

---

# ARCO: THE ITALIAN CULTURAL HERITAGE KNOWLEDGE GRAPH

---

A PREPRINT

**Valentina Anita Carriero**

Semantic Technologies Laboratory  
Institute of Cognitive Sciences and Technologies  
Italian National Research Council  
Via San Martino della Battaglia 44, 00185 Rome, Italy  
valentina.carriero@istc.cnr.it

**Aldo Gangemi**

Digital Humanities Advanced Research Centre  
Department of Classical Philology and Italian Studies  
University of Bologna  
Via Zamboni 32, 00126 Bologna, Italy  
aldo.gangemi@unibo.it

**Maria Letizia Mancinelli**

Central Institute for Cataloguing and Documentation  
Ministry of Cultural Heritage and Activities  
Via di San Michele 18, 00153 Roma  
marialetizia.mancinelli@beniculturali.it

**Ludovica Marinucci**

Semantic Technologies Laboratory  
Institute of Cognitive Sciences and Technologies  
Italian National Research Council  
Via San Martino della Battaglia 44, 00185 Rome, Italy  
ludovica.marinucci@istc.cnr.it

**Andrea Giovanni Nuzzolese**

Semantic Technologies Laboratory  
Institute of Cognitive Sciences and Technologies  
Italian National Research Council  
Via San Martino della Battaglia 44, 00185 Rome, Italy  
andreagiovanni.nuzzolese@istc.cnr.it

**Valentina Presutti**

Semantic Technologies Laboratory  
Institute of Cognitive Sciences and Technologies  
Italian National Research Council  
Via San Martino della Battaglia 44, 00185 Rome, Italy  
valentina.presutti@istc.cnr.it

**Chiara Veninata**

Central Institute for Cataloguing and Documentation  
Ministry of Cultural Heritage and Activities  
Via di San Michele 18, 00153 Roma  
chiara.veninata@beniculturali.it

May 9, 2019

## ABSTRACT

ArCo is the Italian Cultural Heritage knowledge graph, consisting of a network of seven vocabularies and 169 million triples about 820 thousand cultural entities. It is distributed jointly with a SPARQL endpoint, a software for converting catalogue records to RDF, and a rich suite of documentation material (testing, evaluation, how-to, examples, etc.). ArCo is based on the official General Catalogue of the Italian Ministry of Cultural Heritage and Activities (MiBAC) - and its associated encoding regulations - which collects and validates the catalogue records of (ideally) all Italian Cultural Heritage properties (excluding libraries and archives), contributed by CH administrators from all over Italy. We present its structure, design methods and tools, its growing community, and delineate its importance, quality, and impact.

# 1 Bringing the Italian Cultural Heritage to LOD

Cultural Heritage (CH) is the legacy of physical artifacts and intangible attributes of a group or society that is inherited from past generations. It carries aesthetical, social, historical, cognitive, as well as economic power. The availability of linked open data (LOD) about CH has already shown its potential in many application areas including tourism, teaching, management, etc. The higher the quality and richness of data and links, the higher the value that society, science and economy can gain from it.

As of July 2018, UNESCO has designated a total of 1092 World Heritage sites located in 167 different countries around the world, including cultural (845), natural (209), and hybrid (38) sites. According to UNESCO’s list, Italy is the country with the highest number of world heritage sites (54)<sup>1</sup>. UNESCO’s list only indexes the tip of the iceberg of Italian CH, which is managed by the Italian Ministry of Cultural Heritage and Activities (MiBAC). Within MiBAC, ICCD (Institute of the General Catalogue and Documentation) is in charge of maintaining a catalogue of (ideally) every item in the whole Italian CH (excluding libraries and archives), as well as to define standards for encoding catalogue records describing them, and to collect these records from the diverse institutions that administer cultural properties throughout the Italian territory. To date, ICCD has assigned more than 15M unique catalogue numbers to its contributors (cf. Section 2 for further details), and has collected and stored ~2.5M records, ~0.8M of which are available for consultation on its official website. This growing, standardised, curated catalogue is the heart of Italian CH data, and a potential *hub* for a highly reliable and rich *knowledge graph* of Italian CH, and beyond.

