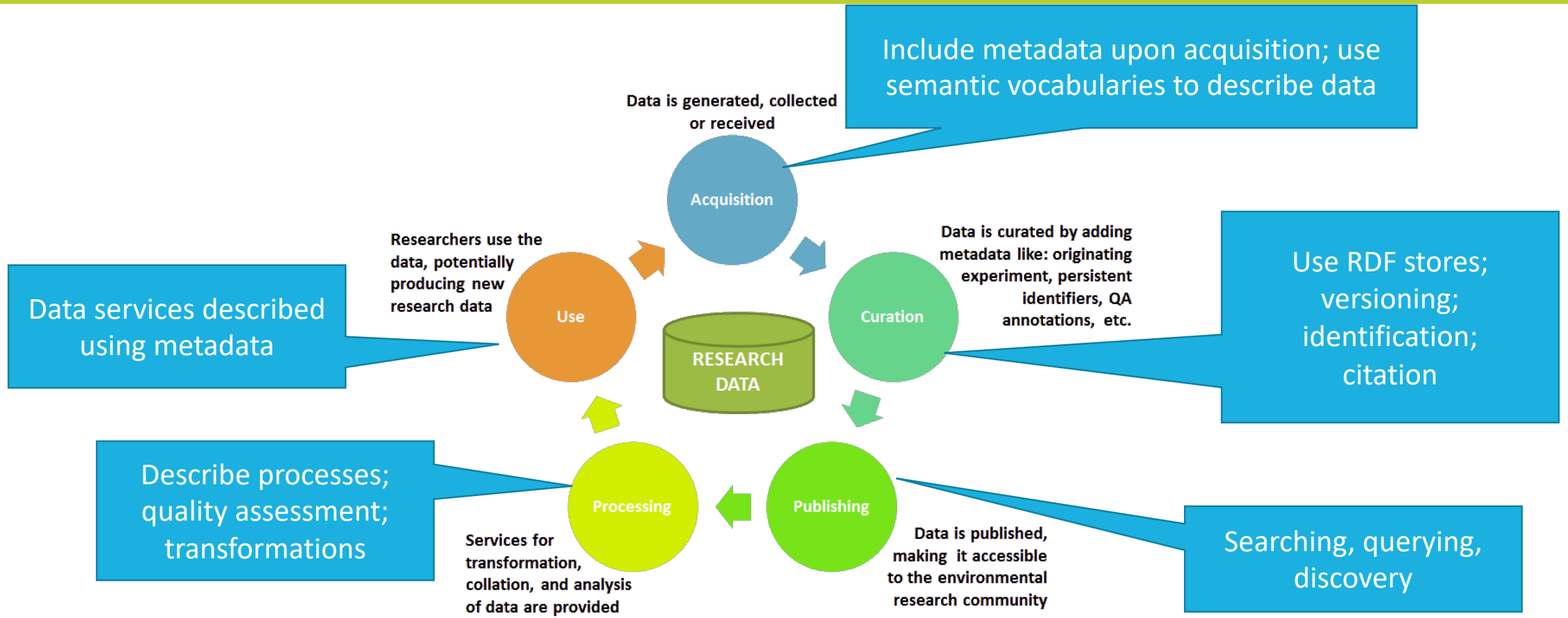
- (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
  - RDF, OWL, JSON-LD…
- (Meta)data use vocabularies that follow FAIR principles
  - Linked Data vocabularies
- (Meta)data include qualified references to other (meta)data
  - Semantic relationships

To be Reusable

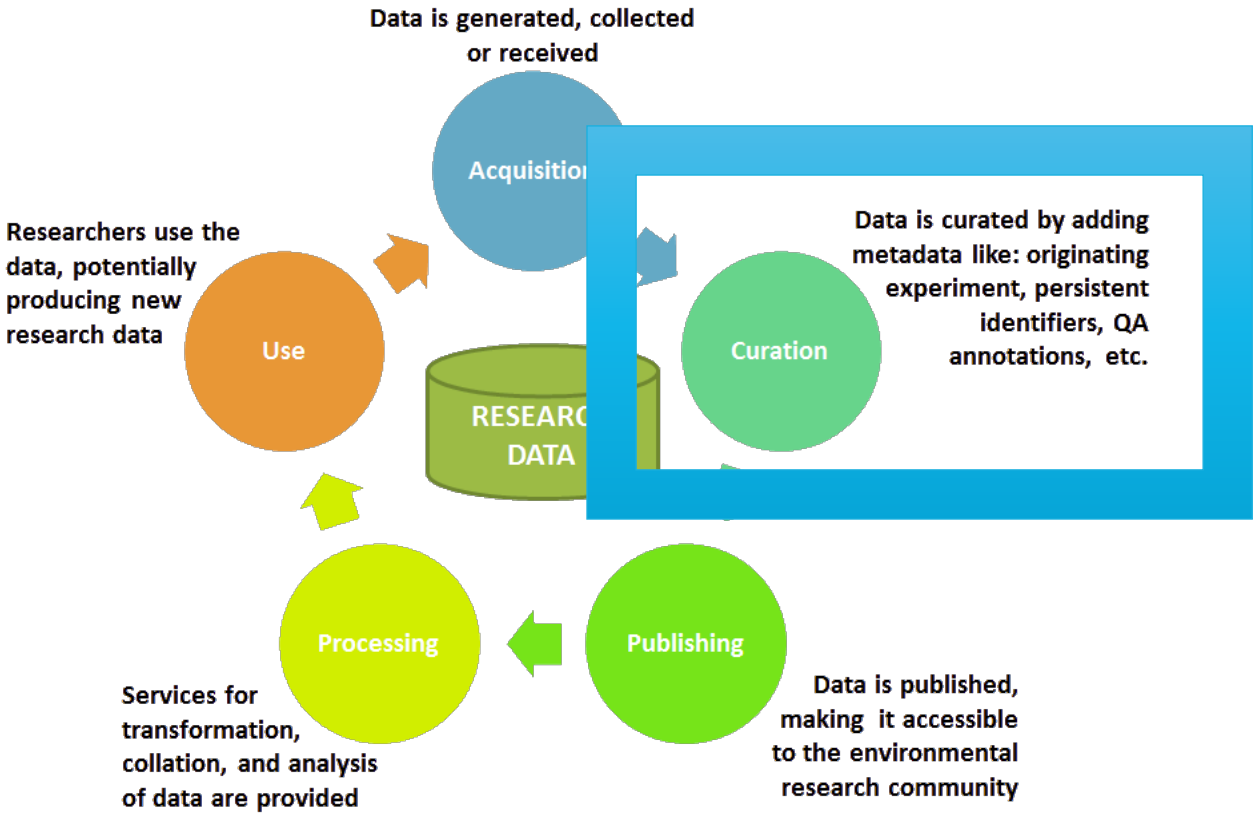
- (Meta)data are richly described with a plurality of accurate and relevant attributes
  - Use of arbitrary number of metadata vocabularies
- (Meta)data are released with a clear and accessible data usage license
  - License ontologies (e.g. CC-RDF)
- (Meta)data are associated with detailed provenance
  - Provenance vocabularies
- (Meta)data meet domain-relevant community standards
  - Domain-specific ontologies/vocabularies

