Web, Data & Provenance

Provenance Research Workshop

Castello di Rivoli Turin

September 26, 2019

Marilena Daquino, marilena.daquino2@unibo.it

About me

RESEARCH ASSOCIATE @unibo/DH.arc

Digital Humanities Advanced Research Centre, University of Bologna

PHD IN LIBRARY AND INFORMATION SCIENCE

Thesis on: Reasoning over argumentations around attributions recorded in art historical photo archives

Consultant for the **Federico Zeri Foundation**

Experience in **Knowledge organisation** and Semantic Web technologies

Currently working on **citation mining** in scholarly data

marilena.daquino2@unibo.it | @emmedaquino
https://marilenadaquino.github.io

Outline

PART I - INTRODUCTION

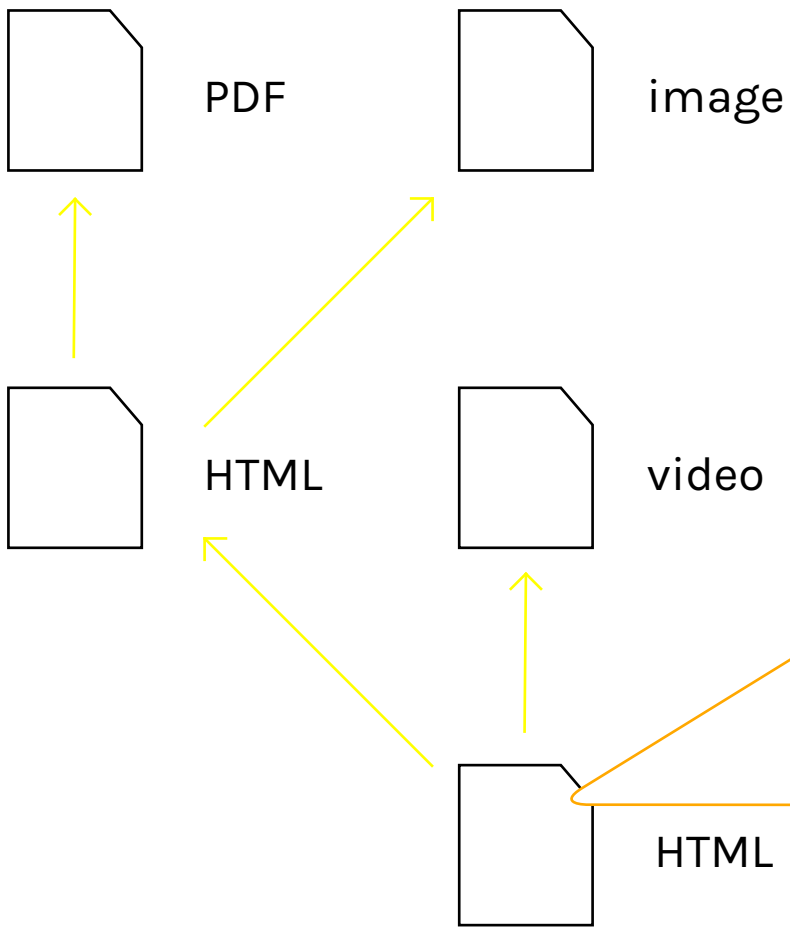
- Semantic Web technologies in nuts
- Organize, Produce, and Query Linked Open Data
- Ontologies and ontology development

PART II - HANDS-ON

- The landscape of Linked Open Data for provenance research
- Working groups

PART I - INTRODUCTION

The web of documents

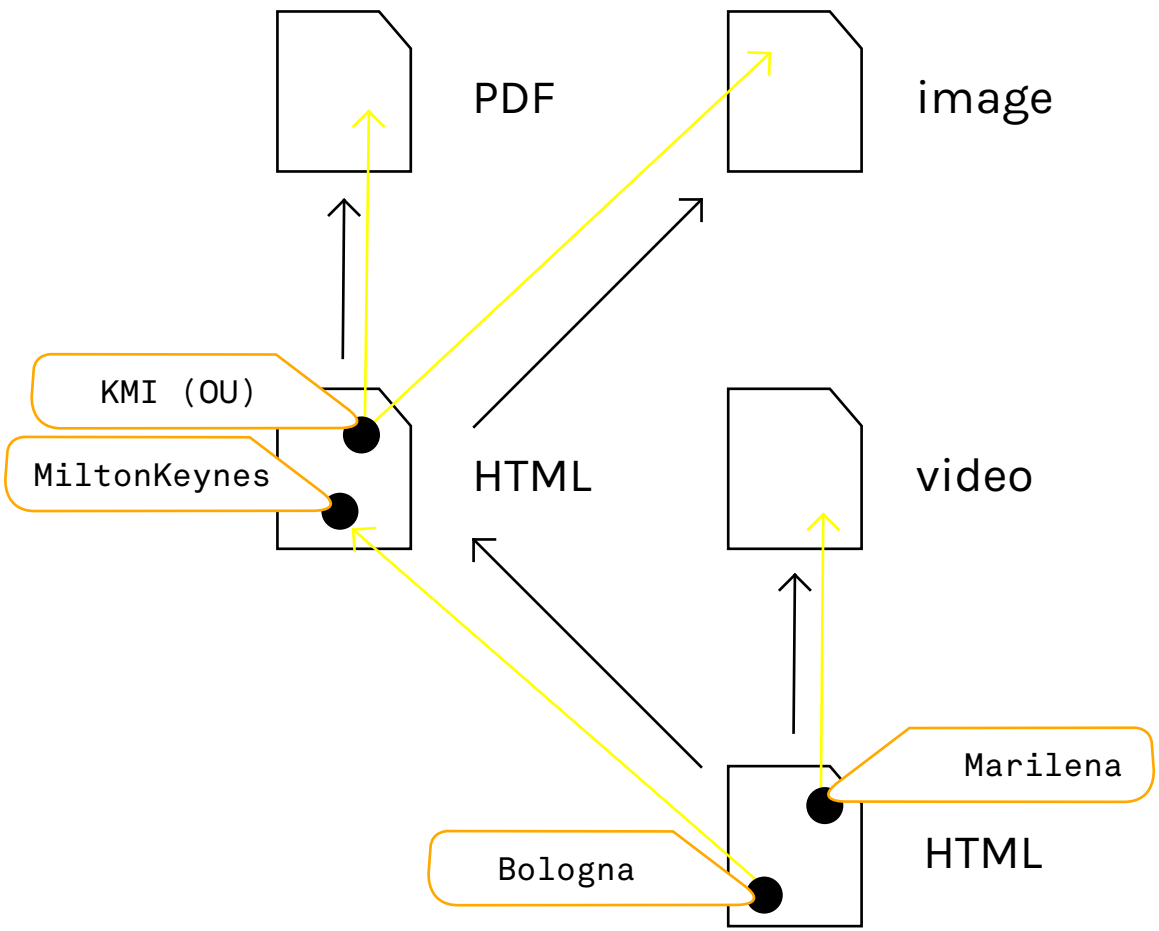


Web documents are linked with each other by means of **hyperlinks.**

The machine cannot interpret the meaning of the link, nor understand what these documents are about.

Marilena Daquino, marilena.daquino2@unibo.it

The web of data (or Semantic Web)



Data included in Web documents describe real objects or concepts. These can be linked with each other by means of links describing **semantic relations.**

Data underlying documents are provided in a **machine-readable format**.

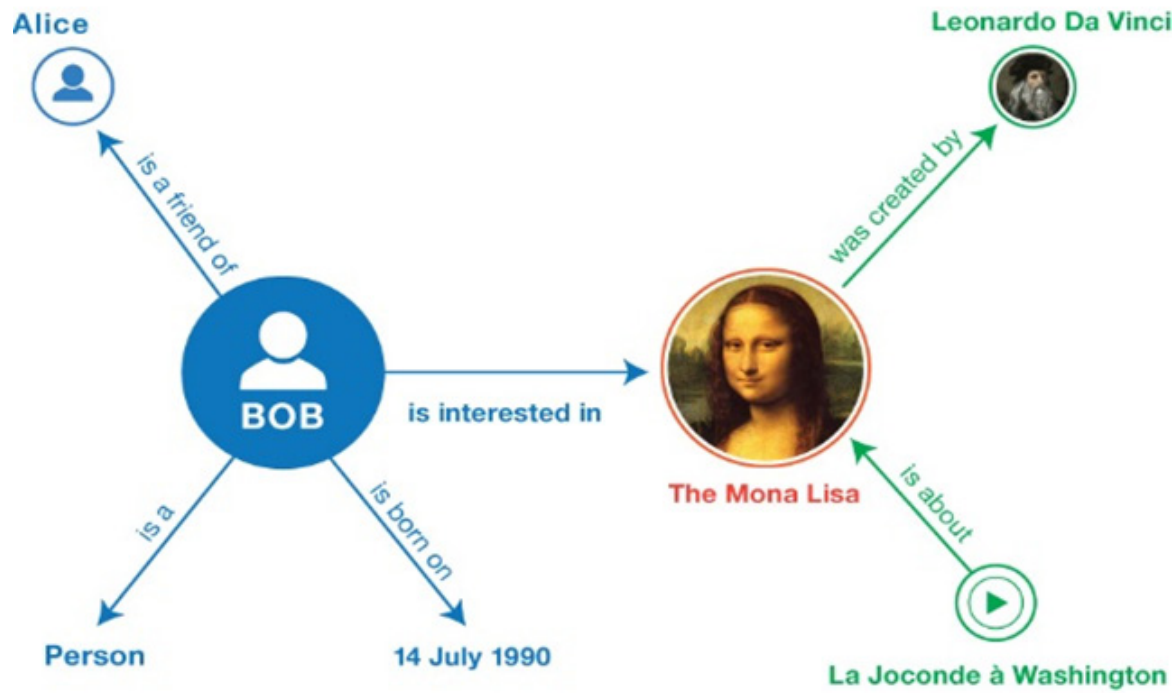
Rather than being something different from the web we know, it is an **extension** of the current web.

What are Linked Open Data?

A **way to publish structured data** underlying documents (web or not) **on the web** and integrate information belonging to **different sources** according to W3C (World Wide Web Consortium) standards.

In order to be linked, data must be freely **accessible** by users and applications, so that they can be reused (i.e. cited) in smart applications.

Linked Data can be incapsulated in web pages (adding annotations) or published as separate documents.