This paper describes ArCo, a new resource that realises this potential. ArCo is openly released with a [CC-BY-SA 4.0 licence](#) both on [GitHub](#)<sup>2</sup> and on the official [MiBAC website](#). It can be downloaded as a docker container and locally installed, or accessed [online](#). ArCo includes:

- a knowledge graph<sup>3</sup> consisting of:
  - a network of ontologies (and ontology design patterns), modeling the CH domain (with focus on cultural properties) at a fine grained level of detail (cf. Section 4);
  - a LOD dataset counting ~169M triples, which describe ~0.8M cultural properties and their catalogue records;
- a software for automatically converting catalogue records compliant to ICCD regulations to ArCo-compliant LOD, which enables automatic and frequent updates, and facilitates reuse;
- a detailed documentation reporting: (i) the ontological requirements, expressed in the form of user stories as well as competency questions (CQs), (ii) the resulting ontological models with diagrams and examples of usage;
- a set of running examples that potential consumers can use as training material. They consist of natural language CQs and their corresponding SPARQL queries, which can be directly tested against ArCo’s SPARQL endpoint;
- a test suite, implemented as OWL files and SPARQL queries, used for validating ArCo knowledge graph (KG). It provides a real-case implementation of an ontology testing methodology, useful to both students, teachers, researchers, and practitioners.
- a SPARQL endpoint to explore the resource, run tests, etc.

It is worth remarking that ArCo data are of highly reliable provenance (cf. Section 2). Its ontology network shows high quality, as resulting both as a project of eXtreme Design (XD), an established methodology [4] based on the reuse of ontology design patterns (ODP) (cf. Section 3), and emerging from an ex-post evaluation described in Subsection 5.1. ArCo ontologies (cf. Section 4) address, and are evaluated against, requirements elicited from both the data provider (ICCD), and a community of independent consumer representatives, including private and public organisations working with CH open data. These requirements have raised the need of new ontology models, which have been developed while reusing or aligning to relevant CH ontologies such as CIDOC-CRM [10] and EDM [15]. ArCo data (cf. Subsection 4.3) links to ~18.7K entities belonging to other LOD datasets, e.g. DBpedia, Wikidata, Geonames. ArCo is reused in a separate project that involves ICCD and [Google Arts & Culture](#), focused on photographic cultural properties. Indeed, there is evidence of a growing community reusing and interested in ArCo as discussed in Section

<sup>1</sup>42 additional sites are currently under review.

<sup>2</sup>The links are hidden by blue clickable words. You can find the complete list of URLs included in this paper at: <https://bit.ly/2VpR5cQ>

<sup>3</sup>There is no consensus on a definition for knowledge graph [7], in this context we refer to linked open data including both OWL and RDF entities, and both schema axioms and factual data.


5. ArCo is published by following FAIR principles (cf. Section 6). It is an evolving project, therefore there are many aspects that can and will be improved: they are briefly discussed in Section 8.

## 2 The official catalogue of Italian Cultural Heritage

ArCo data derive from the [General Catalogue of Italian Cultural Heritage](#) (GC), the official institutional database of Italian CH, maintained and published by ICCD. GC currently contains 2.735.343 catalogue records, 781.902 of which are publicly consultable through the ICCD website. The remaining records may refer to private properties, or to properties being at risk (e.g. items in churches that are not guarded), or to properties that need to be scientifically assessed by accounted institutions, etc. GC is the result of a *collaborative effort* involving many and diverse contributors (currently 487) *formally* authorised by ICCD. These are national or regional, public or private, institutional organisations that administer cultural properties all over the Italian territory. They submit their catalogue records through a collaborative platform named [SIGECweb](#). Submissions undergo an automatic validation phase, aimed to check compliance with cataloguing standards provided by ICCD for all kinds of *cultural properties*. A second scientific validation is performed by appointed experts. The authorisation and validation processes guarantee high standard for quality and provenance reliability of GC data as a source for ArCo.

