Museum Linked Open Data: Ontologies, Datasets, Projects

Vladimir Alexiev [0000-0001-7508-7428]

Ontotext Corp, Sofia, Bulgaria

vladimir.alexiev@ontotext.com

Abstract. The Galleries, Libraries, Archives and Museums (GLAM) sector deals with complex and varied data. Integrating that data, especially across institutions, has always been a challenge. Semantic data integration is the best approach to deal with such challenges. Linked Open Data (LOD) enable large-scale Digital Humanities (DH) research, collaboration and aggregation, allowing DH researchers to make connections between (and make sense of) the multitude of digitized Cultural Heritage (CH) available on the web. An upsurge of interest in semtech and LOD has swept the CH and DH communities. An active Linked Open Data for Libraries, Archives and Museums (LODLAM) community exists, CH data is published as LOD, and international collaborations have emerged. The value of LOD is especially high in the GLAM sector, since culture by its very nature is cross-border and interlinked. We present interesting LODLAM projects, datasets, and ontologies, as well as Ontotext's experience in this domain.

Keywords: semantic technologies, museum data, LODLAM, CIDOC CRM, Wikidata.

1 Introduction

The Galleries, Libraries, Archives and Museums (GLAM) sector deals with complex and varied data. Integrating that data, especially across institutions, has always been a challenge. There is growing consensus in GLAM that Semantic Data Integration and Linked Open Data (LOD) are the best approach to deal with such challenges. LOD enables large-scale Digital Humanities (DH) research, collaboration and aggregation, allowing DH researchers to make connections between (and make sense of) the multitude of digitized Cultural Heritage (CH) available on the web.

An upsurge of interest in semtech and LOD has swept the CH and DH communities. An active Linked Open Data for Libraries, Archives and Museums (LODLAM) community exists, CH data is published as LOD, and international collaborations have emerged. Significant investments were made by the EU (Europeana), US (DPLA), various other countries (e.g. Finland's CultureSampo), international foundations (e.g. Mellon) and important CH institutions (e.g. the Getty Trust) .

Digital Presentation and Preservation of Cultural and Scientific Heritage. Conference Proceedings. Vol. 8, Sofia, Bulgaria: Institute of Mathematics and Informatics – BAS, 2018. ISSN: 1314-4006, eISSN: 2535-0366

Ontotext is a Bulgarian software company that has worked on semantic technologies since 2000, and on CH LOD since 2010. Ontotext has 65 staff (7 PhD, 30 MS, 20 BS, 6 university lecturers) It is part of Sirma Group Holding, the largest Bulgarian public software group, and is a core part of Sirma Strategy 2022 that focuses on cognitive computing. Ontotext works on semantic modelling, data integration and Knowledge Graph creation, semantic repositories (Ontotext GraphDB), semantic text analysis (entity, concept, relation extraction, document classification), machine learning (entity disambiguation, deep learning in graphs), recommendations, sentiment analysis, etc. In addition to numerous commercial projects, Ontotext is one of the most innovative Bulgarian software companies, with over 40 EU-funded research projects (6 currently active) and various innovation awards:

- Innovative Enterprise of the Year 2017
- EU Innovation Radar Prize 2016 nomination
- BAIT Business Innovation Award 2014
- Innovative Enterprise of the Year 2014
- Washington Post "Destination Innovation" Competition 2014 Award
- Pythagoras Award 2010 for most successful company in EU FP6 projects

Ontotext has participated in a number of CH/DH LOD projects:

- ResearchSpace: British Museum, Yale Center for British Art. Largest museum collection converted to CIDOC CRM, semantic search…
- (with Sirma Enterprise) ConservationSpace, Sirma MuseumSpace
- Medieval Cultures and Technological Resources (VCMS) COST action
- Europeana Creative, Europeana Food and Drink, OAI PMH, SPARQL, Europeana members council, 5 work groups, Data Quality Committee
- Initiator of the Bulgariana national aggregator
- Getty Research Institute: vocabularies LOD
- Carnegie Hall LOD
- American Art Collaborative: consulting 14 US museums integrating data using CIDOC CRM
- European Holocaust Research Infrastructure: semantic archive integration
- Canadian Heritage Information Network: consulting the Canadian national aggregator's transition to LOD
- Wikidata: frequent contributions, mostly to authority control
- DBpedia: contributions, association member, data quality/ontology committee
- CLADA BG: key participant in both CLARIN (NLP) and DARIAH (CH/DH)

We present some interesting museum projects, datasets and ontologies, as well as Ontotext's experience in this domain. Other LODLAM domains (archives and libraries) have also made significant progress in LOD adoption, but are out of scope of this paper.

2 Content, Interchange, Metadata Standards

GLAM data is complex and varied: data comes from a variety of systems, the data is not regular (exception is the rule, e.g. there may be several Father relations for a person representing different opinions), and many metadata format variations are in use. To enable efficient interoperation, standardization in several areas is needed: content (what to record about objects), interchange (how to transfer data), metadata schemas (how to encode them in technical formats such as XML). We describe several relevant standards in this section, whereas the next section presents related ontologies.

2.1 CCO

Professional organizations have found it useful to define GLAM **content standards**. These recommends what data to capture about objects, which fields to capture as authority values, the level of detail. They don't formalize how to express the data in machine-readable form.

An important content standard for art, architecture and museums is Cataloguing Cultural Objects (CCO) (Murtha Baca, Patricia Harpring, Elisa Lanzi, Linda McRae, & Ann Baird Whiteside, 2006) by the Getty and Visual Resources Association. It covers the following topics:

- Representing the information in dually: as structured (indexing) elements useful for searching, and as human-readable (display elements) useful to express nuances.
- Object titles and work types
- Creator information, specific contribution: role, place, time, extent (e.g. execution, with additions, design, figures, predella, embroidery, cast, printed, etc)

[for a painting]

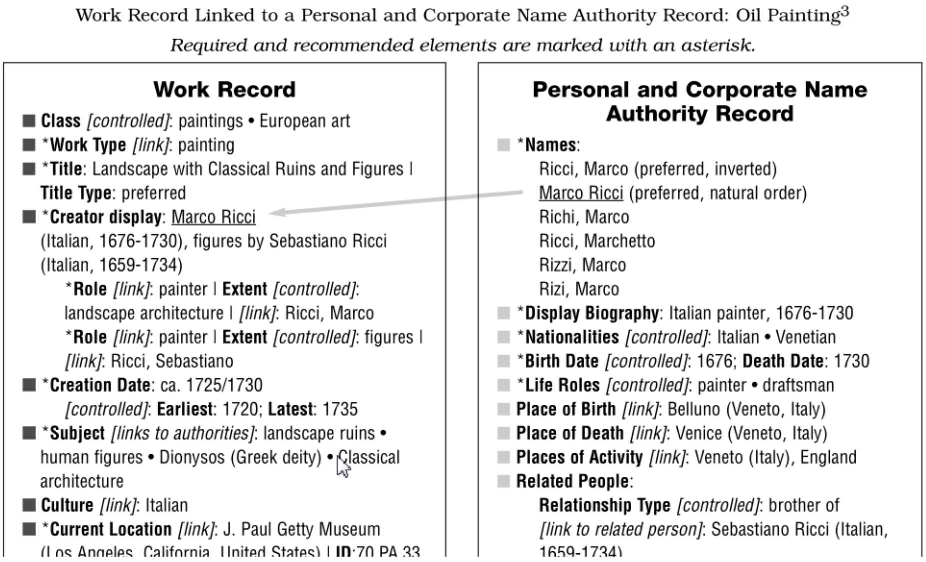
Creator display: figures by Peter Paul Rubens (Flemish, 1577-1640), landscape and still-life objects by Jan Brueghel the Elder (Flemish, 1568-1625)
Controlled fields:
Role: painter
Extent: figures
[link to Personal and Corporate Name Authority]:
Rubens, Peter Paul

Role: painter
Extents: landscape • still life
[link to Personal and Corporate Name Authority]:
Brueghel, Jan, the Elder

Figure 1 CCO Example: Creator Extent

- Physical characteristics: measurements and dimensions (including type, unit, value, object extent, qualifiers); materials, techniques and implements; state and edition (for engravings)
- Stylistic, Cultural, and Chronological Information: style, culture, dates
- Location and geography, different relations between object and place
- Subject information: generic and specific subjects (e.g. iconography, people, places, events)
- Object classification
- Descriptions, notes, sources and contributors

- View (digital assets that capture the object)
- Authorities: persons and corporations, geographic places, concepts, subjects


Work Record Linked to a Personal and Corporate Name Authority Record: Oil Painting[3]

Required and recommended elements are marked with an asterisk.

Work Record

- **Class** *[controlled]*: paintings • European art
- ***Work Type** *[link]*: painting
- ***Title**: Landscape with Classical Ruins and Figures | **Title Type**: preferred
- ***Creator display**: Marco Ricci (Italian, 1676-1730), figures by Sebastiano Ricci (Italian, 1659-1734)
  - ***Role** *[link]*: painter | **Extent** *[controlled]*: landscape architecture | *[link]*: Ricci, Marco
  - ***Role** *[link]*: painter | **Extent** *[controlled]*: figures | *[link]*: Ricci, Sebastiano
- ***Creation Date**: ca. 1725/1730
  - *[controlled]*: **Earliest**: 1720; **Latest**: 1735
- ***Subject** *[links to authorities]*: landscape ruins • human figures • Dionysos (Greek deity) • Classical architecture
- **Culture** *[link]*: Italian
- ***Current Location** *[link]*: J. Paul Getty Museum (Los Angeles, California, United States) | **ID**:70 PA 33

Personal and Corporate Name Authority Record

- ***Names**:
  - Ricci, Marco (preferred, inverted)
  - Marco Ricci (preferred, natural order)
  - Richi, Marco
  - Ricci, Marchetto
  - Rizzi, Marco
  - Rizi, Marco
- ***Display Biography**: Italian painter, 1676-1730
- ***Nationalities** *[controlled]*: Italian • Venetian
- ***Birth Date** *[controlled]*: 1676; **Death Date**: 1730
- ***Life Roles** *[controlled]*: painter • draftsman
- **Place of Birth** *[link]*: Belluno (Veneto, Italy)
- **Place of Death** *[link]*: Venice (Veneto, Italy)
- **Places of Activity** *[link]*: Veneto (Italy), England
- **Related People**:
  - **Relationship Type** *[controlled]*: brother of *[link to related person]*: Sebastiano Ricci (Italian, 1659-1734)

Figure 2 CCO Example: Artwork and Creator Record

The most important part of CCO are the examples that are extremely useful for data modelers to decide how to map data to various models. CCO was the basis of developing the XML metadata standards described next.

2.2 OAI PMH, ResourceSync

Metadata could be exchanged by file, email or a sharing service such as Google Drive, OneDrive, DropBox (commercial) or b2drop (hosted by EUDAT and used by European research projects such as EHRI). There are also metadata transfer protocols that enable automatic interchange, harvesting and aggregation. They allow a client to discover from a server new, updated or deleted records and to transfer the records.

- Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) (Carl Lagoze & Herbert Van de Sompel, 2015) is a venerable protocol that is widely used by digital libraries and other repositories. It is also used by Europeana in client mode for record ingestion, and Ontotext implemented an OAI server so one can obtain all Europeana records (Vladimir Alexiev & Dilyana Angelova, 2015)
- ResourceSync (NISO & Open Archives Initiative, 2017) is a newer standard based on SiteMaps, which allows simpler implementation because no server is necessary. Instead, one can use scripts to compute manifests of new and modified metadata files, which are returned to the client in predefined formats. This standard is used e.g. by some EHRI partners (e.g. USHMM) to transfer archival records to the EHRI portal.

2.3 RNC

Below we present several GLAM-related XML schemas. While they are traditionally represented in W3C XSD format, a much more readable way exists: **Relax NG Compact** (RNC). A Github repository (Vladimir Alexiev, 2015-2018) is available, including tools for working with RNC and all the schemas described below.

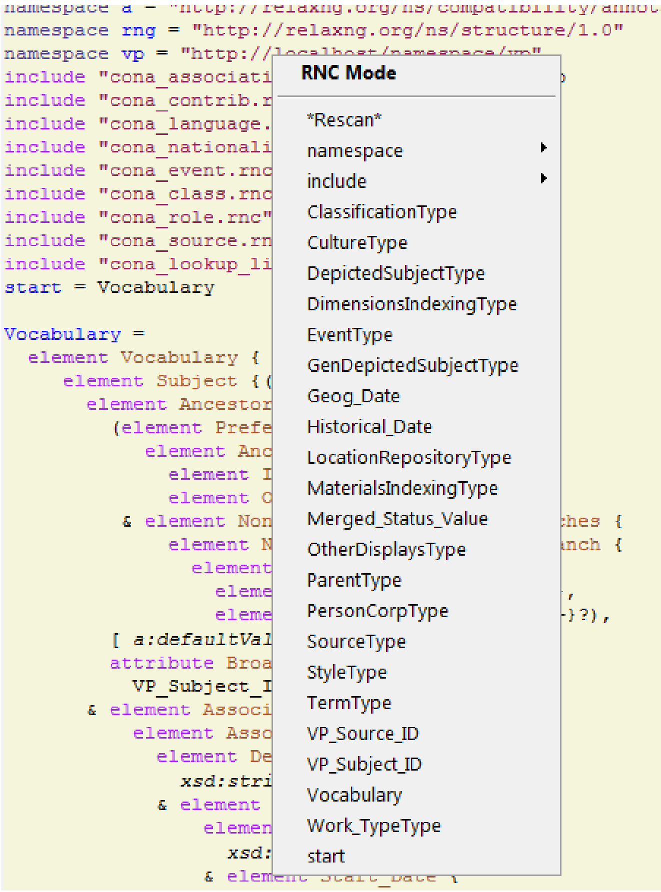


Figure 3 RNC "matryoshka" schema and its "table of contents" in Emacs editor using imenu

2.4 CDWA

While content standards describe what to catalog, metadata (Data Format) standards define how to encode data in a machine-readable way to facilitate information exchange. Here we describe popular XML-based standards for museum metadata.

Categories for the Description of Works of Art (CDWA) (Murtha Baca & Patricia Harpring, 2016) elaborates CCO and covers the following categories of object data (where ♦ indicates a "core" i.e. mandatory category):

- Object/Work ♦
- Classification ♦
- Titles or Names ♦
- Creation ♦
- Styles/Periods/Groups/Movements
- Measurements ♦
- Materials and Techniques ♦
- Inscriptions/Marks
- State
- Edition
- Facture
- Orientation/Arrangement
- Physical Description
- Condition/Examination History
- Conservation/Treatment History
- Subject Matter ♦
- Context
- Descriptive Note
- Critical Responses
- Related Works
- Current Location ♦
- Copyright/Restrictions
- Ownership/Collecting History
- Exhibition/Loan History
- Cataloging History
- Related Visual Documentation
- Related Textual References ♦

It also covers authority control:

- Person/Corporate Body Authority ♦
- Generic Concept Authority ♦
- Place/Location Authority ♦
- Subject Authority ♦

Greek Vase

Object/Work♦	***Catalog Level:♦*** item ***Type:♦*** Panathenaic amphora	Controlled list Authority
	Components/Parts-Quantity: 1 ***Type:*** amphora ***Components/Parts-Quantity:*** 1 ***Type:*** lid	Controlled format Authority
Classification♦	***Terms:♦*** ceramics Greek and Roman art	Controlled list
Titles or Names♦	***Text:♦*** Prize Vessel from the Athenian Games ***Preference:*** preferred ***Type:*** repository ***Text:♦*** Panathenaic Prize Amphora and Lid ***Preference:*** alternate ***Type:*** former	Free text Controlled list
Creation♦	***Creator Description:♦*** attributed to the Painter of the Wedding Procession as painter (Greek vase painter, active ca. 360s BCE); signed by Nikodemos as potter (Attic potter, active ca. 362 BCE)	Free text
	Qualifier: attributed to ***Identity:♦*** Painter of the Wedding Procession ***Role:♦*** painter ***Identity:♦*** Nikodemos ***Role:♦*** potter	Controlled list Authority Authority

Figure 4 CDWA Cataloging Example

Most importantly, CDWA defines a XML schema called CDWA Lite (Getty Research Institute, 2013) that captures the data to be exchanged.

2.5 CONA

Cultural Object Names Authority (Getty Research Institute, 2017) is an aggregation of art object data that realizes the ideas of CCO and CDWA. Although there are still relatively few records (under 50 thousand), it sets an important example for comprehensive object description.

ID: 700006395 **Record Type:** Movable Work
Page link: http://vocab.getty.edu/page/cona/700006395
Images: 1

A Battle from the Trojan War (detached leaf; Unknown French (French Scribe); First Master of the Bible hi...; about 1390 - 1400; J. Paul Getty Museum at the Getty Center (Los Angeles, Los A...; 83.MP.146.3.recto)
Note: The battle, which historically took place in ancient Greece, is here depicted in the guise of a medieval tournament. The raging battle takes place before the fortified walls of a medieval town, whose central tower thrusts beyond the confines of the miniature's golden frame. The artist portrayed the Trojans and Greeks as medieval knights on horseback equipped with armor of the 1300s, and the composition recalls contemporary illustrations in chivalric manuscripts
Titles:
A Battle from the Trojan War (preferred,C,U,English-P,U,U)
Leaf from Ancient History up to the Reign of Caesar (C,U,English,U,U)
Histoire ancienne jusqu'à César (C,U,VL,French,U,U)
Catalog Level: item
Work Types:
detached leaf [300379890] (preferred)
..... (Objects Facet, Components (hierarchy name), components (objects parts), <components by specific context>, information form components, gathered matter components, leaves (gathered matter))

Classifications:
manuscripts (preferred)

Creation Date: about 1390 - 1400

Creator Display:
Unknown French (French Scribe); First Master of the Bible historiale of Jean de Berry (French, active about 1390 - about 1400) (French Illuminator) [preferred,JPGM]
scribe unknown French (French cultural designation) [500125235]
First Master of Bible Historiale of Jean de Berry (French illuminator, active ca. 1390-ca. 1400 in France) [500021767]
Locations:
Current: J. Paul Getty Museum at the Getty Center (Los Angeles, Los Angeles county, California, United States) [500329598] J. Paul Getty Museum, J. Paul Getty Trust, Corporate Bodies (Corp. Body)
Repository Numbers: 83.MP.146.3.recto

Figure 5 CONA Object Record (only 1/4 is shown)

CONA objects are harvested using its own CONA schema (similar to CDWA). It has about 210 elements and is more complex than CDWA but less complex than LIDO (see below). The main elements (**Vocabulary** meaning aggregation of artworks and **Subject** meaning one artwork) are organized hierarchically and refer to smaller elements that also have a hierarchical structure.

```
Vocabulary =
  element Vocabulary {
    element Subject {
      (element Ancestors {
         (element Preferred_Ancestor_Branch {
            element Ancestor {
              element Immediate {ParentType},
              element Other {ParentType}*}}
          & element Non-Preferred_Ancestor_Branches {
              element Non-Preferred_Ancestor_Branch {
                element Ancestor {
                  element Immediate {ParentType},
                  element Other {ParentType}*}}+}?),
         [a:defaultValue = "300000000"]
         attribute Broad_Parent_ID {VP_Subject_ID}?}
       & element Associative_Relationships {
           element Associative_Relationship {
             element Description {xsd:string}?
             & element AR_Date {
                 element Display_Date {xsd:string}
                 & element Start_Date {Geog_Date}
                 & element End_Date {Geog_Date}}?
             & element Historic_Flag {associative_rels_historic_flag}?
             & element Relationship_Type {ar_code}
             & element Related_Subject_ID {
                 element VP_Subject_ID {VP_Subject_ID}
                 | element Contrib_Subject_ID {xsd:string}}}*}?
       & element Descriptive_Notes {
           element Descriptive_Note {
             element Note_Text {xsd:string}
             & element Note_Language {language_code}
             & element Note_Sources {
                 element Note_Source {
                   element Source {SourceType}
                   & element Page {xsd:string}?}+}}*}?
       & element Record_Type {sub_record_type}?
       & element Sort_Order {xsd:int}
       & element Special_Project {xsd:string}?
       & element Subject_Contributor {
           element Contributor_id {contrib_code}
           & element Contrib_subject_id {xsd:string}?
           & element Contrib_Note {xsd:string}?
           & element Editor_name {xsd:string}
           & element Editor_email {xsd:string}}
       & element Subject_Sources {
           element Subject_Source {
             element Source {SourceType}
             & element Page {xsd:string}?}+}?
       & element Terms {
           element Preferred_Term {TermType},
```

```
ClassificationType =
  element Class_ID {class_code}
  & element Preferred {class_rels_preferred}

CultureType =
  element Nationality_Code {nationality_code}
  & element Display_Order {xsd:unsignedShort}
  & element Preferred {depicted_sub_rels_preferred}

DepictedSubjectType =
  element Place_Number {VP_Subject_ID}?
  & element Person_Corp_Number {VP_Subject_ID}?
  & element AAT_Number {VP_Subject_ID}?
  & element CONA_Number {VP_Subject_ID}?
  & element Iconography_ID {VP_Subject_ID}?
  & element Preferred {depicted_sub_rels_preferred}
  & element Display_Order {xsd:unsignedShort}
  & element Indexing_Type {depicted_sub_rels_idx_type}?
  & element Subject_Extent {depicted_sub_rels_sub_extent}?
  & element Depicted_Term_ID {empty}?

DimensionsIndexingType =
  element Value {xsd:decimal}
  & element Unit {dimensions_idx_unit}
  & element Dimensions_Type {dimensions_idx_dim_type}
  & element Dimensions_Extent {dimensions_idx_dim_ext}?
  & element Dimensions_Qualifier {dimensions_idx_dim_qual}?
  & element Scale_Type {dimensions_idx_scale_type}?
  & element Format {dimensions_idx_format}?
  & element Shape {dimensions_idx_shape}?

EventType =
  element Event_ID {event_code}
  & element Display_Order {xsd:unsignedShort}
  & element Place {xsd:string}?
  & element Event_Date {
      element Display_Date {xsd:string}
      & element Start_Date {Historical_Date}
      & element End_Date {Historical_Date}}?

GenDepictedSubjectType =
  element Preferred {depicted_sub_rels_preferred}
  & element Display_Order {xsd:unsignedShort}
  & element General_Depicted {general_depicted_subject_gendep}
  & element Indexing_Type {depicted_sub_rels_idx_type}?
  & element Subject_Extent {depicted_sub_rels_sub_extent}?

Geog_Date = xsd:integer {maxInclusive = "9999"}

Historical_Date = xsd:integer {maxInclusive = "9999"}
```

Figure 6 CONA RNC Schema

2.6 LIDO

Lightweight Information Describing Objects (LIDO) (Erin Coburn, Richard Light, Gordon McKenna, Regine Stein, & Axel Vitzthum, 2010) is an XML schema that evolved from CDWA, a similar German standard called museumdat, and was inspired by CONA and the CIDOC CRM ontology (see below). LIDO covers descriptive and administrative elements and adds object-related events.

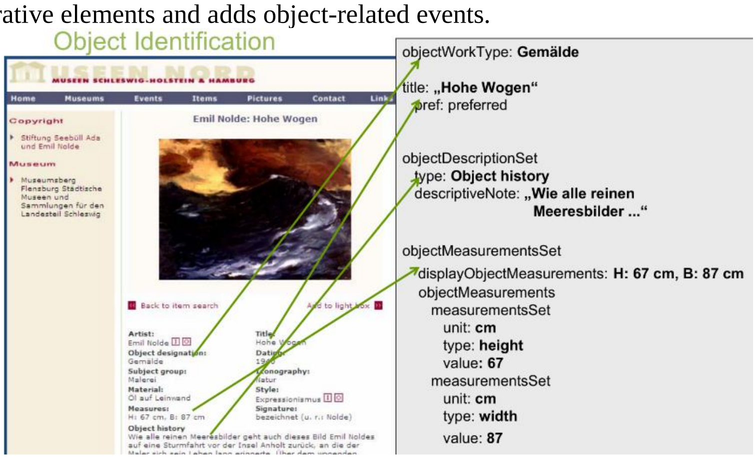


Figure 7 LIDO: Example Object and LIDO XML Representation

LIDO inherits from CDWA the dual representation of Display (viewing) vs Indexing (structured) elements. It is more complex than CDWA, e.g. when referring to a related object, you can provide almost as much detail as for the main object.
A LIDO Terminology (ICOM CIDOC LIDO Working Group, 2017) was developed that standardizes important concepts used in LIDO exchange, such as event types (creation, exhibition, acquisition, etc) and refers to available LOD sources as much as possible. The LIDO Terminology also presents an example of proper semantic publication, including semantic resolution and content negotiation. However, LIDO does not leverage sufficiently opportunities for linking between objects: objects are not required to have a web URL.

2.7 SPECTRUM

SPECTRUM (Kevin Gosling & Gordon MacKenna, 2017b, 2017a) is a UK museum standard (also used in other countries) that defines museum processes and data (Units of Information). SPECTRUM 4 defines a rather comprehensive data schema with 595 elements, of which the majority (490) are about **Object** (artwork). It covers the following areas, where we have also listed just a few of the sub-areas:

- Acquisition: AccessionDate, Funding, Price, Method (e.g. auction or gift), Owner, TitleTransfer
- Associations: General and Specific, including AssociatedActivity, RelatedEvent, RelatedObject
- Condition: Assessor, Method, TechnicalAttributes, Completeness, DamageLoss
- Conservation: Material, Treatment, Conservator
- Deaccession: Disposal, Reason, Recipient
- Description: Age, Colour, Components, Content: Activity, Concept, Event, Keyword, Name, Dimension, Form, Inscription (Content, Interpretation, Translation, Transliteration), Material, TechnicalAttribute, Technique
- Entry: Depositor, Packing, Reason, EnvironmentalCondition, FieldCollection: Collector, Place, Position, StratigraphicUnit (archaeological information)
- Exit: Destination, OrganizationCourier
- Exhibition: Begin/End, CatalogueNumber
- Identification: Name, Number, Title, OtherNumbers
- Indemnity: MinumumLiabilitySum
- Insurance: Insurer, PolicyNumber, RenewalDate
- Loan: In (Conditions, Lender), Out (Conditions, Borrower, Status, Venue)
- ObjectLocation: Address, Condition, Movement, Shipment: Shipper, PackingRecommendations
- ObjectView: a subjective Viewers experience, not digital depictions of the object
- Ownership: Current, Previous; AcquisitionMethod, Note, Price
- Procedure: Reason, Request, Manager, Cost, Process, Begin/End
- Production: Date, Person, People, Organization, Style (SPECTRUM allows both a specific author, "peoples" i.e. culture/ethnicity, or an organization such as a workshop or artist collective)
- Reference: AuthorEditor, Publication Date/Place

- Rights: CreditLine, Licenses, Begin/EndDate, Holder
- RiskManagement: Hazard, SalvagePriority
- Usage: DisplayStatus, Label, Recommendations (Display, Handling, Security), Research
- Valuation: Date, Valuer, Value
- Smaller areas such as: AccessCategory, Audit, Despatch

```
Object =
  element AccessCategory {
    element Date {ExactDate}?,
    element Name {xsd:string},
    element Note {xsd:string}?}?,
  element Acquisition {
    element AccessionDate {ExactDate}?,
    element Authorisation {
      element Authoriser {xsd:string},
      element Date {ExactDate}?}*,
    element Date {Date}?,
    element Funding {
      element Provisos {xsd:string}?,
      element Source {xsd:string}?,
      element Value {xsd:string}?}*,
    element GroupPrice {xsd:string}?,
    element Method {xsd:string}?,
    element Note {xsd:string}*,
    element OfferedPrice {xsd:string}*,
    element OfferPrice {xsd:string}*,
    element Owner {Person}*,
    element Price {
      element LocalValue {xsd:string},
      element OriginalCurrency {xsd:string}?,
      element OriginalValue {xsd:string}?}?,
    element Provisos {xsd:string}*,
    element Reason {xsd:string}*,
    element Reference {xsd:string}*,
    element Source {empty}*,
    element TitleTransfer {
      element Number {xsd:string}}?}?,
  element Associations {
    element GeneralAssociations {
      element Concept {xsd:string}?,
      element CulturalAffinity {xsd:string}?,
      element HistoryNote {xsd:string}?}*,
    element SpecificAssociations {
      element Association {
        element AssociatedActivity {
          element Activity {xsd:string}?,
          element AssociatedActivityNote {xsd:string}?}*,
        element AssociatedObject {
          element AssociatedObject {Object}?,
          element AssociatedObjectType {xsd:string}?}*,
        element AssociationNote {xsd:string}?,
        element Date {Date}?,
        element Organisation {Organisation}?,
        element People {People}?,
        element Person {Person}?,
        element Place {Place}?,
        element RelatedEvent {
```

```
People =
  element Culture {xsd:string}*,
  element Group {xsd:string}*,
  element LinguisticGroup {xsd:string}*,
  element PeoplesAssociation {xsd:string}?

Person =
  element Address {Address}*,
  element BiographicalNote {xsd:string}*,
  element Birth {
    element Date {Date}?,
    element Place {Place}?}?,
  element Death {
    element Date {Date}?,
    element Place {Place}?}?,
  element Gender {xsd:string}?,
  element Group {xsd:string}*,
  element Name {
    element Additions {xsd:string}?,
    element Forename {xsd:string}*,
    element Initials {xsd:string}?,
    element Note {xsd:string}*,
    element Surname {xsd:string}},
  element Nationality {xsd:string}*,
  element Occupation {xsd:string}*,
  element PersonsAssociation {xsd:string}?,
  element PersonTitle {xsd:string}*,
  element Reference {xsd:string}?,
  element Salutation {xsd:string}?,
  element SchoolStyle {xsd:string}*

Place =
  element Coordinates {
    element Qualifier {xsd:string}?,
    element Type {xsd:string}?,
    element Value {xsd:string}}*,
  element EnvironmentalDetails {xsd:string}*,
  element Feature {
    element Date {Date}?,
    element Keyword {xsd:string},
    element PlaceTypeSystem {xsd:string}?,
    element Type {xsd:string}?}*,
  element Name {xsd:string}?,
  element NameType {xsd:string}?,
  element Note {xsd:string}*,
  element Owner {empty}*,
  element PlaceAssociation {xsd:string}?,
  element Reference {
    element Number {xsd:string},
    element Type {xsd:string}?}?,
  element Status {xsd:string}?
```

Figure 8 SPECTRUM Units of Information Schema (part)

While there are several SPECTRUM-recommended systems ("SPECTRUM Partners"), we are unfortunately not aware of any system that can process SPECTRUM XML.