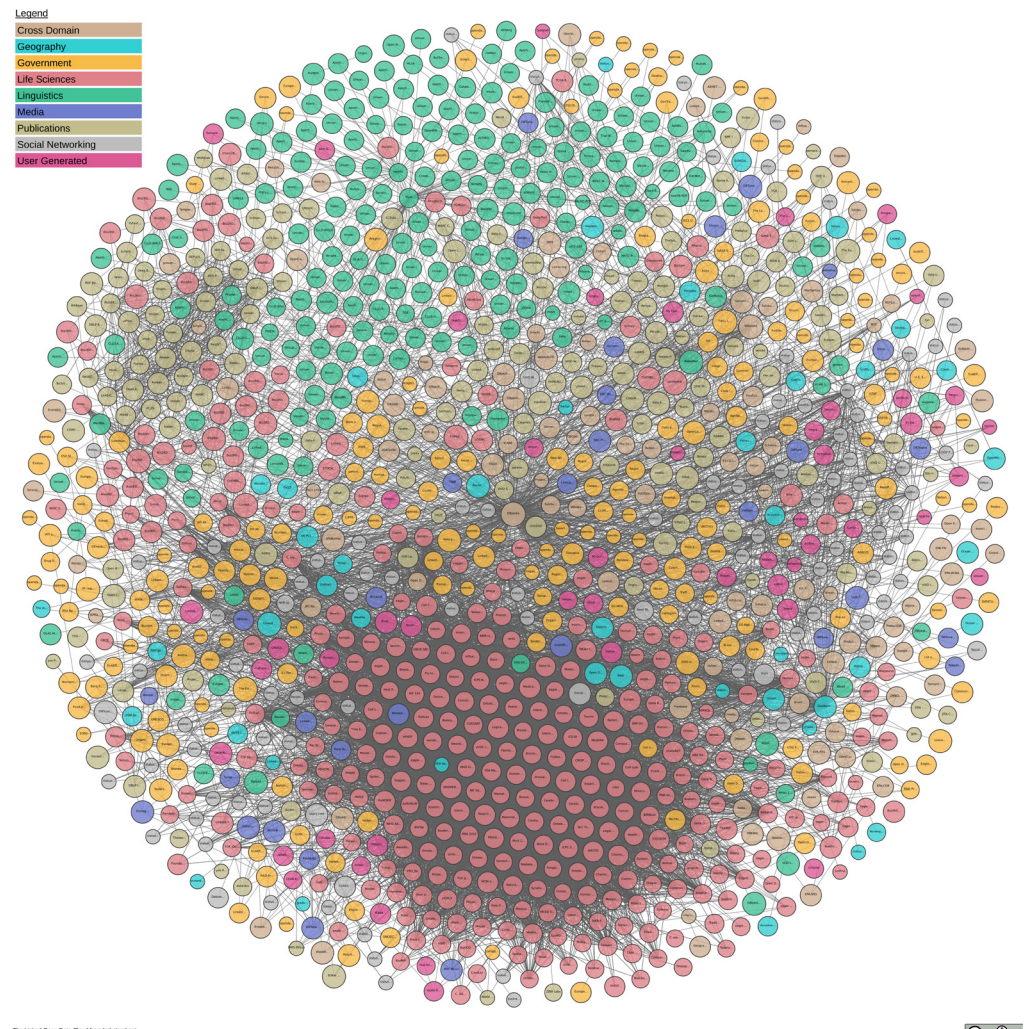


Source: https://www.w3.org/TR/rdf11-primer/

Who is using Linked Data?

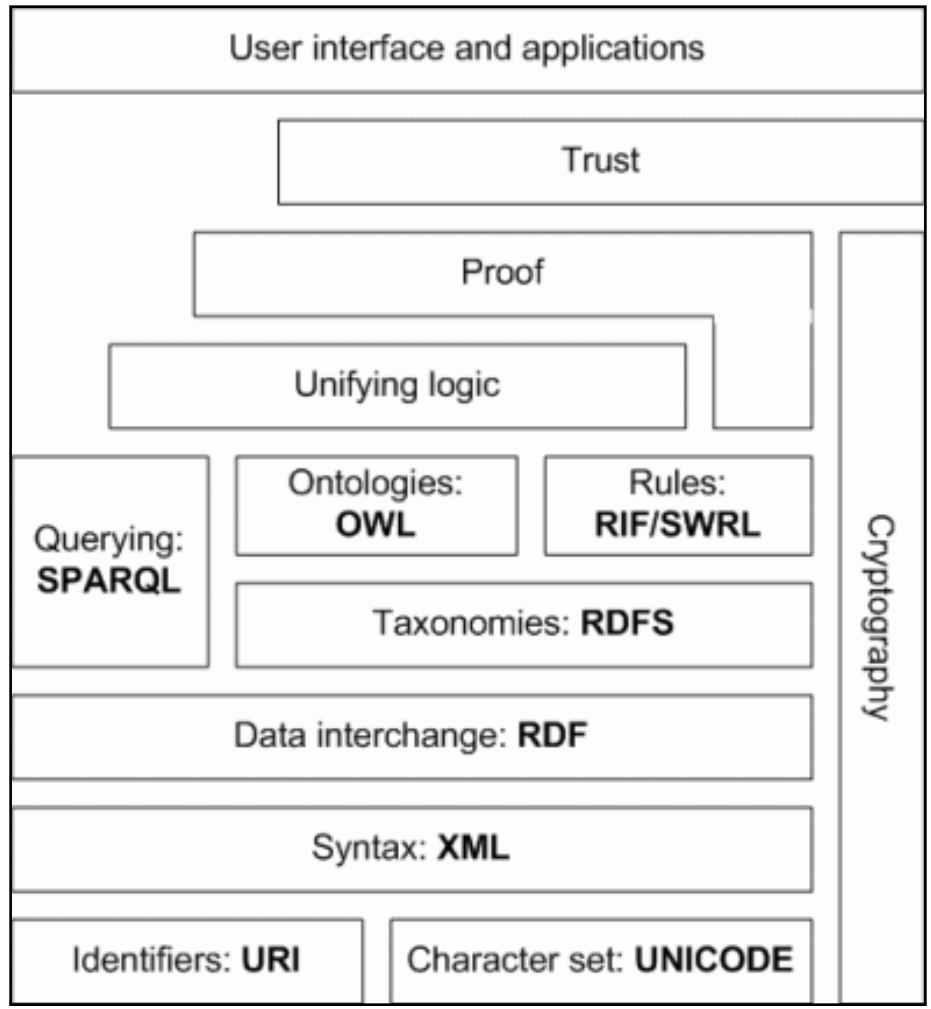
The LOD Cloud includes about 1400 datasets, belonging to diverse domains (healt, government, geography) or crossdomain (such as DBpedia, Google, Yahoo).

Cultural institutions, such as libraries (Library of Congress, BNF), museums (Rijksmuseum, Smithsonian), archives (Yale Center of British Art, Vatican) publish LOD as well so as to enable users to perform research on their data.



Marilena Daquino, marilena.daquino2@unibo.it

Semantic Web technologies stack



UNICODE

Machines basically read numbers: Unicode associates to every character or glyph (in around 150 modern and historic languages) a unique number.

It has several implementations, among which there is **UTF-8**, the most used in the web.

The standard is at the basis of many **programming languages** and web browsers.

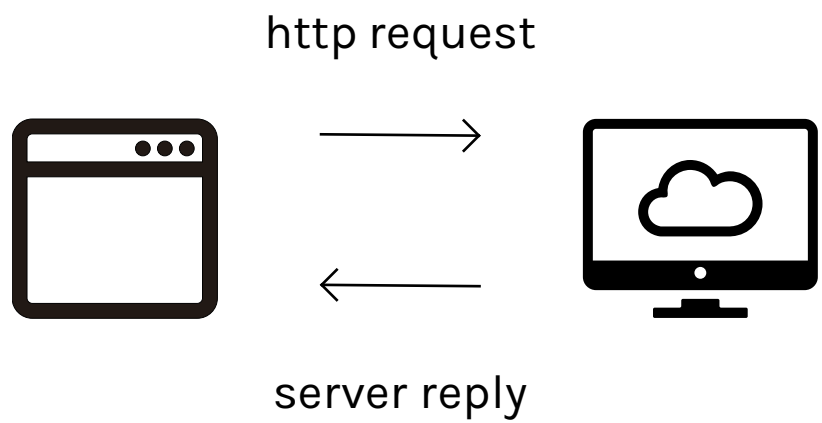
U+0061	U+0062	U+0063	U+0064	U+0065	U+0066	U+0067	U+0068	U+0069
a	b	c	d	e	f	g	h	i
129	130	131	132	133	134	135	136	137
U+006A	U+006B	U+006C	U+006D	U+006E	U+006F	U+0070	U+0071	U+0072
j	k	l	m	n	o	p	q	r
145	146	147	148	149	150	151	152	153
U+007E	U+0073	U+0074	U+0075	U+0076	U+0077	U+0078	U+0079	U+007A
~	s	t	u	v	w	x	y	z
161	162	163	164	165	166	167	168	169

HyperText Transfer Protocol (HTTP)

HTTP is the protocol for exchanging information on the web.

A client (a browser) asks for a **resource** to a server (a machine where the resource is stored). The resource is requested by means of its **Uniform Resource Locator (URL)**.

The server replies with a document (if found), generally an **HTML document**.



Marilena Daquino, marilena.daquino2@unibo.it

Uniform Resource Identifier (URI)

URL identify the **location** of a resource on the web.

https://en.wikipedia.org/wiki/Uniform_Resource_Identifier

URN identify the **name** of a resource, but not its location.

urn:isbn:0-486-27557-4

URI identify **resources** (either documents or concepts) on the web, including both their location and name.

http://dbpedia.org/resource/Robert_Capa

IRI identify resources regardless the **language**.

https://en.wiktionary.org/wiki/῾Ρόδος

Uniform Resource Identifier (URI)

Several URI, minted by different institutions, can identify the same concept, which have to be explicitly **declared as the same** (by creating an equivalence link).

http://dbpedia.org/resource/Robert_Capa
owl:sameAs
http://viaf.org/viaf/54145320

URI identifying concepts are different from URL of pages describing concepts. Usually, URI are **dereferenced**, that is, when a browser requests it, a web document can be returned, while applications can retrieve data about the concept.

Resource Description Model (RDF)

RDF is a **model** (not a language) for:

1) describing resources

    Robert Capa **is a** Person

2) representing relations between data as qualified links

    Robert Capa **has wife** Gerda Taro

3) exchanging structured data on the web in a standard way

RDF triple

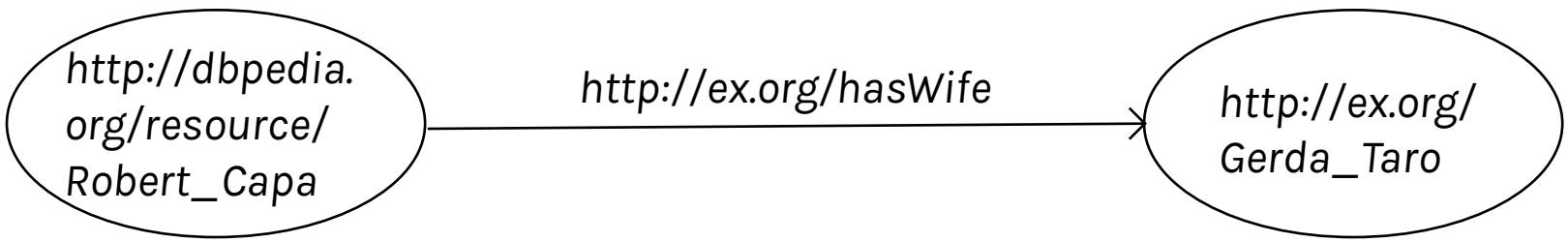
The smallest unit of information is represented by means of a **triple**, including a subject, a predicate, and an object.