Applying Semantics to the research data lifecycle

ENVRI RM Research Data Lifecycle



Let's focus on curation



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)

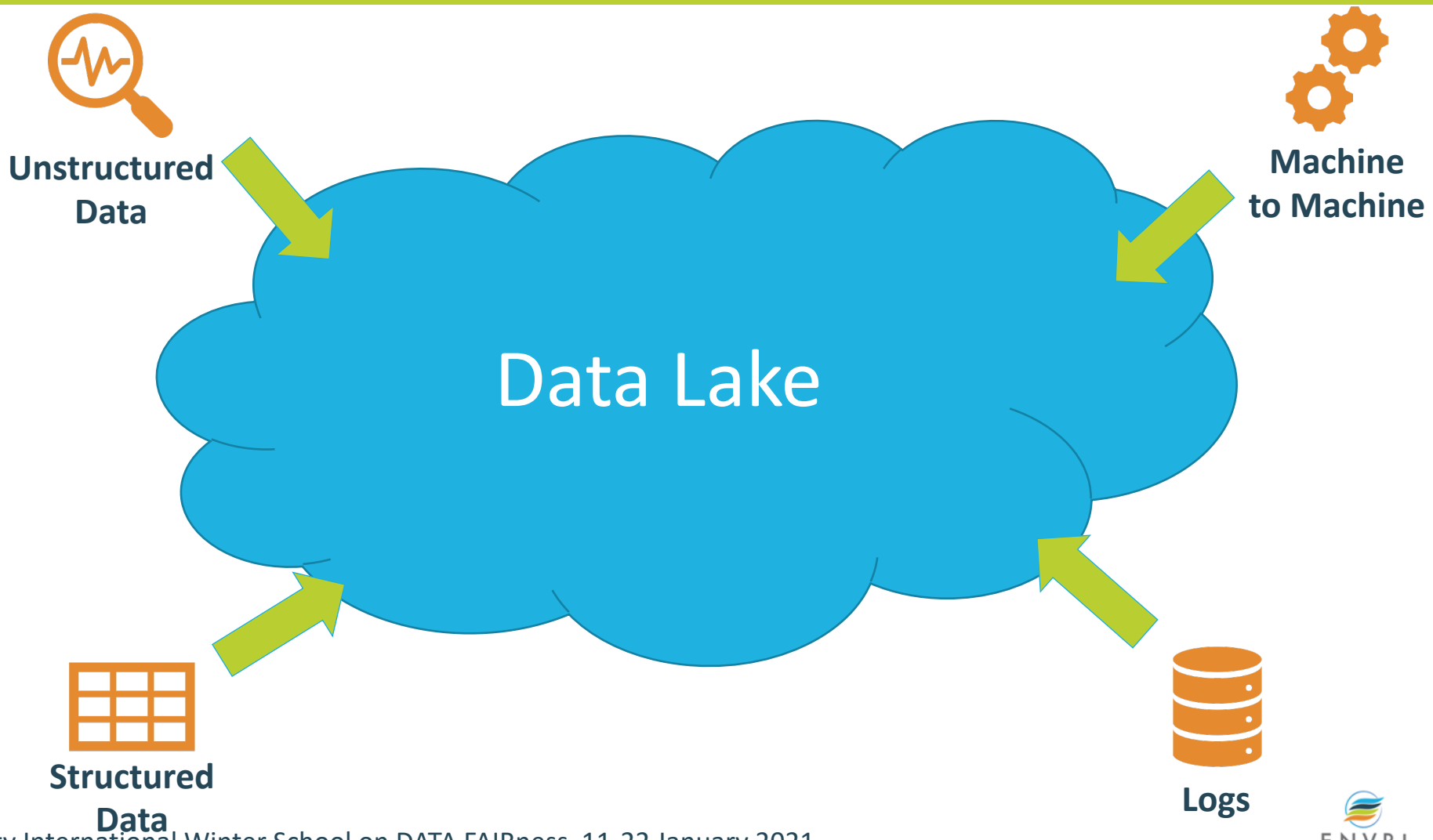
*"Digital curation involves maintaining, preserving and adding **value** to digital research data throughout its **lifecycle**"*