A mapping from LIDO to SPECTRUM (Gordon McKenna, 2018) underscores the complexity of LIDO.

Table 1 Mapping of LIDO to SPECTRUM (part)

Spectrum	**LIDO elements path in the MINT system - element containing data in bold**
[Name of organisation]	[Object Identification] ⇨ repositoryWrap ⇨ repositorySet [with **type** attribute = ‘Current’] ⇨ repositoryName ⇨ legalBodyName ⇨ **appellationValue**
Object number	[Object Identification] ⇨ repositoryWrap ⇨ repositorySet ⇨ **workID** [with **type** attribute = ‘Current’]
Brief description	[Object Identification] ⇨ objectDescriptionWrap ⇨ objectDescriptionSet ⇨ **descriptiveNoteValue**
Number of objects	[Object Identification] ⇨ objectDescriptionWrap ⇨ objectDescriptionSet [with **type** attribute = 'Number-of-objects'] ⇨ **descriptiveNoteValue**
Object name	[Object Classification] ⇨ objectWorkTypeWrap ⇨ objectWorkType ⇨ **term**
Title	[Object Identification] ⇨ titleWrap ⇨ titleSet ⇨ **appellationValue**

3 Ontologies and Semantic Projects

While XML schemas enable the exchange of information, they carry a lot of syntactic baggage (there are many different ways to structure the same information) and do not enable global information sharing (objects are not required to have URLs). RDF and semantic technologies eliminate these shortcomings, enabling the global accumulation and reuse of museum and authority data LOD.

In my opinion, currently there is no dominant and commonly accepted ontology for describing artworks and museum objects. There is tension between several communities. Below are the strongest candidates per my subjective opinion:

- **CIDOC CRM**. Pros: strong foundational ontology, used by numerous projects especially in Europe. Cons: many consider it complicated, some shortcomings for describing relations between people and between objects, not friendly for integrating with other ontologies, the community (SIG) is slow to adopt practically important issues, few application profiles for specific kinds of objects (e.g. coins vs paintings).
- **linked.art**. Pros: a simplified CRM profile created under the moniker "Linked Open **Usable** Data (LOUD)", more developer friendly through an emphasis on JSONLD, used by some projects especially in the US. Cons: various simplifications that are not vetted by the CRM SIG, rift with European CRM developments.
- **Schema.org**. Pros: supported by the major search engines thus ensures semantic SEO and findability, used by the largest amount of LOD (on billions of websites), pragmatic and collaborative process for data modeling with a lot of examples, possible extensions as exemplified by bibliographic (SchemaBibEx) and archival extension. Cons: not yet proven it is sufficient to represent
- **Wikidata**. Pros: universal platform for data integration, richer model than RDF (but also exposed as RDF), pragmatic and versatile collaborative process for data modeling (property creation) with a lot of examples and justifications, used by some GLAMs and crowd-sourced projects (e.g. Authority Contorl,Sum of All Paintings, Wiki Loves Monuments). Cons: institutional endorsement is not yet

strong enough, concerns of institutions how they can be masters of "their own" data.

But I open this section with two ontologies that are not limited to CH, and nevertheless are used widely in CH applications.

3.1 Web Annotation

Web Annotation (Open Annotation, OA) is an important W3C standard that covers all kinds of interactions between users and resources: bookmarking, commenting, editing, highlighting, sharing, making relations between resources, etc. Together with ontologies for advanced citation (the SPAR ontologies), it is by now considered crucial for supporting structured scholarly collaboration on the web, and used widely in life sciences, CH and DH. It is also the foundation of advanced IIIF applications such as Shared Canvas, see below.

OA has been in development for over 8 years. It incorporated results from several initiatives:

- The Open Annotation Collaboration project funded by the Mellon Foundation lasted from 2009 to 2013. It developed core use cases and data models and included many participants from the GLAM domain.
- Annotation Ontology and several annotator softwares (e.g. DOMEO) inspired by the Life Sciences and scholarly domain

The W3C Open Annotation Community Group merged these efforts, and the Web Annotation Working Group evolved the specifications to Recommendation status

OA Specifications The most recent specifications (Feb 2017) include the following.

- Web Annotation Data Model: description of the ontology, different use cases and combinations
- Web Annotation Protocol: protocol that defines the interaction between annotation servers and annotation clients
- Selectors and States: how to select part of a resource (e.g. section of a HTML document, a particular sentence, rectangle from a PNG image, structural part of a SVG image, page of a PDF) or specify a particular version of a resource as it existed at a certain time. The specification can be done as RDF triples or as URL "fragment selectors" (e.g. "#page=100" for a PDF or "#xywh=100,100,300,300" for an image). This is extracted from the main specification, so it can be reused by other ontologies.
- Embedding Web Annotations in HTML. The examples in RDFa are derived from dokieli, which is a HTML+RDFa template and tool for next-generation scholarly writing and communication.

OA Resources

Perhaps the best way to learn about OA are the slideshares of Rob Sanderson and Paolo Ciccarese. See for example

- Open Annotation Core Data Model Tutorial
- Open Annotation Specifiers and Specific Resources Tutorial
- Multiplicity and Publishing in Open Annotation Tutorial. (Multiplicity is when a **group** of resources is related to another resource, or another group).

While the specifications are rather dry, there are excellent illustrations in:

- The Open Annotation Collaboration site
- The above-mentioned slideshares
- The Open Annotation Cookbook

A further development of these ideas are the Linked Data Platform and Linked Data Notifications specifications, which allow fully distributed notification and storage, creating the foundation for distributed social applications and Linked Research (see e.g. dokieli). Solid and CrossCloud are two current projects to decentralize the web, involving Tim-Berners Lee (creator of the web and the semantic web).

Annotation has always been an interesting topic of development, starting with the W3C Annotea project (2001-2003)

- An older list of Annotation Projects & Resources (compiled by Hypothes.is) includes: A.nnotate, AnnotateIt, Annozilla, BounceApp, CritiqueIt, Crocodoc, dbunkr, Diigo, DocumentCloud, Filesquare, FinalsClub, Findings, Goodreader, Goozy, Highlighter.com, HowTru, iCorrect, Insight, JournalTalk, LEMO, Marginize, Markup.io, Mendeley, My WOT, NotableApp, Orseis, Shareflow, Substance.io, WebKlipper, YouSticker.
- Within ResearchSpace, Ontotext implemented old versions of OA for Data and Image Annotation (with Deep Zoom)

The availability of an open and stable OA specification has spurned renewed interest, and a large number of implementation efforts. Some interesting examples (mostly from the GLAM domain):

- Annotorious image and text annotator by Austrian Institute of Technology, developed as part of the EuropeanaConnect project
- Lorestore server and Annotator OA client by University of Queensland, Australia
- OACVideoAnnotator by UMD MITH and Alexander Street Press
- The LombardPress annotator of ancient manuscripts that works over canonic text representations in the Scholastic Commentaries and Texts Archive (based on Linked Data Notifications)
- Annotopia by MIND Informatics group, Massachusetts General Hospital

To reach Recommendation status, every W3C specification requires test suites, and a certain number of independently developed conforming implementations. Tests for a number of web technologies have now been centralized on the Web Platform Tests site. Compliant implementations listed on the Annotation Model testing report include:

- Reference Implementation of an Annotation protocol server that implements the new Collection and Page portions of the annotation data model.
- Conquering Corsairs (MangoServer) by Rob Sanderson
- Emblematica Online by University of Illinois Library
- Hypothes.is, perhaps the largest OA project and development community, funded by the Knight, Mellon, Shuttleworth, Sloan, Helmsley, and Omidyar foundations and supported by OKFN. It implements the core AnnotatorJS project. A number of tools, plug-ins and integrations are available, including Drupal, WordPress and Omeka integrations. Omeka is a popular light-weight CMS and virtual exhibition system
- Europeana Annotation Server

- Mirador client (a well-known IIIF viewer, see below) with MangoServer
- Wellcome Quilt, funded by the Wellcome Trust
- Pundit by Net7, developed through several EU projects (e.g. SemLib, DM2E)
- Image Annotator by KANZAKI Masahide
- Page Notes
- Re-narrations and SWeeT Web (source)

We expect the list of implementations to grow quickly, e.g. a new one is:

- Annotation module for Omeka-S, the new generation of Omeka implemented over JSONLD RDF. It allows various annotations (tag, comment, rate, highlight, draw, etc)

Two examples of OA RDF data models are shown below.

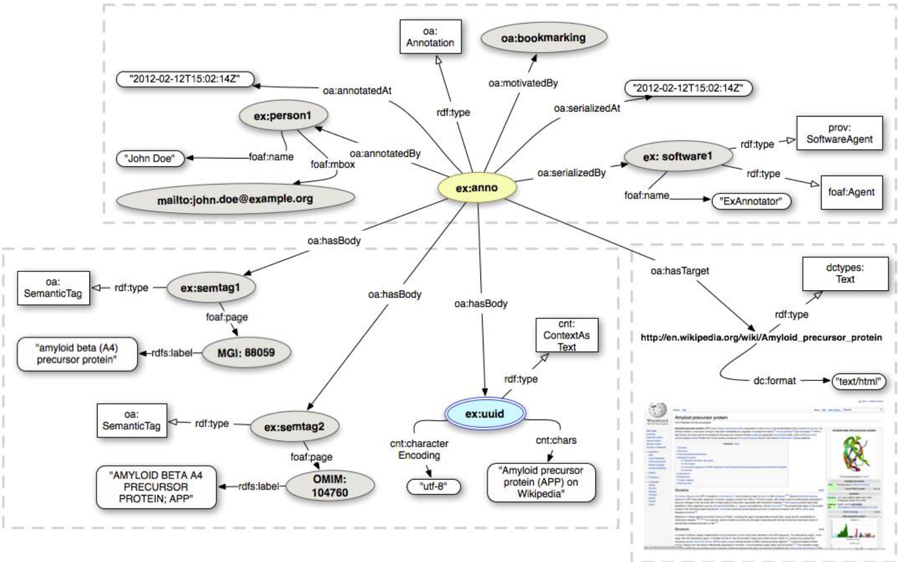


Figure 9 OA Cookbook: Bookmarking and Semantic Tagging

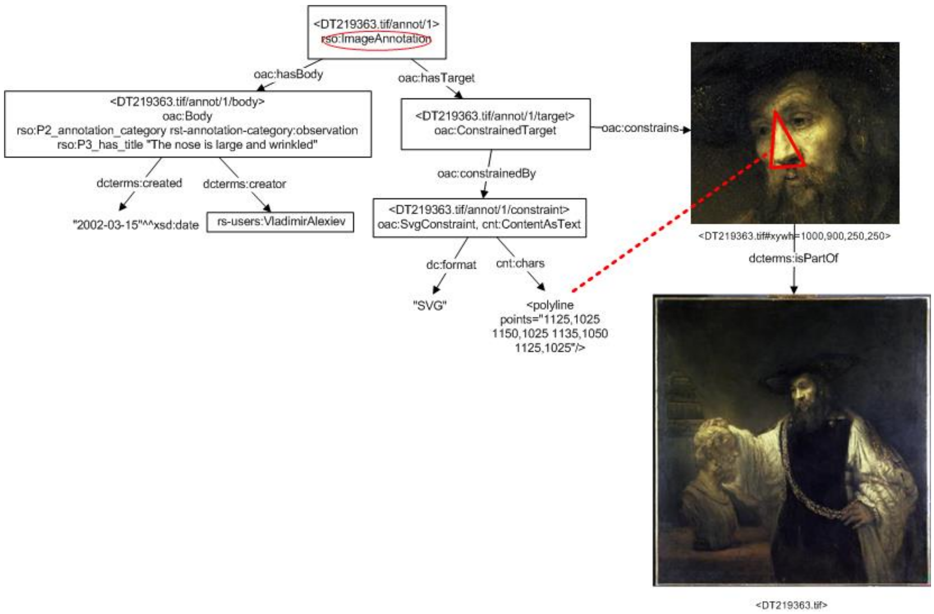


Figure 10 ResearchSpace Image Annotation: Annotating Part of Image with SVG

3.2 IIIF

The International Image Interoperability Framework (IIIF, http://iiif.io/) enables handling of deep zoom (very large resolution) images and applications based on them: book viewers, image composition, image annotation, etc. By defining a client-server protocol, it enables interoperability between image servers (Digital Asset Management) and clients (viewers, annotators). It specifies 4 APIs:

- Image: semantic description of images (available resolutions, features, credit line, conformance level, etc) and serving features such as zooming, gray-scaling, cropping, rotation, etc
- Presentation (Shared Canvas): laying images side by side, assembling folios and books (using so-called IIIF Manifests), image annotation. This has been very popular for virtual reconstruction of manuscripts, book viewers, etc
- Authentication: describes modes or interaction patterns for getting access to protected resources (e.g. Login, Click-through, Kiosk, External authentication)
- Search: search of full-text embedded or related to image resources (e.g. OCRed or manually annotated text of some old book).

Various open source IIIF clients are available, most are based on Javascript and HTML5:

- Diva.js, especially suited for use in archival book digitization initiatives
- IIPMooViewer, for image streaming and zooming

- Mirador, implementing a workspace that enables comparison of multiple images from multiple repositories, widely used for manuscripts
- OpenSeadragon, enabling smooth deep zoom and pan
- Leaflet-IIIF, a plugin for the Leaflet framework that also includes display of geographic maps
- Universal Viewer, widely used by CH institutions

Examples of IIIF servers include:

- Cantaloupe, enabling on-demand generation of image derivatives
- IIPImage Server, fast C++ server also used for scientific imagery such as multispectral or hyperspectral images
- Loris, a server written in Python
- ContentDM, a full-featured digital collection management (DAM) system
- Djatoka, a Java-based image server
- Digilib, another Java-based image server

Below we show two examples of IIIF applications:

- Biblissima is the French national manuscript library, based on CIDOC CRM and FRBRoo metadata and IIIF digital asset handling. An IIIF Mirador Viewer is configured to view and compare manuscript images of mermaids from various sources.
- Europeana can search for CHO with IIIF representations by using the search term **sv_dcterms_conformsTo:*iiif***. This returns 2.5M objects.

Figure 11 IIIF Mirador Viewer at Biblissima, the French manuscript library

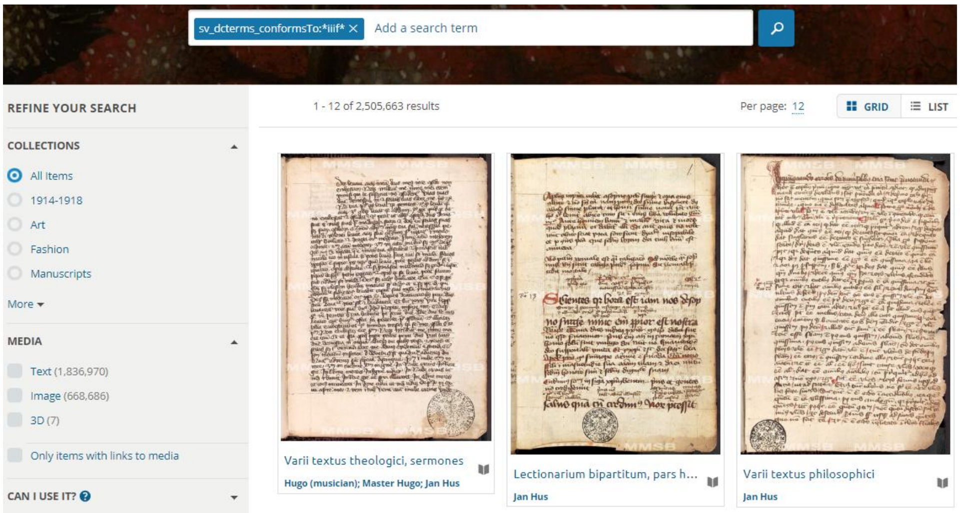


Figure 12 Search for IIIF Images on Europeana

3.3 Europeana and EDM

Europeana is a large-scale CH aggregation that covers CH from institutions in Europe (not only EU member states), Israel and some other countries. It includes artefacts from all over the world (not limited to Europe). It started in 2008 and has aggregated 53M objects at present, described using the Europeana Data Model (EDM), an RDF ontology. Europeana has a general search and display mechanism. The search is not semantic (e.g. won't catch different multilingual names, unless they are included in enriched object data) and includes a set of fixed facets (including image characteristics). Below we show a search for objects, and one particular object (manuscript). The same object is depicted as an RDF graph at the end of this section.

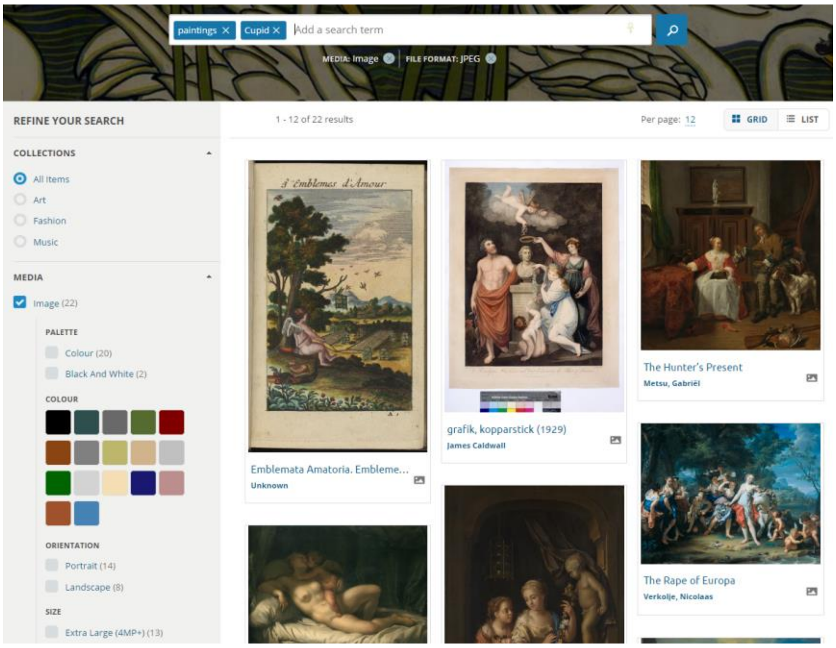


Figure 13 Europeana Search: Paintings of Cupid

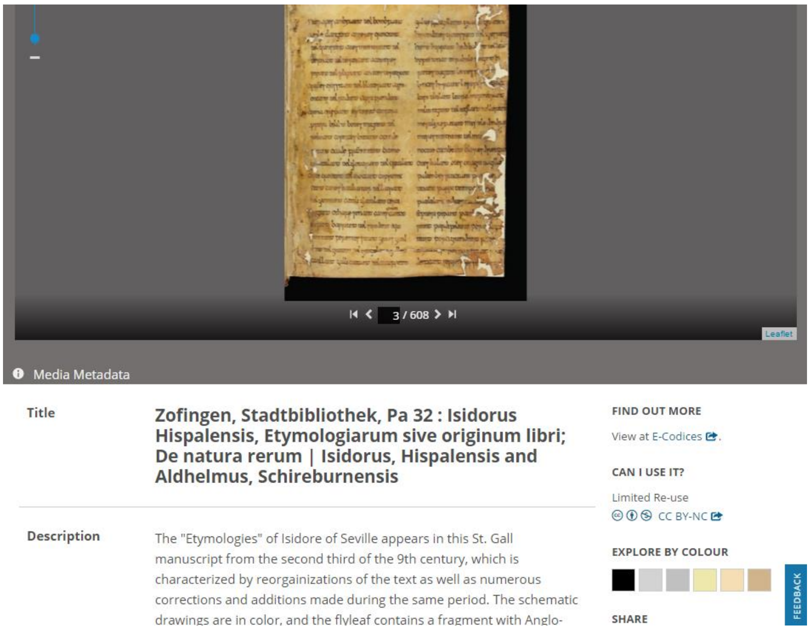


Figure 14 Example Europeana Object: a Swiss Manuscript

Europeana has been criticized for providing a similar look to all kinds of objects, thus not respecting provider wishes and established practices in different domains. So Europeana has created several Thematic Collections (Art, Fashion, Music) that have their own look and features.

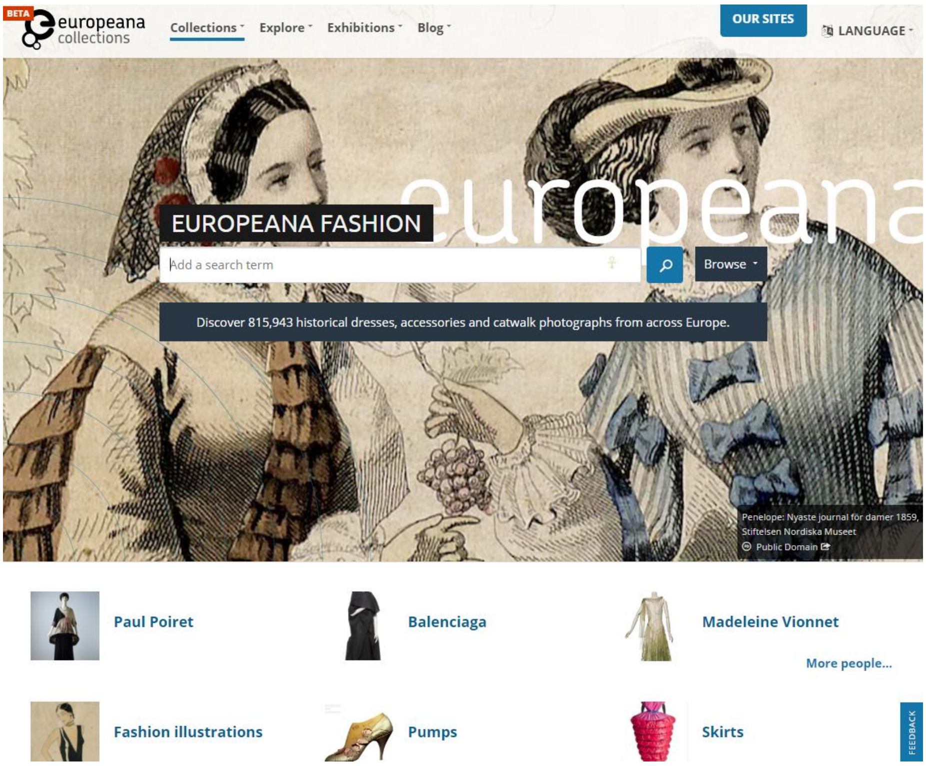


Figure 15 Europeana Fashion Thematic Collection

Europeana is a long-term program (over 10 years), with perhaps 50 associated projects that aggregated data in particular domains, e.g.:

- APE and APEx aggregated archival information (see below)
- Europeana Regia collected royal illuminated manuscripts
- DM2E (Digital Manuscripts to Europeana) contributed medieval manuscripts, and developed an EDM extension for manuscripts
- PartagePlus collected Art Nouveau
- Europeana Fashion collected artefacts about fashion and garments, which resulted in the establishment of a professional association to continue the project
- Europeana Judaica collected artefacts related to the Judaic tradition
- ECLAP aggregated objects describing performance art
- Europeana Inside developed connectors for several popular Collection Management Systems (CMS) to ease the aggregation of Europeana objects.
- Europeana Creative developed several creative applications, paving the way for reuse of Europeana data by the creative industries.

- Europeana Sounds collects music and other audio, and developed an EDM extension for music.
- Europeana Food and Drink collected food and drink related heritage and developed several applications, including a semantic app (Vladimir Alexiev, Andrey Tagarev, & Laura Tolosi, 2016). It includes semantic hierarchical facets for food and drink topics (based on Wikipedia categories) and places (based on Geonames).

Figure 16 Europeana Food and Drink (EFD) Semantic Application

In addition, various national aggregators have emerged, e.g.:
- Collection Trust established CultureGrid in the UK
- The German Digital Library (DDB) is the aggregator for DE
- DigitaleCollectie is the aggregator for NL
- The Varna library established the first BG aggregator, and Ontotext established Bulgariana, an aggregator with a more technological orientation. E.g. Bulgariana submitted a BG traditional recipes collection to EFD, including semantic enrichment.

Many Europeana satellite projects have faced sustainability problems, i.e. inability to continue collecting and updating objects after the project finishes. Good exceptions are Apex and Europeana Fashion that have established respective associations to continue the work. Not coincidentally, these projects (especially Apex) often collect richer metadata and submit a subset of it as EDM to Europeana. Even some national aggregators faced sustainability problems.

Aggregating a collection often takes a long time by Europeana (several weeks) because of slow iteration cycles of test ingestion, previews, checking object quality. Europeana has changed several aggregation approaches and software: SIP ingest, Unified Ingest Manager (UIM); Europeana Inside (connectors to various Collection Management Systems, CMS); Operation Direct (announced at Europeana's 2016 AGM): an API-based ingestion approach, where a CMS can submit and update individual objects directly to Europeana, and it adds them to the search index incrementally; and is currently working on Metis.

The Europeana API allows applications to search for objects, using a large selection of search fields. However, it does not allow complex queries (e.g. across objects, result aggregation such as count or sum and group by, searching by author characteristics such as nationality, by concept or place hierarchy, etc). Although EDM is an RDF ontology, semantic technologies are not used in the core of Europeana. Instead, it uses SOLR to index all search fields.