Robert Capa	has wife	Gerda Taro
subject	*predicate*	*object*

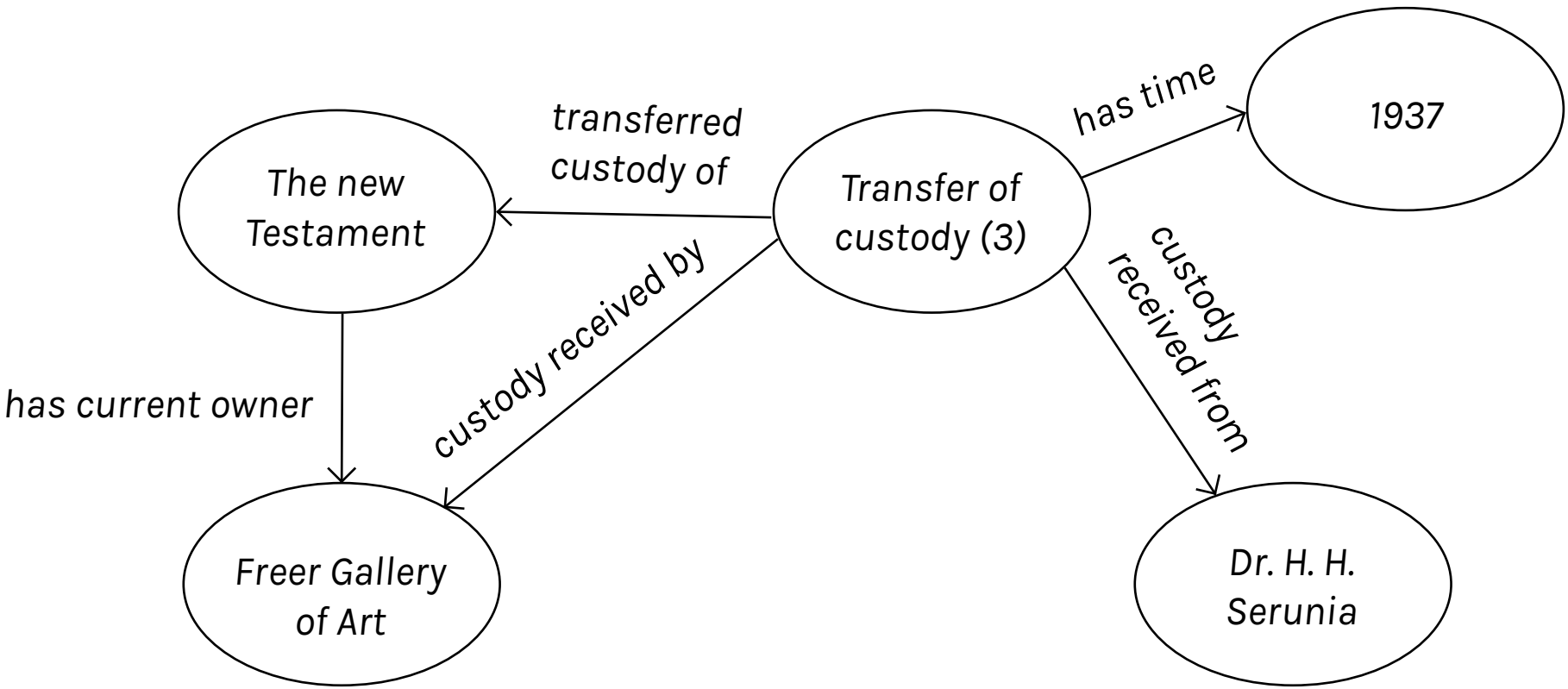
In the Semantic Web URI/IRI are used to identify both resources and the **relations** between them.

Robert Capa	*http://dbpedia.org/resource/Robert_Capa*
has wife	*http://ex.org/hasWife*
Gerda Taro	*http://ex.org/Gerda_Taro*

Resources can be graphically represented as **nodes** and relations as **arks**.



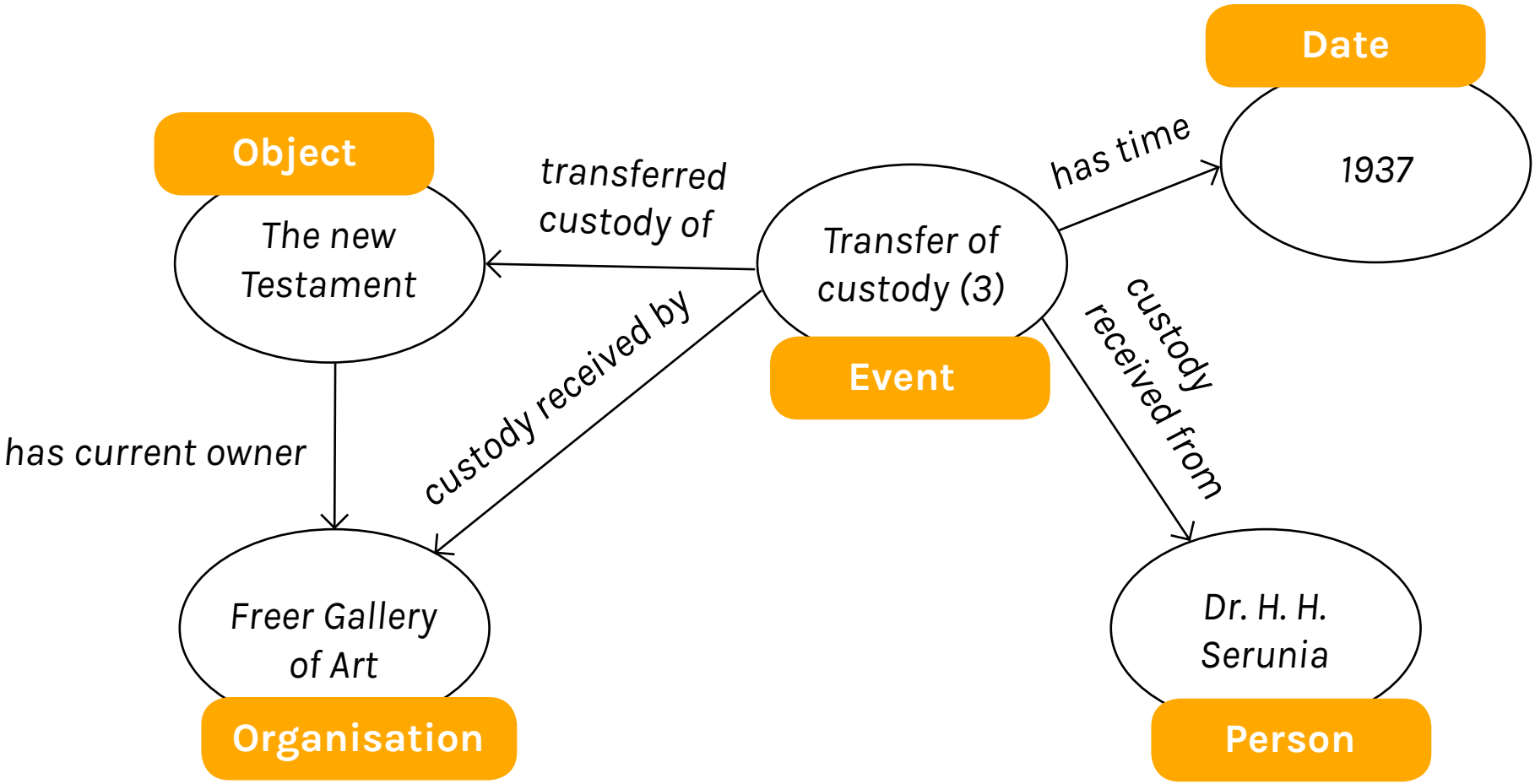
RDF graph



Example elaborated from *David Newbury, Art Tracks: Using Linked Open Data For Object Provenance In Museums, Carnegie Museum of Art. 2017.* https://mw17.mwconf.org/paper/art-tracks-using-linked-open-data-for-object-provenance-in-museums/

Marilena Daquino, marilena.daquino2@unibo.it

RDF graph



Example elaborated from *David Newbury, Art Tracks: Using Linked Open Data For Object Provenance In Museums, Carnegie Museum of Art. 2017.* https://mw17.mwconf.org/paper/art-tracks-using-linked-open-data-for-object-provenance-in-museums/

Marilena Daquino, marilena.daquino2@unibo.it

RDF serializations

RDF can be serialised (i.e. written) by means of several languages, e.g. XML, Turtle, JSON-LD. **Namespaces** can be used to shorten URIs.

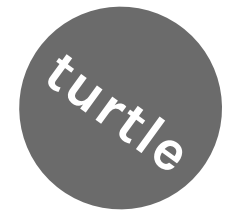
turtle

```
freer_object:10173 a crm:E22_Man-Made_Object ;
  rdfs:label "The New Testament";
  crm:P52_has_current_owner freer_constituent:3326.

_:10173_prov_3 a crm:E8_Acquisition, crm:E10_Transfer_of_Custody;
  crm:P2_has_type mprov_acq:purchase;
  crm:P120i_occurs_after _:10173_prov_2;
  crm:P30_custody_transferred_of freer_object:10173;
  crm:P23_transferred_title_from freer_constituent:4274;
  crm:P29_custody_received_by  freer_constituent:3326;
  crm:P3_has_note "From 1937 Freer Gallery of Art, purchased from Dr. H. H.
       Serunian, Worcester, Massachusetts.";
  crm:P4_has_time-span _:10173_prov_3_timespan.

_:10173_prov_3_timespan a crm:E52_Time-Span;
  crm:P82a_begin_of_the_begin "1937";
  rdfs:label "From 1937".
```

Source: http://www.museumprovenance.org/reference/standard/

Data types

In RDF data can be represented as URI, but also as strings, associated to URIs. Several types of data types exist, such as literals or dates.




```
freer_object:10173 a crm:E22_Man-Made_Object ;
  rdfs:label "The New Testament";
  crm:P52_has_current_owner freer_constituent:3326.

_:10173_prov_3 a crm:E8_Acquisition, crm:E10_Transfer_of_Custody;
  crm:P2_has_type mprov_acq:purchase;
  crm:P120i_occurs_after _:10173_prov_2;
  crm:P30_custody_transferred_of freer_object:10173;
  crm:P23_transferred_title_from freer_constituent:4274;
  crm:P29_custody_received_by  freer_constituent:3326;
  crm:P3_has_note "From 1937 Freer Gallery of Art, purchased from Dr. H. H.
       Serunian, Worcester, Massachusetts.";
  crm:P4_has_time-span _:10173_prov_3_timespan.

_:10173_prov_3_timespan a crm:E52_Time-Span;
  crm:P82a_begin_of_the_begin "1937";
  rdfs:label "From 1937".
```

Source: http://www.museumprovenance.org/reference/standard/

Web Ontology Language (OWL)

Data, relations, and constraints on relations are described by means of vocabularies of terms (taxonomies, thesauri, vocabularies, ontologies).