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

Data curation lifecycle

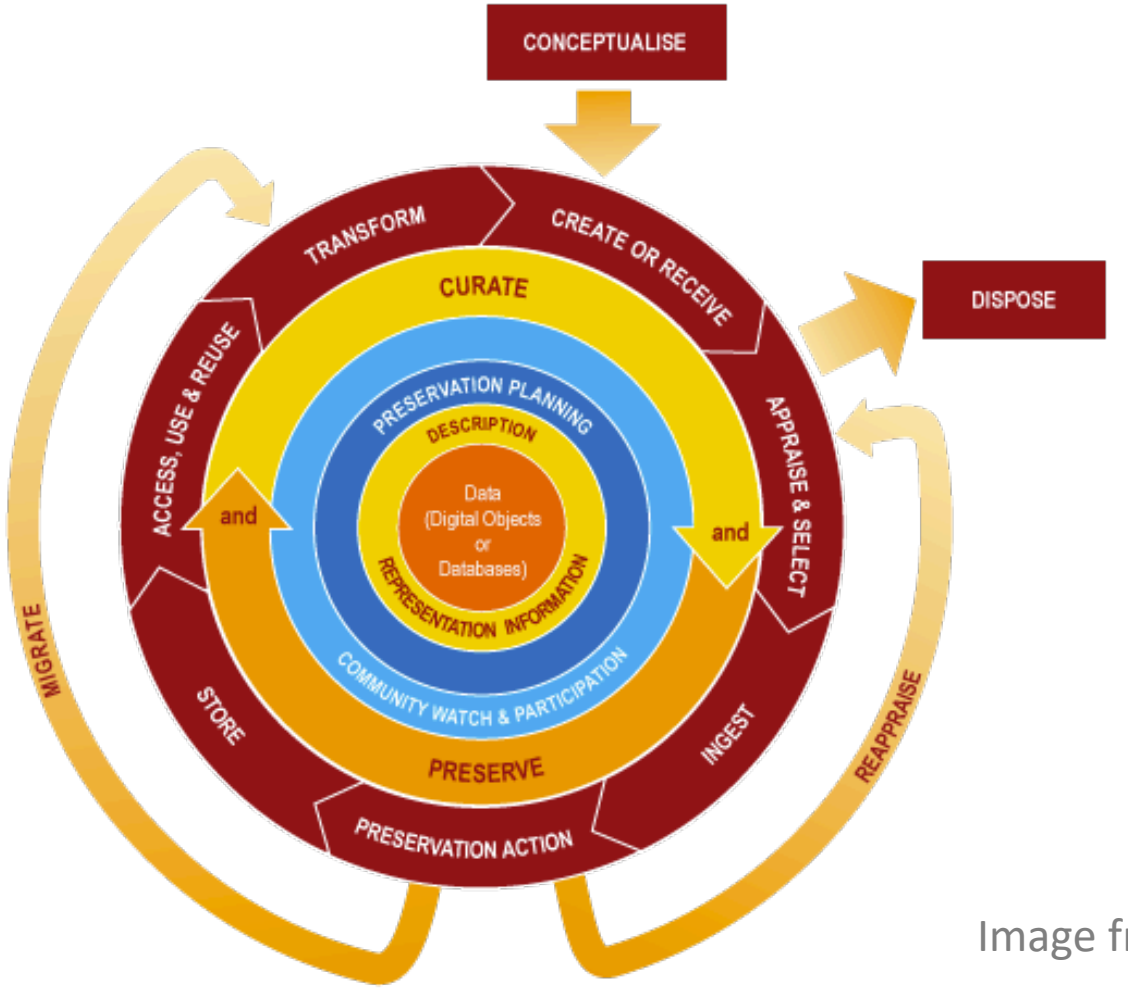
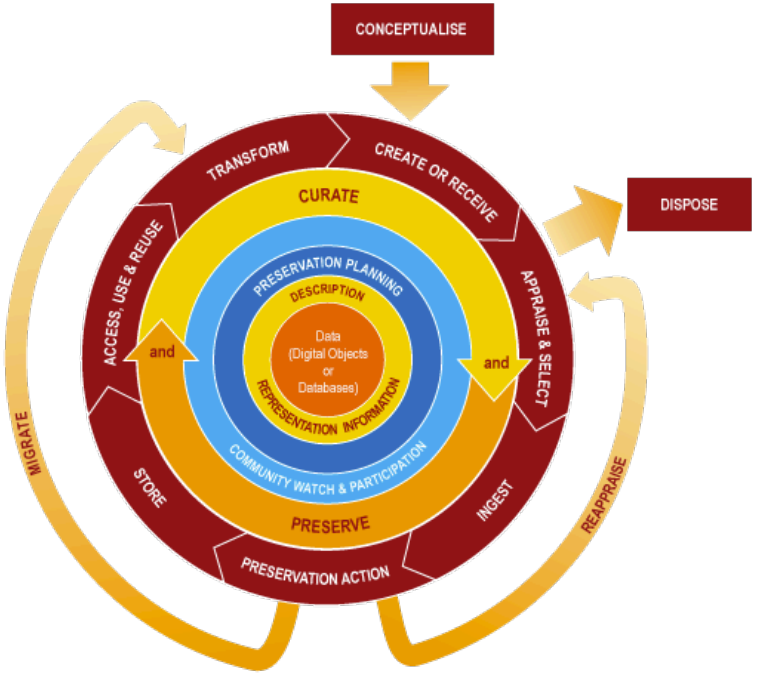