Europeana Labs provides a gallery of datasets and apps. Several Europeana projects (starting with Europeana Creative and Food and Drink) have organized competitions, provided prizes and start-up support, in an effort to increase creative reuse of CH materials.

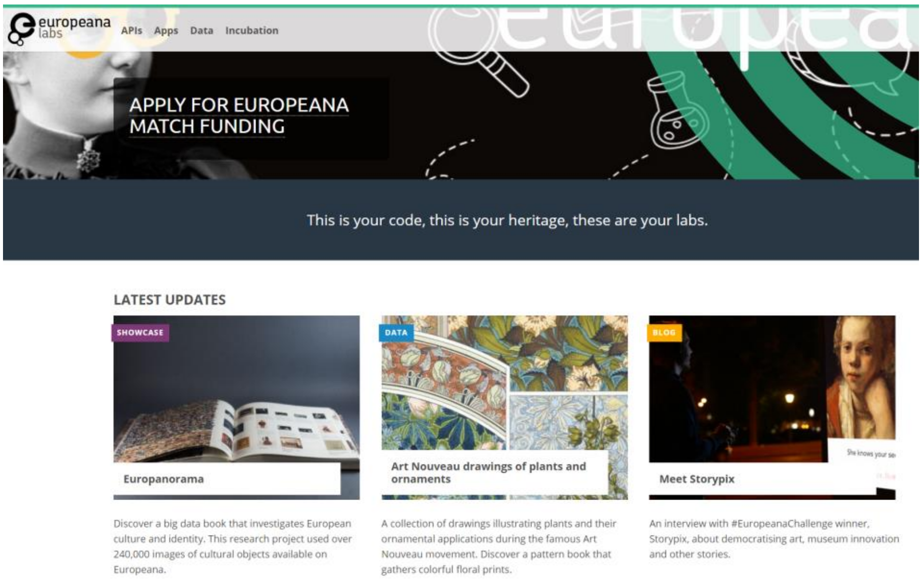


Figure 17 Europeana Labs Galleries of Apps and Datasets

Europeana uses the OAI PMH protocol to aggregate content from aggregators. In 2015 it also established an OAI PMH server developed by Ontotext (Vladimir Alexiev & Dilyana Angelova, 2015) to allow mass-downloading of metadata. Ontotext also created the Europeana SPARQL endpoint allowing complex queries, which was later replaced by an open source RDF repository. However, the SPARQL endpoint is not supported well (there is a google group with little traffic) and is not widely used.

Europeana is currently funded by the EC as a Digital Service Infrastructure (DSI) under the Connecting Europe Facility (CEF). Although the funding is smaller than in previous years. This ensures Europeana's longevity. Recent targeted funding includes projects for creating more collections, e.g.

- 2016-EU-IA-0101 Migration in the Arts and Sciences
- 2016-EU-IA-0093 Rise of Literacy in Europe
- 2016-EU-IA-0094 BYZART - Byzantine Art and Archaeology

Critiques For a long time Europeana has focused on quantity rather than quality, which led to:

- Low metadata quality of some of the collected objects: poor or incomplete metadata, mistakes in metadata structure, broken links, etc.
- Uneven content selection criteria. For example, AskAboutIreland contributed yellow pages (phone books) from 1975, every page as a separate object; LGMA contributed photos of common foods like carrots and jelly, etc
- Aggregation through one-off projects, leading to inability to update the aggregated collections (provide new content) and low availability of images and institutional websites

EDM has been criticized by some in the CIDOC CRM community (Dominic Oldman, Martin Doerr, Gerald de Jong, Barry Norton, & Thomas Wikman, 2014) for being a least-common-denominator model that shoe-horns CH institutions into providing a poorer version of their metadata. Since aggregation initiatives are expensive, data should be aggregated in a rich format to begin with, and the Synergy Reference Model is proposed to that end. While EDM allows richer modelling such as events, this is not supported by Europeana and many existing metadata collections have little more than Dublin Core.

In the last two years Europeana has put Data Quality in the middle of its Strategic agenda. In particular:

- Two task forces have focused on Enrichment, since semantic enrichment of metadata is one of the ways to increase the value of metadata.
- A Data Quality task force (May 2015) took account of the situation and outlined problems.
- A permanent Data Quality Committee was formed to define and validate quality rules (e.g. using mechanisms such as RDF Shapes) and measure metadata coverage.
- The Europeana Publishing Framework established tiers of participation, where some institutions can benefit more by providing higher-quality collections, better-resolution images, and richer metadata.

Despite the progress, a lot of work remains to make Europeana objects most useful for consumers and researchers.

Pros One of the most important achievements of Europeana is increasing the level of networking of CH institutions in Europe. Europeana has also been very strong in user engagement, developer engagement (hackathons and Europeana Lab), lobbying for digitization and CH in Europe.

Europeana has a strong distributed organization. It operates through several interconnected groups:

- About 3500 CH institutions contribute content through a network of Aggregators, reducing the load on the Europeana office.
- Funding is sought by and provided through the Europeana Foundation, which is hosted by the KB (Dutch national library)
- The Europeana Association is a voluntary organization with about 3000 individual members. It meets yearly at the Europeana AGM (the travel of one member per

organization is funded by Europeana). It elects a Members Council, which elects a Management Board that participates in setting Europeana strategy, selecting task forces, etc

- Task Forces (e.g. Enrichment Strategy, Enrichment and Evaluation, Open Source Software, etc) are temporary groups assembled to elaborate and make recommendations on issues of importance, usually finishing with a Final report. Working Groups are more permanent (e.g. Data Quality).

Europeana has set some technological examples (e.g. the EDM) that have been followed by DPLA. Also, Europeana is cooperating with DPLA and other organizations on license standardization (RightsStatements.org), IIIF images, schema.org representation for better findability by search engines, etc.

The Europeana Data Model (EDM) (Europeana, 2017) is an RDF ontology used by Europeana for harvesting and managing cultural heritage objects (CHO). EDM builds upon:

- Dublin Core (DC): descriptive metadata
- OAI ORE (Open Archives Initiative Object Reuse & Exchange): organizing object metadata and digital representations (WebResources)
- SKOS (Simple Knowledge Organization System): contextual objects (concepts, agents, etc)

EDM is inspired by CIDOC CRM (see below): events, some relations between objects. EDM describes:

- CHOs (ProvidedCHO)
- The real-world things related to them (Non-Information Resources, also called contextual entities or contextual objects).
- Metadata records (aggregations)
- Associated media (WebResources)

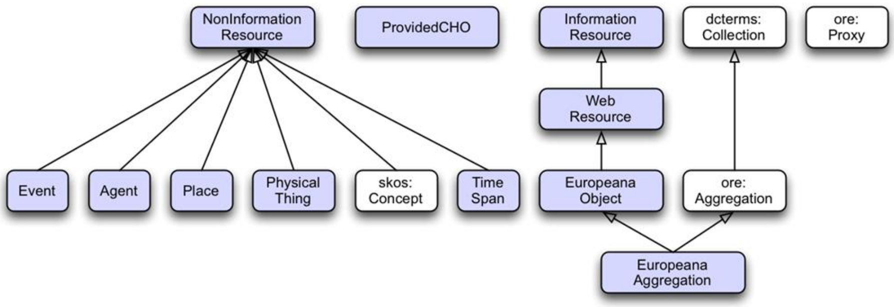


Figure 18 EDM Class Hierarchy

EDM includes two auxiliary classes from ORE, which are used to split the information into clearly delineated nodes:

- Proxy carries object information, as provided by a certain agent (the data Provider or Europeana)
- Aggregation carries information about the provider, collection, metadata rights, etc

EDM has two flavors:

- **External** as served by the Provider (aggregator). It has only 2 nodes, ProvidedCHO and Aggregation.
- **Internal**: after Europeana ingests the object, it splits the object info to the Provider Proxy and adds extra info in the Europeana Aggregation and Proxy. The ProvidedCHO node itself does not carry information.

A typical EDM graph is shown below, highlighting the nodes centered around CHO.

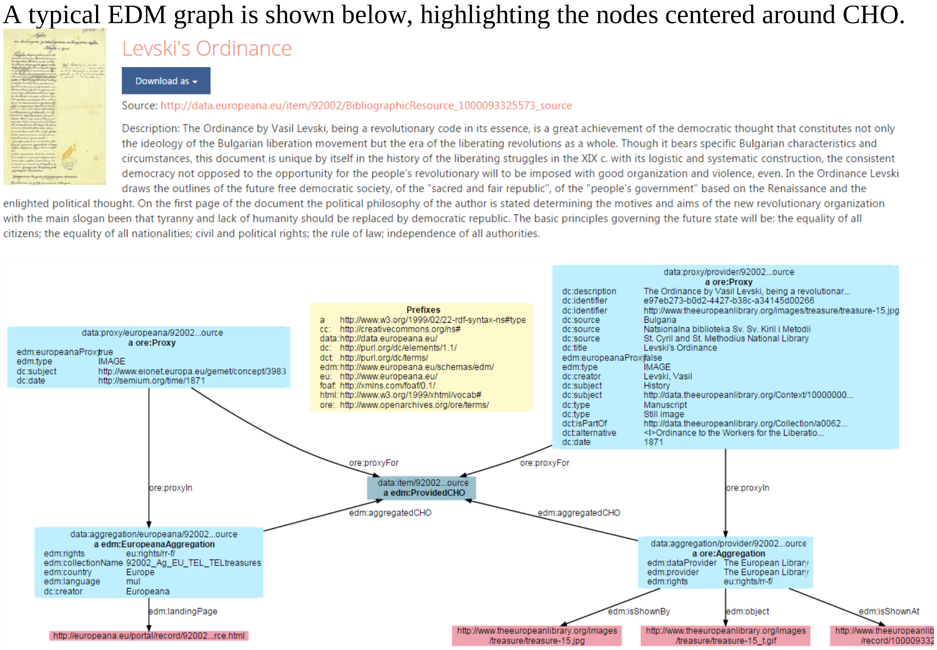


Figure 19 Typical EDM Graph (from Ontotext's Europeana endpoint)

Despite the complicated graph structure, typical EDM objects used to have little more than DC information. In particular, they had few if any references to global authorities, rather providing mere strings. However, Europeana has started providing more enrichments against authorities such as Geonames, DBpedia, Getty AAT. The Europeana Entity Base copies relevant authorities from LOD sources (only resources that appear in CHOs or are widely used) and equivalences to the original URLs in those datasets. Also, Europeana implements and promotes the use of IIIF for deep zoom images. Consider the following CHO describing a manuscript from e-codices (Switzerland).

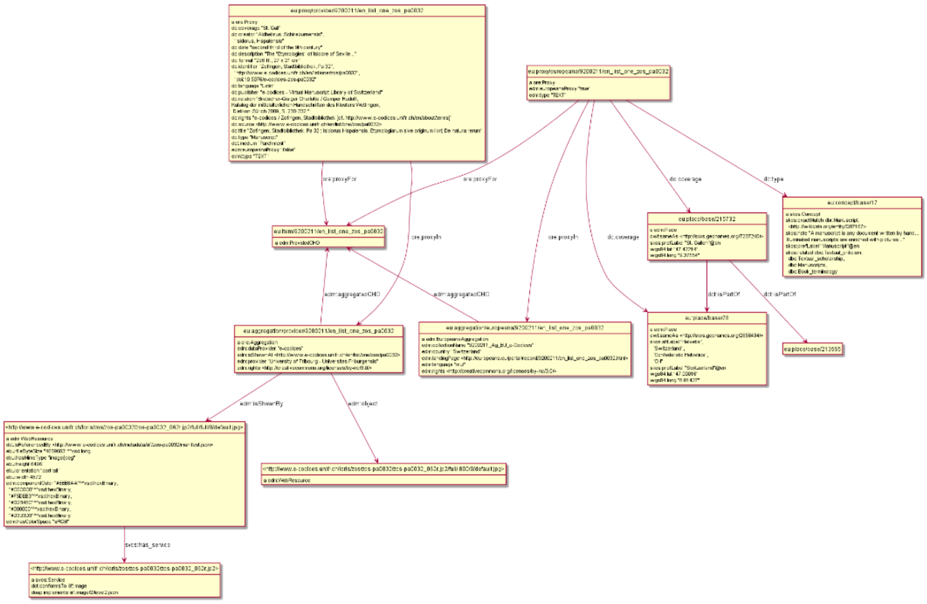


Figure 20 EDM Example Graph: Manuscript with IIIF Representation and Contextual Entities

This CHO has the following features:

- The Provider Proxy (original information) is very rich: it provides info about creators, related pages and digital object (DOI), bibliography (dc:relation), license (dc:rights), source page. Nevertheless, all these are strings not things: e.g. the creators don't refer to a global authority like VIAF, the rights URL is hidden in a string, the bibliographic entry doesn't link to the referenced book and its authors.
- The Provider Aggregation links to the rich IIIF description (edm:isShownBy), and to a simpler thumbnail (edm:object).
- The IIIF description includes link to the IIIF manifest (json) that further describes the deep-zoom image; some properties from an EBU (European Broadcasting Union) ontology: max width and height, byte size, orientation; EDM extensions for describing colors (extracted by Europeana), and link to the IIIF service URL that implements deep zoom for that image (IIIF Image API), also stating the IIIF profile implemented (level2).
- The Europeana Aggregation provides rich contextual entities
  - The CHO is a manuscript (concept/base/17), with multilingual labels (skos:prefLabel) and definitions (skos:note), only EN is shown for simplicity, links to DBpedia (dbr:) and Wikidata, links to Wikipedia categories (dbc:)
  - The CHO is related (dct:coverage) to two places: Switzerland and St.Gallen, provided with coordinates; multilingual labels: preferred (e.g. "Switzerland") and alternative (e.g. "Helvetia", "CH"); hierarchical links (place/base/21355 is the second-level administrative division ADM2 Wahlkreis St. Gallen); links to Geonames.

All of this information and links allow the creation of useful applications, such as semantic and hierarchical faceted search, e.g. the Europeana Food and Drink semantic application. But few Europeana objects have this level of detail and quality.
Several EDM extensions and profiles have been proposed:

- Describing Hierarchical Objects, such as books
- Extending EDM with properties from FRBRoo
- EDM Profile for Sound

The EDM mappings, refinements and extensions task force published a report (2014) on various extensions and extension approaches. The EDM Extensions workshop (2015) developed directions for future extensions.

An important EDM feature was identified by the Europeana Data Quality Committee as required to improve the precision of describing author contributions to artworks: edm:Event with dc:type being Production or a specific "business sub-type" such as design, gilding, decoration, translation, etc. Although EDM includes such class, it is not implemented in the Europeana portal and consequently is not used by data providers. This class can be used to express exactly who did what and when to contribute to the creation of the artwork (inspired by CRM class crm:E12_Production). Some providers have such data in their databases, but currently cannot transmit it to Europeana.

3.4 CIDOC CRM

The CIDOC Conceptual Reference Model (CRM) (Patrick Le Boeuf, Martin Doerr, Christian Emil Ore, & Stephen Stead, 2018) is a foundational ontology for history, archeology and art. It is developed by ICOM, CIDOC (International Committee for Documentation), CRM Special Interest Group and to some extent the Documentation Standards Working Group. It has been in development for 17 years (since 1999) and standardized as ISO 21127:2006 in 2006. The ontology continues to evolve: the current version with RDF representation is CRM 6.2.1 (Oct 2015), the version in progress is CRM 6.2.3 (May 2018). It has about 85 classes and 285 properties (about 140 object properties and their inverses, and a few that don’t have inverses).

The CIDOC CRM site (http://www.cidoc-crm.org) has a new design, but the old site should also be consulted sometimes, since some important resources are missing or have changed URL, e.g. http://old.cidoc-crm.org/comprehensive_intro.html. Many resources are available to learn CIDOC CRM, e.g.:

- Video Tutorial (2008) that explains the logic of CIDOC CRM, especially the event orientation.
- Graphical Representation: presents "typical situations" or CRM constructs. Includes a comprehensive property and class index that allows you to lookup a certain ontology element in all typical situations. E.g. below we see that E36 Visual Item appears in Documentation and References, Image Information Objects and Carriers, and Mark and Inscription Information (the subclasses of Visual Item are Image, Mark and Inscription). Further CRM Graphical examples are used below.

1. CIDOC CRM Class Hierarchy
2. Acquisition Information E1 E2 E5 E7 E8 E18 E39 E52 E55 E62 E78 E87; P2 P3 P3.1 P4 P14 P14.1 P22 P23 P24 P51 P52 P147
3. Appellation Information E1 E15 E35 E39 E41 E42 E44 E49 E51 E52 E53 E71 E75 E82 E90; P1 P37 P38 P48 P78 P87 P102 P131 P139
4. Attribute Assignement E1 E5 E7 E13 E55 E62; P2 P3 P3.1 P140 P141
5. Changing Thing E11 E12 E18 E24 E63 E64 E77 E79 E80 E81; P31 P92 P93 P108 P110 P111 P112 P113 P123 P124
6. Collection Information E7 E18 E24 E39 E53 E78 E79 E80 E87; P14 P14.1 P46 P49 P50 P53 P109 P110 P111 P112 P113 P147
7. Condition Information E1 E2 E3 E5 E7 E14 E18 E29 E39 E52 E55 E62; P2 P3 P3.1 P4 P5 P14 P14.1 P20 P32 P33 P34 P35 P44
8. Deaccession and Disposal Information E1 E2 E5 E6 E7 E8 E10 E18 E39 E52 E55 E62; P2 P3 P3.1 P4 P13 P14 P14.1 P22 P23 P24 P28
9. Description Information E1 E62 P3 P3.1
10. Documentation and References E1 E29 E31 E32 E33 E36 E55 E73 E89 E90; P67 P67.1 P70 P71 P106 P129 P138 P138.1 P148
11. Existence Information E6 E12 E18 E21 E24 E28 E63 E64 E65 E66 E67 E68 E69 E74 E77 E81; P13 P92 P93 P94 P95 P96 P97 P98 P99
12. Group Dynamics E5 E7 E39 E74 E77 E85 E86; P11 P12 P143 P144 P145 P146
13. Image Information Objects and Carriers E1 E24 E36 E38 E41 E73 E84 E89 E90; P1 P62 P62.1 P65 P67 P67.1 P106 P128 P129 P138
14. Institution Information E20 E21 E39 E40 E41 E44 E45 E51 E53 E74 E77 E82 E90; P74 P76 P87 P107 P131
15. Location Information E4 E5 E7 E9 E12 E18 E19 E24 E44 E45 E46 E47 E48 E53 E55 P7 P8; P20 P21 P25 P26 P27 P53 P54 P55 P58 P
16. Mark and Inscription Information E1 E7 E11 E18 E19 E22 E24 E25 E26 E33 E34 E36 E37 E38 E53 E55 E56 E57 E62 E70 E73; P2 P
17. Material and Technique Information E7 E11 E12 E18 E24 E29 E39 E55 E57; P14 P14.1 P31 P32 P33 P45 P68 P69 P69.1 P108 P129
18. Measurement Information E1 E7 E13 E16 E29 E54 E55 E58 E60 E70; P16 P16.1 P32 P33 P39 P40 P43 P90 P91 P125 P140 P141
19. Object Association Information E2 E4 E5 E7 E39 E52 E53 E55 E70 E71 E77; P4 P7 P11 P12 P16 P16.1 P19 P19.1 P101 P103
20. Object Collection Information E1 E2 E4 E7 E10 E18 E39 E41 E52 E53 E55 E62; P1 P2 P3 P3.1 P4 P7 P14 P14.1 P21 P28 P29 P30 P49
21. Object Entry Information E1 E2 E5 E7 E10 E18 E39 E41 E52 E55 E62; P1 P3 P3.1 P4 P14 P14.1 P20 P21 P28 P29 P30 P49 P50
22. Object Name and Classification Information E1 E7 E13 E15 E17 E32 E41 E42 E55; P2 P37 P41 P42 P71 P140 P141
23. Object Number Information E1 E2 E5 E7 E13 E15 E39 E41 E42 E52; P4 P14 P14.1 P20 P37 P38 P48 P140 P141 P142
24. Object Production Information E2 E4 E5 E7 E11 E12 E24 E39 E52 E53 E55 E57 E70 E77; P4 P7 P11 P12 P14 P14.1 P16 P16.1 P21 P
25. Object Title Information E28 E33 E35 E41 E55 E56 E70 E71 E73; P72 P73 P102 P130 P130.1
26. Part and Component InformationE11 E18 E19 E26 E57 E59 E60 E79 E80; P45 P46 P56 P57 P111 P113
27. Person Nationality Information E4 E21 E44 E45 E46 E47 E48 E53 E55 E67 E74; P2 P7 P10 P87 P98 P107
28. Planned Activities (design, purpose, use)E1 E5 E7 E11 E29 E55 E57 E70 E71 E83; P2 P15 P16 P17 P19 P20 P21 P32 P33 P68 P101 P1
29. Recorder Information (Attribute Assignment) E1 E2 E7 E13 E28 E31 E39 E52 E65 E89 P4 P14 P14.1; P67 P70 P94 P140 P141
30. Reference Information E1 E24 E28 E31 E89; P62 P62.1 P67 P67.1 P70 P129
31. Reproduction Rights Information E1 E30 E39 E55 E62 E70 E72 E89; P2 P3 P3.1 P75 P104 P105 P129
32. Spatial - Temporal Relationship E2 E4 E5 E52 E53 E54; P1 P4 P7 P83 P84 P88 P89 P114 P115 P116 P117 P118 P119 P120 P121 P122
33. Subject Depicted Information E1 E18 E21 E24 E27 E53 E55 E89; P53 P62 P62.1 P67 P67.1 P148
34. Taxonomic Discourse E1 E7 E17 E28 E55 E65 E83; P2 P41 P42 P94 P127 P135 P136 P137 P137.1
35. Time-Span Information E1 E2 E3 E4 E5 E41 E49 E50 E52 E53 E59 E61 E62 E77; P1 P4 P5 P7 P9 P10 P78 P79 P80 P81 P82 P83 P84

Figure 21 CRM Graphical: Index Searching for E36 Visual Item

- The CRM Primer (Dominic Oldman & Donna Kurtz, 2014) presents CRM in brief by presenting the representation of typical museum information.

Most CRM classes fall in the following fundamental divisions (see red lines in the following figure:

E77 Persistent (endurant): whenever it exists, it exists with all its parts simultaneously. This does not preclude changes in time (e.g. part additions/removals)

- **E18 Physical**. Includes physical things such as objects, features (e.g. scratches, marks, inscriptions), collections, and even persons.
- **E28 Conceptual**. Includes ideas, text, images, formulas and other "information" entities that can be easily copied communicated in many different formats, with some variations of physical rendition that still keep them recognizable. Includes information found on museum objects (e.g. inscriptions, text, images) but also museum documentation info such as titles, identifiers, types, languages, etc.
- **E39 Actor**. Please note that E21 Person has two super-classes. If you study the actions of a person, that corresponds to his role as Actor. But if you study his remains, that would be under his role as E19 Physical Thing.

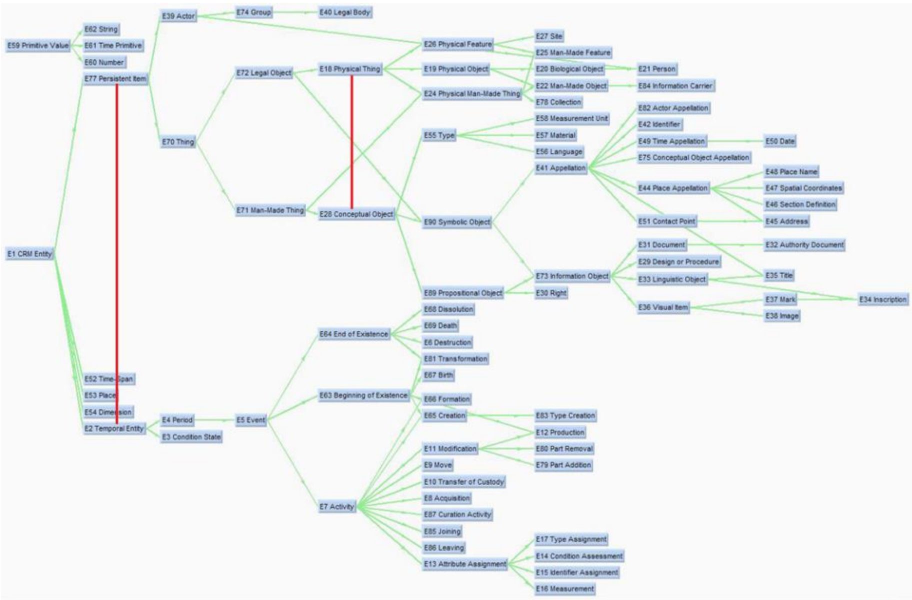


Figure 22 CRM Class Hierarchy

E2 Temporal (perdurant): progresses through time. This includes such large temporal entities like **E4 Period** (a whole cultural period), shorter specific **E5 Events**, and **E7 Activity** (which is caused by an actor). Specific events/activities include:

- **Beginning of Existence**: **Birth** of a person, **Formation** of a group, **Production** of a physical object, **Creation** of a conceptual object.
- **End of Existence**: **Death** of a person, **Dissolution** of a group, **Destruction** of a physical object. Conceptual objects cannot be destroyed, since they exist separately from any and all physical carriers.
- **Transformation**, which is both the End of an old object, and the Beginning of a new one
- **Move, Acquisition** (Transfer of Ownership), **Transfer of Custody**. CRM distinguishes between owner and custodian (keeper/curator).
- **Modification** (of an object), **Part Addition, Part Removal** (of an object or collection); **Joining/Leaving** (of a group)
- Activities related to museum documentation: **Attribute Assignment** and its subclasses. E.g. **Measurement** records the details of how a Dimension was obtained, **Identifier Assignment** records when an identifier or title started to be used (assignment) and stopped to be used (deassignment).

A few classes outside these branches:

- **Primitive Value** and its subclasses are not used in RDF. Instead, appropriate RDF literals are used (e.g. xsd:string, rdf:langString, xsd:decimal, xsd:date, xsd:gYearMonth, xsd:gYear)
- **Place**: can be a place on Earth or on an object, identified through a "Section Definition"

- **Dimension**: some dimension of an object, comprising type, unit and value
- **Time-Span**: temporal info (see below)

CRM Time

Designations like "18TH CENTURY AD", "EARLY 18TH CENTURY" carry some cultural baggage: Gregorian calendar, Christian year-count, Earth locality. Nevertheless, they are not E4 Periods but mere time intervals. Such time intervals are expressed in CRM using E52 Time-Span, which allows fuzzy intervals and comprises:

- A label (e.g. "started circa 1520, finished no later than 1610")
- Duration (minimum, maximum): **P83 had at least duration, P84 had at most duration**
- Up to 4 dates (see below) that are refinements of **P82 at some time within, P81 ongoing throughout**. These define the outer and inner bounds of the interval.