Ontologies (or vocabularies) are defined as

> *a formal explicit specification of a shared conceptualization of a domain of interest. (Gruber)*

The standard language for expressing ontologies is OWL (likewise RDF it has several serialisations).

individual

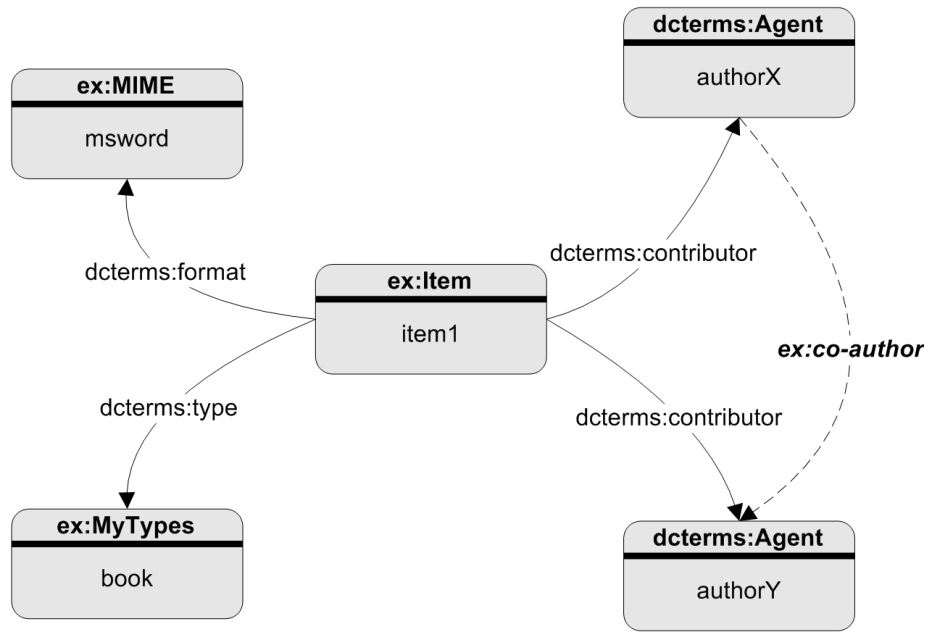
```
freer_object:10173 a crm:E22_Man-Made_Object ;
  rdfs:label "The New Testament";
  crm:P52_has_current_owner freer_constituent:3326.
```

class
data property
object property

Classes and predicates are uniquely identified by URIs as well.

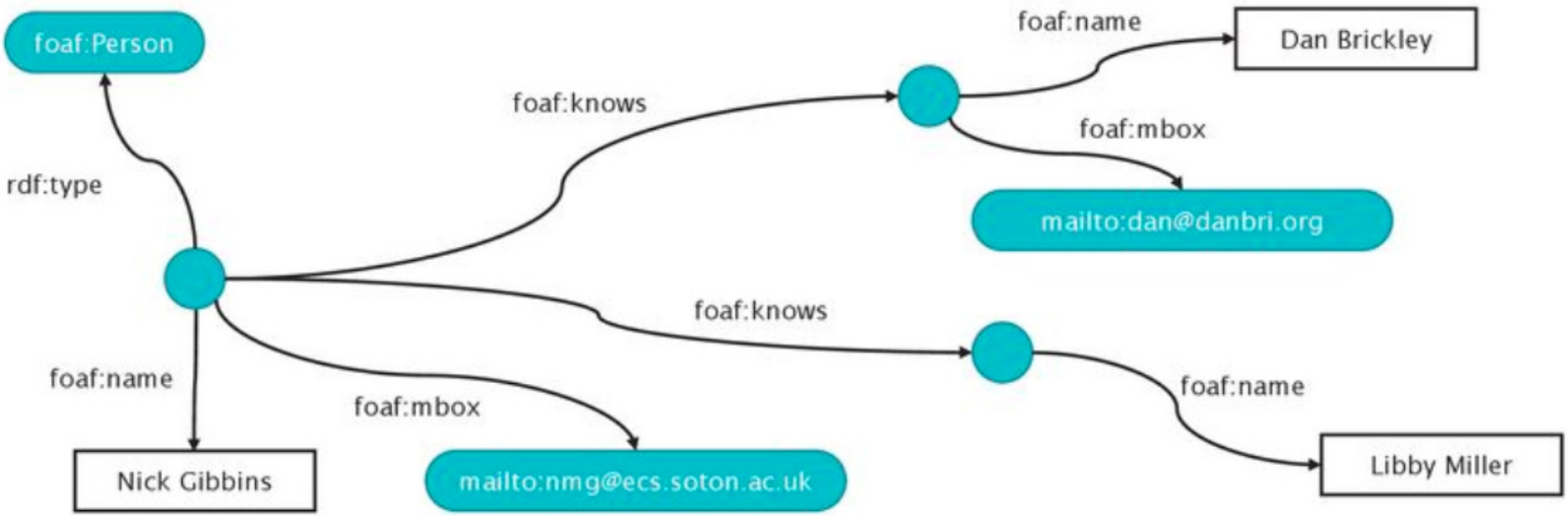
Dublin Core Terms (DCTERMS)

An ontology for describing digital resources and their cataloguing metadata



Friend Of A Friend (FOAF)

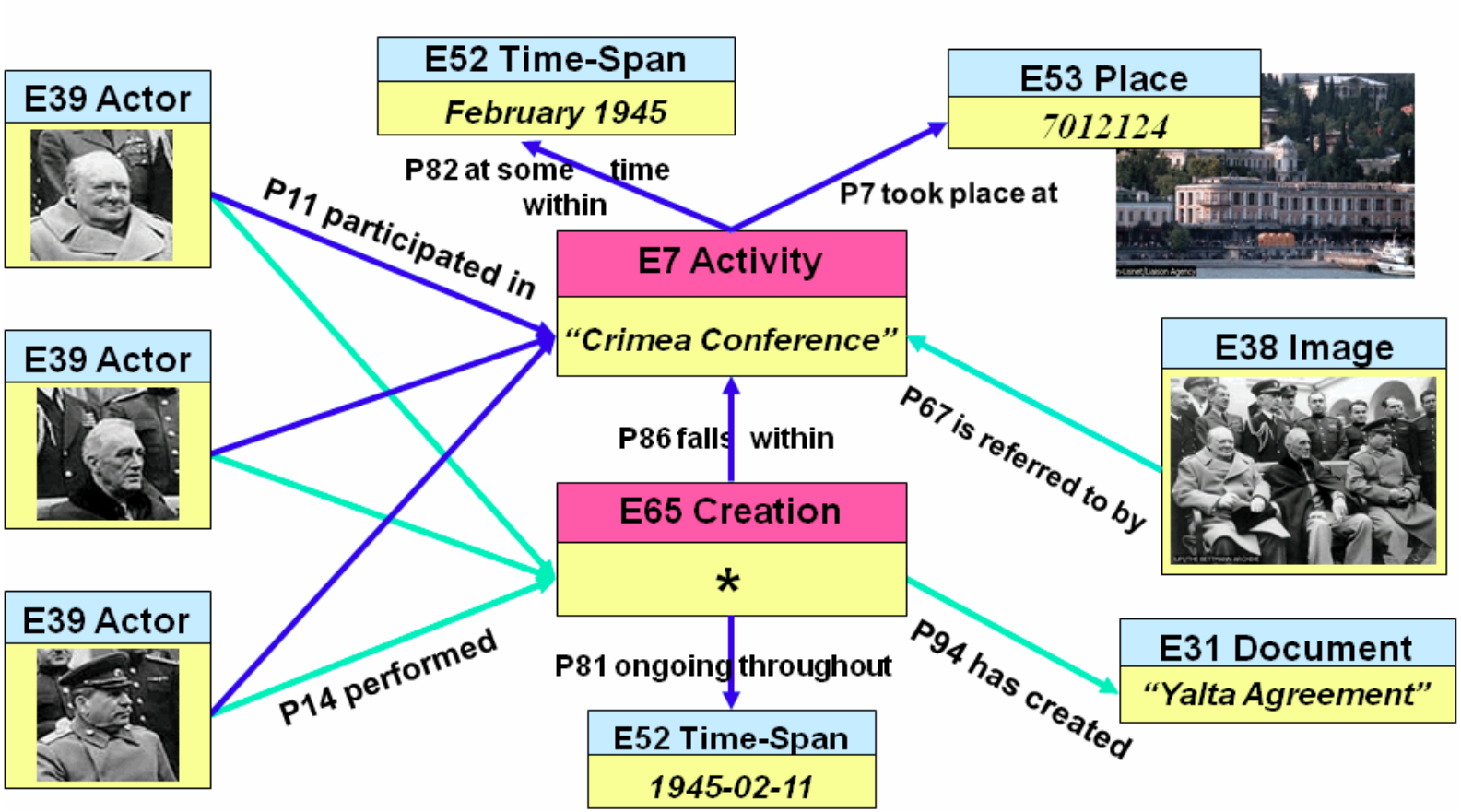
An ontology for describing persons and their relationships