Image from the Digital Curation Centre, UK

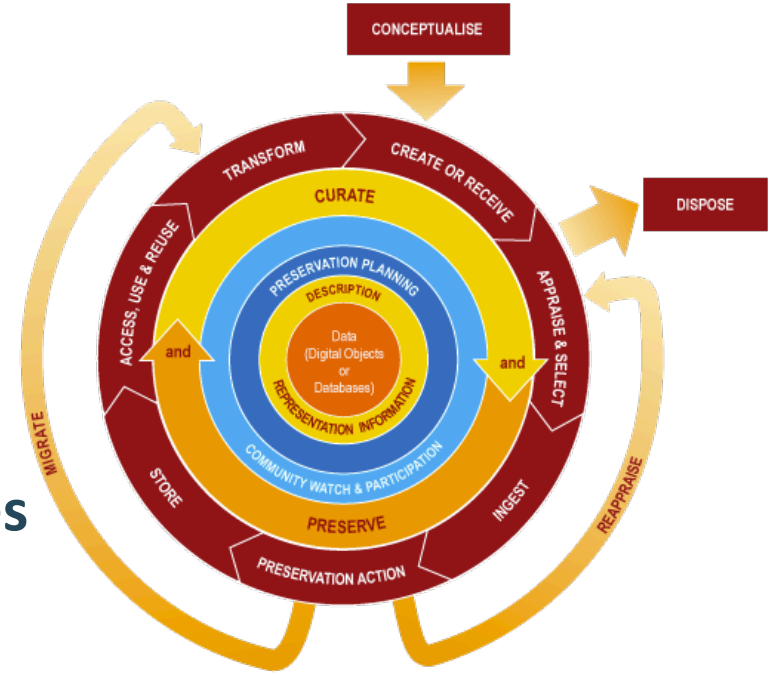
Data curation generic activities

- 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



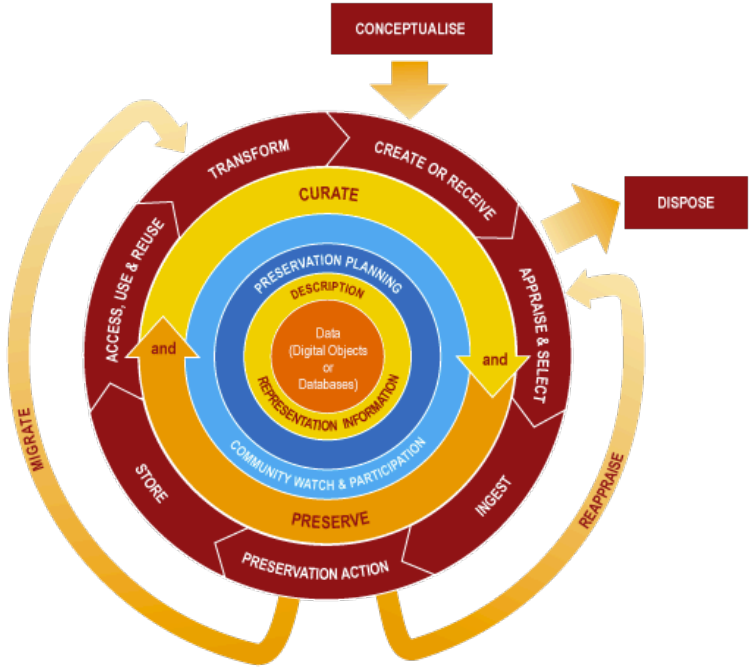
Data curation sequential activities (I)

- 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



The need for semantics in data curation

- Metadata
  - Data that provides information about other data
- Semantic annotations of data (and datasets) to provide context
- Useful along different activities of the curation process
- Providing building blocks for FAIRness during the rest of the research data lifecycle

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

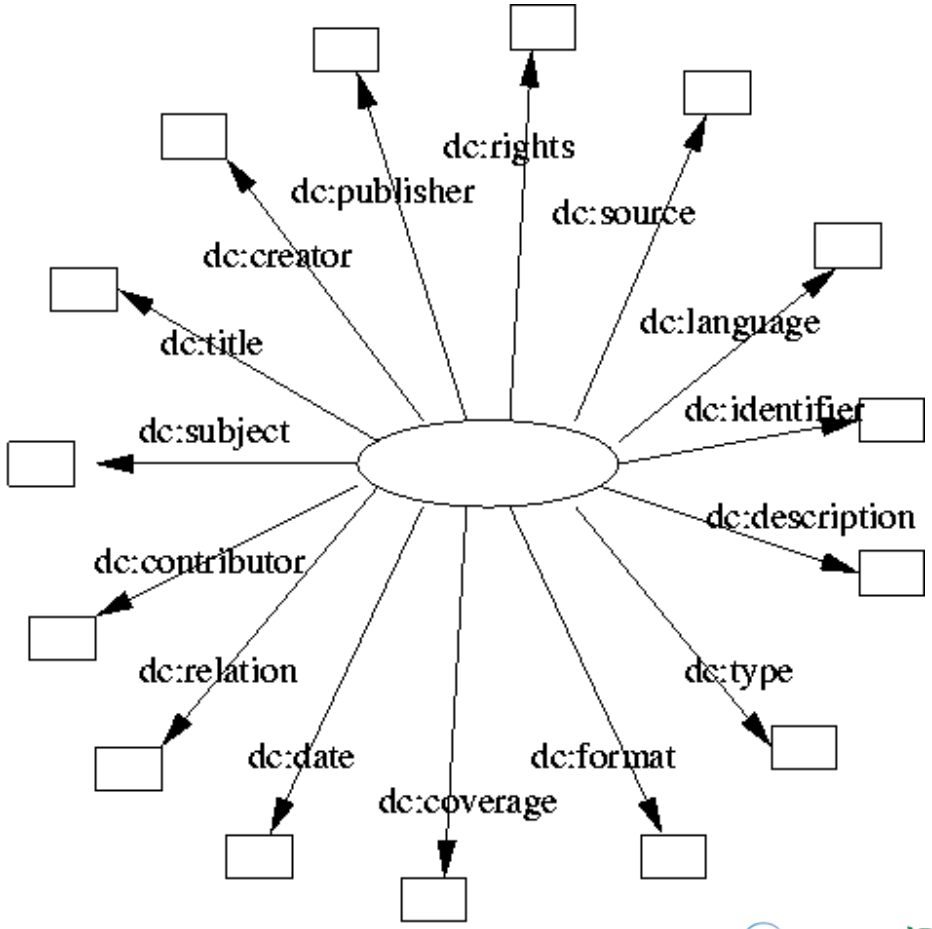
- Different from data in mode of use
- Not just for data, but also for users, services, computing resources…
- Not just for description and discovery, but also for contextualization and interlinking
- Must be machine and human understandable
- Management (meta)data is also relevant (research proposals, funding, projects…)

Persistent Identifiers

- Fundamental for data curation
- Uniqueness
- Reusable between datasets
- Helps long-term preservation
- URIs, DOIs, ORCID