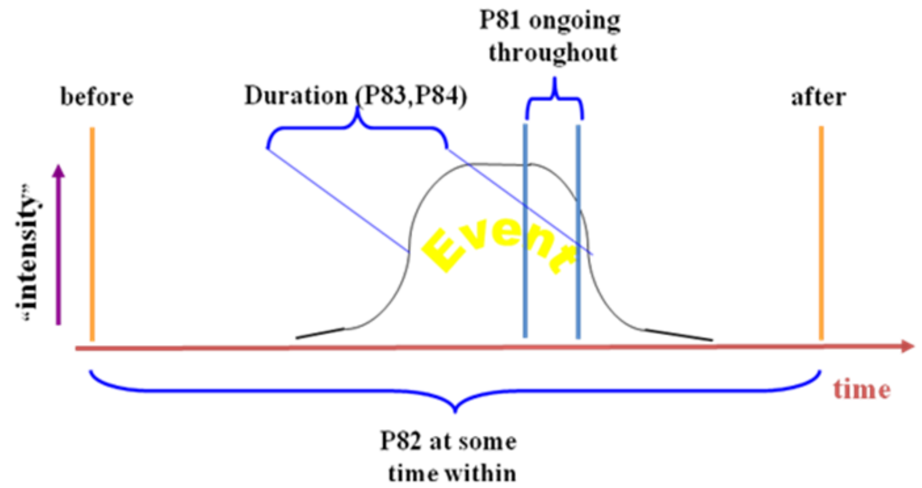


Figure 23 How to Implement CRM Time in RDF

Table 2 CRM Time-Span Bounds

CRM property	Meaning	Latin phrase	Meaning
P82a_begin_of_the_begin	started after this moment	terminus post quem	limit after which
P81a_end_of_the_begin	started before this moment	terminus a quo	limit from which
P81b_begin_of_the_end	finished after this moment	terminus ad quem	limit to which
P82b_end_of_the_end	finished before this moment	terminus ante quem	limit before which

Representing Objects and Features Museum objects are mapped to **E22 Man-Made Object** (or **E19 Physical Object** if they are natural such as a rock). Further distinctions are introduced with **P2_has_type** which points to a thesaurus (**E55 Type** or **skos:Concept**); this is a universal property that applies to any **E1 Entity**. This underlies the universality of CIDOC CRM.

It may be tempting to define more specific classes like Painting or Sculpture. But museums hold all kinds of weird and wonderful things (e.g. the Getty AAT Object hierarchy has 20k concepts). What classes would be needed to describe e.g. a "cake inkjet-printed portrait/sculpture with box" (Sarah Lucas, 2001)?

CRM has sufficient universal constructs to model more specialized domains. E.g. consider Numismatics. Coins use specific dimension types (e.g. die-axis, o'clock) that can be modeled with P2_has_type, referring to a specialized thesaurus (e.g. AAT or BM thesaurus). We need to describe separately the images and inscriptions on the Obverse and Reverse sides of the coin. To model this, consider the CRM Graphical diagram below (double arrows show sub-class and sub-property relations, single arrows are properties). We model Coins as follows:

- E22_Man-Made_Object (with standardized P2_has_type Coin) P56_bears_feature E25_Man-Made_Feature (with standardized P2_has_type Obverse or Reverse). These classes can be related by P56 because they are sub-classes of E19 respectively E26, which are the defined domain & range of P56
- E25_Man-Made_Feature (obverse/reverse) P65_carries_visual_item E38_Image (e.g. of a ruler) or E34_Inscription (some text). These classes can be related by P65 because they are subclasses of E24 respectively E36, which are the defined domain & range of P65.
- E38_Image P138_represents (some ruler, e.g. from ULAN). You can find this relation on graphical diagram Image Information Objects and Carriers
- E34_Inscription P3_has_note "the text" and P72_has_language (some language from a thesaurus, e.g. Latin from AAT). We could also record P73_has_translation to another node (Linguistic Object), e.g. a translation to English

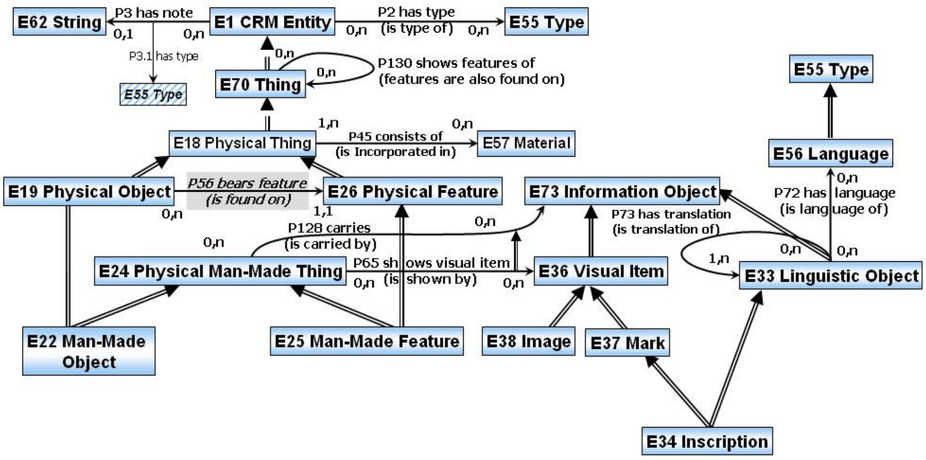


Figure 24 CRM Graphical: Mark and Inscription Information (part 1)

Since Features are considered Things, one can represent these situations (consider the following diagram):

- Represent a wax seal on a parchment, or an ink stamp or signature on a paper document, and use P45_consists_of to designate the material

- Record the specific technique (e.g. incised) or creator of a mark or inscription by using E12 Production or E11 Modification, recording P32_used_general_technique and P14_carried_out_by

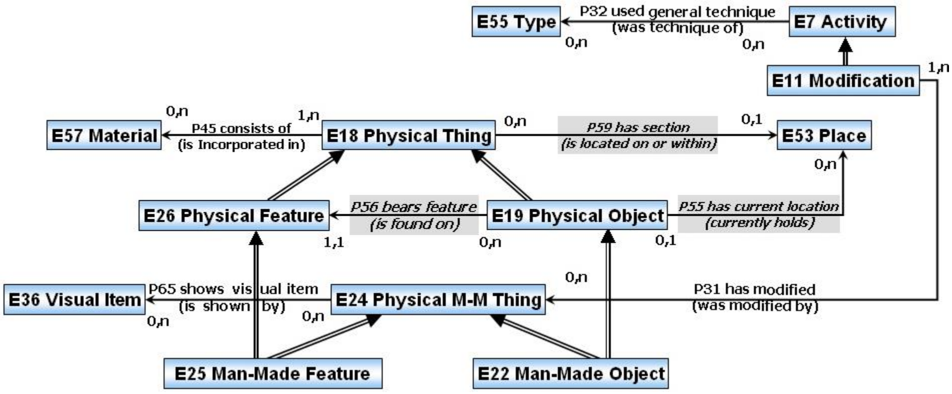


Figure 25 CRM Graphical: Mark and Inscription Information (part 2)

CRM has "part of" relations for various entities (physical object, conceptual object, place, temporal object including event, actor). It has title/ identifier/ image (representation) for objects; who (actor)/ when (time span)/ where (place) for events/activities. CRM includes limited object relations (shows features of, motivation/influence), and it has been criticized for that. CRM is strongly event-oriented. One cannot attach person, place and date information to an object directly: there are no simple properties like "creator", "created on", "created at": one must create Events, e.g. Production. But this allows richer representation of more complex cases, e.g. different kinds of contribution as production sub-events, Attribution Qualifiers (workshop of, circle of, attributed to), etc.

CRM FOL Recent CRM releases include a formalization in First Order Logic (FOL). For example, P46_is_composed_of has the following axiom:

$P46(x,y) \supset (\exists uzw)\, [E93(u) \wedge P166\,(x,u) \wedge E52(z) \wedge P164(u,z) \wedge E93(w) \wedge P166\,(y,w) \wedge P164(w,z) \wedge P10(w,u)]$

Decoding of the terms: E93_Presence, P166_presence_of, E52_Time-Span, P164_restricted_by, P10_falls_within. We can depict the axiom as follows.

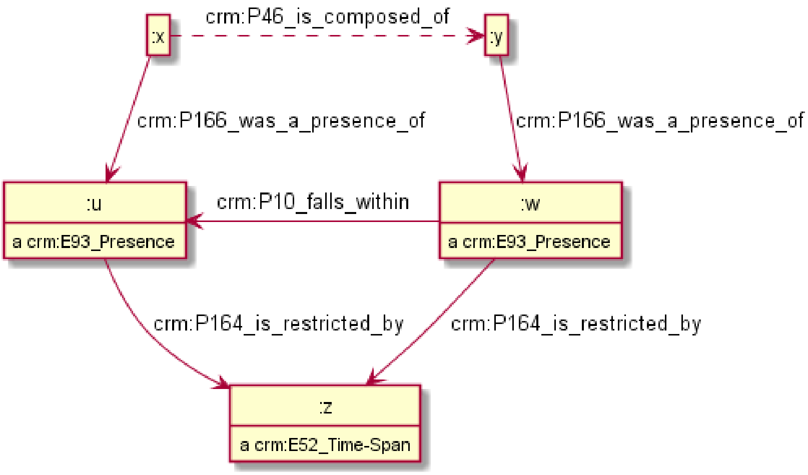


Figure 26 CIDOC CRM First Order Logic diagram

We can read the axiom in two ways:

- If **y** is part of **x** (**x** is composed of **y**) then there must exist presences **u** and **w** (temporal snapshots of E92 Spacetime Volume) considered over the same time-span **z**, such that **w** falls within **u** (i.e. the spatial extent of **y** over the given time-span falls within the spatial extent of **x**)
- If there are two presences **w** and **u** over the same time-span **z**, and corresponding objects **y** and **x**, and if **w** falls within **u**, then **y** must be a part of **x**.

Both of these readings express the notion that for physical objects, if **y** is part of **x** then the spatial extent of **y** always falls within **x**.

CRM Short Cuts and Long Paths The above illustrates an important CRM notion: **short-cuts vs long-paths**. The short-cut information **x-P46-y** could be elaborated in the long-path **x-P166-E93-P10i-E93-P166i**, or the even longer path going through the node **z**. While this example is not very practical, Measurement Information shown below is very practical: while Dimension records the direct info about a Thing, the Measurement node also allows to record extra info about that data: who did the measurement, when, what tools were used, what was the precision, etc. E13_Attribute_Assignment is the prototypical class that participates in long-paths, and activities such as Measurement, Type Assignment are sub-classes thereof.

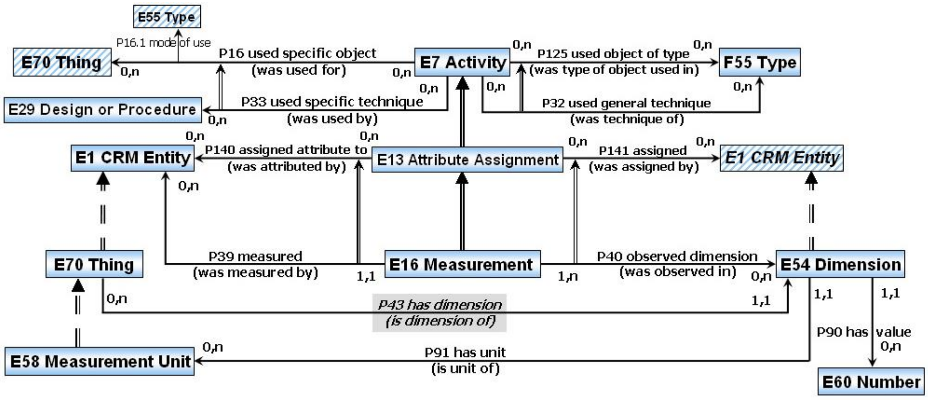


Figure 27 CRM Graphical: Measurement Information

There are several CRM RDF definitions, the two most important being:

- CRM SIG: RDFS. It defines classes, properties (with multiple language translations), and sub-class and sub-property relations.
- Erlangen CRM: OWL-DL. It tracks the official definition and adds inverse and transitive property declarations and class restrictions (owl:Restriction). It is developed on github and full version history is available. Since some of these additions (especially the restrictions) are controversial, I provided a script ecrm-simplify.xq that can generate CRM "application profiles", e.g. leave only the inverse declarations, which are an innate feature of CRM.

CRM Extensions The CIDOC CRM specification, section Modelling principles: Extensions, defines principles how to extend CRM in a compatible way, so that an application that understands only the core ontology, still can consume data conforming to the extension. Basically the guidance is to create extension properties and classes as sub-properties and sub-classes of the core. The following CRM extension ontologies have been developed. See (Martin Dörr, 2018) for an overview:

- FRBRoo: bibliographic information following the FRBR principles (Work-Expression-Manifestation-Item), artistic performances and their recordings
- PRESoo: periodic publications
- DoReMus: music and performances
- CRMdig: digitization processes and provenance metadata
- CRMinf: statements, argumentation, beliefs
- CRMsci: scientific observations
- CRMgeo: spatiotemporal modeling by integrating CRM to GeoSPARQL
- Parthenos Entities: research objects, software, datasets
- CRMeh (English Heritage): archeology
- CRMarchaeo: archeology, excavation, stratigraphy
- CRMba: buildings
- CRMx: proposed extension for museum objects, including simple properties such as main depiction of an object, preferred title, extent, etc

Benefits

- Provides a strong ontological foundation
- Being event-based, it is well suited for representing deeper details, such as separate contributions to an artwork, object parts, etc.
- Used by a large number of (especially European) projects, e.g. UK Claros, UK ResearchSpace, H2020 Gravitate, H2020 Parthenos, etc
- Has extensions in various domains, most importantly archeology and bibliography

Cons

- Somewhat complicated
- Some CRM SIG members are somewhat theoretical, with little regard for practical implementation
- Most collaboration happens in face to face meetings (not so strong electronic collaboration)
- Overly deep class hierarchy with a lot of abstract and not so useful classes
- Strict (monomorphic) domains and ranges, which leads to modeling complications

3.5 UK ResearchSpace (British Museum)

CRM has been used in projects since about 2000 (e.g. CLAROS-Net at Oxford started in 2009). But the first large-scale CRM-based effort was ResearchSpace (RS). It is a Mellon-funded project that started in 2010 and is ongoing. The purpose of the project is to develop a web-based Virtual Research Environment (VRE) where art researchers can collaborate on different projects, import and interlink semantic data, coreference thesauri, use semantic search, annotate data and images, etc. The project is strongly based on CIDOC CRM and has provided CRM consulting and mapping advice in various summer schools and other fora. RS is led by the British Museum and works with both dedicated project staff, and collaboration with external groups. Different groups that the project has worked with include:

- Seme4 (University of Southampton) did the first mapping of BM data to CRM. However, this mapping used too many custom classes and properties, and was not an appropriate CRM extension. For example, it had properties like "gilded at", "gilded by", whereas the proper way is to use the CRM E12_Production event and P2_has_type to describe the specific production process (gilding).
- Ontotext was involved from 2011 to 2014 and implemented the first RS working prototype, mapping of BM data to CRM, semantic search based on Fundamental Relations, data and image annotation using Deep Zoom and SVG for arbitrary annotation shapes (Parvanova, Alexiev, & Kostadinov, 2013).
- ICS FORTH is the main developer of CIDOC CRM and has worked with RS to develop the XML mapping formalism X3ML and the mapping memory manager (3M) tool.
- Delving collaborated on implementing the mapping tool.
- MetaPhacts developed the current version of RS semantic search

RS is also partner in various EU projects, e.g. the H2020 Gravitate project whose aim is to devise semantic representations and create software tools to allow archaeologists and curators to reconstruct shattered or broken CH objects, to identify and re-unify parts of a CH object that has been separated across collections and to recognize associations between CH artefacts.

British Museum Data as CIDOC CRM. As part of the RS project, the British Museum data was mapped to CRM and published semantically. In Oct 2015 the Open Data Institute and NESTA organized the Heritage+Culture Open Data Challenge, and as part of that initiative released a Data Guide and a comparison of CH open datasets. In that comparison, the BM SPARQL Endpoint received a perfect score, for depth of data representation and other indicators.

- The BM SPARQL endpoint was hosted on Ontotext GraphDB (formerly OWLIM) from 2012 until 2017, though using an old version and with no support.
- Currently it is hosted on BlazeGraph using the Metaphactory platform

Mapping documentation (Oldman, Mahmud, & Alexiev, 2013) is very comprehensive but is monolithic and has imprecisions. There is a lot more technical information at the Ontotext ResearchSpace confluence. This model of mapping museum data to CIDOC CRM has been followed by some US museums: Yale Center for British Art (YCBA) and Smithsonian American Art Museum (SAAM).

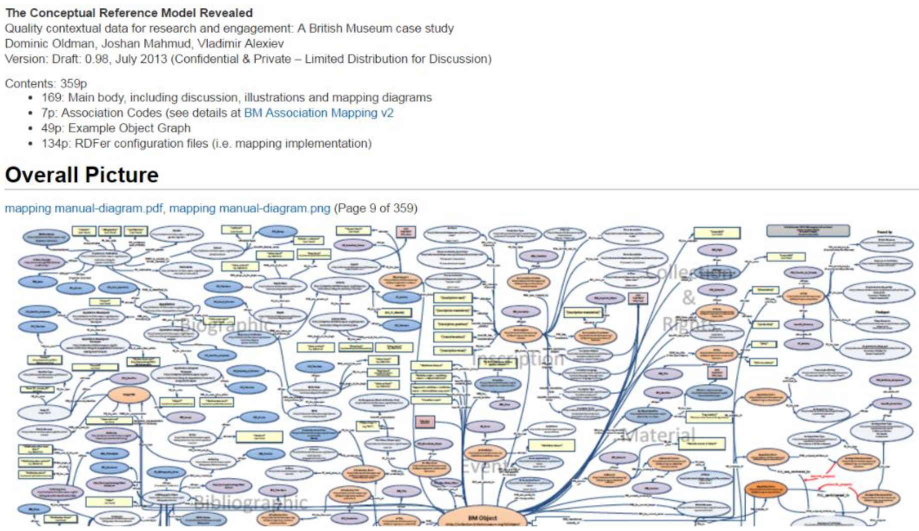


Figure 28 ResearchSpace British Museum Mapping to CIDOC CRM

CIDOC CRM Semantic Search. RS implemented semantic search based on CRM Fundamental Relations (FR). It was based on GraphDB Rules and is an example of large-scale reasoning over CH data (Alexiev, 2012; Alexiev, Manov, Parvanova, & Petrov, 2013), with 4.7x reasoning expansion ratio and 900M statements. FR (Katerina Tzompanaki & Martin Doerr, 2012) is an approach of creating a set of "indexing" relations that abstract over complex CIDOC CRM networks. A number of FRs are defined across 5 types of Fundamental Classes (what, who, where, when).

Table 3 CRM Fundamental Classes and Fundamental Relations

Domain (select)	Range(query parameter)				
	Thing	Actor	Place	Event	Time
Thing	8.has met 9.refers to or is about 10.is referred to by 3.has part 7.is similar or same with 5. from 4.is part of was made from	8.has met 5.from 9.refers to or is about 10.is referred to by 12.by Used by Created by Modified by Found or acquired by	9.refers to 10.is referred to at 5.from Used at Created at Found or acquired at Was created/produced by person from Is/was located at	9.refers to 10.is referred to by 5.from Destroyed in Created in Modified in Used in	5.from Destroyed on Created on Modified on Used on
Actor	8.has met 6.is owner or creator of 9. refers to 10.is referred by	4.is member of 3.has member 8. has met 5.has generator 6.is generator of 9.refers to 10.is referred by	8.has met 5.from 9.refers to 10.is referred to at	9.refers to 10.is referred to by 5.from 8.has met Brought into existence at Taken out of existence at Performed action at Influenced	9.refers to 5.from 8.has met Brought into existence at Taken out of existence at Performed action at Influenced
Place	8.has met 6.Is origin of 9.refers to or is about 10.is referred by	8.has met 6.Is origin of 9.refers to or is about 10.is referred by 8.has met	4.is part of 3.has part 11.borders or overlaps with	9.refers to 10.is referred by 8.has met	5.from 10.refers to 8.has met
Event	6.is origin of 10.is referred by 9.refers to or is about 8.has met created destroyed modified used	12.by 10.is referred by 9.refers to or is about 8.has met brought into existence took out of existence	9.refers to or is about 10. is referred to at 5.from	9.refers to or is about 10.is referred by 3.has part 5.from	9.refers to or is about 5.from starts ends has duration

As an example, the FR Thing From Place codifies the notion that a Thing may have its origin at Place if the thing (or a part of it) was used for an important activity at place, or was created at place, or was made by someone born at place, or who had its residence at place, etc.

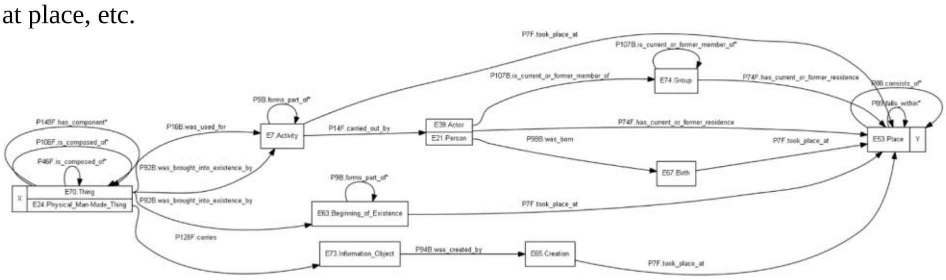


Figure 29 CRM Fundamental Relation: Thing From Place

The first version of RS semantic search implemented 23 FRs, all of them about Thing (e.g. the above one is called rso:FR7_from_place). It also included several semantic hierarchical facets: object type, creator, place, date created, etc. It implemented a natural-language-like interface for defining the query. It employs query expansion across hierarchical thesauri, e.g. searching for "Mammal" finds drawings of horses and pigs.

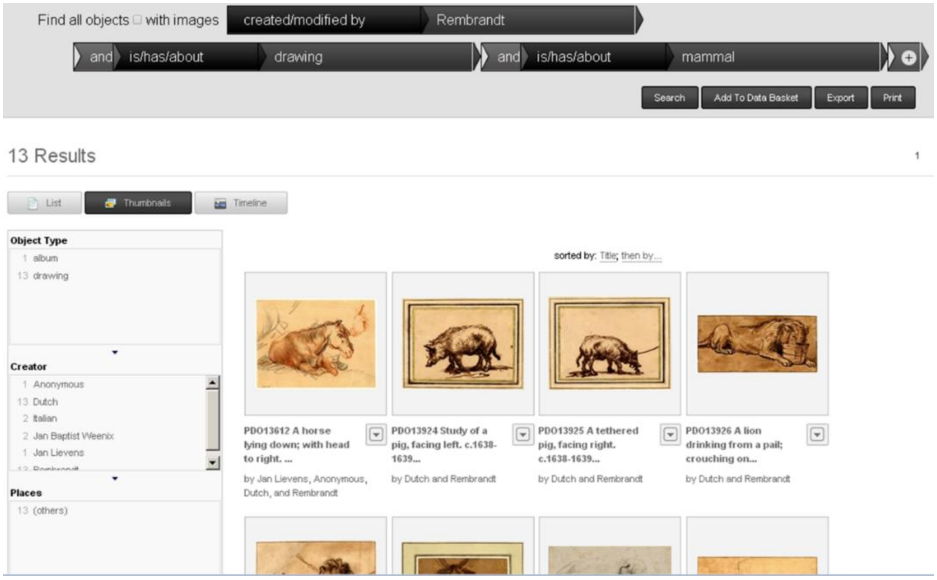


Figure 30 RS First Semantic Search: Hierarchical Query Expansion

The current version of RS semantic search implements a lot more FRs, "stored queries", ability to join against such queries, and a nicer user interface.

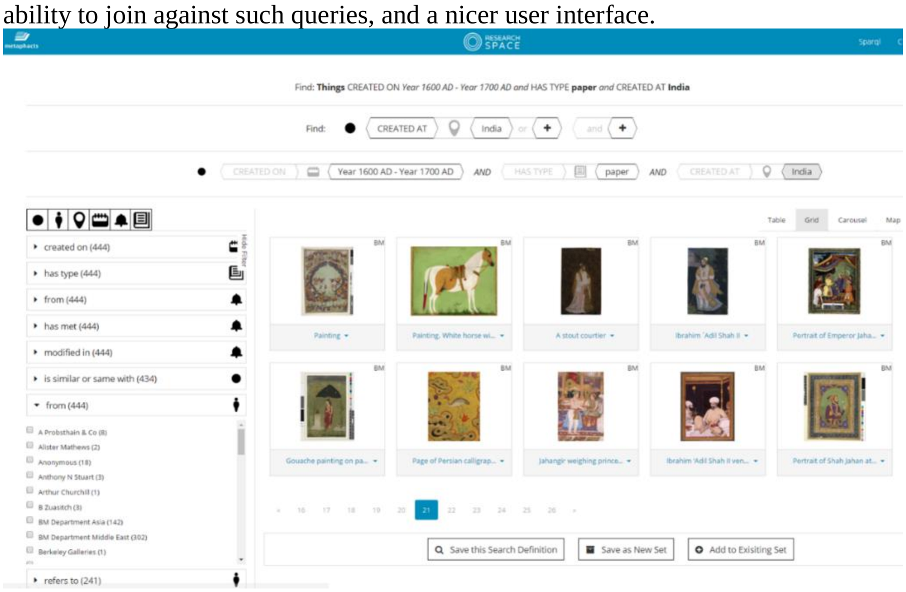


Figure 31 RS Current Semantic Search: More Relations, Saved Queries, Better UI

Pros: RS has pioneered several novel approaches in CH: CIDOC CRM representation, powerful semantic search, image annotation, saved searches, data basket, etc. It intends to be a generic art research system that can be adapted for various needs and projects.

Cons: RS still has very few production users to use the system on a daily basis.

3.6 US ConservationSpace, Sirma MuseumSpace

Like RS, ConservationSpace is a Mellon-funded project that started in 2009 and spent about 3 years defining requirements, creating UI mockups, designs and RFP documentation (see old project site). The project is led by the US National Gallery of Art and includes a strong consortium. The goal is to create a system for conservation specialists, including tasks such as object examination, image annotation, process and work flows, intelligent documents, etc

Development started in 2013 and a production system was completed in 2015-16. Noblis developed functional requirements, Design for Context facilitated requirements workshops and created UI mockups. Several Bulgarian companies were involved: Sirma Enterprise developed the system, Lucrat conducted usability testing, and Ontotext provided semantic database and semantic consulting. ConservationSpace is now in production in all partner institutions, several other deployments are in progress, and it is being adopted for two MS programs in conservation.

ConservationSpace features include:

- The ability to import data from collection and digital asset management systems
- Storing data in a semantic database (Ontotext GraphDB)
- Generating user interfaces from ontologies and declarative descriptions
- Flexible access control and user rights model
- Cloud-based deployment (Software as a Service) with a full multi-tenant model (each tenant institution operates completely independently from the others)
- Capabilities that facilitate both enterprise-level and user-level customization of system object templates and code lists
- Role-based security management controls, specific to each institution’s standards
- System object security controls permitting controlled access to sensitive documentation or data
- Adoption of image annotation standards in conformance with established protocols such as the International Image Interoperability Framework (IIIF)
- Extended Mirador viewer for working with images
- Dashboard customization capabilities for individual users
- Full workflow management capabilities to support the unique business processes of each institution
- Capabilities to support the use of locally preferred terminology by institutions
- Version management and rollback capabilities for key system objects
- Cultural and digital object record management and search/retrieval independent from the project/case/task system object hierarchy
- Reports on system status and activity
- Intelligent Documents (iDoc) which incorporate data entry forms and can query information from the system.
- Ability to print and export iDoc-based documents

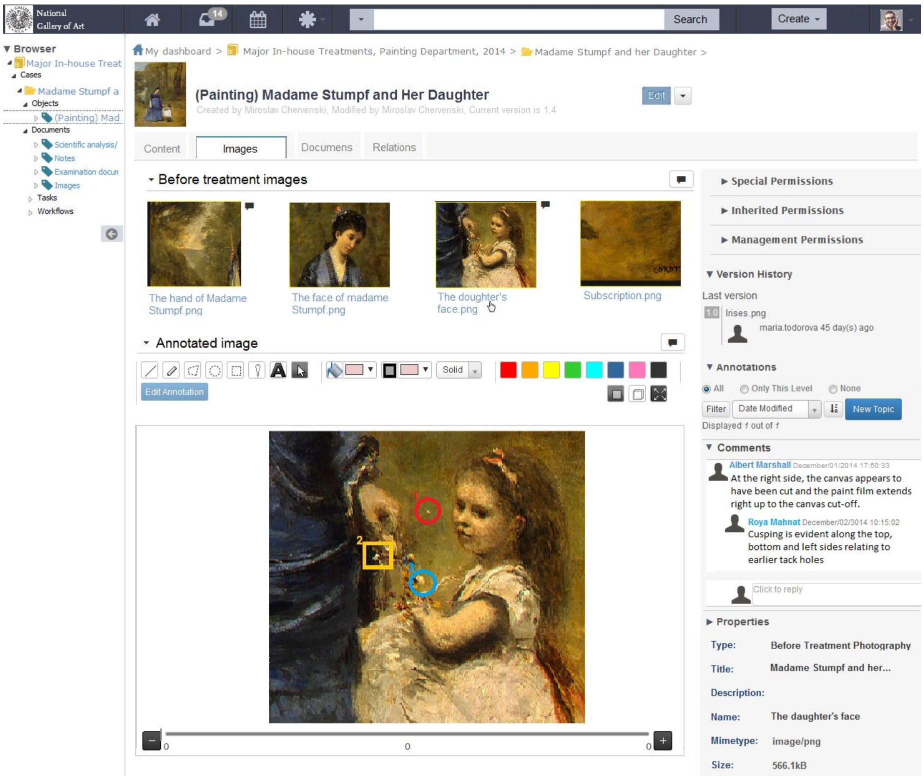


Figure 32 ConservationSpace Painting Examination

ConservationSpace spent significant time on user workshops and requirements definition, ensuring the applicability and longevity of the project. Several institutions use the system in production, and there is a thriving user community. Deployment for a new institution involves defining specific objects, work-flows and customizations, but little programming.