Source: Gibbins, Social Linked Data. https://slideplayer.com/slide/16802641/

CIDOC-CRM

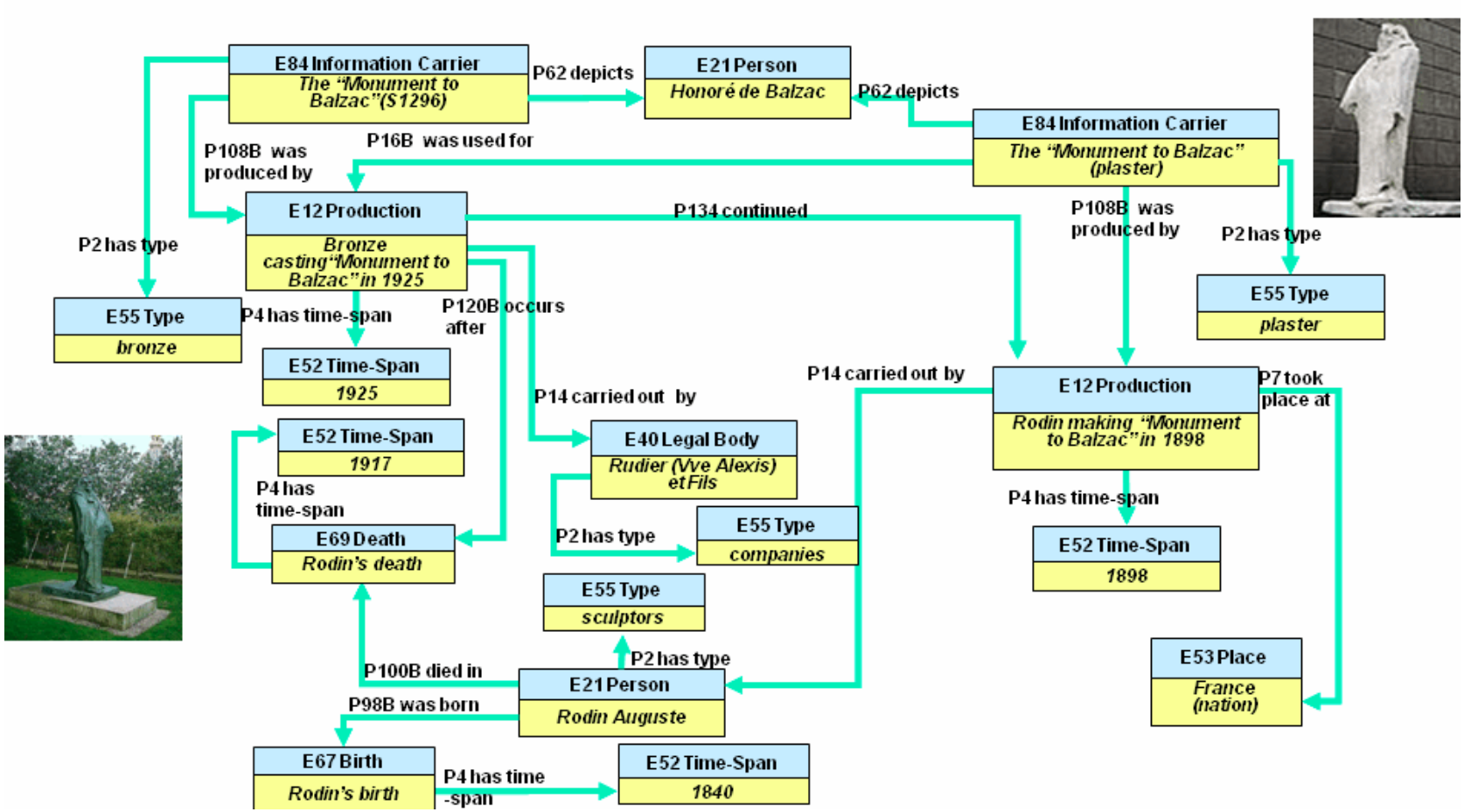
An event-driven ontology for describing cultural heritage artefacts and their life-cycle



Marilena Daquino, marilena.daquino2@unibo.it

CIDOC-CRM

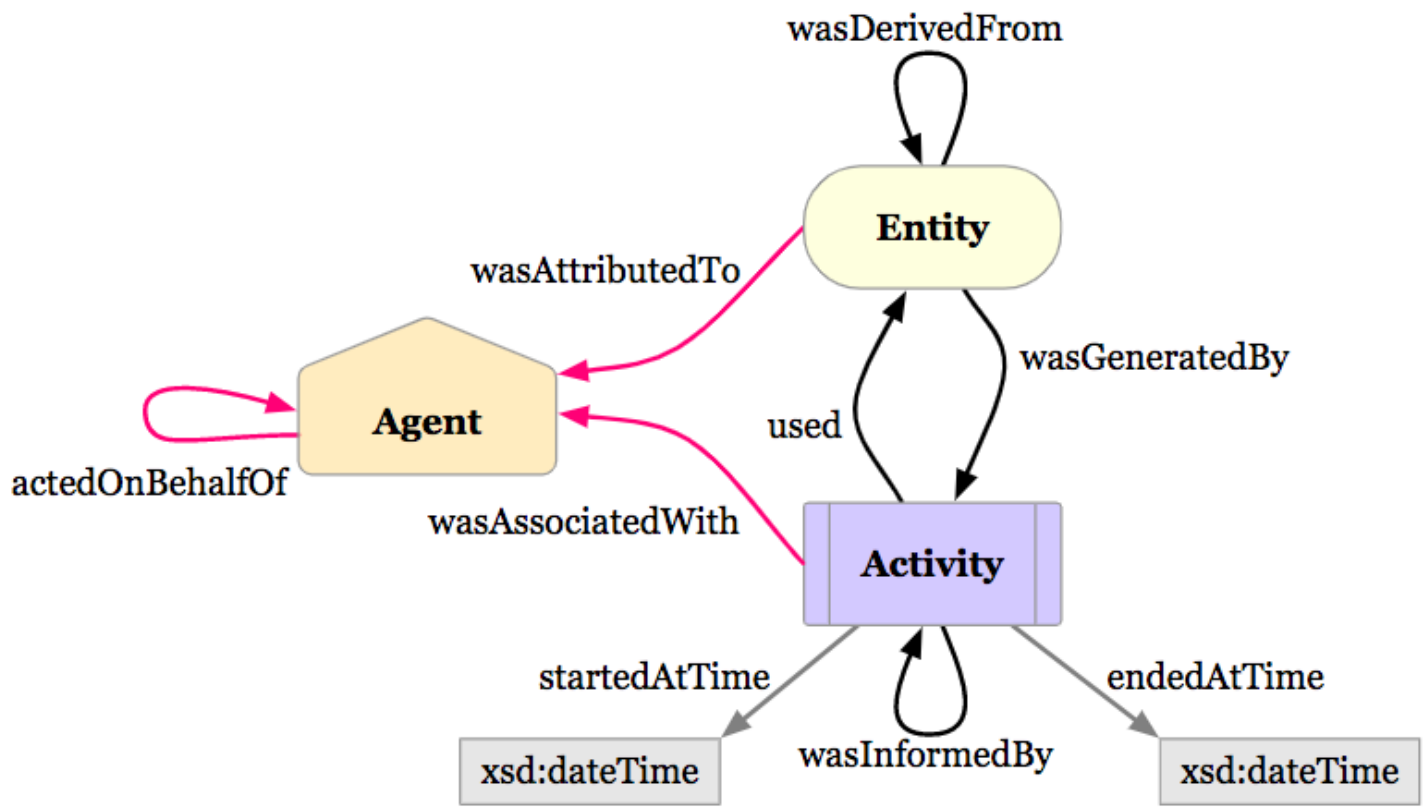
An event-driven ontology for describing cultural heritage artefacts and their life-cycle



Marilena Daquino, marilena.daquino2@unibo.it

PROV

An ontology for describing provenance of information. Includes agents, entities, and activities.



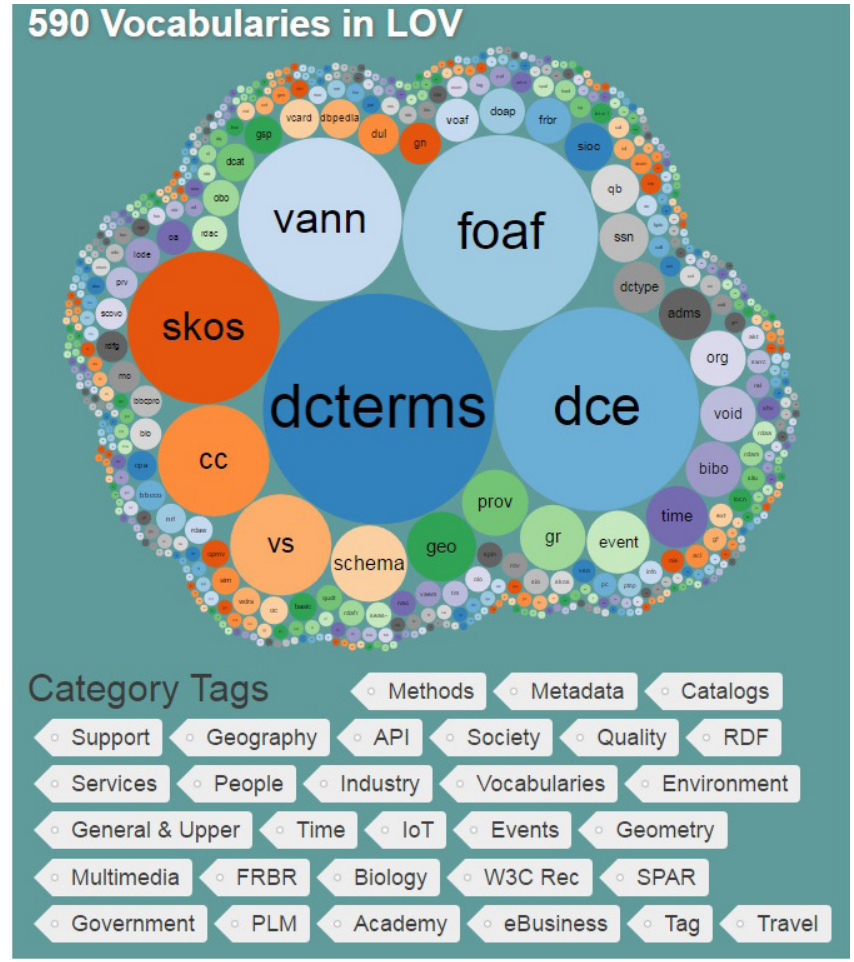
Marilena Daquino, marilena.daquino2@unibo.it

Where to find ontologies?

Ontology reuse is considered a best practise in Semantic Web.

LOV is the registry of vocabularies, including their documentation for facilitating discovery and reuse.

Ontologies can be mixed, merged, or aligned.



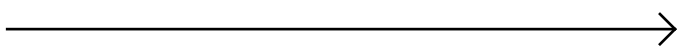
Marilena Daquino, marilena.daquino2@unibo.it

SPARQL Protocol and RDF Query Language

SPARQL is the RDF query language, that is, a semantic query language for graph databases.

e.g. Which artefacts have been transferred in 1937?

```
PREFIX crm: <http://www.cidoc-crm.org/cidoc-crm/>
PREFIX freer_object: < http://asia.si.edu/collections/freer/object/>

SELECT ?artefact

WHERE {
    ?custody    crm:P30_custody_transferred_of ?artefact ;
                crm:P4_has_time-span ?time .
    ?time       crm:P821_begin_of_the_begin "1937".
}
```



```
freer_object:10173
```

Semantic Web Rule Language (SWRL)

The OWL language is not able to express everything.