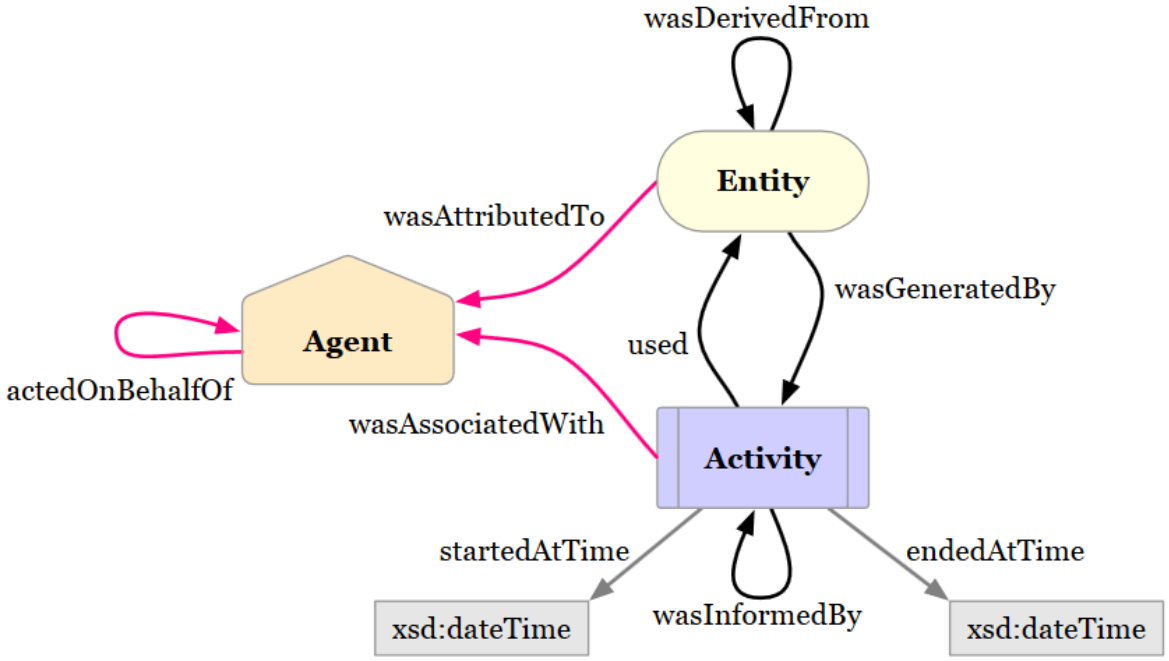
Generic vocabularies for metadata

- Dublin Core
- Friend-of-a-friend
- …



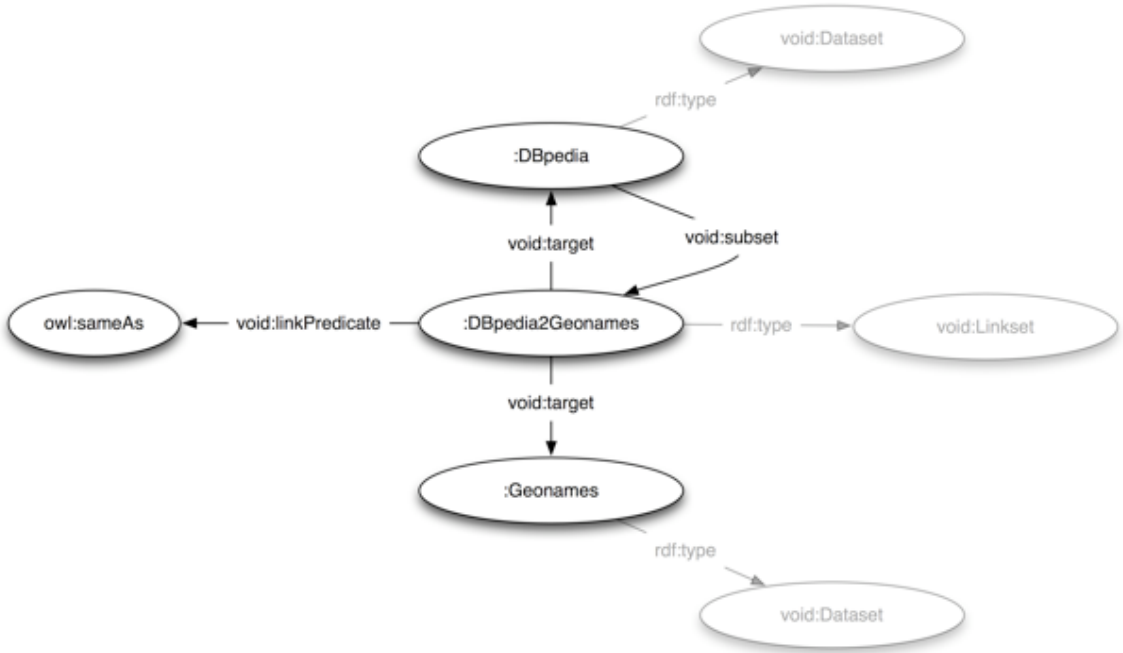
Provenance

- Information about entities, activities and agents involved in producing something
- Chronology of ownership, custody, location, transformation
- PROV framework from W3C