ConservationSpace is based on the **Sirma Enterprise Platform**, a flexible software solution that includes semantic data modelling, process management, work flow, (BPMN process definition), collaboration (contextual comments, email notifications, etc). It was deployed in a variety of domains, such as:

- CH: Sirma MuseumSpace (to be demonstrated at DiPP 2018) includes modules for curation/collection management, exhibition and loan management, conservation management, etc.
- Semantic integration, enrichment and publication of CH data
- Digital Asset Management
- Thesaurus Management
- Paper-less office (Sirma GO Digital)
- Contract management

- ISO 9001 QMS document management, with document workflows and records management

3.7 US AAC (American Art Collaborative) and linked.art

The publication of semantic data by the BM, YCBA and SAAM generated enough interest, so the American Art Collaborative (AAC, http://americanartcollaborative.org) was established as a 2-year project (from Oct 2015 to Nov 2017) with Mellon Foundation funding. 14 US museums and galleries participate in this collaboration to publish their data in RDF. Although the Getty Trust is not formally affiliated with AAC, it had a crucial role, as the project was started by the former founder of the Getty Vocabulary Program, and two Getty staff (the semantic architect and data architect) had core involvement in developing the data model.

A lot of the technical work was done by external consultants: data conversion mostly by USC ISI students using the ISI Karma tool. Design for Context created UI mockups and implemented the Browse and Mapping/Review apps. Vladimir Alexiev and Stephen Stead provided CIDOC CRM mapping advice. Vladimir Alexiev provided detailed bug reports and semantic data publishing advice. Towards the end of the project a couple of the institutions took charge of their transformations, aiming to establish their own sustainable RDF publication.

The project did a lot of its work in the open (http://github.com/american-art/): this site has 26 repositories with common tools, data and issues (e.g. aac-alignment, aac_mappings, AAT-Term-Mappings, Semantic-UI, AAC-Instructions, linking, semantic-hosting, pubby) per-museum repositories (e.g. **DMA** for Dallas Museum of Art, **npg** for National Portrait Gallery) with source data, Karma mapping models and converted data. Semantic resources at http://data.americanartcollaborative.org/:

- Per-museum data, e.g. http://data.americanartcollaborative.org/npg
- Per-agent data, e.g. http://data.americanartcollaborative.org/npg/person-institution/44424
- Per-object data, e.g. http://data.americanartcollaborative.org/npg/object/29
- SPARQL endpoint http://data.americanartcollaborative.org/sparql. There is no editor there, so it's best to use YASGUI. E.g. http://yasgui.org/short/H1u89bnJG is a query that returns NPG objects and their images

Mapping/review app: http://review.americanartcollaborative.org/. This tool uses Ontotext's rdfpuml visualization tool (Vladimir Alexiev, 2016) and is used to both define the desired mapping and check certain semantic URLs for conformance. E.g. here is the mapping for Actor Gender

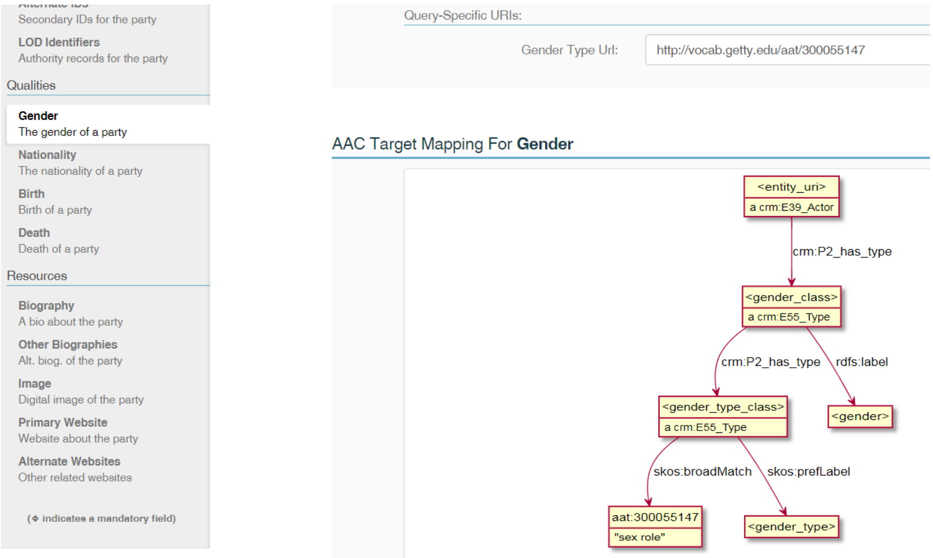


Figure 33 AAC Target Mapping for Gender

The web browse app http://browse.americanartcollaborative.org shows an overview of aggregated collections, simple full-text search, individual object pages, artist pages, and statistics about number of objects per artist across collections.

(Craig Knoblock et al., 2017) describe project challenges, volumetrics and semantic conversion experience. (Fink, 2018) describes lessons learned and an overview of good practices.

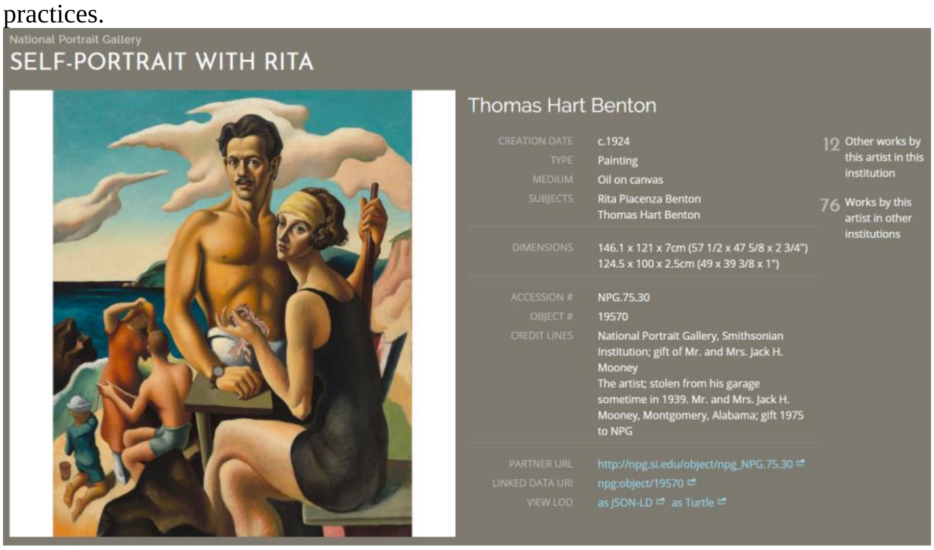


Figure 34 AAC Artwork View

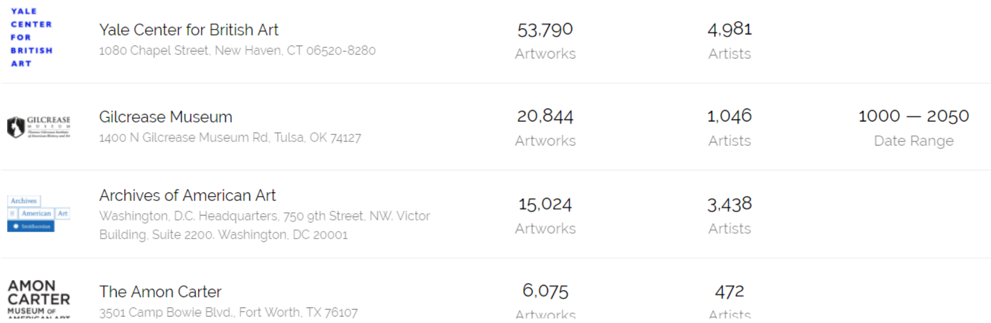


Figure 35 AAC Browse App: Partner Institutions Statistics

Pros

- The project aggregated artwork data from 14 institutions: 233,666 Objects, 28,882 Artists and 20,446 other agents (Related Parties), comprising about 15M triples. (For comparison, the British Museum semantic data comprises 2.5M objects and 960M triples.)
- Used a harmonized data model so the data can be shown together.
- Harmonized not only data models but also value sets. AAC standardized on using Getty AAT concepts for "business classification" of various aspects as the value of crm:P2_has_type, e.g. http://vocab.getty.edu/aat/300055147 for "Gender". Furthermore, an USC ISI tool was used successfully by the institutions for linking artists to ULAN (though later a comparison to Wikidata Mix-n-Match showed that tool could have been used to better effect).
- Raised LOD awareness with the target institutions and a wider audience and mobilized inter-institutional collaboration.
- Towards the end of the project a lot of IT people and data curators from the institutions became deeply involved in the details of the semantic representation. Some of the institutions took charge of their transformations to establish a sustainable LOD publication process.
- The project created excellent use cases and UI mockups for browsing and exploration, e.g. comparing artists by style, material and genres; artwork timelines, etc.

Cons

- The project started mapping without having a proper mapping specification. As a result, some mappings were reworked up to 6 times (Craig Knoblock et al., 2017). A lot of bugs were filed (total 592 issues). A lot of these are still open (107 open issues as of Mar 2018). Some of the issues were postponed for a future version, and then closed without being implemented, i.e. dismissed (e.g. mapping Exhibitions). Many issues were replicated between the different institutions, so had to be posted and fixed several times. Perhaps the most important lesson learned was that one should not attempt a massive mapping effort without having an agreed data model and strong mapping specifications (prototypical mappings): bug reports are no substitute for a proper specification.
- Some data submitted by the institutions was left unmapped and therefore not published semantically (e.g. Exhibitions, Publications/bibliographic info, Videos, etc).

A lot of the use cases and mockups could not be implemented because of data omissions or insufficient harmonization of the data.

- Various details were glossed over, e.g. the Actor Image mapping disregards the SAAM flag PrimaryDisplay. This means that when an artist has many images (e.g. a photo and a self-portrait), a random one needs to be selected to display just one image (e.g. in search results). But even the old SAAM mapping had that, e.g. see Ivan Albright at SAAM: it has two links P138i_has_representation but only one of them is PE_has_main_representation.
- The mapping specification omits important details, such as URL patterns. As a result, many conversion implementers (ISI students) have made mistakes.
- Since the adopted data model (linked.art) was derived post-factum, various problems still remain.

E.g. regarding Title Types I have posted the following github issues (aac_mappings/48 and cbm/58):

- an object may have several titles of the same type, in which case their labels get mixed together
- all title types of an object are mixed together
- commonality of title types across objects is not captured
- there is no relation from title type to AAT (whereas now AAT has related concepts such as "Group Title")
- The use of aat:300404670 "preferred name" is wrong, e.g. for Group Title
- Use crm:P48_has_preferred_identifier instead of crm:P1_is_identified_by for the title id
- No need to use aat:300404012 Unique Identifier for the title id
- The title type mixes SKOS and CRM in an undisciplined way
- DisplayOrder of titles is not captured
- Group Title reflects a collection of objects so it should be modelled as crm:E78_Collection

http://linked.art emerged from the AAC effort as an application profile for CRM, i.e. a particular way of using CRM. It was created out of frustration with the complications of applying CRM (Robert Sanderson, 2016) and is promoted under the moniker Linked Open **Usable** Data (LOUD). linked.art steps on the following principles:

- CIDOC-CRM as the core ontology, giving an event-based paradigm
- The Getty Vocabularies (see next) as core sources of identity, i.e. specific object types (e.g. painting), activity types (e.g. book binding, gilding, etching), title types (e.g. artists vs repository title), etc etc
- JSON-LD as the primary RDF serialization. Being JSON, it is more developer-friendly than other serializations.

linked.art makes a number of simplifying assumptions, defining "a streamlined profile of CRM for better consistency and comprehension". Its CRM Class Analysis uses 28 of the CRM classes, dismisses about 60 classes (under headings Overly Abstract, Overly Specific, Datatypes, Ineffective, Unnecessary, Incomprehensible), and introduces 7 new classes. It similarly dismisses a number of CRM's properties.

One of the most useful features of linked.art is the large number of examples (model components) that guide the semantic representation of museum data. The count of examples (Aug 2018) per area is: 42 activity, 1 concept, 2 group, 2 identifier, 2 legal, 1 name, 46 object, 12 person, 6 place, 7 set, 11 text, 2 value. E.g. below is one such example.

```json
"@context": "https://linked.art/ns/v1/linked-art.json",
"id": "https://linked.art/example/activity/15",
"type": "Activity",
"label": "Example Travelling Exhibition at Two Museums",
"classified_as": ["aat:300054773"],
"timespan": {
  "id": "https://linked.art/example/time/9",
  "type": "TimeSpan",
  "begin_of_the_begin": "1980-10-01",
  "end_of_the_end": "1981-08-14"
},
"part": [
  {
    "id": "https://linked.art/example/activity/16",
    "type": "Activity",
    "label": "Exhibition at Museum 1",
    "classified_as": ["aat:300054766"],
    "timespan": {
      "id": "https://linked.art/example/time/10",
      "type": "TimeSpan",
      "begin_of_the_begin": "1980-10-01",
      "end_of_the_end": "1981-03-01"
    }
  },
  {
    "id": "https://linked.art/example/activity/17",
    "type": "Activity",
    "label": "Exhibition at Museum 2",
    "classified_as": ["aat:300054766"],
    "timespan": {
      "id": "https://linked.art/example/time/11",
      "type": "TimeSpan",
      "begin_of_the_begin": "1981-03-14",
      "end_of_the_end": "1981-08-14"
```

```turtle
@base          <https://linked.art/example/>.
@prefix aat:     <http://vocab.getty.edu/aat/> .
@prefix aat:     <http://vocab.getty.edu/aat/>.
@prefix crm:     <http://www.cidoc-crm.org/cidoc-crm/> .
@prefix dc:      <http://purl.org/dc/elements/1.1/> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix foaf:    <http://xmlns.com/foaf/0.1/> .
@prefix la:      <https://linked.art/ns/terms/> .
@prefix rdf:     <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
@prefix rdfs:    <http://www.w3.org/2000/01/rdf-schema#> .
@prefix schema:  <http://schema.org/> .
@prefix skos:    <http://www.w3.org/2004/02/skos/core#> .
@prefix xml:     <http://www.w3.org/XML/1998/namespace> .
@prefix xsd:     <http://www.w3.org/2001/XMLSchema#> .

<activity/15> a crm:E7_Activity ;
    rdfs:label "Example Travelling Exhibition at Two Museums" ;
    crm:P2_has_type aat:300054773 ;
    crm:P4_has_time-span <time/9> ;
    crm:P9_consists_of <activity/16>, <activity/17> .

<activity/16> a crm:E7_Activity ;
    rdfs:label "Exhibition at Museum 1" ;
    crm:P2_has_type aat:300054766 ;
    crm:P4_has_time-span <time/10> .

<activity/17> a crm:E7_Activity ;
    rdfs:label "Exhibition at Museum 2" ;
    crm:P2_has_type aat:300054766 ;
    crm:P4_has_time-span <time/11> .

<time/10> a crm:E52_Time-Span ;
    crm:P82a_begin_of_the_begin "1980-10-01" ;
    crm:P82b_end_of_the_end "1981-03-01" .

<time/11> a crm:E52_Time-Span ;
    crm:P82a_begin_of_the_begin "1981-03-14" ;
    crm:P82b_end_of_the_end "1981-08-14" .

<time/9> a crm:E52_Time-Span ;
    crm:P82a_begin_of_the_begin "1980-10-01" ;
    crm:P82b_end_of_the_end "1981-08-14" .
```

Figure 36 linked.art Traveling Exhibition: JSON-LD (left) and Turtle (right)

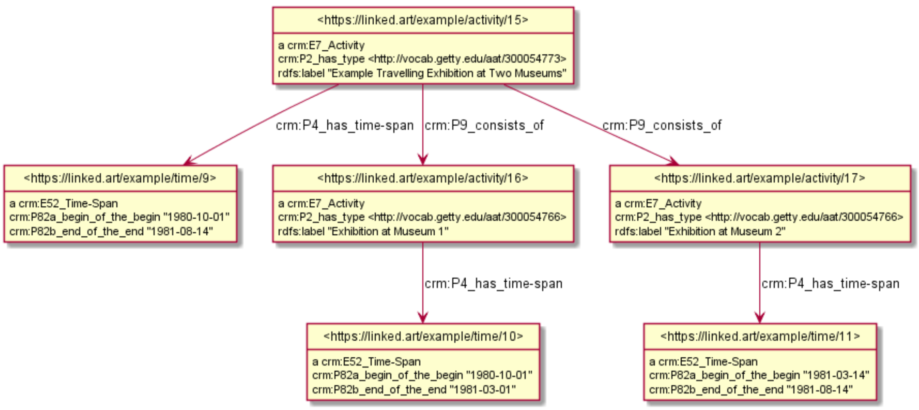


Figure 37 linked.art Representation of Traveling Exhibition: rdfpuml Diagram

Pros: linked.art is used in a number of US-based projects: AAC (post-factum), Getty Provenance Index, Getty Museum data mapping (upcoming), Pharos.net photographic consortium.

Cons: the linked.art simplifications are controversial and have not been accepted by the "mainstream" CRM SIG. Therefore, it creates a rift in the CRM community: European projects using full CRM, and US projects using linked.art.

3.8 US GVP (Getty Vocabulary Program)

The Getty Research Institute (part of the Getty Trust) manages the Getty Vocabulary Program (GVP), which publishes some of the core and most respected CH thesauri. They sit in the center of a CH LOD diagram:

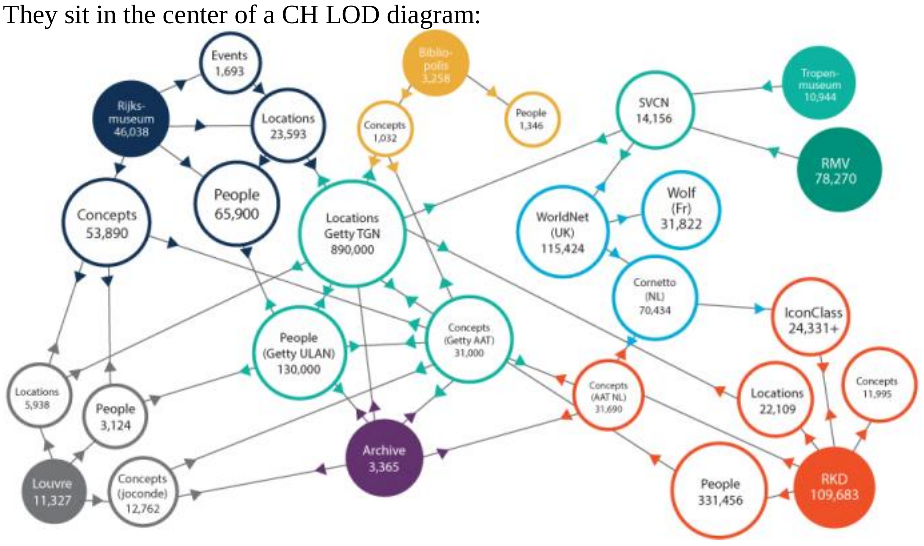


Figure 38 CH LOD Diagram (after M.Hildebrand, 2012)

Getty's vision for the GVP thesauri includes the following elements:

- The thesauri are interconnected with each other. For example, TGN uses AAT for place types, ULAN uses AAT for artist types (roles) and event types, and TGN for places of birth, death, etc.
- The thesauri provide shared data for Getty's own databases and systems: Arches (see next section), Provenance Index, Getty Museum. Getty site-wide web search, AATA Online (bibliography of art and architecture).
- AAT is translated internationally to Dutch, Spanish, German, Chinese, and work on a Swiss Art and Architecture thesaurus is pending. The International Terminology Working Group (ITWG) coordinates this work.
- The thesauri are coreferenced to other relevant thesauri, for example LCSH and VIAF (ULAN is completely incorporated in VIAF, though with a narrower scope of data, e.g. VIAF doesn't include artist relations).

- The thesauri are used by various external databases, projects, search engines and Collection Management Systems.

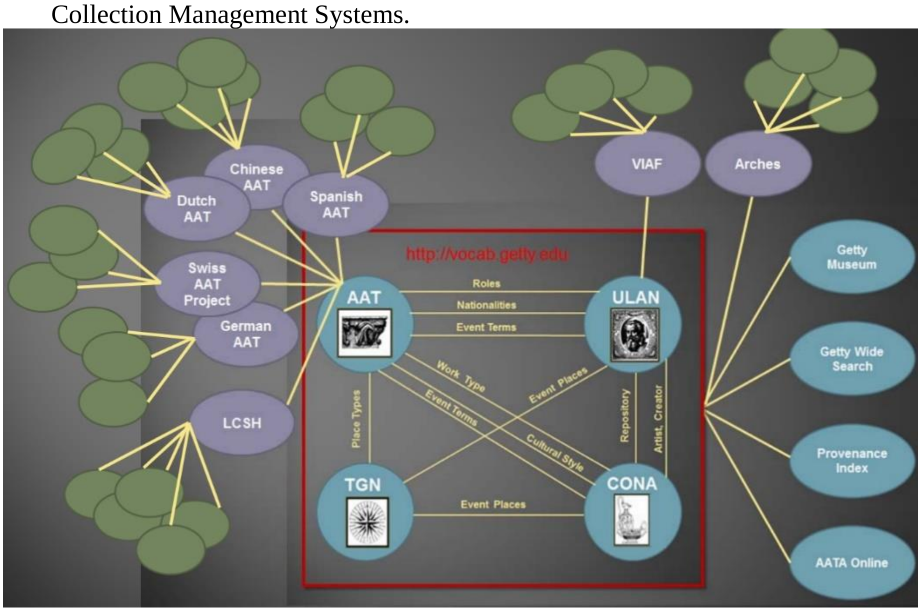


Figure 39 Getty Vocabulary Dependencies and Collaborations (J.Cobb, 2014)

GVP started a LOD publication program in 2013. To date it has published the following thesauri as LOD at http://vocab.getty.edu, sharing the same basic semantic representation, and publicized in blog posts by J.Cuno, CEO of the Getty Trust.

- Art and Architecture Thesaurus (AAT): 2014-02
- Thesaurus of Geographic Names (TGN): 2014-08
- Union List of Artist Names (ULAN): 2015-03

GVP LOD was presented at the CIDOC Congress in Dresden in 2014 (Vladimir Alexiev, 2014). GVP Training Materials include more LOD presentations. Ontotext provided the following services as part of this project:

- Semantic/ontology development (Alexiev, 2015b)
- Contributed to the ISO 25964 ontology, which is the latest standard on thesauri. Provided implementation experience, suggestions and fixes. Published on varieties of Broader relations (Vladimir Alexiev, Jutta Lindenthal, & Antoine Isaac, 2015)
- Complete mapping specification. Helped implement R2RML scripts working off Getty's Oracle database, contribution to Perl implementation (RDB2RDF), R2RML extension (rrx:languageColumn)
- Worked with a wide External Reviewers group (people from OCLC, Europeana, ISO 25964 working group, etc.)
- GraphDB semantic repository, clustered for high-availability
- Semantic application development (customized user interface), technical consulting

- SPARQL 1.1 compliant endpoint, comprehensive documentation (Vladimir Alexiev, Joan Cobb, Gregg Garcia, & Patricia Harpring, 2015), sample queries (Alexiev, 2015a). Per-entity export files, explicit/total data dumps.
- Help desk / support on twitter and google group (continuing until now)

GVP Ontologies: GVP LOD uses SKOS and a number of additional ontologies to represent all data present in the thesauri:

- bibo: Bibliography Ontology: representation of Sources
- bio: Biography Ontology: representation of ULAN person events
- dc: Dublin Core Elements: common data
- dct: Dublin Core Terms: common data (e.g. dct:modified)
- foaf: Friend of a Friend ontology: representation of Contributors
- iso: ISO 25946 ontology: ThesaurusArray, BTG, BTP, BTI
- luc: Ontotext GraphDB's built-in Lucene indexing: Full Text Search
- ontogeo: Ontotext GraphDB geo-spatial extensions, e.g. Places Within Bounding Box, Places Near Each Other
- prov: Provenance Ontology: revision history
- ptop: Ontotext PROTON ontology: used in Extended Property Constructs
- schema: Schema.org common properties
- sesame: rdf4j (Sesame): special predicate directSubPropertyOf
- skos: Simple Knowledge Organization System: basic thesaurus data
- skosxl: SKOS Extension for Labels: to record label sources, contributors, changes
- wgs: W3C Geo ontology: TGN place coordinates

In addition, it features the GVP LOD ontology that has 10 classes and 177 properties. The ontology is documented with Parrot and registered in Linked Open Vocabularies to facilitate discovery.

Figure 40 GVP LOD Ontology

The GVP ontology captures specific Getty classes and properties that are not available in SKOS, SKOS-XL and ISO 25964. Nevertheless, it maps to these established ontologies, so one can also consume the data using only these ontologies.

- Includes these specific node types: gvp:Facet, gvp:Hierarchy, gvp:GuideTerm, gvp:Concept, gvp:ObsoleteSubject. These are implemented as subclasses of skos:Concept, skos:Collection, iso:ThesaurusArray.
- Most of the properties are GVP Associative Relations, defined as sub-properties of skos:related. These were described by GVP domain experts in Excel, and we generated the ontological definitions from that.
- The inference from GVP custom properties to standard properties is shown below (blue=standard relation, black=GVP relation, bold=transitive closure, red=restriction)

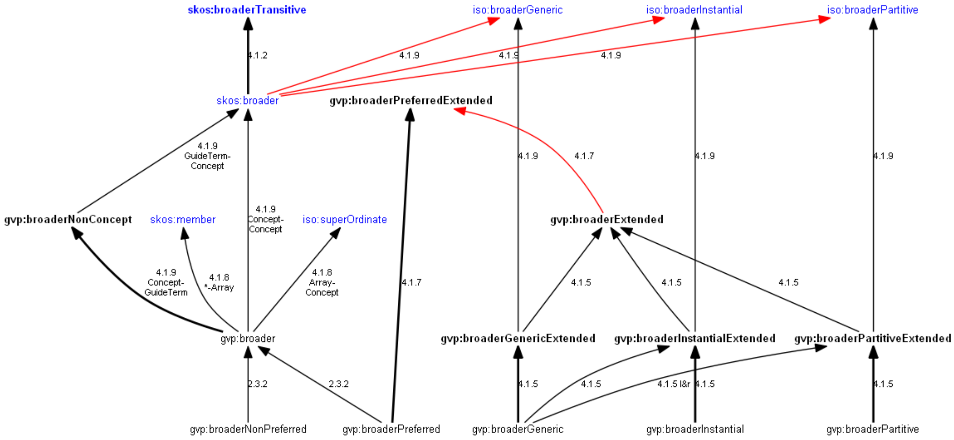


Figure 41 GVP Hierarchical Relation Inference

GVP uses ISO 25964 for attaching non-concepts (the so-called Guide Terms) under concepts, and representing them as iso:ThesaurusArray. In addition, these are represented as skos:OrderedCollection since often the child order is significant.