For instance, it cannot express the relation *nephew of father's brother*, because there is no way in OWL 2 to express the relation between individuals with which an individual has relations.

SWRL allows to add **rules** to an ontology.

```
hasParent(?me,?dad) => hasBrother(?dad,?uncleTom) => hasUncle(?me,?uncleTom)
```

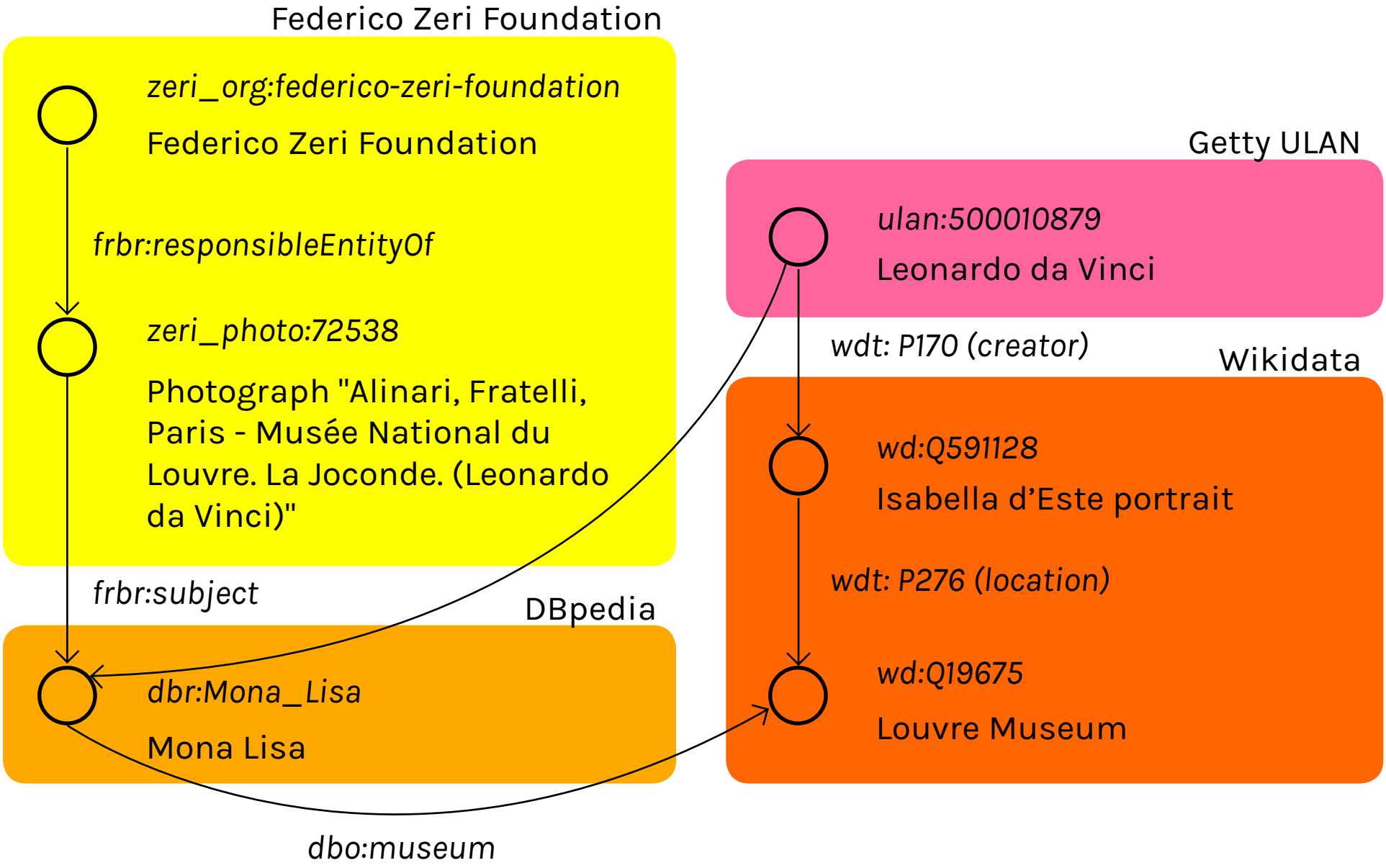
Proof and Trust

At this stage in the development of the Semantic Web, Proof and trust mechanisms are still to be developed.

Most applications build proof evidences according to some fairly constrained rules (e.g. third party opinions, users' rating).

Most of the work done in the **web of trust** is based on **provenance of information**.

An example of Linked Open Data



How to cook Linked Open Data?

A typical workflow for creating and publishing LOD includes the following activities:

KNOWLEDGE ORGANIZATION
- study of the domain
- ontology reuse or development

DATA REEINGEENIRING OR CREATION
- from structured, semi-structured or unstructured data

DATA STORAGE AND PUBLICATION
- dereferencing
- data storage and access points
- dataset annotation

Ontology development

Ontology reuse is considered the best practise in Semantic Web for **semantic interoperability** purposes.

Methodologies for **ontology development** require diverse skills to be involved - domain experts (DE), ontology engineers (OE) - and include a number of steps for validating consistency (with regard to the terminological part - TBox - and the assertion part - ABox) and vocabulary completeness.

Protégé is one of the most common visual interfaces for developing ontologies (https://protege.stanford.edu/)

SAMOD

https://essepuntato.it/samod/

Simplified Agile Methodology for Ontology Development (SAMOD) is a agile methodology for the development of ontologies by means of small steps of an **iterative workflow** that focuses on creating well-developed and documented models starting from exemplar domain descriptions.

1. OEs and DEs work together to write down a motivating scenario
2. Given a motivating scenario, OEs and DEs should produce a set of informal competency questions CQ
3. OEs and DEs write down a glossary of terms GoT
4. The OE develops a modelet according to the MS, the informal CQs and the glossary of terms, starting from a graphical representation written in a proper visual language.
5. The OE runs a model test on modelet.
6. The OE creates an exemplar dataset.
7. The OE writes as many formal queries SQ as the informal CQs and runs a query test.
8. If everything succeeds, the test case is added to a bag of test cases.
9. Terms of the modelet are refactored according to existing ontologies. Go back to step 1.

Marilena Daquino, marilena.daquino2@unibo.it

SAMOD: an example

MS.1

Identify artworks currently part of a collection, and which have been transferred elsewhere.

CQ.1

Description

Retrieve all the artworks that currently belong to the collection.

Example

The artworks currently owned by the Grassi collection in Rome and the artworks that have been transferred/acquired by other actors.

https://w3id.org/zericatalog/grassi-roma

Expected output

The list of artworks, artists, dates of creation

Reengineer or create data

FROM RELATIONAL DATABASES
- mapping languages (R2RML)
- graph visualisation of databases (D2RQ)

FROM WEB CONTENTS
- RDFa and microformats for embedding semantics in HTML pages

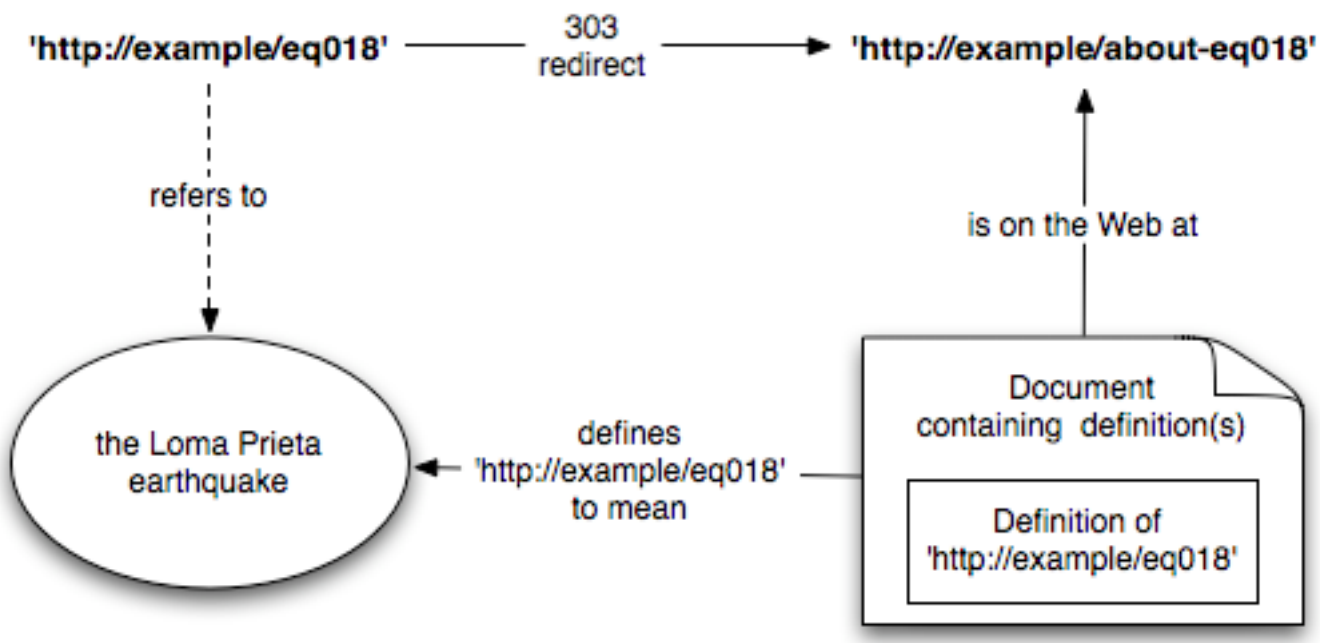
FROM XML, CSV OR OTHER DATA FORMATS
- remodelling and reengineering
    - find resources, datatypes of cells
    - assign identifiers
    - map columns to predicates
    - iterate over rows