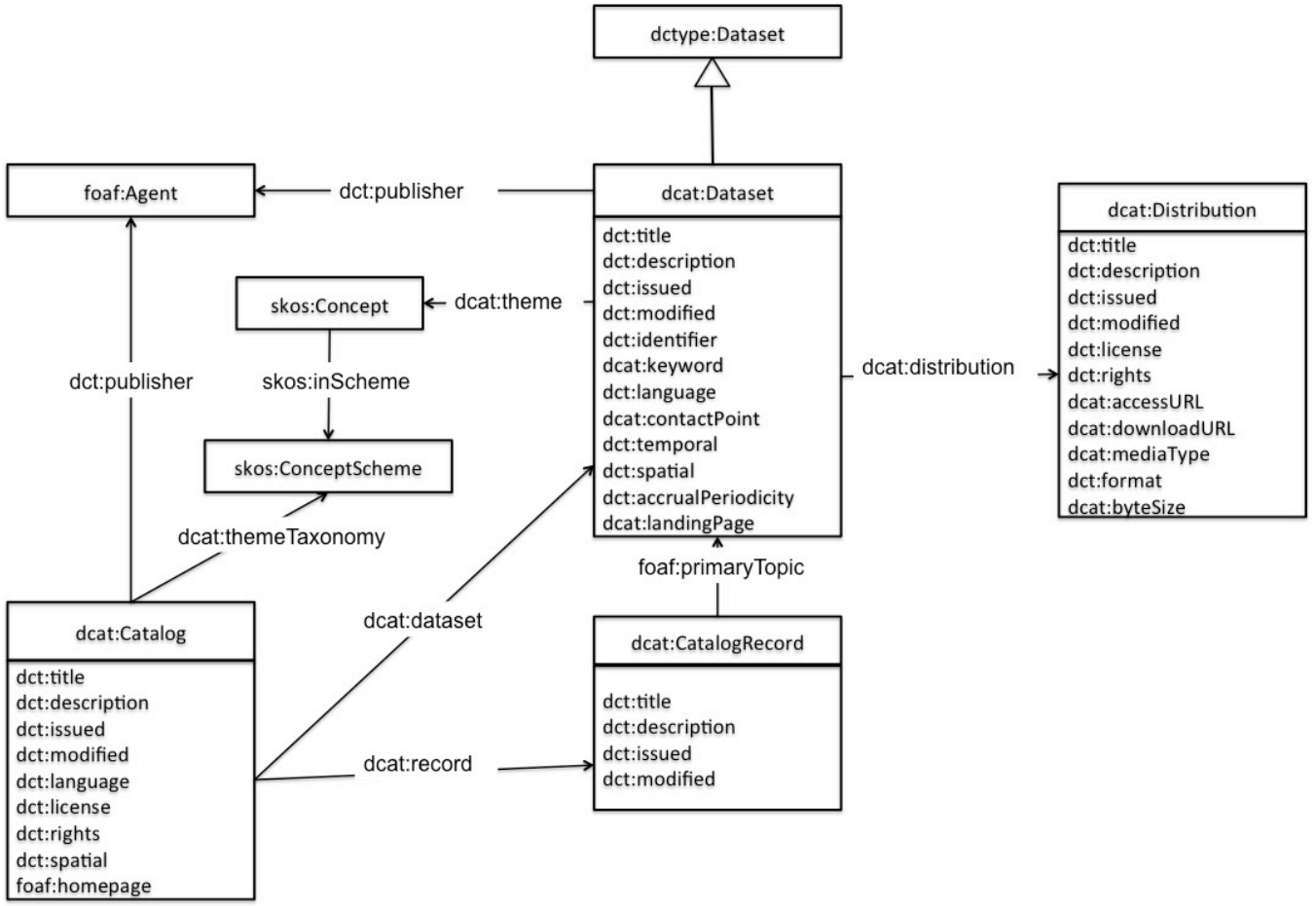
Dataset description

- Vocabulary of Interlinked Datasets - VoID (W3C)
- General dataset metadata (license, subject, features…)
- Access (endpoints, distribution…)
- Dataset structure (examples, vocabularies used, links)



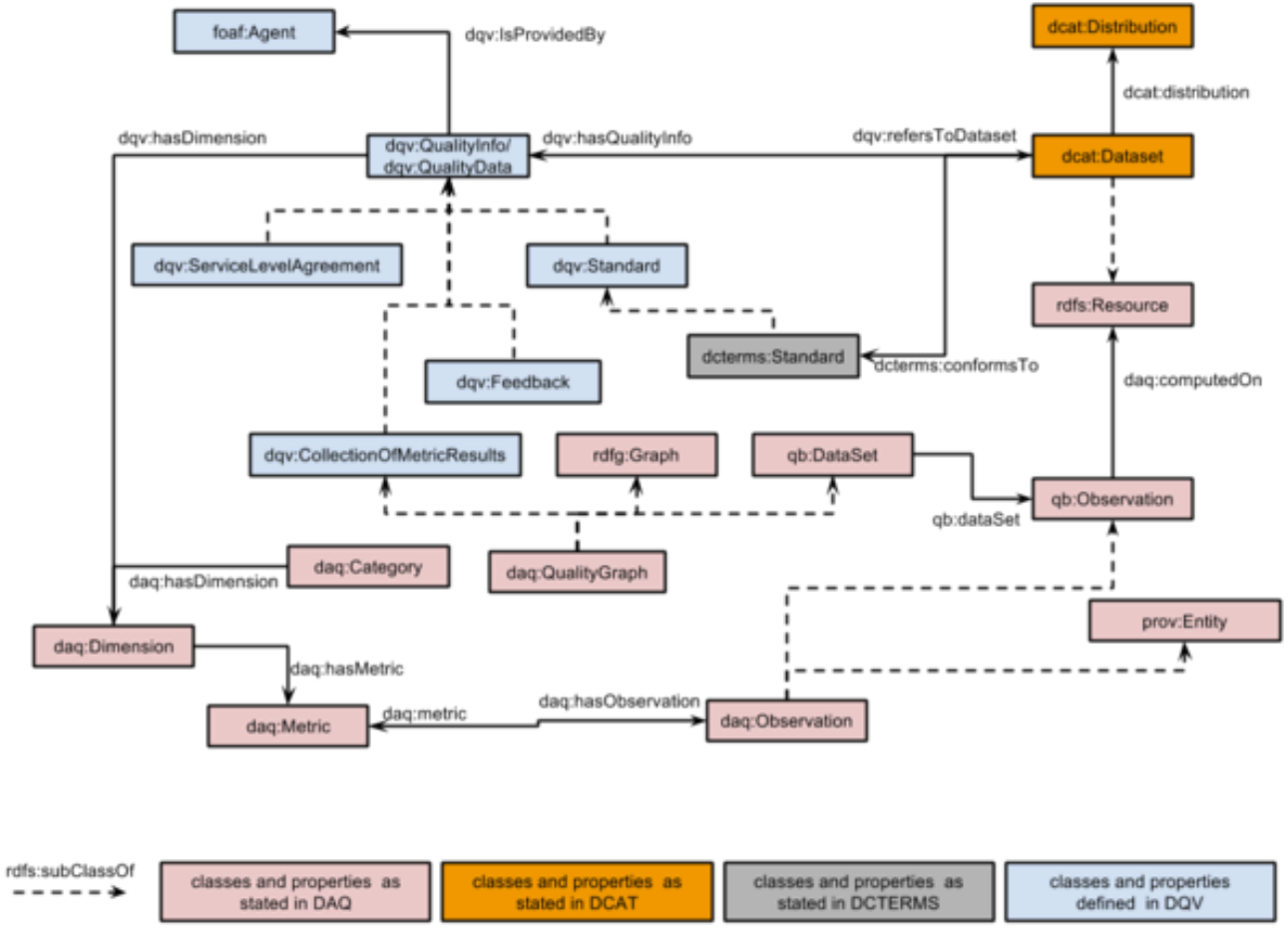
Data catalogs description

- Data Catalog Vocabulary – DCAT (W3C)
- Facilitate interoperability between data catalogs