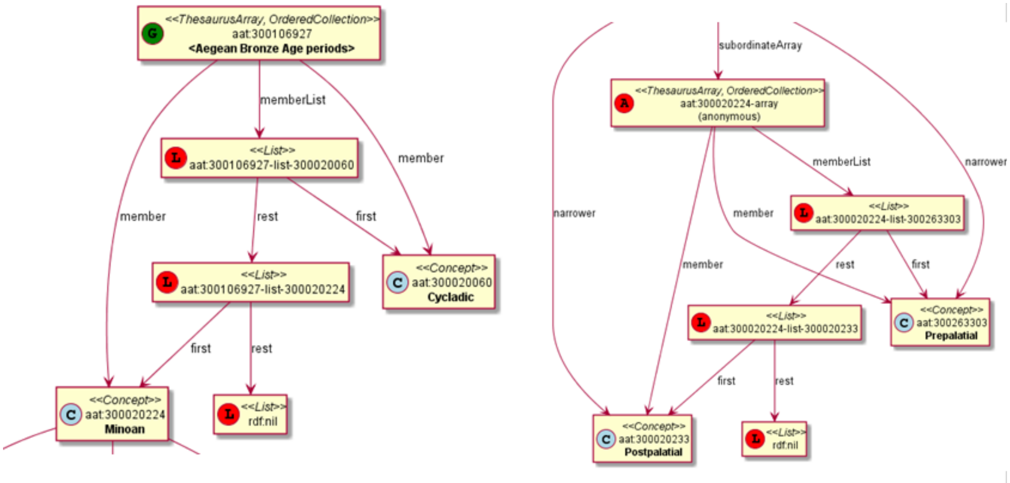


Figure 42 Example of GVP Ordered Guide Term, represented as iso:ThesaurusArray

Documentation, Sample Queries, Support. GVP LOD has set best-practice standards for good quality CH LOD semantic publishing.

- Comprehensive documentation (100 pages) that describes all aspects of representation, semantic resolution, URLs, content negotiation, It is kept up to date, with complete revision notes.

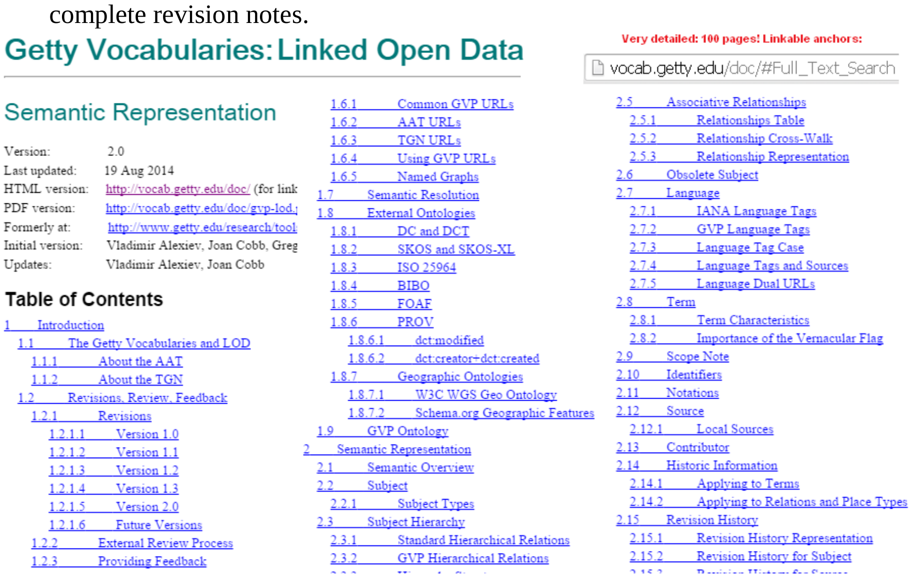


Figure 43 GVP LOD Documentation: Table of Contents

- There are about 100 sample queries, covering topics such as Full-Text Search (many external systems use GVP LOD for auto-completion), getting various kinds of information, TGN and ULAN specific queries (e.g. by geographic proximity), language-related queries, making graphs and charts, etc. There is a special Sample Queries UI that shows the outline of queries (TOC), the description of each query, and allows the user to easily select and execute the query.

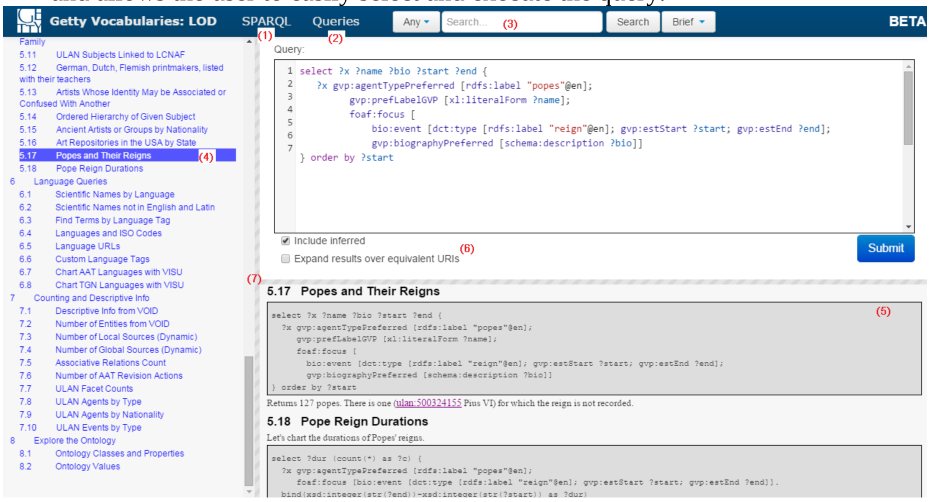


Figure 44 GVP LOD Sample Queries UI

- The GVP UI includes other convenient features, such as full-text search, exploring data, download in a variety of semantic formats (RDF/XML, NTriples, RDF/JSON, Turtle, JSONLD), bidirectional links between LOD and the traditional website. There is a community support group that is monitored regularly, questions are answered, additional queries are added, and issues are resolved.
- GVP has a comprehensive URL strategy that covers all objects and sub-objects. The stability and permanence of URLs is guaranteed by Getty and doesn't change over time (e.g. with new versions), which is extremely important for the consumers of this data (CH institutions that embed GVP thesaurus references in their own data). Obsolete concepts are not deleted for 5 years, rather they are marked as obsoleted, with potentially a dct:isReplacedBy link to the new concept.
- The various semantic formats can be downloaded by extension or through content negotiation (Accept header with appropriate MIME type). All URLs have proper semantic resolution, which was validated with Vapour. GVP provides per-entity download, which includes not just the immediate triples but all nodes and triples of the "business object" (see next figure). In addition, complete downloads (dumps) per thesaurus are available.

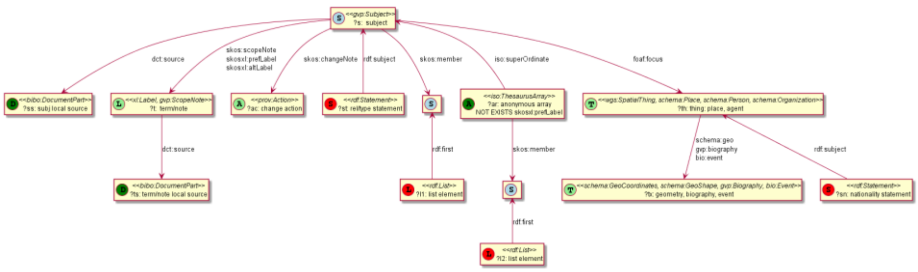


Figure 45 GVP CONSTRUCT Query Returning Complete Business Object

- Dataset Description: GVP LOD uses a number of external ontologies for machine-readable description of the dataset, SPARQL endpoint, preferred prefix, used vocabularies, number of triples per property, number of entities per class, etc.

Table 4 Ontologies Used for GVP Dataset Semantic Description

Prefix	Ontology	Used for
adms:	Asset Description Metadata Schema	Dataset description
cc:	Creative Commons Rights Expressions	License rights
dcat:	Data Catalog Vocabulary	Dataset description
dctype:	DCMI Type Vocabulary	Dataset class
fmt:	RDF formats used in datasets	Formats of data dumps
sd:	SPARQL Service Description	SPARQL endpoint capabilities
vaem:	Vocabulary Attaching Essential Metadata	Not used yet
vann:	Vocabulary for annotating vocabularies	Preferred namespaces and prefixes
vcard:	vCard (contact info)	Contact info
vdpp:	Vocabulary for Dataset Publication Projects	Not used yet
voaf:	Vocabulary of a Friend	Linked Open Vocabularies (LOV)
voag:	Vocabulary Of Attribution and Governance	Frequency of publication
void:	Vocabulary of Interlinked Datasets	Basis description, LOD registration
wdrs:	Protocol for Web Description Resources	"Described by" from dataset to doc
wv:	A vocabulary for waivers of rights	License rights

We used several descriptive ontologies to cater to different kinds of software agents, enabling dataset discovery and crawling. For example, the datasets of each vocabulary are declared void:Dataset, dct:Dataset, dcat:Dataset, adms:Asset, cc:Work, dct:Collection. There is good agreement between the conceptual models of the main descriptive ontologies (VOID, DCAT, ADMS), which makes this possible. Complete licensing info, keywords, subjects, crawling entry points (void:rootResource) are described.

GVP LOD Uses: GVP LOD has found a wide variety of uses in the CH community: over 50 actual and potential uses. The thesauri are used by many CH institutions (including the Google Cultural Institute) and CH-related software (including Gallery Systems TMS, which is widely used in the US). The reliability of the GVP SPARQL endpoint is such that many use the thesauri directly, without a need to copy them locally.

Below are a couple of figures illustrating such use. Above we saw that AAT is used crucially in the linked.art semantic profile, to describe specific semantic types (e.g. painting, gilding, author's title, etc)

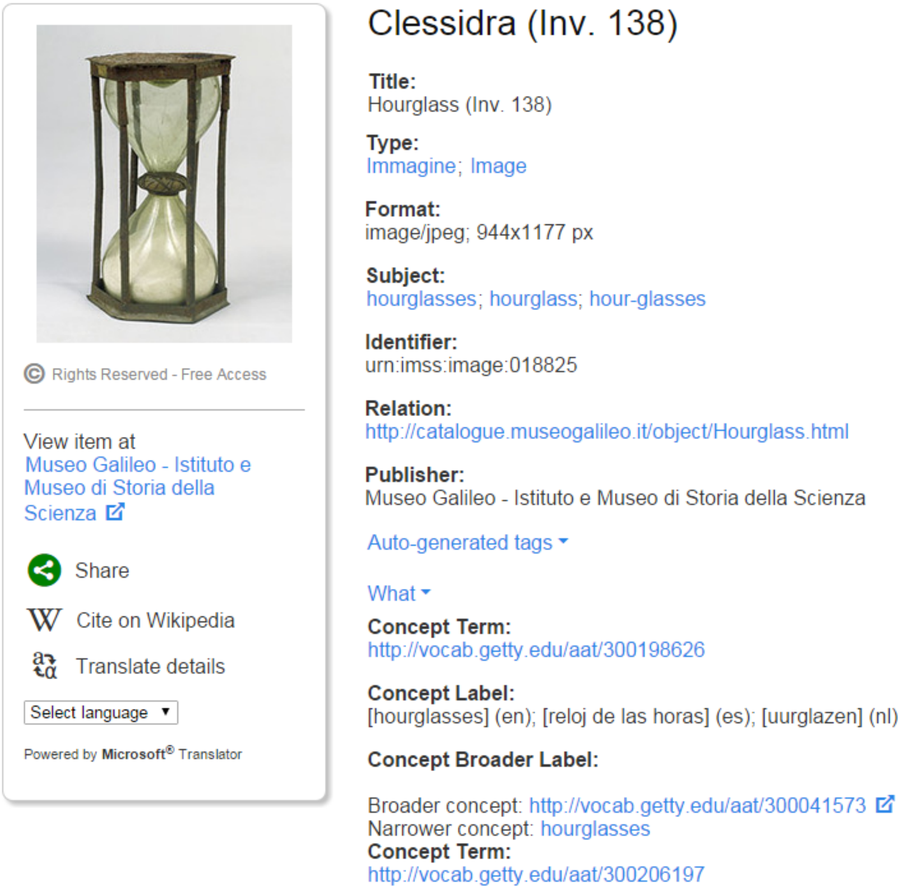


Figure 46 GVP Use in Europeana

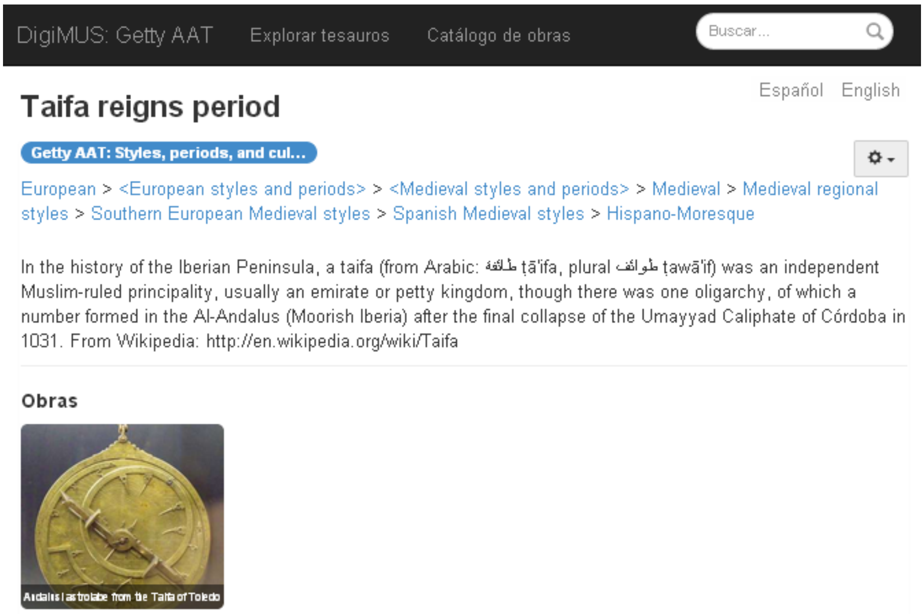


Figure 47 GVP Use in the Spanish DigiMus

Pros: GVP data has been modelled comprehensively, and auxiliary aspects were taken into account such as proper semantic resolution, serving useful entities, licensing, dataset description. This LOD publication has been praised as a comprehensive example to be followed by other CH publications. Getty took care of all aspects of documentation, hosting and support, so GVP LOD is used widely.

Cons: Some people find the representation too complex, since it exposes all aspects of the data. Thus, Getty is considering serving different profiles of the data, e.g. simple SKOS that conflates the difference between guide terms and concepts, without label metadata, etc.

3.9 Cultural Periods and Styles

Dealing with cultures and periods is of prime importance in art research. Getty AAT considers culture, peoples, ethnic groups, historic periods, art movements, and even religion in a uniform way, since any of these can generate related artworks. Some examples: Stone Age, Christianity, Alhambra style, Reign of the Knight Templars in Malta, Impressionism, Nazism

CRM's E4 Period is a complex cultural phenomenon that has spatial and temporal extent, a cultural/historic dimension, and may be dis-continuous (see more at the CRM Tutorial). Two co-extensive periods are not necessarily the same. E.g. the Nazi occupation of France and the French resistance movement are co-extensive, but these are distinct, opposing cultural phenomena. There are a few projects/datasets that try to build databases of periods:

- Getty AAT Periods and Styles: 5569 ethnic and artistic styles, using this query. Does not include date info.
- British Museum thesauri: over 6000. Does not include date info.
- Wikidata: only about 396 but see discussion on WikiProject Visual arts: Item structure: Art_movements: Matching Periods and Styles for trying to bring AAT, BM and WD together.
- PeriodO: A gazetteer of period definitions for linking and visualizing data.
- STAR.Timeline: treatment of archeological time periods. Has a UI demo and REST API returning JSON. Searching by date-range returns only "correlated" periods, using a measure of closeness that considers relation and the duration of query and found period. E.g. below: enter 1701-1800, it returns "18TH CENTURY AD", and 9 other periods. One of them is "NAPOLEONIC WARS", which does not intersect with the 18th century, but is right after it, so is considered related.

Timeline service test client. Enter start and end years and press 'search'. Click on any result to search again using that start / end year. The graphical display shows the periods relative to each other on the timeline.

From year: 1701 To year: 1800 Search

Service call: http://reswin1.isd.glam.ac.uk/rest_ws/getRelatedPeriods?startYear=1701&endYear=1800

Relationship	Label	Start year	End year
equals	18TH CENTURY AD	1701	1800
overlaps	GEORGIAN	1714	1830
during	POST MEDIEVAL	1540	1901
includes	GEORGE II	1727	1760
finished by	LATE 18TH CENTURY	1767	1800
started by	EARLY 18TH CENTURY	1701	1732
overlapped by	WILLIAM III	1689	1702
before	NAPOLEONIC WARS	1803	1815
after	LATE 17TH CENTURY	1667	1700

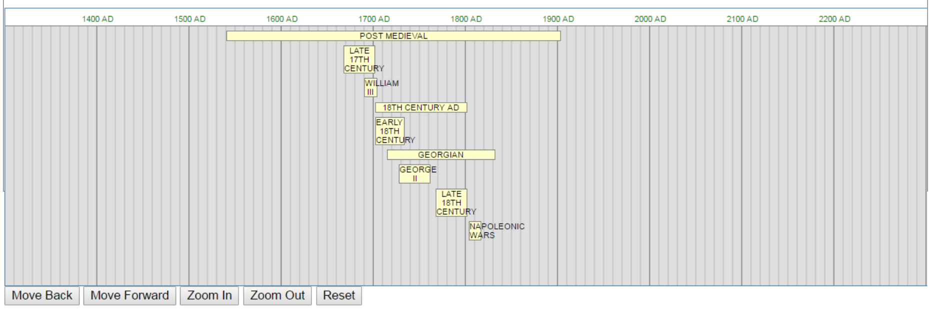


Figure 48 STAR Timeline Periods: Demo Client

3.10 Iconography

Iconography studies the identification, description, and interpretation of the content of images: the subjects depicted, the particular compositions and details used to do so, and other "standard" elements that are distinct from artistic style. Thus iconography is the art and science of capturing subjects that often appear in artworks. Two iconographic datasets are available.

Iconclass. This is a well-known Dutch iconographic effort maintained by RKD. Iconclass includes three sets of data:

- **Classification System**: 28,000 hierarchically ordered definitions divided into ten main divisions. Each definition consists of an alphanumeric classification code (notation) and the description of the iconographic subject (textual correlate). The definitions are used to index, catalogue and describe the subjects of images represented in works of art, reproductions, photographs and other sources. Example of a biblical topic:

```
7 Bible
71 Old Testament
71H story of David
71H7 David and Bathsheba (2 Samuel 11-12)
71H71 David, from the roof (or balcony) of his palace, sees Bathsheba bathing
71H713 Bathsheba receives a letter from David
71H7131 Bathsheba (alone) with David's letter
```

- **Alphabetical Index**: 14,000 keywords used for locating the notation and its textual correlate needed to describe and/or index an image.
- **Bibliography**: 40,000 references to books and articles of iconographical interest (not yet online).

Iconclass has a comprehensive and complicated notation system including "auxiliaries" that allow a huge number of combinations (about 1.3 million notations with all keys and children fully expanded):

- **Bracketed text**, e.g. 25G41(**ROSE**) meaning "rose"
- **Key** (+digits), e.g. 25F23(LION)(**+12**) meaning "heraldic lion"
- **Queuing of keys** (catenating +digits), e.g. 25FF241(+5**11**) meaning "unicorn with nose or tusk in an unusual place"
- **Doubling of letter** to modify the meaning, e.g.
    - Animals: 25F Animals vs 25F**F** fabulous animals
    - The (nude) human figure: 31A male vs 31A**A** female
    - Wedding feast/meal: 42D25 indoors vs 42D**D**25 out of doors
- **Structural digit**: indicates important episodes in a character's lifetime, e.g.
    - For saints, 2 means early life, e.g. 11H(FRANCIS)**2** "**early life** of St. Francis of Assisi"
    - For classical gods, 2 means love-affairs, e.g. 92B3**2** "**love-affairs** of Apollo"

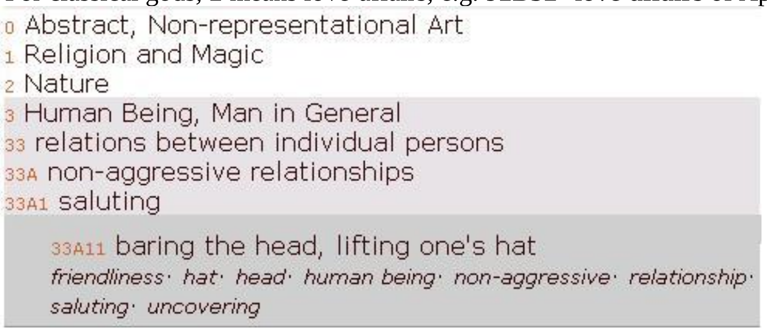


Figure 49 Example of Subject Classification with Iconclass

Iconclass is available in numerous languages (Dutch, English, French, German, Italian, Finnish) through:

- Iconclass Browser, e.g. http://www.iconclass.org/rkd/94L/ is Hercules
- Iconclass LOD, e.g. http://iconclass.org/94L is the semantic URL for Hercules. It's available as RDF and JSON (but not JSON-LD)
- FINTO (the Finnish Thesaurus and Ontology Service) has an excellent Iconclass browser with alphabetical and hierarchical browsing. E.g. Hercules is at https://finto.fi/ic/en/page/94L, and that page offers RDF/XML, TURTLE and JSON-LD downloads

There are several art search systems based on Iconclass. Brill Arkyves is a commercial database, treasure trove and toolbox for the History of Culture. It is a single access point for thematic searches across a wide variety of cultural heritage collections.

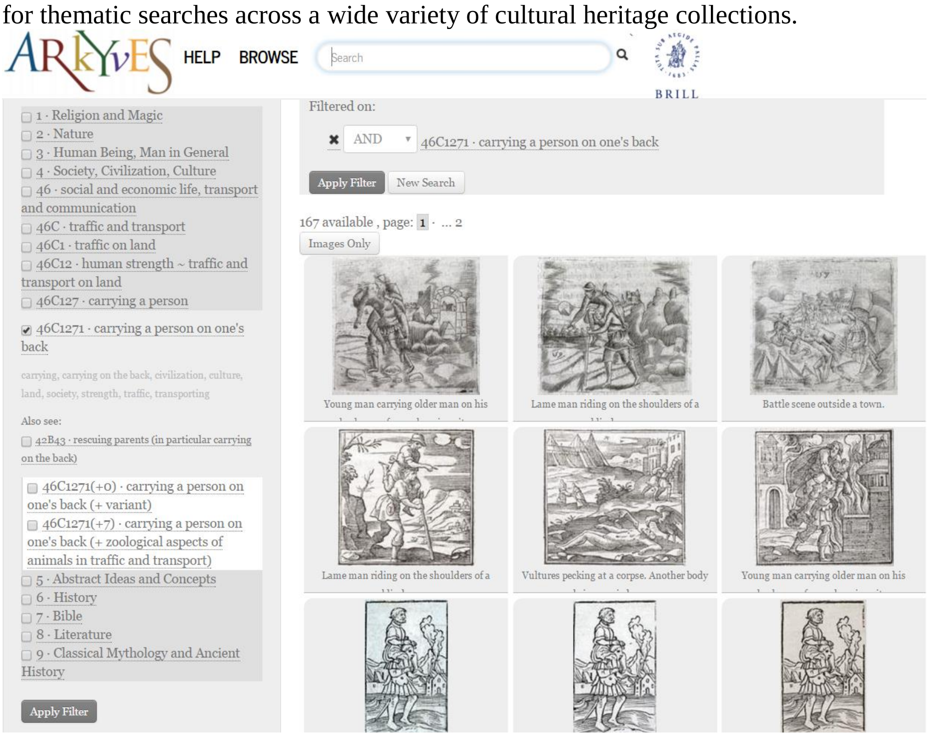


Figure 50 Brill Arkyves System Using Iconclass 46C1271 "Carrying a person on one's back"

Getty Iconography Authority. Getty is developing IA (Patricia Harpring, 2016) as a sub-project of CONA (so it's called CONA IA). IA includes all subjects except those that belong to AAT (general concepts), TGN (real places), ULAN (real agents) or CONA (real artworks). The scope of IA includes:

- Character, Fictional Person, Named Animal, Event/Narrative, Fictional Place, Allegory/Symbolism, Fictional Built Work, Fictional Literature, Religion/Mythology/Legend (as described in CONA section 3.6.3.18)
- Person (character), animal (character), event, imaginary place (as described in CCO section A.4.2.2.5.2)

While Iconclass is well developed but focuses on ancient mythology and Christian religious iconography, IA is in development and has wider remit. IA includes the following data aspects:

- Multilingual labels and descriptions
- IA hierarchical organization, including Root Record, Facets, Guide Terms
- Associative relations within IA
- Relations from IA to the other Getty vocabularies

Let's compare the data about Hercules in the two authorities:

iconclass:94L

- Labels in English, French, German, Italian, Finnish
- No description
- Allows many combination (or expansion) terms
- Shallower hierarchy: 9 Classical Mythology and Ancient History> 94 the Greek heroic legends (I)> 94L (story of) Hercules (Heracles)
- Indexing terms: Greek legend · Hercules · ancient history · classical antiquity · hero · heroic legend · history · legend · mythology
- Numerous narrower concepts: 94L1 early life, prime youth of Hercules; 94L2 love-affairs of Hercules; 94L3 most important deeds of Hercules: the Twelve Labours; ...
- Labels in Latin, English, Greek, Russian, Italian, Etruscan
- Language and source for every label
- Description
- No indexing terms
- The hierarchy is not yet shown, since IA is in development

The diagram below illustrates a LOD mapping for the following facts about Hercules. As you can see, a lot of info is packed into this graph!