FROM NATURAL LANGUAGE TEXTS
- Natural Language Processing (NLP) techniques

Store and publish data

ENABLE DEREFERENCING OF URIS

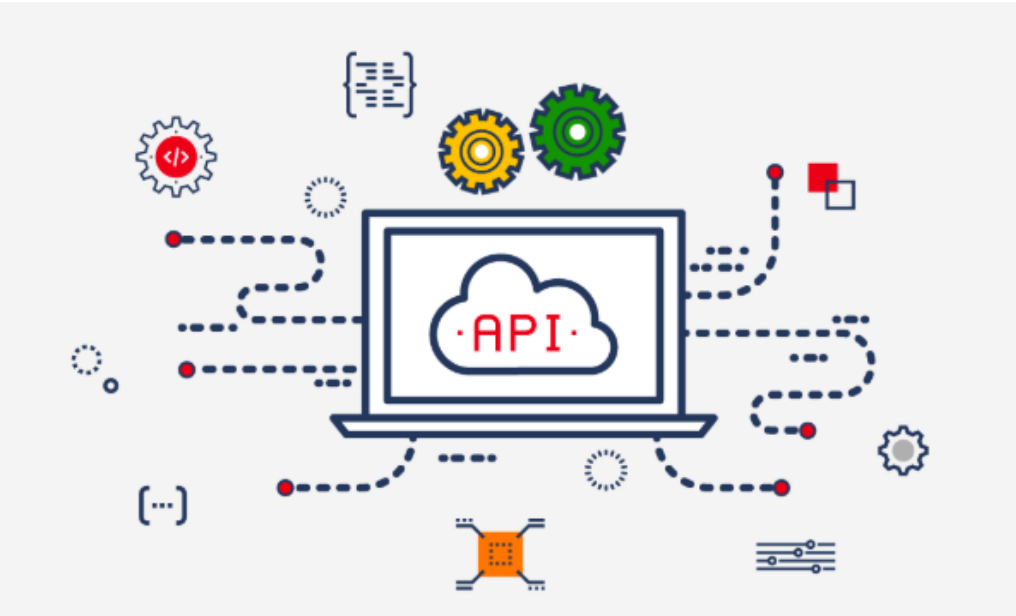
- setup content negotiation (303 redirect)



Store and publish data

DATA STORAGE AND ACCESS POINTS

- upload data on a triplestore, a graph database
- provide a SPARQL endpoint
- release data dumps
- provide access through application program interface (APIs)



Marilena Daquino, marilena.daquino2@unibo.it

Store and publish data

DATASET ANNOTATION

- VoID vocabulary (that reuses DCterms) for adding cataloguing information to datasets

```
:DBpedia a void:Dataset;
    dcterms:title "DBPedia";
    dcterms:description "RDF data extracted from Wikipedia";
    dcterms:contributor :FU_Berlin;
    dcterms:contributor :University_Leipzig;
    dcterms:contributor :DBpedia_community.
```

So what? Benefits of SW for humanists

IDENTIFY RESOURCES
- align records in different catalogues/data sources
- reconcile multilingual terms
- disambiguation of homonymous
- data citation

CONSISTENCY ACROSS SYSTEMS

INTEGRATE DATA
- enrich data
- facilitate data curation
- data aggregation

IMPROVED DATA QUALITY

CREATE SMART APPLICATIONS
- follow your nose
- mashup
- recommendation
- inference

ANSWER COMPLEX QUESTIONS

So what? How to contribute

1 KNOWLEDGE ORGANISATION

2 IDENTIFY CONCEPTS, RELATIONS AND INDIVIDUALS WITH URI

3 TRANSFORM DATA TO RDF

4 PUBLISH RDF DATA

5 LINK RDF DATA TO OTHER DATA SOURCES

6 CONSUME DATA IN WEB APPLICATIONS

So what? How to contribute

1 KNOWLEDGE ORGANISATION

2 IDENTIFY CONCEPTS, RELATIONS AND INDIVIDUALS WITH URI

3 TRANSFORM DATA TO RDF

4 PUBLISH RDF DATA

5 LINK RDF DATA TO OTHER DATA SOURCES

6 CONSUME DATA IN WEB APPLICATIONS

PART II

Linked Open Data for Provenance Research

By LOD for Provenance Research we mean all of those data sources that include fundamental information for tracing provenance paths of artworks, namely:

- the names of the owners
- places of ownership
- dates of ownership
- methods of transfer between owners

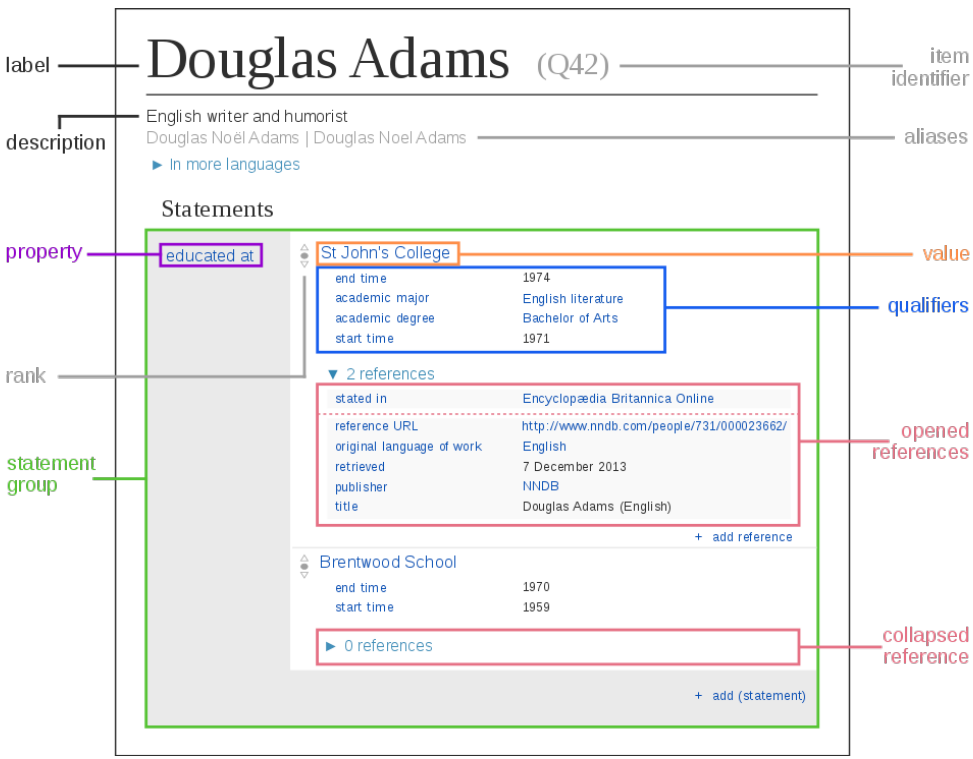
See:
Schiuma, Giovanni, and Daniela Carlucci. Big Data in the Arts and Humanities: Theory and Practice. Auerbach Publications, 2018.

Digital Provenance Symposium 2016
https://cmoa.org/art/art-tracks-digital-provenance-project/digital-provenance-symposium-2016/

Marilena Daquino, marilena.daquino2@unibo.it

Wikidata

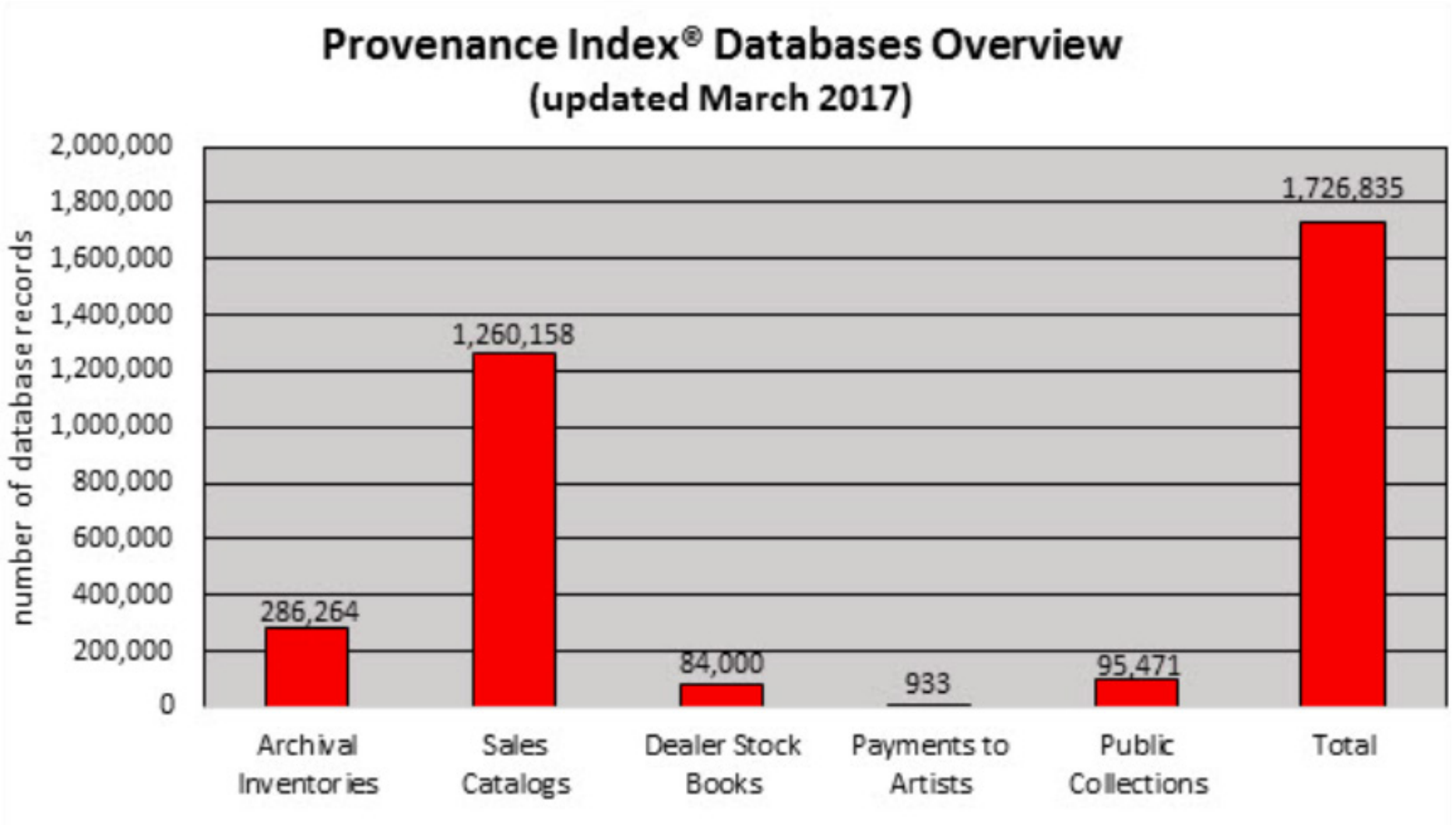
Wikidata is the knowledge graph for the structured data of its Wikimedia sister projects including Wikipedia, Wikivoyage, Wiktionary, Wikisource, and others.