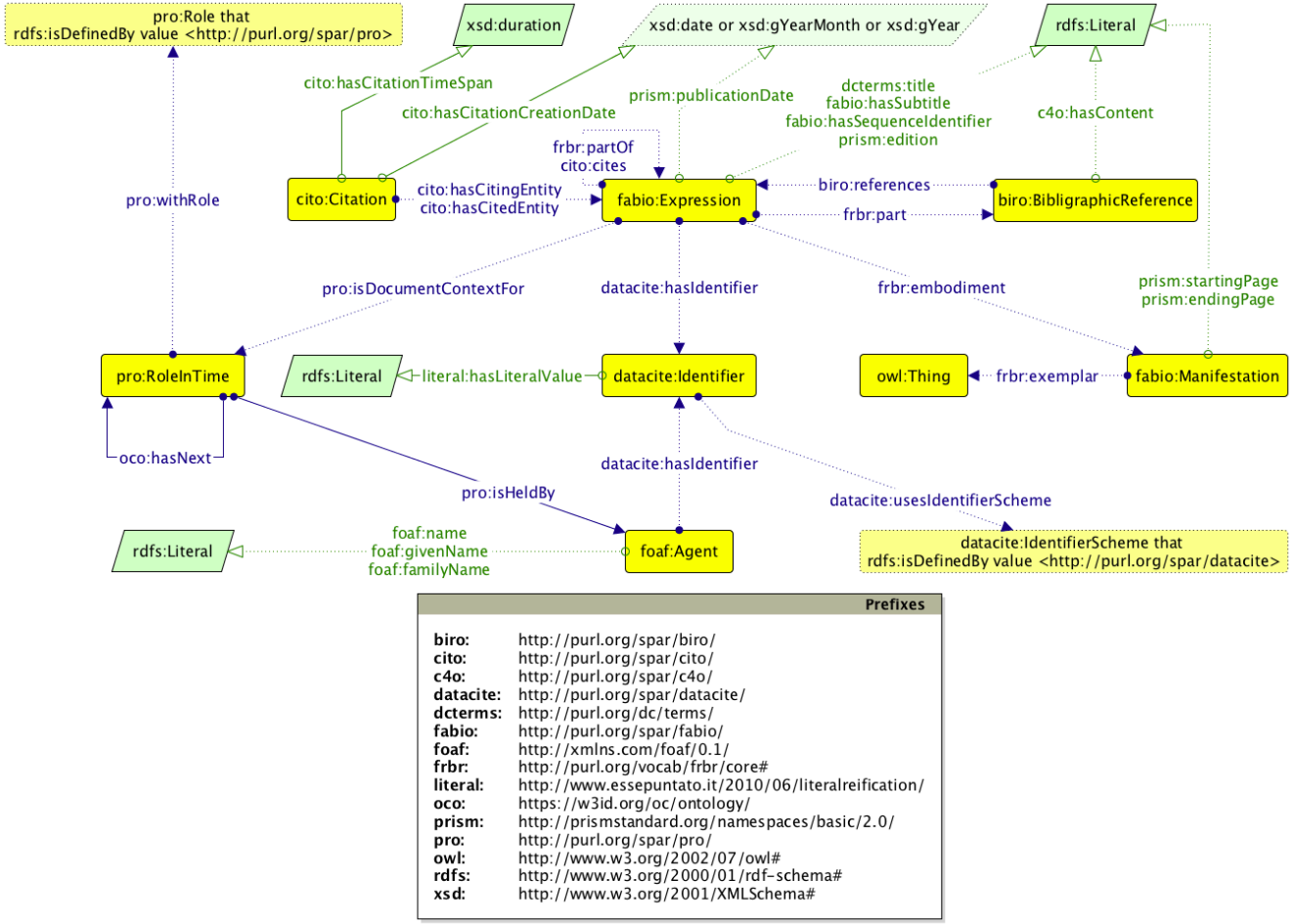


Quality assurance

- Data Quality Vocabulary (W3C)
- Applied to the dataset

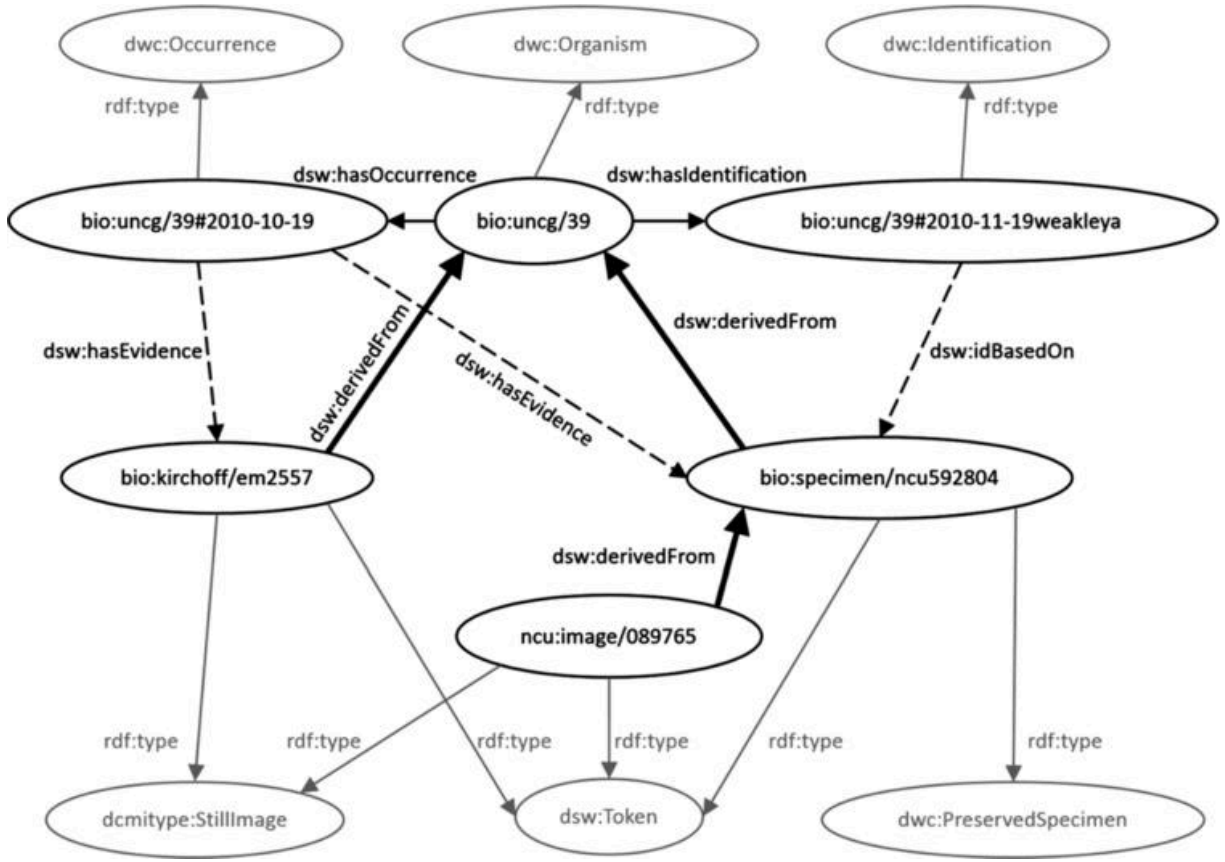


OpenCitations

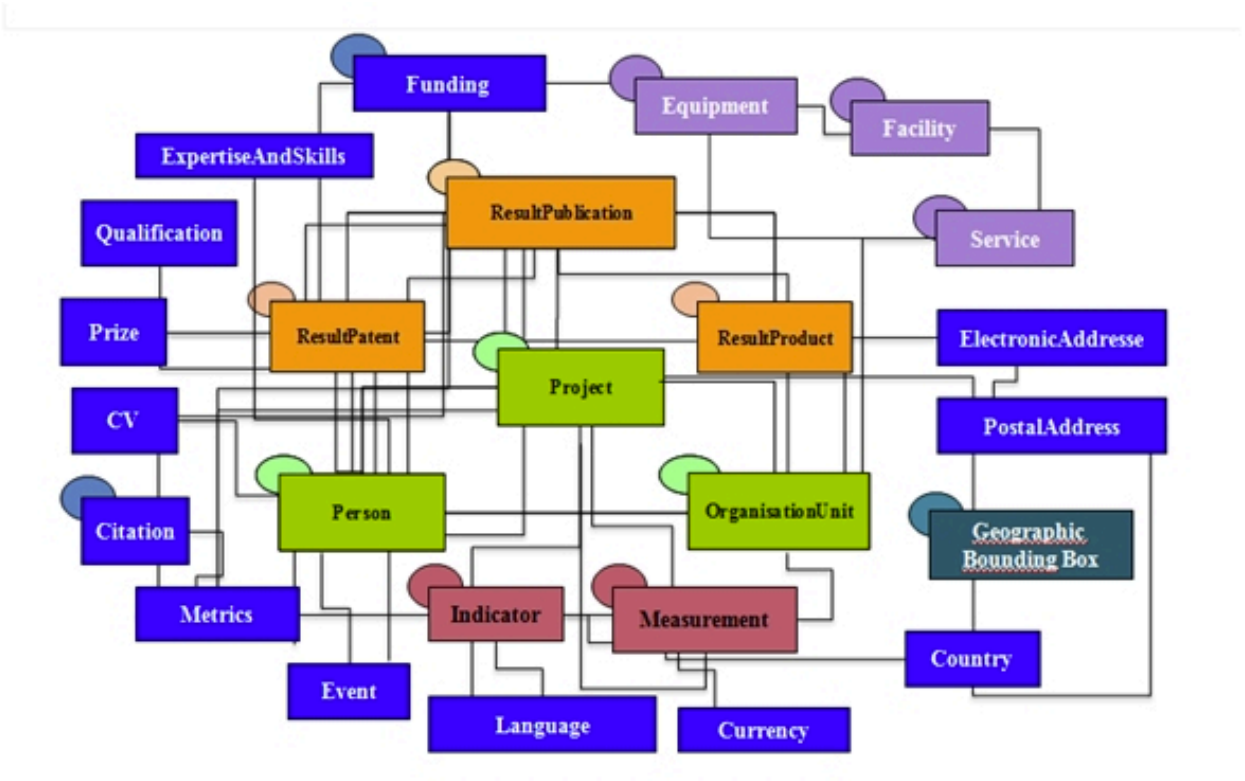


Metadata for environmental sciences and RI

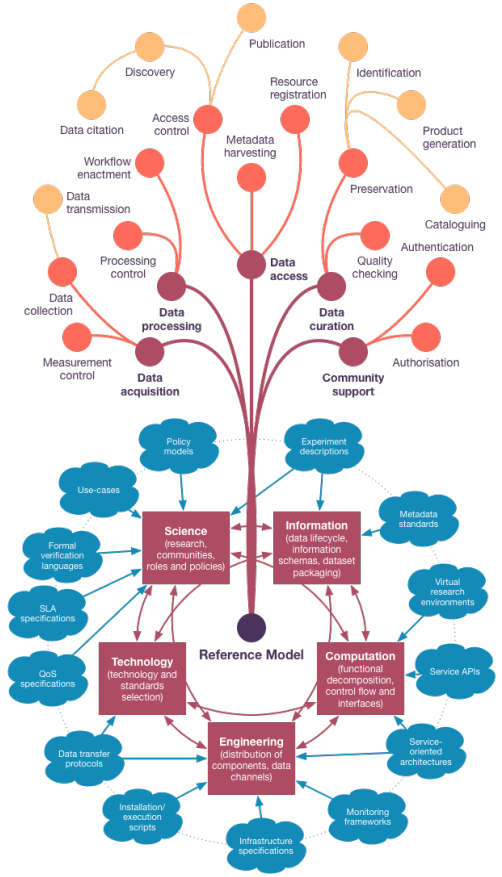
- Darwin Core
- Ecological Metadata Language (EML)
- Environmental Monitoring Facility (EMF)



Metadata for environmental sciences and RI



Common European Research Information Format (CERIF)



ENVRI-RM

Tooling support for data curation

- Annotation tools
- Assign PIDs to every object
- Customize the metadata schema
- Interlinking support
- Versioning
- Quality assessment

Security and additional aspects

Hands-on session

Hands-on session

- We will use WebProtege to collaboratively develop some ontologies
- https://webprotege.stanford.edu/
- Form groups of ~6 members (the more diverse background the better)
- Decide on some research use case that you have data
- Create classes, properties, and instances (individuals) that correspond to the concepts used in your use case
- (Optionally) Reuse some (meta) data ontologies and vocabularies
- Interactions and group formation at Wonder.me
  - https://www.wonder.me/r?id=1eb7bc05-9894-422e-a139-00f86d9fd007
  - Password: ***WinterS21****

ENVRI

FAIR

envri.eu/envri-fair

@envri_fair

company/envri-fair

facebook.com/ENVRIcomm