- Part of: IA Thesaurus
- Record Type: Religion/Mythology/Legend
- Concept sources and Locators in those sources
- Same As: iconclass:94L
- Labels (names) and their Sources
- Description: "Probably based on an actual historical figure, a king of ancient Argos. The legendary figure was the son of Zeus and Alcmene ..."
- Hierarchical position: Classical Mythology> Greek heroic legends> Story of Hercules
- Birth place: tgn:7010720 Argos, Associated place: tgn:7029383 Thebes. Notice that mythological characters may be related to real historic places
- Father: Zeus (Greek god), with comment: "was his favorite son"
- Mother: Alcmene (Greek heroine)
- Role: Greek hero, king
- Participated in event: Labors of Hercules, Clean the stables of King Augeas
- Additional associative relations:
    - "Zeus" has spouse "Alcmena"

- "Labors of Hercules" has subevent "Clean the stables of King Augeas" (i.e. we enumerate the 12 labors of Hercules as separate events, and connect them using this relation)
- "Augeas" participated in event "Clean the stables of King Augeas" (presumably Augeas asked/motivated Hercules to perform this labor)

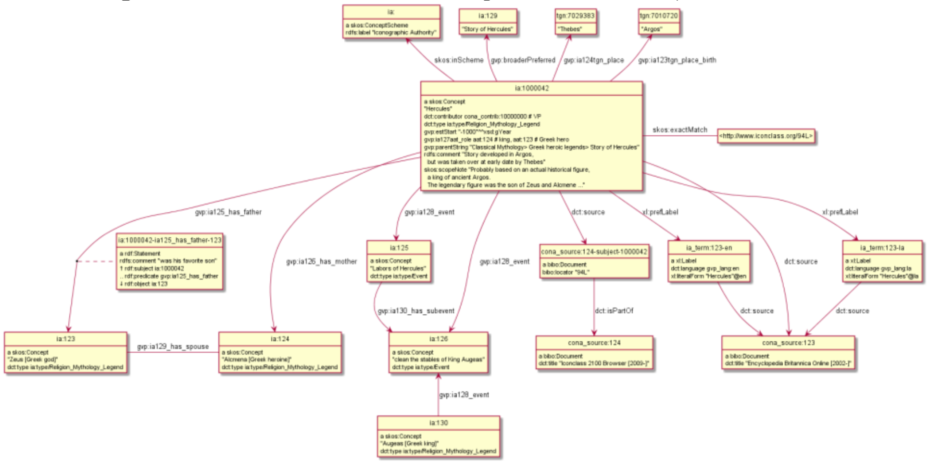


Figure 51 Semantic Representation of Hercules Info in Getty IA

3.11 Wikidata for CH Data Integration

Wikidata (WD) has emerged as a very important LODLAM data integration resource, especially for Authority Control. In the recent International Linked Data Survey for Implementers (OCLC Research, 2018) WD gains 10 places compared to 2015:

Table 5 CH LOD Data Sources (OCLC Research, 2018)

Linked data source	2018 Rank	2015 Rank
id.loc.gov	1	3
VIAF (Virtual International Authority File)	2	1
DBpedia	3	2
GeoNames	4	3
Wikidata	5	15

WD deals with the same sort of data as DBpedia (DBP), namely facts in Wikipedia, but has taken a totally opposite approach:

- DBP extracts facts from Wikipedia using sophisticated IT approaches, a small editorial community, and sadly lacking editorial processes.
- WD curates facts to be used by Wikipedia, using a variety of import/export tools ("bots"), a very large editorial community, and strong editorial processes.

Thus, the quality of data in Wikidata is typically higher than in DBpedia. Wikidata covers all entities from all language editions of Wikipedia (i.e. same as the potential union of all language editions of DBP), plus additional entities created for infrastructural reasons, or by aggregation from external vocabularies (e.g. AAT) for which no encyclopedic article exists. Wikidata has lots of information on all kinds of entities, including GLAM institutions, artists, artworks, etc. But for some entities (especially less notable ones, WD still has less info than DBP.

Reasonator provides a nice display of Wikidata information. Note in particular the auto-generated description and the large variety of authority control references (external-ids) on the right side.

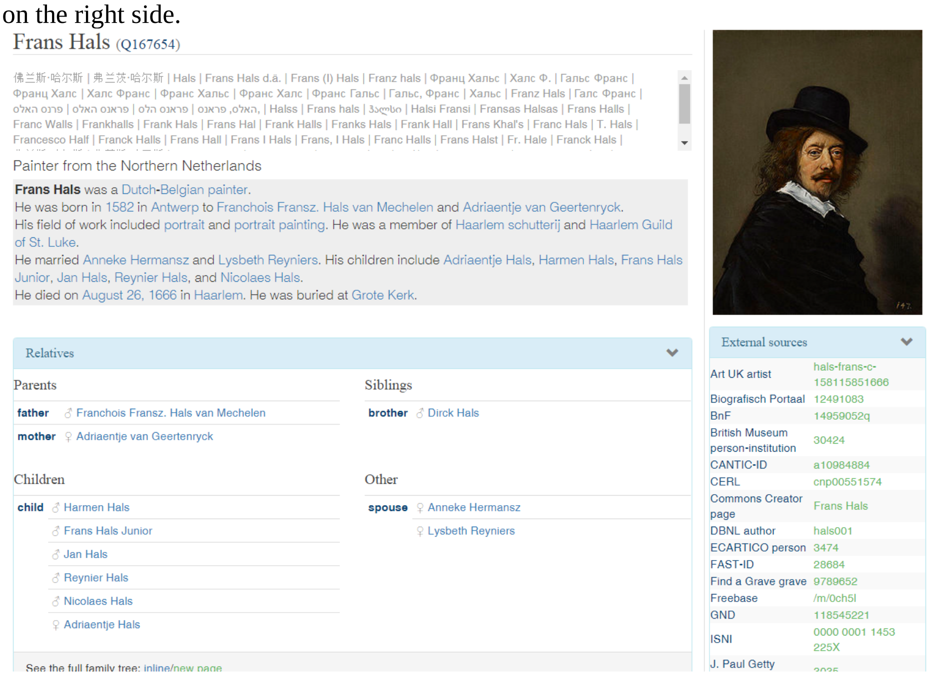


Figure 52 WD Info about Frans Hals on Reasonator

Wikidata and GLAM. Wikidata sees increasing use by the LODLAM and wider semantic communities, e.g.:

- Europeana recommends data providers to use Wikidata as target entities for semantic enrichment (Alexiev, Valentine Charles, & Hugo Manguinhas, 2015)
- Schema.org has recently adopted a similar recommendation: that web entities marked up with Schema should relate to Wikidata as a global "inventory of senses".
- Various GLAMs and GLAM projects have started contributing or integrating their data with Wikidata (Alexiev, 2015c). This gives a CH institution an excellent way to obtain multilinguality, globalize its reach, and leverage LOD.
- For example, a group of Flemish museums and art collections have published on Wikidata metadata 30,000 artworks from their collections (S. Fauconnier, B. Lemmens, & B. Dierickx, 2017).
- The GLAM Wiki Facebook group is a hot bed of related activities

- The National Library of Wales published detailed linked data set for a complete art collection (5000 artworks)

Wikidata EHRI Example. The European Holocaust Research Infrastructure (EHRI) project had collected information about concentration camps and ghettos, but the information was very poor (just label and links subcamp-main camp). E.g. EHRI1 knew very little about the "Maly Trostinec" camp, merely the label.

Wikipedia knows a lot more. The info is not structured, but references for many of the facts are provided:

- names: Maly Trostinets, Maly Trastsianiets, Trasciane, Малы Трасцянец, Maly Tras'tsyanyets, Малый Тростенец, Maly Trostinez, Maly Trostenez, Maly Trostinec, Klein Trostenez
- location: outskirts of Minsk
- admin district: Reichskommissariat Ostland
- established: 10 May 1942
- victim countries: predominantly Belarus (inferred, not explicitly stated). Also Austria, Germany, Czech Republic
- victim places: Minsk, Berlin, Hanover, Dortmund, Münster, Düsseldorf, Cologne, Frankfurt am Main, Kassel, Stuttgart, Nuremberg, Munich, Breslau, Königsberg, Vienna, Prague, Brünn, Theresienstadt
- known victims: Vincent Hadleŭski (Wincenty Godlewski): arrested in Minsk on December 24, 1942 and shot at Trascianiec the same day. Norbert Jokl (debated). Margarete Hilferding (in transit to the camp from Terezín). Grete Forst. Cora Berliner (most likely)
- killing grounds: Blagovshchina (Благовщина) forest, Shashkovka (Шашковка) forest
- perpetrators (and their fate): lead: SS Unterscharführer Heinrich Eiche (fled to Argentina after the war and all trace of him was lost). Eduard Strauch (died in Belgian prison in 1955). Rottenführer Otto Erich Drews (1968: the Court in Hamburg sentenced to life imprisonment). Revieroberleutnant Otto Hugo Goldapp (1968: the Court in Hamburg sentenced to life imprisonment). Hauptsturmführer Max Hermann Richard Krahner (1968: the Court in Hamburg sentenced to life imprisonment). Heinrich Seetzen (committed suicide in a British POW camp). Gerhard Maywald (settled after the war in West Germany; On August 4, 1977 sentenced to 4 years imprisonment). Jewish Sonderkommando 1005

WD knows the following **structured** info:

- names and Wikipedia links in the following languages: Беларуская, Беларуская (тарашкевіца), Čeština, Dansk, Deutsch, Español, Français, Frysk, Italiano, עברית, Nederlands, Norsk bokmål, Polski, Português, Русский, Српски / srpski, Suomi, Svenska, Українська, 中文
- additional aliases, eg Vernichtungslager Maly Trostinez, KZ Maly Trostinez, Blagowschtschina
- country: Belarus
- location: 53°51'3"N, 27°42'17"E
- Authority IDs: Geonames, VIAF, Freebase

DBP knows the following **structured** info:

- links to Wikidata, Geonames, Freebase, different Wikipedias
- coordinates
- a few more aliases: Maly Tras'tsyanyets, Maly Tras'tsyanyets camp, Maly Tras'tsyanyets concentration camp, Maly Tras'tsyanyets extermination camp
- the fact that it is DeathPlace of the following people. This comes from the articles about these people (i.e. inverse links): dbr:Margarete_Hilferding, dbr:Grete_Forst,dbr:Vincent_Hadleŭski
- categories that can be used to find other relevant articles: dbc:World_War_II_sites of_Nazi_Germany, dbc:Geography_of_Minsk, dbc:History_of_Belarus_(1939–1945), dbc:History_of_Minsk, dbc:Maly_Trostenets_extermination_camp (super-categories of this category lead to Nazi_extermination_camps, Nazi_concentration_camps_that allow to discover all concentration camps), dbc:The_Holocaust_in_Belarus, dbc:World_War_II_sites_in_Belarus, dbc:Belarus_in_World_War_II

Because of the richer info, EHRI2 decided to use WD as a semantic integration platform to both find, contribute data about camps and ghettos, and re-ingest it back to EHRI for use in tasks such as semantic text enrichment, geo mapping of Holocaust victim "life trajectories", etc.

First we created or enriched Wikidata entities for relevant sources, e.g. United States Holocaust Memorial Museum, USHMM Holocaust Encyclopedia, USHMM encyclopedia of camps and ghettos, 1933-1945, Yad Vashem Encyclopedia of the Holocaust, EHRI Project, CIA Nazi camps list (a primary document from 7 May 1945 named Axis Concentration Camps And Detention Centres Reported As Such In Europe. Basic Handbook KLs Konzentrationslager).

Then we proposed relevant external-id properties, such as USHMM Holocaust Encyclopedia ID, USHMM person ID, Yad Vashem Encyclopedia of the Ghettos ID (others already existed, e.g. Oorlogsmonument ID)

Then we undertook a laborious data matching task:

- Identify EHRI camps and ghettos on Wikidata using various matching strategies
- Collect more candidate ghettos from searches on Wikipedia, DBpedia, and the USHMM places list
- When there was no entry for a ghetto but only for the corresponding town, create a separate entry and link it to the town
- Ensure every entry has coordinates, and the coordinates between Wikidata and other sources agree (at least approximately)
- Ensure that sufficient labels exist for the entry
- Ensure the entry has appropriate type
- Provide additional coreferencing identifiers (external-ids)

The work is completed for ghettos and is in process for camps. As a result, EHRI has contributed records for some 2000 ghettos to Wikidata. The maps below compare ghettos (blue) and camps (orange): before the contribution (left) and after the contribution (right):

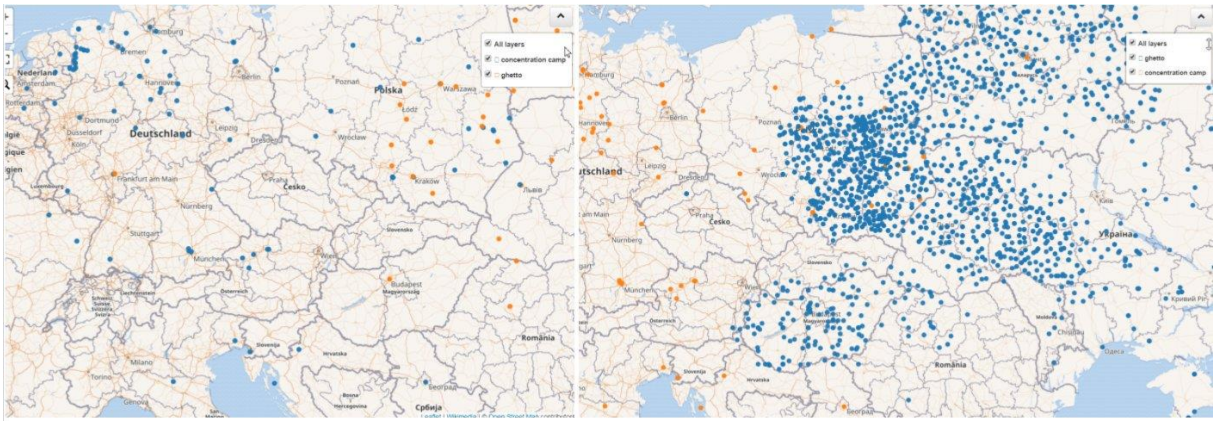


Figure 53 Nazi ghettos and camps on Wikidata, 15 Mar vs 30 June 2017

WD SPARQL Query. WD has a powerful SPARQL query service. The map above is produced with the following query:

```
#defaultView:Map
select distinct ?place ?placeLabel ?location ?layer {
  {?place wdt:P31/wdt:P279* wd:Q152018; wdt:P625 ?location.
    bind("ghetto" as ?layer)} union
  {?place wdt:P31/wdt:P279* wd:Q328468; wdt:P625 ?location.
    bind("concentration camp" as ?layer)}
  SERVICE wikibase:label {bd:serviceParam wikibase:language "en,de,ru,nl"}
}
```

Although Wikidata queries include cryptic numbers, they are not hard to write since Wikidata includes excellent explanation and auto-completion features

- Hover over any term and a tooltip shows its meaning
- Type "#" at the start of the query and the different defaultViews are suggested. These include image grid, map, tree, etc etc
- Type wd: and part of the name (or alias) of some entity (e.g. "wd:ghetto") and press control-Space: an auto-complete box is shown so you can select the right term
- Type wdt: and part of the name of some property (e.g. "wdt:coord" or "wd:location") and press control-Space: an auto-complete box is shown so you can select the right property
- Type "serv" and control-Space: a box to select a special query service is shown. In this case we used the wikibase:label service to automatically fetch the name of a camp/ghetto in one of several languages in order of preference.

Names or aliases can be used equally, e.g. both "wdt:artwork" and "wdt:work of art" will complete to wd:Q838948. wdt:P31/wdt:P279* is a commonly occurring idiom that reads "instance of, followed by any number of subclass of", which means we're looking for entities of a given class or any of its subclasses.

As another example, the following query returns some paintings from Canadian museums:

```
#defaultView:ImageGrid
select ?item ?itemLabel ?typeLabel ?image ?collectionLabel {
```

```
  ?item wdt:P495|wdt:P17 wd:Q16.
  ?item wdt:P31/wdt:P279* wd:Q3305213.
  ?item wdt:P31 ?type.
  ?item wdt:P18 ?image.
  optional {?item wdt:P195 ?collection}
  SERVICE wikibase:label { bd:serviceParam wikibase:language "en,fr". }
} limit 100
```

Reading the query line by line:

- Display as image grid
- Select several variables bound by the query. Variables ?itemLabel ?typeLabel ?collectionLabel are "magically computed" from the corresponding resources ?item ?type ?collection by the wikibase:label SERVICE
- Country of origin (P495) or Country (P17), being Canada (Q16)
- The item should have class Painting (Q3305213) or one of its subclasses
- Get the item's type for display
- Get the item's image for display
- Get the item's collection (museum) for display
- Use the wikibase:label SERVICE to automatically pick labels (artwork title etc) in English or else in French (a fallback mechanism)
- Limit to 100 results

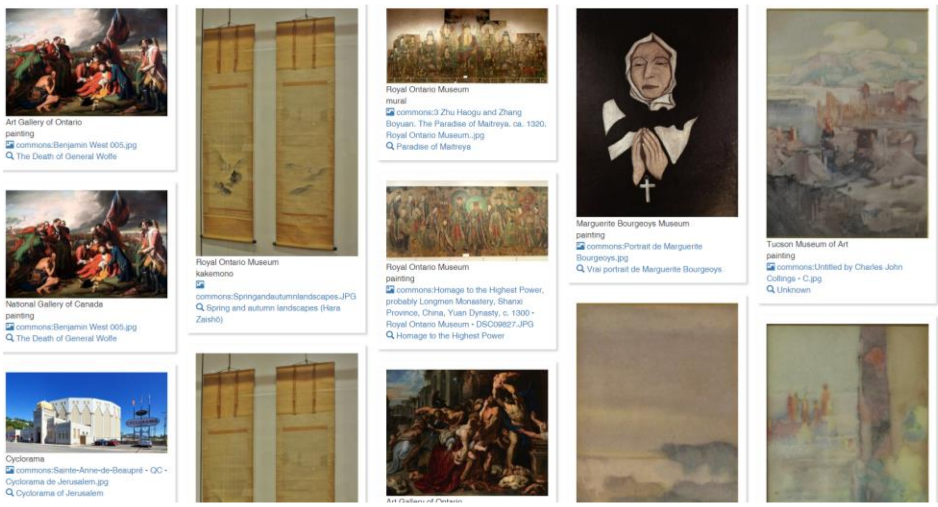


Figure 54 WD Image Grid of Paintings in Canadian Museums

It is easy to make queries and display interesting results, and the twitter user https://twitter.com/WikidataFacts is always ready to help creating more.

Sum of All Paintings. WD development is funded by the WikiMedia Foundation and is mainly performed by WM Germany (WMDE). However, many data development tasks are performed by informal communities called "WikiProjects". One such is the Wikidata Project Sum of All Paintings, whose goal is to collect information about most paintings (and some other artworks) in the world. As Oct 2016, it had info about 140k paintings. Of course, this is a small part of all paintings in the world (e.g. the J.

Paul Getty museum alone has over 800k artworks), so data donations by GLAMs are highly appreciated.

This data can be used for showcasing paintings held by a particular country (see the paintings in Canadian institutions query above), or works by a painter across collections (catalogue raisonné).

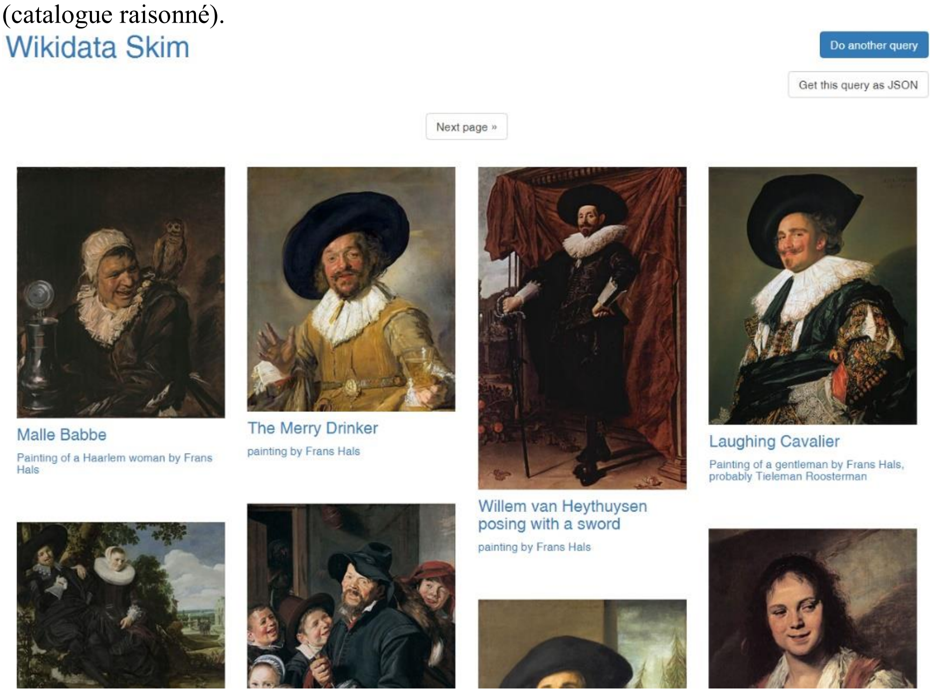


Figure 55 WD Paintings by Frans Hals on Skim (a display application)

Crotos is another nice image search application. It shows links to Wikidata, Wikimedia Commons (the WikiMedia collection of multimedia materials), and the original website.

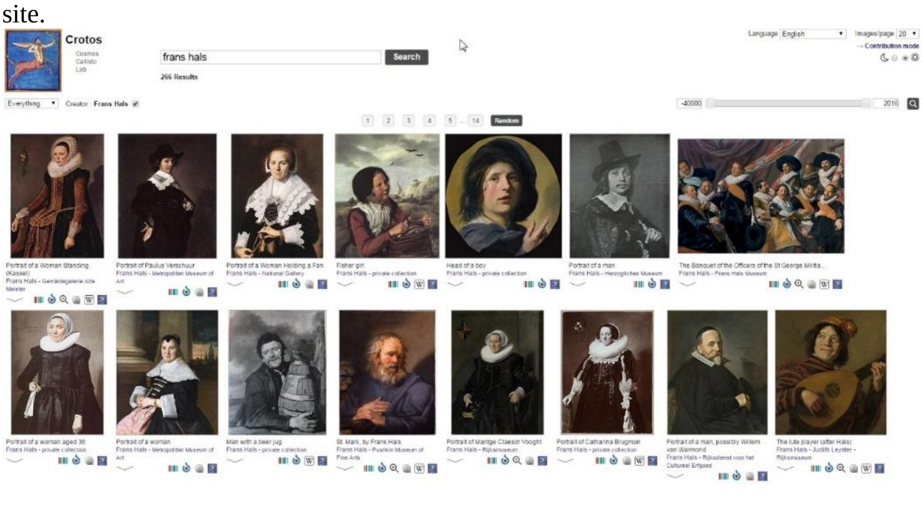


Figure 56 WD Paintings by Frans Hals on Crotos

Wikidata Editing. WD is easy to edit and has excellent auto-completion features. Anyone can help with data maintenance tasks. For example, the call "Hunting for missing inventory numbers" asks people to add inventory numbers for paintings, which is important because the pair <collection, inventory number> is used to identify the painting and avoid duplicate creation on new data ingests. About 9.9k of 140k paintings don't have inventory number. There are sub-lists broken down by country (e.g. US) and institution (e.g. Getty Museum).

label	description	image	creator	inventory number	collection
An Italianate Landscape with Travelers on a Path	painting by Jan Both		Jan Both		J. Paul Getty Museum
Porträt des Julien de La Rochenoire			Édouard Manet		J. Paul Getty Museum

Figure 57 WD Paintings From the Getty Museum Without Inventory Number

It is easy to find the info on Getty's site and add it to Wikidata like this:

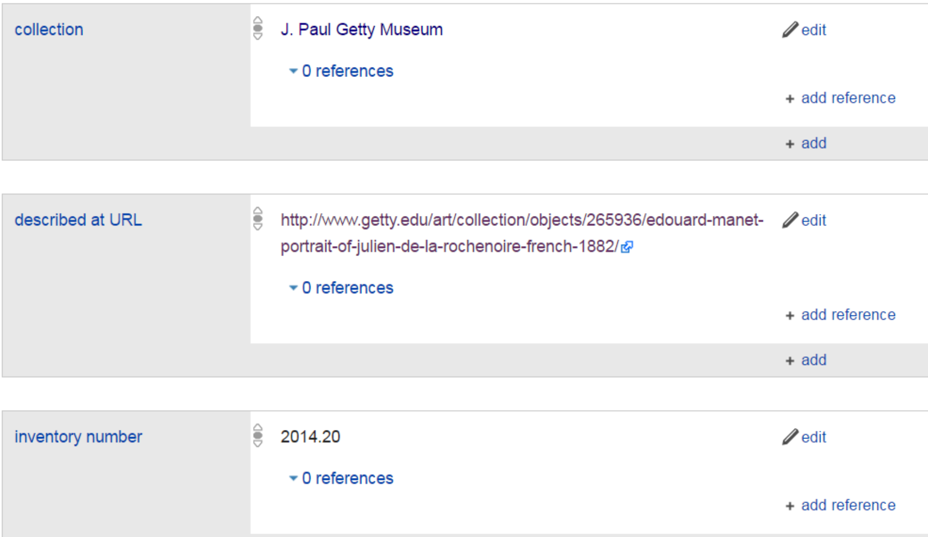


Figure 58 Adding a Painting Inventory Number on WD

Wikidata Batch Editing. For bulk updates and other specific data tasks, people often write "bots", use some of the existing tools (see the extensive list Wikidata External Tools), or implement other automated methods.

- Quick Statements allows adding entities or making changes through a simple tab-delimited format
- PetScan (formerly AutoList 2) allows search by Wikipedia categories, Wikidata properties, etc; and facilitates a variety of maintenance tasks.

For example, below I found entities in the category Bulgarian soccer (football) player that don't have a corresponding Wikidata statement (claim) that their profession is football player, and then automatically added such statement. This added 1597 statements, including for Bulgarian premier Boyko Borisov, who is indeed a 3rd league football player.

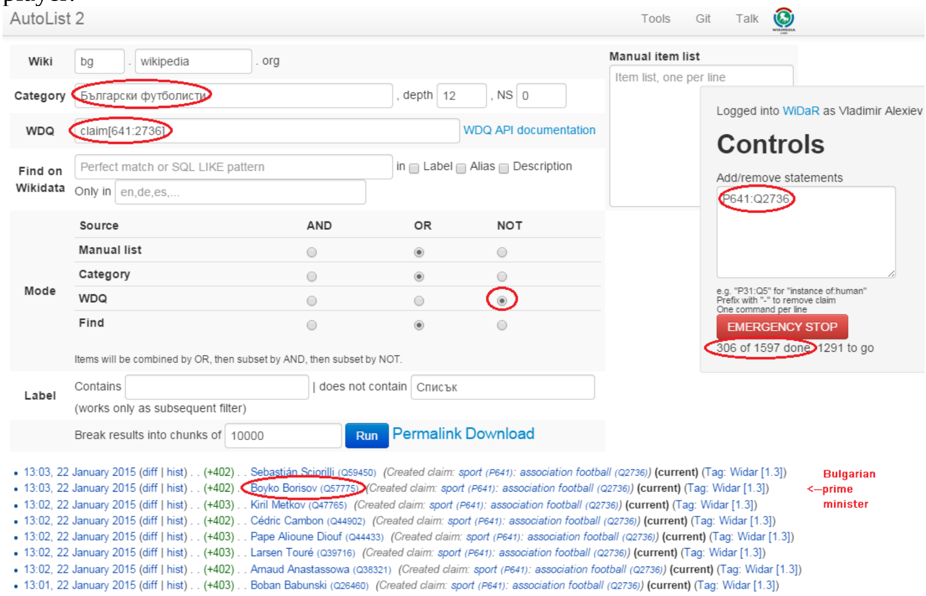


Figure 59 WD Bulk Addition of Statements re Bulgarian Football Players with AutoList

Processing Tabular Data and Adding with Quick Statements. The gist (Vladimir Alexiev, 2017a) explains how to use SPARQL to find J. Paul Getty Museum objects without inventory number (museum ID), how to create such values from property "described at URL" (museum webpage URL), and how to transform the results to WD's Quick Statements format, then insert the IDs directly to Wikidata.

Histropedia is a nice application where users have created all kinds of interesting timelines, most of them based on data from Wikidata.