Art dealers in wikidata: http://tinyurl.com/y95d7bfj
Source: https://www.openartdata.org/p/blog-page_11.html

Marilena Daquino, marilena.daquino2@unibo.it

Getty Provenance Index Database

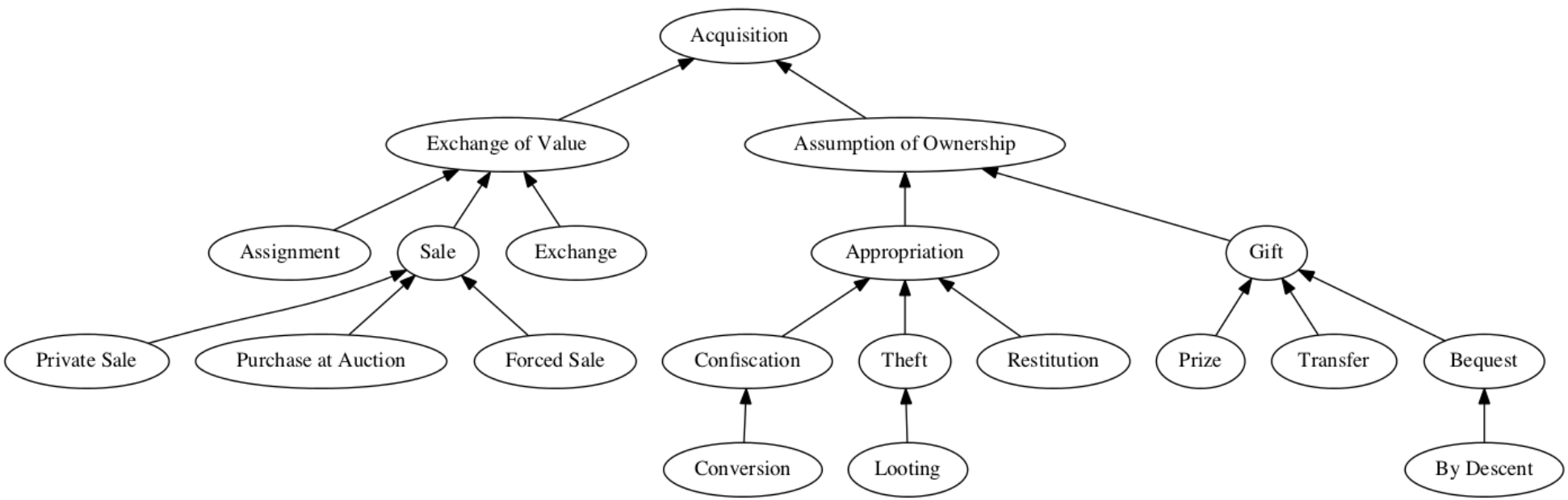


Marilena Daquino, marilena.daquino2@unibo.it

Carnergie (CMOA)

The objective of the Art Tracks project is to generate a digital model for storing and capturing the data within provenance in a machine-readable format.

http://www.museumprovenance.org/reference/standard/

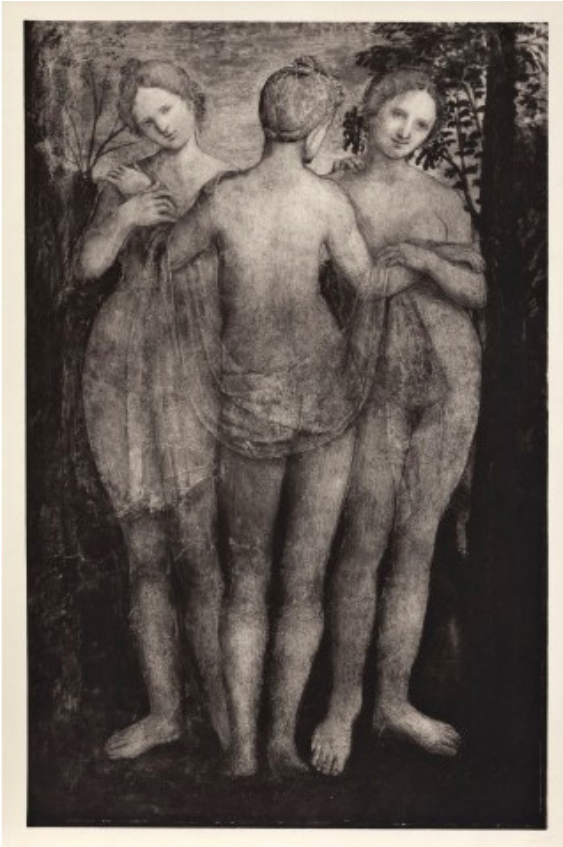


Marilena Daquino, marilena.daquino2@unibo.it

PHAROS

http://pharosartresearch.org

A consortium of 14 art historical photo archives that aim at creating a common platform where to share LOD according to CIDOC-CRM about artwork and photography data



Federico Zeri Photo Archive

An example of provenance in Linked Open Data:

Perruzzi Baldassarre, The three Graces
https://w3id.org/zericatalog/artwork/39794

- Bibliotheca Hertziana Max-Planck-Institut für Kunstgeschichte
- Courtauld Institute of Art
- Deutsches Dokumentationszentrum für Kunstgeschichte – Bildarchiv Foto Marburg
- Federico Zeri Foundation
- Frick Art Reference Library
- Getty Research Institute
- Institut National d'Histoire de l'Art
- Kunsthistorisches Institut in Florenz
- National Gallery of Art
- Paul Mellon Centre for Studies in British Art
- RKD – Netherlands Institute for Art History
- Villa I Tatti - The Harvard University Center for Italian Renaissance Studies
- Warburg Institute
- Yale Center for British Art

Marilena Daquino, marilena.daquino2@unibo.it

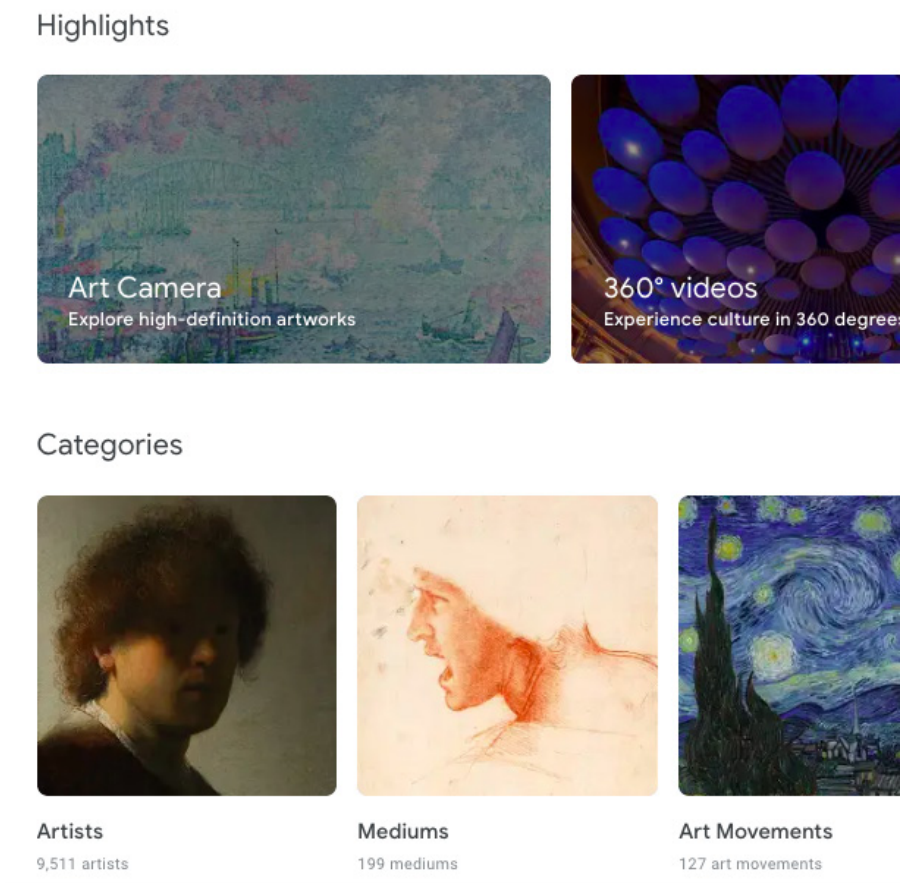
Google Art

An aggregator of images for virtual exhibitions

Data provided are not publicly available.

Metadata requested to
partner cultural institutions

https://support.google.com/
culturalinstitute/partners/answer/7574684



Marilena Daquino, marilena.daquino2@unibo.it

Ontologies for Provenance Research

PROVENANCE ONTOLOGY (PROV)
An ontology for representing agents, entities, and activities.

Linked.art
An application profile of CIDOC-CRM

CMOA STANDARD
An application profile of CIDOC-CRM explicitly devoted to provenance information

Marilena Daquino, marilena.daquino2@unibo.it

Blockchain for art and provenance

FRESCO
https://fresco.work

ARTORY
https://www.artory.com

VERISART
https://verisart.com

BLOCKCHAIN ART COLLECTIVE
https://blockchainartcollective.com

Marilena Daquino, marilena.daquino2@unibo.it

What is the problem?

WHAT?

- benchmarking of projects, data sources, ontologies
- define motivating scenarios and research questions provided by experts
- discover resources available and coverage
    (what are the gaps in terms of 1. content, 2. data formats, services for querying, and interoperability, 3. licenses for reuse)

WHY?

- a comprehensive literature review of digital sources
for provenance research is missing
- Lack of LOD sources
- Lack of semantic interoperability
- Restrictive licenses for data reuse (Blockchain art, Google)

FOR WHOM?

- Linked.art, CIDOC-CRM-focus groups etc. for modelling purposes;
- researchers for having an exhaustive catalogue of online resources

Marilena Daquino, marilena.daquino2@unibo.it

Hands-on exercises

FIRST TASK: STUDY OF THE DOMAIN

General discussion

- address Motivating Scenarios in art provenance research and identify Competency Questions
- fill a table describing them

https://tinyurl.com/art-provenance

Table MS*

Hands-on exercises

SECOND TASK: BENCHMARKING

Working groups

- look for relevant datasets/catalogues online: include both LOD and other formats of open data (catalogues, databases, repertoires)
    N.B. **NOT** scholarly publications, e.g. articles, books, posts
- fill a table describing them

Table Datasets

Hands-on exercises

THIRD TASK: IDENTIFY RESOURCES

Working groups

Extend MS* tables:
- list sources that could be used and potentially linked to others in order to answer a CQ (if applicable)

General discussion
- refine or add new motivating scenarios/CQs

Hands-on exercises

FOURTH TASK: MODELLING

General discussion

Working groups
- draw a model for describing your CQs

Use mindmup

https://www.mindmup.com

Hands-on exercises

FIFTH TASK: MAPPING

Working groups

- identify what data are included in datasets (e.g. people, places, roles)
- map data sources to the model (add child nodes to the mind map)

Future plans

1 write a white paper on the topic
2 create a online catalogue of existing resources

Marilena Daquino, marilena.daquino2@unibo.it

Thanks

Marilena Daquino, marilena.daquino2@unibo.it