In addition to GC data, ArCo's input includes requirements deriving from consumers' elicited use cases (cf. Section 3), and [ICCD cataloguing standards](#) that define many types of *cultural properties*, precisely: archaeological, architectural and landscape, demo-ethno-anthropological, photographic, musical, natural, numismatic, scientific and technological, historical and artistic properties.

Figure 1 depicts a painting by [Albert Friscia](#) with some excerpts from its XML catalogue record, from GC. The first



<p><b>definition «painting»</b>  &lt;OGTD hint="Definizione"&gt;dipinto&lt;/OGTD&gt;  <b>title</b>  &lt;SGTT hint="Titolo"&gt;Self portrait&lt;/SGTT&gt;</p> <p><b>measurements</b>  unit, height, width  &lt;MIS hint="MISURE"&gt;  &lt;MISU hint="Unità"&gt;cm&lt;/MISU&gt;  &lt;MISA hint="Altezza"&gt;50&lt;/MISA&gt;  &lt;MISL hint="Larghezza"&gt;40&lt;/MISL&gt;  &lt;/MIS&gt;</p> <p><b>conservation status «good»</b>  &lt;STCC hint="Stato di conservazione"&gt;buono&lt;/STCC&gt;</p> <p><b>dating «20<sup>th</sup> century»</b>  &lt;DTZG hint="Secolo"&gt;sec. XX&lt;/DTZG&gt;</p> <p><b>material, technique</b>  canvas/ oil-painting  &lt;MTCI hint="Materiali, tecniche"&gt;tela/  pittura a olio&lt;/MTCI&gt;</p> <p><b>inscription</b>  technique «painting»  position «bottom right»  transcript «A. Friscia '47»  &lt;ISR hint="ISCRIZIONI"&gt;  &lt;ISRS hint="Tecnica"&gt;a pennello&lt;/ISRS&gt;  &lt;ISRP hint="Posizione"&gt;in basso a destra&lt;/ISRP&gt;  &lt;ISRI hint="Trascrizione"&gt;A. Friscia '47&lt;/ISRI&gt;</p>	<p><b>author</b> unique code, name «Albert Friscia»  biographical data «active from 1935 to 1989»  &lt;AUT hint="AUTORE"&gt;  &lt;NCUN hint="Codice univoco ICCD"&gt;00001499&lt;/NCUN&gt;  &lt;AUTN hint="Nome scelto"&gt;Albert Friscia&lt;/AUTN&gt;  &lt;AUTA hint="Dati anagrafici"&gt;Attivo dal 1935 al 1989&lt;/AUTA&gt;</p> <p><b>acquisition</b> type «donation», previous owner «Di Bello, Lidia»  date, location  &lt;ACQ hint="ACQUISIZIONE"&gt;  &lt;ACQT hint="Tipo acquisizione"&gt;donazione&lt;/ACQT&gt;  &lt;ACQN hint="Nome"&gt;Di Bello, Lidia&lt;/ACQN&gt;  &lt;ACQD hint="Data acquisizione"&gt;10/ 02/ 2000&lt;/ACQD&gt;  &lt;ACQL hint="Luogo acquisizione"&gt;RM/Roma/Via Muratte,94&lt;/ACQL&gt;</p> <p><b>location type «previous location»</b>  <b>site</b> type «private house»  name «house of Mr and Mrs Friscia»  address, specifications «inside, 2° floor»  &lt;TCL hint="Tipo localizzazione"&gt;luogo di provenienza&lt;/TCL&gt;  &lt;PRCT hint="Tipologia"&gt;casa privata&lt;/PRCT&gt;  &lt;PRCD hint="Denominazione"&gt;casa dei coniugi Friscia&lt;/PRCD&gt;  &lt;PRCU hint="Denominazione via"&gt;Via Muratte,94 Roma&lt;/PRCU&gt;  &lt;PRCS hint="Specifiche"&gt;interno&lt;/PRCS&gt;</p> <p><b>photographic documentation</b>  category «attached», type «color»  owner agency  &lt;FTA hint="DOCUMENTAZIONE FOTOGRAFICA"&gt;  &lt;FTAX hint="Genere"&gt;documentazione allegata&lt;/FTAX&gt;  &lt;FTAP hint="Tipo"&gt;fotografia colore&lt;/FTAP&gt;  &lt;FTAE hint="Ente proprietario"&gt;SPSAE MT&lt;/FTAE&gt;</p>
---	--