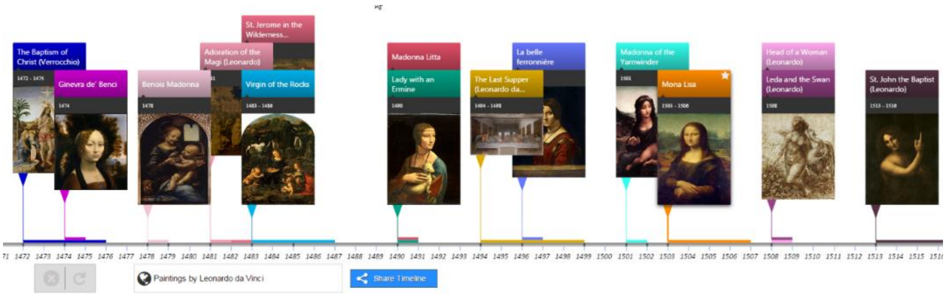


Figure 60 Histropedia timeline: paintings by Leonardo da Vinci

3.12 Wikidata for Global Authority Control

The librarian's dream of global all-encompassing authority control may be coming to fruition. There is a number of massive efforts:

- GND (Gemeinsame Normdatei) is the Integrated Authority File of Germany and Austria.
- VIAF (Virtual International Authority File) is an international effort where data is contributed by 20 national libraries and 15 additional contributors such as Getty ULAN, ISNI, Wikidata. OCLC uses advanced matching and clustering algorithms to find corresponding records between these sources

Since 2013, Wikidata is turning into a hotbed of authority coreferencing activities. In Jan 2015, the Wikidata Project Authority Control was initiated by Ontotext to coordinate and energize such activities. The related WikiProject Biographical Identifiers also includes nice info about tools to use with WD.

A comparative study of Person/Organization datasets was published for Europeana Creative (Alexiev, 2015d). The conclusions of this study are:

- The best datasets to use for name enrichment are VIAF and Wikidata
- There are few name forms in common between the "library-tradition" datasets (dominated by VIAF) and the "LOD-tradition datasets" (dominated by Wikidata)
- VIAF has more name variations and permutations, Wikidata has more multilingual names (translations)
- VIAF is much bigger: 35M persons/orgs. Wikidata has 2.7M persons and maybe 1M orgs
- Only 0.5M of Wikidata persons/orgs are coreferenced to VIAF, with maybe another 0.5M coreferenced to other datasets, either VIAF-constituent (e.g. GND) or not-yet in VIAF (e.g. RKDartists)

A lot can be gained by leveraging coreferencing across VIAF and Wikidata. For example, RKDartists (the most important Dutch art authority list with 475k Dutch and other painters) can be incorporated in VIAF, leveraging the partial matchings to Wikidata Wikidata has great tools for crowd-sourced coreferencing

As part of the study, the number of name variations of Lucas Cranach the Elder was analyzed across 7 LOD datasets. There is a total of 153 name forms that are distributed across datasets as follows.

Table 6 Number of Name Forms of Lucas Cranach the Elder in LOD Datasets

Dataset	**dbpedia**	**freebase**	**ISNI**	**ULAN**	**VIAF**	**wikidata**	**yago**
Count	43	33	51	25	71	70	37
Unique	0	2	1	0	17	24	1

The commonality of name forms across datasets is shown on the following Venn diagram (see interactive diagram in the report for better understanding). Interestingly, the "LOD tradition" datasets (Wikidata: Freebase, DBpedia, Yago) and the "library tradition" datasets (VIAF: ISNI, ULAN) have few forms in common, while each cluster has a lot of forms in common.

This means that to maximize name matching across collections and languages, one should use both VIAF and Wikidata for coreferencing of persons. We can also leverage coreferences from one dataset to another and form an even bigger union authority list, e.g. bring RKDartists into VIAF.

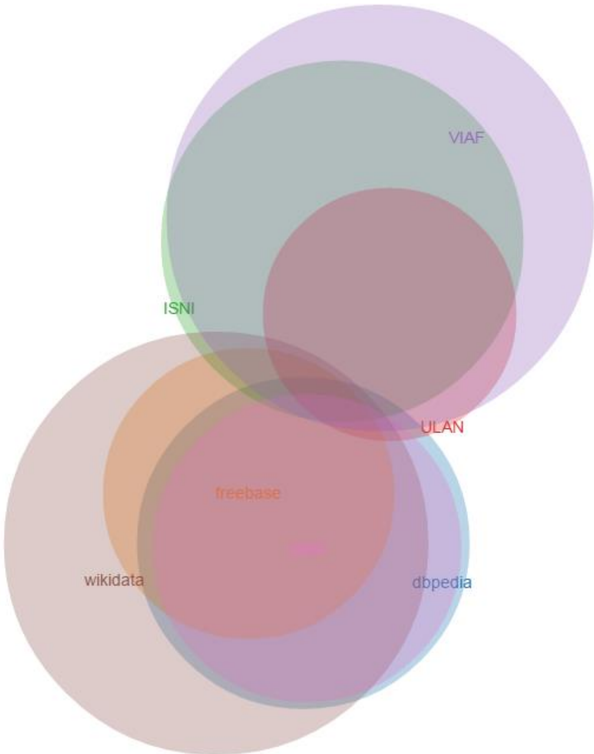


Figure 61 Venn Diagram of Name Forms of Lucas Cranach Across 7 LOD Datasets

VIAF	id in VIAF	Wikidata	id in Wikidata
viafID	49268177	VIAF	49268177
BAV	ADV10197613		
BNC	.a10853637		
BNE	XX907273		
BNF	cb12176451h	BNF	12176451h
DNB	118522582	GND	118522582
ISNI	0000000121319721	ISNI	0000 0001 2131 9721
JPG	500115364	ULAN	500115364
LC	n50020861	LCCN	n50020861
LNB	LNC10-000002573		
NDL	00436834		
NKC	jn20000700335		
NLA	000035031951		
NLI	000035532,001445575,001448179		
NLP	a16828161		
NTA	068435312	NTA PPN	068435312
NUKAT	vtls000190728		
SELIBR	182422		
SUDOC	028710010		
WKP	Lucas_Cranach_the_Elder	Many Wikipedias	
IMAGINE	T7238,T267474	Cantic	a10853637
		Commons Creator	Lucas Cranach (I)
		Commons category	Lucas Cranach d. Ä.
		Freebase	/m/0kqp0
		RKDartists	18978
		SIMBAD	CRANACH, Lucas the Elder
		Your Paintings	lucas-the-elder-cranach

Figure 62 VIAF-WD Identifier Coreferences for Lucas Cranach

WD Coreferencing with Mix-n-Match. The WD MnM tool is the center of authority coreferencing activity in Wikidata. As of Sep 2017, about 507 authority databases (catalogs) are loaded on MnM (growing from 234 as of Oct 2016). The total number of external-id properties has reached almost 2000, with more than 1500 classified as “properties for authority control”. The list of databases:

- Includes categories: Art, Biography, Biology, Encyclopedia, Food, General, Geography, Infrastructure, Language, Location, Organisation, Religion, Science.
- Includes catalogs such as: Getty AAT, TGN, ULAN; RKD artists, RKD works; LoC Authorities; VIAF (not in M-n-M but on Wikidata); British Museum persons; BBC YourPaintings; Artsy, etc, etc.

In the figure below, the colored bars show percent completion (green are matched and manually verified, blue are auto-matched awaiting verification, red are authority entries not present on Wikidata, and gray are not yet matched). Most Catalogs have a corresponding external-id property that can be used to record the correspondence in a Wikidata entry (those with a red bar are an exception).

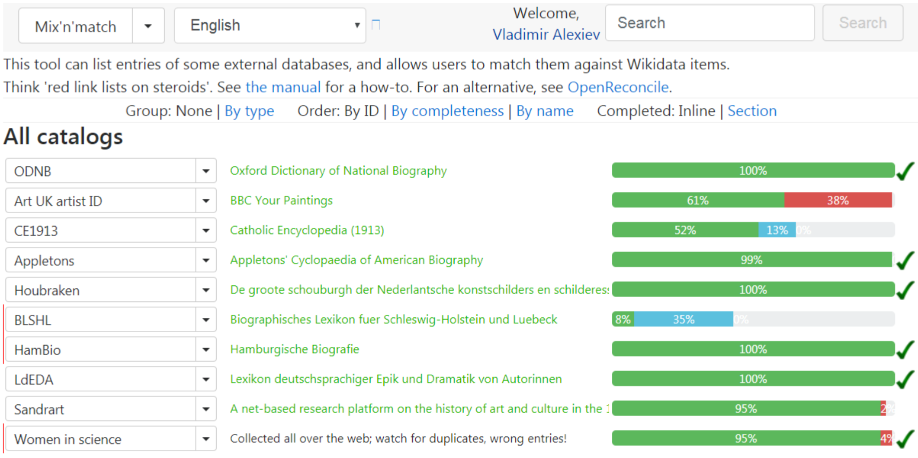


Figure 63 WD Mix-n-Match Home Screen

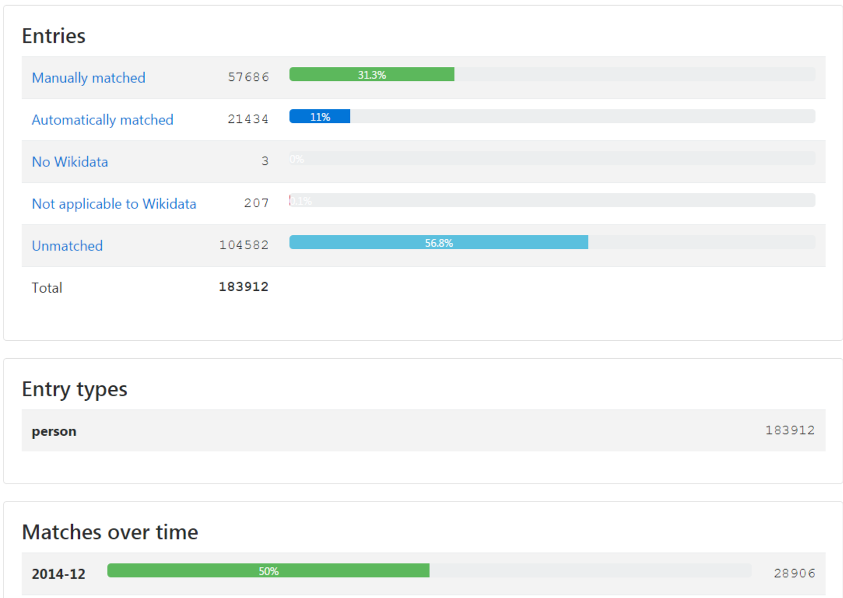


Figure 64 Getty ULAN Catalog on WD Mix-n-Match

MnM has a fairly good automatic matching algorithm. Then doing manual verification is an easy and intuitive task. One can confirm the match, change the WD entry (Q number) after searching in WD, declare there is no matching WD entry, or that it does not belong on WD. Because WD has names from all Wikipedias, it is sometimes amazing how it can match a name.

Matches can be fetched with a simple URL query from the BEACON service, or with a Wikidata authority file mapping tool "wdmapper".

Figure 65 Working on ULAN Entries in WD Mix-n-Match

There is also a "Visual game" that picks entries needing verification and opens the two entries side by side for easier examination, easy construction of web crawlers to harvest an institutional catalog, etc.

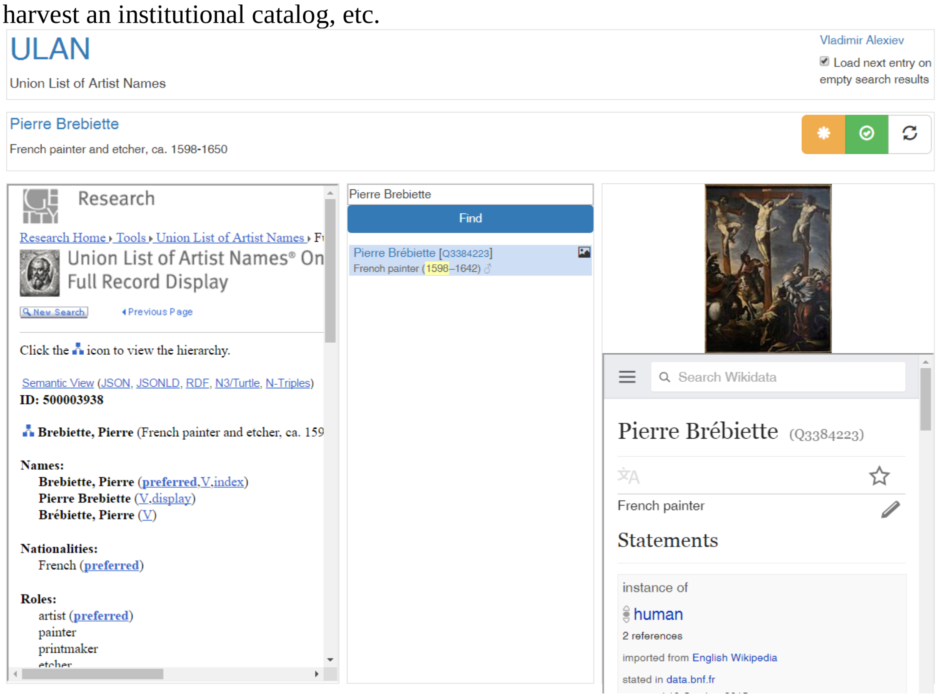


Figure 66 WD Visual Matching Game

Batch Authority Matching. In addition to MnM, there are various tools for matching of thesauri or other instance datasets, amongst them OpenRefine and CultuurLink.

The gist (Vladimir Alexiev, 2017b) describes a custom tool based on advanced use of Google Sheets. It matches 3k AAT concepts to WD, WordNet and BabelNet (it restored an old mapping to Wordnet, retrieved it from BabelNet, mapped to Wikipedia). For each row, it uses the following Google sheet formula (column C) to query the Wikipedia API and get the corresponding Wikidata ID (wikibase_item):

```
=ImportXml(concat("https://en.wikipedia.org/w/api.php?action=query&prop=pageprops&
ppprop=wikibase_item&redirects=1&format=xml&titles=",G1),"//@wikibase_item")
```

It asks for results in XML format, and the part "//@wikibase_item" is a XPath that fetches the WD ID from the resulting XML. Making 3k API calls is slow, so Google sheet initially shows "Loading…" for all rows, and gradually "materializes" the WD IDs (Qnnnn) as they come in. I have periodically sorted the column and used "Edit> Paste special> Values only" for the "materialized" IDs in order to fix them and not cause re-fetching next time when I open the google sheet.

Columns C,D,E are specially formatted to produce the required QuickStatements tab-delimited format. E.g. for row 62 AAT "patrol_wagons" (corresponding to Wikipedia "Police_van") this statement links the WD ID to AAT ID:

```
Q1023646	P1014	"300212831"
```

The benefit of google sheets is that they allow easy addition of columns and convenient facilities for manual tasks:

- Collaborative editing by several people at once.
- Tracking which rows are checked, etc (Column A)
- Using filters to find rows of interest. E.g. check=1 means rows that are manually checked and ready for insertion to QuickStatements. After insertion, I change it to check=2 to mark it as already inserted.
- Tracking already existing WD IDs (Column B)
- Conditional formatting to colour existing WD IDs (column B) that differ from my idea what is the matching WD ID (column C) and therefore must be checked.

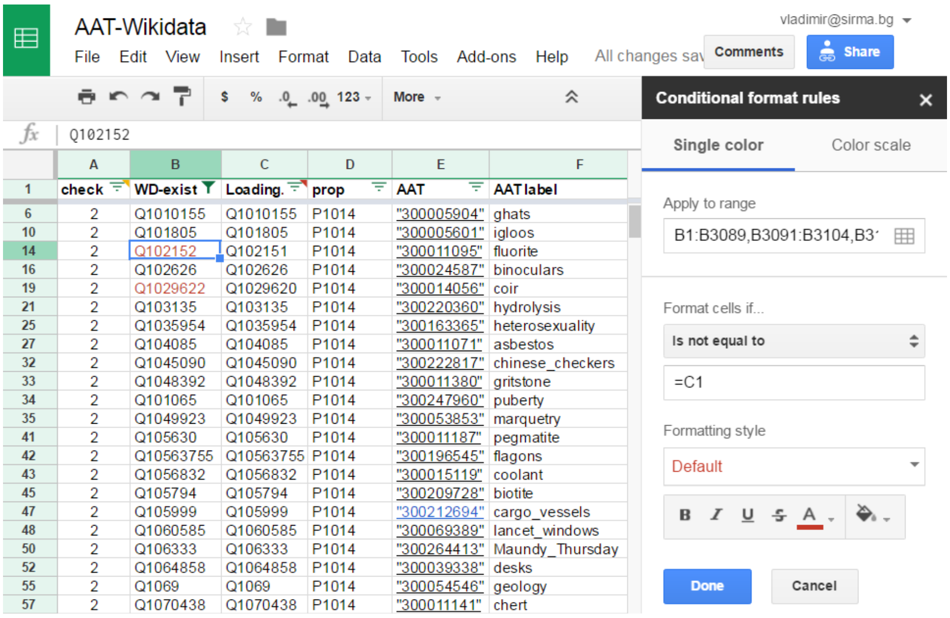


Figure 67 Conditional Formatting in Google Sheets to Highlight Differences

As a result, I was able to match 3000 AAT entries to WD in a short time.

Conclusions

We presented a comprehensive overview of CH ontologies, datasets and semantic projects. Following earlier researcher communities in Life Sciences, the CH and DH communities have come to the conclusion that semantic data integration is the key to interlinking CH data across time and borders, future-proofing it, and enabling DH research based on BigData, semantic linking, semantic text enrichment, inference, network analysis, network visualization, etc.

CH LOD (also called LODLAM) remains an exciting area of research as more and more institutions publish their data in a semantic way, enabling new modes of consumption.

Acknowledgements

Part of this work was done in a consulting project with the Canadian Heritage Information Network (CHIN) to create a strategy for transitioning the Canadian national museum data aggregation to LOD.

References

Alexiev, V. (2012). Implementing CIDOC CRM Search Based on Fundamental Relations and OWLIM Rules. In *Workshop on Semantic Digital Archives (SDA 2012), part of International Conference on Theory and Practice of Digital Libraries (TPDL 2012)*. Paphos, Cyprus: CEUR WS Vol.912. Retrieved from http://ceur-ws.org/Vol-912/paper8.pdf

Alexiev, V. (2015a). *Getty Vocabularies: LOD Sample Queries* (3.2). Getty Research Institute. Retrieved from http://vocab.getty.edu/doc/queries/

Alexiev, V. (2015b). *Getty Vocabulary Program (GVP) Ontology* (3.2). Getty Research Institute. Retrieved from http://vocab.getty.edu/ontology

Alexiev, V. (2015c). *GLAMs Working with Wikidata*. Athens, Greece.

Alexiev, V. (2015d). *Name Data Sources for Semantic Enrichment* (Deliverable No. Part of Deliverable D2.4). Europeana Creative project. Retrieved from http://vladimiralexiev.github.io/CH-names/README.html

Alexiev, V., Manov, D., Parvanova, J., & Petrov, S. (2013). Large-scale Reasoning with a Complex Cultural Heritage Ontology (CIDOC CRM). In *Workshop Practical Experiences with CIDOC CRM and its Extensions (CRMEX 2013) at TPDL 2013*. Valetta, Malta. Retrieved from http://www.ontotext.com/sites/default/files/publications/CRM-reasoning.pdf

Alexiev, V., Valentine Charles, & Hugo Manguinhas. (2015). *Wikidata, a Target for Europeana's Semantic Strategy*. The Hague. Retrieved from http://vladimiralexiev.github.io/pubs/GLAMwiki2015.ppt

Carl Lagoze, & Herbert Van de Sompel. (2015). *Open Archives Initiative Protocol for Metadata Harvesting*. Retrieved from http://www.openarchives.org/OAI/openarchivesprotocol.html

Craig Knoblock, Pedro Szekely, Eleanor Fink, Duane Degler, David Newbury, Robert Sanderson, … Yixiang Yao. (2017). Lessons Learned in Building Linked Data for the American Art Collaborative. In *International Semantic Web Conference (ISWC)*. Retrieved from http://usc-isi-i2.github.io/papers/knoblock17-iswc.pdf

Dominic Oldman, & Donna Kurtz. (2014). *CRM Primer v1.1*. Retrieved from http://www.cidoc-crm.org/sites/default/files/CRMPrimer_v1.1_1.pdf

Dominic Oldman, Joshan Mahmud, & Alexiev, V. (2013). *The Conceptual Reference Model Revealed. Quality contextual data for research and engagement: A British Museum case study* (p. 359 pages). ResearchSpace Project. Retrieved from http://confluence.ontotext.com/display/ResearchSpace/BM+Mapping

Dominic Oldman, Martin Doerr, Gerald de Jong, Barry Norton, & Thomas Wikman. (2014). Realizing Lessons of the Last 20 Years: A Manifesto for Data Provisioning & Aggregation Services for the Digital Humanities (A Position Paper). *D-Lib Magazine*, *20*(7). https://doi.org/10.1045/july2014-oldman

Erin Coburn, Richard Light, Gordon McKenna, Regine Stein, & Axel Vitzthum. (2010). *LIDO - Lightweight Information Describing Objects*. ICOM CIDOC. Retrieved from http://network.icom.museum/cidoc/working-groups/lido/lido-technical/specification/

Europeana. (2017). *Definition of the Europeana Data Model v5.2.8*. Retrieved from https://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation//EDM_Definition_v5.2.8_102017.pdf

Fink, E. E. (2018). *American Art Collaborative (AAC) Linked Open Data (LOD) Initiative: Overview and Recommendations for Good Practices*. American Art Collaborative.

Getty Research Institute. (2013, October 7). CDWA Lite. Retrieved July 27, 2018, from http://www.getty.edu/research/publications/electronic_publications/cdwa/cdwalite.html

Getty Research Institute. (2017). Cultural Objects Name Authority. Retrieved July 27, 2018, from http://www.getty.edu/research/tools/vocabularies/cona/

Gordon McKenna. (2018). *Mapping of LIDO to Spectrum 5.0*. Collections Trust. Retrieved from https://collectionstrust.org.uk/resource/mapping-of-lido-to-spectrum-5-0/

ICOM CIDOC LIDO Working Group. (2017). LIDO Terminology. Retrieved July 27, 2018, from http://network.icom.museum/cidoc/working-groups/lido/lido-technical/terminology/

Katerina Tzompanaki, & Martin Doerr. (2012). *Fundamental Categories and Relationships for intuitive querying CIDOC-CRM based repositories* (No. TR-429). ICS-FORTH. Retrieved from http://www.cidoc-crm.org/docs/TechnicalReport429_April2012.pdf

Kevin Gosling, & Gordon MacKenna (Eds.). (2017a). *Spectrum 5.0. Appendix - Information requirements* (Version 5.0). London: Collections Trust.

Kevin Gosling, & Gordon MacKenna (Eds.). (2017b). *Spectrum 5.0: The UK museum collections management standard* (Version 5.0). London: Collection Trust. Retrieved from https://collectionstrust.org.uk/product/spectrum-5-0-main-volume-and-appendix/

Martin Dörr. (2018, May). *CRM Family of Models*. Presented at the 2nd Data for History workshop, Lyon France. Retrieved from http://dataforhistory.org/sites/default/files/dfh20180525_doerr.pdf

Murtha Baca, & Patricia Harpring. (2016). *Categories for the Description of Works of Art (CDWA)*. Getty Research Institute and College Art Association. Retrieved from http://www.getty.edu/research/publications/electronic_publications/cdwa/index.html

Murtha Baca, Patricia Harpring, Elisa Lanzi, Linda McRae, & Ann Baird Whiteside. (2006). *Cataloging Cultural Objects: A Guide to Describing Cultural Works and Their Images*. Chicago: American Library Association. Retrieved from http://cco.vrafoundation.org/index.php/toolkit/cco_pdf_version/

NISO, & Open Archives Initiative. (2017, February 2). ResourceSync Framework Specification. Retrieved July 27, 2018, from http://www.openarchives.org/rs/1.1/resourcesync

Parvanova, J., Alexiev, V., & Kostadinov, S. (2013). RDF Data and Image Annotations in ResearchSpace. In *Collaborative Annotations in Shared Environments: metadata, vocabularies and techniques in the Digital Humanities (DH-CASE 2013). Collocated with DocEng 2013*. Florence, Italy. Retrieved from

http://www.ontotext.com/sites/default/files/publications/Parvanova2013-SemanticAnnotation.pdf
Patricia Harpring. (2016, June). Getty Iconography Authority: Introduction and Overview. Getty Vocabulary Program. Retrieved from http://www.getty.edu/research/tools/vocabularies/cona_ia_in_depth.pdf
Patrick Le Boeuf, Martin Doerr, Christian Emil Ore, & Stephen Stead. (2018). *Definition of the CIDOC Conceptual Reference Model v6.2.3*. Retrieved from http://www.cidoc-crm.org/releases_table
Robert Sanderson. (2016, December). *Community Challenges for Practical Linked Open Data*. Presented at the Linked Pasts keynote, Madrid, Spain. Retrieved from https://www.slideshare.net/azaroth42/community-challenges-for-practical-linked-open-data-linked-pasts-keynote
S. Fauconnier, B. Lemmens, & B. Dierickx. (2017). *Manual on open data management by cultural institutions on Wikidata*. Retrieved from https://commons.wikimedia.org/wiki/File:Flemish_art_collections,_Wikidata_and_Linked_Open_Data_-_Manual_20160331.pdf
Sarah Lucas. (2001). *Boots: Inkjet on iced fruitcake plus cardboard box*. Retrieved from https://collections.britishart.yale.edu/vufind/Record/1670022
Vladimir Alexiev. (2014). *Getty Vocabulary Program LOD: Ontologies and Semantic Representation*. Dresden, Germany. Retrieved from http://vladimiralexiev.github.io/pres/20140905-CIDOC-GVP/GVP-LOD-CIDOC.pdf
Vladimir Alexiev. (2016). RDF by Example: rdfpuml for True RDF Diagrams, rdf2rml for R2RML Generation. In *Semantic Web in Libraries 2016 (SWIB 16)*. Bonn, Germany. Retrieved from http://vladimiralexiev.github.io/pres/20161128-rdfpuml-rdf2rml/index-full.html
Vladimir Alexiev. (2016). *rnc: RelaxNG Compact schemas for some GLAM-related XML schemas*. Retrieved from https://github.com/VladimirAlexiev/rnc (Original work published 2015)
Vladimir Alexiev. (2017a, January). How to Add Museum IDs to Wikidata. Retrieved from https://gist.github.com/VladimirAlexiev/e51e256be18870ac5033901197ee8277
Vladimir Alexiev. (2017b, March). How to use Google Sheets to Manage Wikidata Coreferencing. Retrieved from https://gist.github.com/VladimirAlexiev/8201d614a819cb7d4023ce9aa315af65
Vladimir Alexiev, Andrey Tagarev, & Laura Tolosi. (2016). *Europeana Food and Drink Semantic Demonstrator Extended* (Deliverable No. D3.20d). Europeana Food and Drink project. Retrieved from http://vladimiralexiev.github.io/pubs/Europeana-Food-and-Drink-Semantic-Demonstrator-Extended-(D3.20d).pdf
Vladimir Alexiev, Cobb, J., Gregg Garcia, & Patricia Harpring. (2015). *Getty Vocabularies Linked Open Data: Semantic Representation* (3.2). Getty Research Institute. Retrieved from http://vocab.getty.edu/doc/
Vladimir Alexiev, & Dilyana Angelova. (2015, July). *O is for Open: OAI and SPARQL interfaces for Europeana*. Presented at the Europeana Creative Culture Jam, Vienna, Austria. Retrieved from http://vladimiralexiev.github.io/pubs/O_is_for_Open_(CultJam_201507)_poster.pdf

Vladimir Alexiev, Jutta Lindenthal, & Antoine Isaac. (2015). On the composition of ISO 25964 hierarchical relations (BTG, BTP, BTI). *International Journal on Digital Libraries*, 1–10. https://doi.org/10.1007/s00799-015-0162-2