Figure 1: An example of XML data from ICCD Catalogue. Each snippet is translated to English.

issue with XML data is that all information in the records is expressed as strings. In order to build an RDF KG, good practice suggests to produce individuals for every element (or set of elements), whose value (or set of values) refers to anything that may participate as a subject in a triple, or can be linked to, or from, external resources. For example, we want to create an individual for Albert Friscia (the author), as well as an individual for the technical status of this painting. As part of the modeling process (cf. Section 4), we define ArCo classes by abstracting from sets of fields and we define rules for creating URIs for their individuals (cf. Section 3).

### 3 Using eXtreme Design for developing ArCo Knowledge Graph

ArCo knowledge graph (KG) is developed by following eXtreme Design (XD), focused on ontology design patterns (ODPs) reuse [4]. XD is iterative and incremental, implementing a feedback loop cycle by involving different actors: (i) a *design team*, in charge of selecting and implementing suitable ODPs as well as to perform alignments and linking; (ii) a *testing team*, disjoint from the design team, which takes care of testing the ontology; (iii) a *customer team*, who elicits the requirements that are translated by the design team and testing team into ontological commitments (i.e. competency questions and other constraints) that guide the ontology development.

Figure 2 depicts as XD applied to ArCo, jointly with the tools used in the process, e.g. GitHub, Protégé, etc. The remainder of this section provides a detailed explanation of how each phase is implemented.

**Ontology project initiation.** The design team and the customer team (initially composed of experts from ICCD) have shared their knowledge about the domain, the data and the method, and have agreed on a release plan and on communication means.

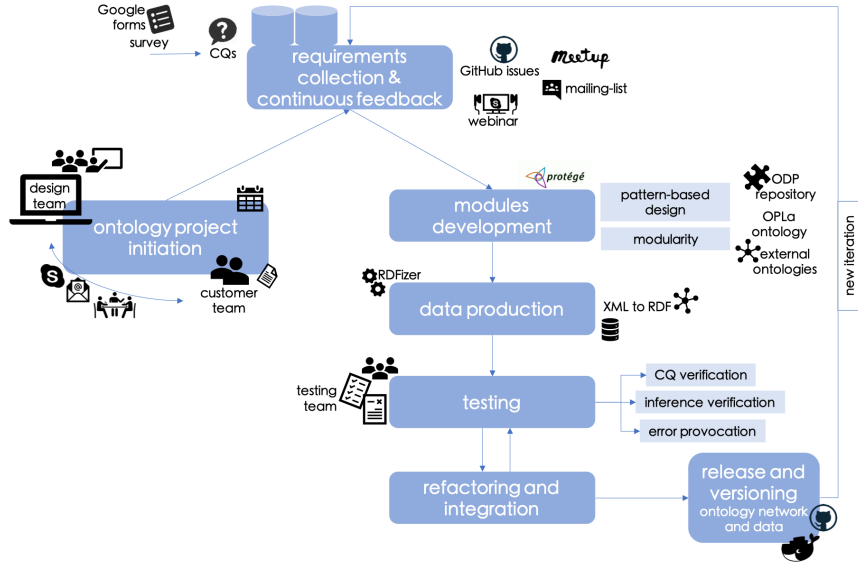


Figure 2: The XD methodology as implemented for the ArCo knowledge graph.

**Requirements collection and continuous feedback.** ArCo’s requirements are collected in the form of small stories (according to XD). They are then reformulated as **Competency Questions** (CQs, cf. [4]), and used for ODP selection by the design team, as well as in the testing phase, by the testing team (more in the remainder of this section). Stories are submitted by the customer team to a **Google Form**. In order to capture a wider perspective on requirements than the institutional and regulatory ones, we extended the customer team by involving a number of representative stakeholders such as private companies and public administrations working with CH data, in addition to the data owner (ICCD). Improvement proposals and bugs are collected as **issues through GitHub**. As ArCo is published with incremental releases (cf. Releases and versioning), the customer and the testing teams can contribute continuous and updated feedback, which allows the design team to early detect new emerging requirements and errors, and schedule them for next releases. A growing community, involving interested stakeholders and consumers, interacts *via* a dedicated **mailing-list**, as well as by participating at meetups and webinars.

**Module design.** Ontology design patterns [12] play a central role in ArCo’s design as recommended by XD. We adopt both *direct reuse* (i.e. reusing patterns from other ontologies by embedding their implementations in the local ontology) and *indirect reuse* (i.e. reusing patterns from other ontologies as templates, and adding alignment axioms to them). We reuse patterns from online repositories, e.g. **ODP portal**, and from existing ontologies, e.g. CIDOC-CRM. For details about indirect and direct reuse, the reader is invited to consult [18]. In some cases we have developed new ODPs, as in the case of modeling recurrent events. All (re)used ODPs in ArCo are annotated with OPLa ontology [13], which facilitates future reuse of ArCo as well as matching to other resources. XD encourages and supports a modular design, where each ontology module addresses a subset of requirements and covers a coherent sub-area of the domain. Therefore, ArCo ontology network consists of seven modules, each with its own namespace.

**Data production.** ArCo RDF data are produced with [RDFizer](#). Its core component is *XML2RDF Converter*, which takes two inputs: an XML file compliant with [ICCD cataloguing standards](#), and XSLT stylesheets specifying how to map XML tags to RDF. Its output is an RDF dump used to feed a triplestore.

**URI production.** Let us consider generating the URI for the author *Albert Friscia*, referring to Figure 1. ArCo base URI for individuals is <https://w3id.org/arco/resource/> with prefix `data:`. Every individual’s ID is preceded by the name of its type, e.g. `Agent`. For each type, we manually identify a set of elements that constitute a possible key (e.g. AUTN, the author’s name). We remove punctuation from the values of these elements (which are strings), convert them in lower case, concatenate them and sort them in alphabetical order, e.g. `albert-friscia`. We compute an MD5 checksum on the resulting string, which is used as the URI’s ID e.g. [data:Agent/dcd4ca7b54dd3d7dac083dd4c54a9eef](https://w3id.org/arco/resource/data:Agent/dcd4ca7b54dd3d7dac083dd4c54a9eef). Some types have a unique identifier, e.g. cultural properties. In those cases we directly use it as the URI’s ID. In order to minimise duplication, we then perform an entity linking step on the resulting individuals by using LIMES, with the same approach used for linking to external datasets (cf. Subsection 4.3).

**Testing.** To detect any incoherence in the ontologies, we regularly run a reasoner, [HermiT](#), during the modeling phase. Then, to evaluate the appropriateness of the ontologies against requirements, we follow the methodology described in [5], which focuses on testing an ontology against its requirements, intended as the ontological commitment expressed by means of CQs and domain constraints. All testing activities and resulting data (OWL files complying to the ontology described in [5]) are documented in a specific section of [ArCo’s GitHub repository](#). The testing activity is iterative, and goes in parallel with the modeling activity (XD is test-driven). The testing team performs iterative testing, applying three approaches: *CQ verification*, *inference verification*, *error provocation*. *CQ verification* consists in testing whether the ontology vocabulary allows to convert a CQ, e.g. “When was a cultural property created, and what is the source of its dating?” to a SPARQL query. *CQ verification* allows to detect any missing concept or gap in the vocabulary (e.g. whether the class for representing the source of a *dating* has been modeled). *Inference verification* focuses on checking expected inferences. For example, if a `:ComplexCulturalProperty` is defined as a `:CulturalProperty` that has one or more `:CulturalPropertyComponents`, an axiom stating that a `:CulturalProperty` has a `:CulturalPropertyComponent` would suffice to infer that the property is complex, even if it is not explicitly asserted. But if the reasoner does not infer this information, this means that the appropriate axiom (in this case an equivalence axiom) is missing from the ontology. *Error provocation* is intended to “stress” the knowledge graph by injecting inconsistent data. E.g. when characterising a `:CulturalProperty`, an individual belonging to both `a-cd:AuthorshipAttribution` and `a-cd:Dating` classes, which are supposed to be disjoint, should result as inconsistent. If the reasoner does not detect the injected error, it means that the appropriate (disjointness) axiom is missing.

**Refactoring and integration.** Problems spotted during the testing phase are passed back to the design team as issues. The design team refactors the modules and updates the ontology after performing a consistency checking. The result of this step is validated again by the testing team before including the model in the next release.

**Releases and versioning.** Incremental versions of ArCo KG are periodically and openly released. Every release has a version number, and each ontology module is marked with its own *version number* and *status*, the latter being one of: (i) *alpha* if the module has partly passed internal testing, (ii) *beta* if internal testing has been thoroughly performed, and testing based on external feedback is ongoing and partly fulfilled, (iii) *stable* when both internal and external testing have been thoroughly done. Releases are published as Docker containers on [GitHub](#) and [online](#).

## 4 ArCo Knowledge Graph

ArCo’s main component is a knowledge graph (KG), intended as the union of the ontology network and LOD data. Nevertheless, ArCo KG is released as part of a package including accompanying material (documentation, software, online services) that support its consumption, understanding and reuse. In this Section, we firstly detail what an ArCo release contains, and then we provide details about ArCo KG.

### 4.1 How to use ArCo

Each release of ArCo consists of a *docker* container available on [GitHub](#), and its running instance [online](#) - both English and Italian versions. Each release contains:

- **User guides** for supporting users in understanding the content of each release, with [Graffoo](#) diagrams and narrative explanations of every ontology module.
- **Ontologies**, including their source code and a human-readable HTML documentation created with [LODE](#).



- A **SPARQL endpoint** storing ArCo KG. The SPARQL endpoint also includes LOD data about [cultural institutes or sites](#) and [cultural events](#), extracted from “DB Unico 2.0”. We use [Lodview](#) as RDF viewer.
- **Examples of CQs** (cf. Section 3) that ArCo KG can answer, with their corresponding SPARQL queries. This helps users to have a quick understanding of what is in ArCo ontologies and data, and how to use it.
- A **RDFizer tool** converting XML data represented according to [ICCD cataloguing standards](#) to RDF complying to ArCo ontologies.

## 4.2 ArCo Ontology Network

ArCo ontology network consists of 7 ontology modules connected by `owl:imports` axioms (cf. Figure 3). Two modules – **arco** and **core** – include top-level concepts and cross-module generic relations respectively. The **catalogue** module is dedicated to catalogue records: not only do ArCo ontologies represent cultural properties, but also their ICCD catalogue records, in order to preserve the provenance and dynamics of the data. The remaining four modules (**cultural-event**, **denotative-description**, **location**, **context-description**) focus on cultural properties and their features.

The network base namespace is <https://w3id.org/arco/ontology/>, which is shared by all modules (for each module we indicate its specific namespace).

ArCo ontologies define 327 classes, 379 object properties, 154 datatype properties, 176 named individuals (at the schema level). ArCo *directly* reuses [Cultural-ON](#) and [OntoPiA](#) (the ontology network for Italian Public Administrations), while it *indirectly* reuses [CIDOC-CRM](#), [EDM](#), [BIBFRAME](#), [FRBR](#), [FaBiO](#), [FEntry](#), [OAEntry](#).

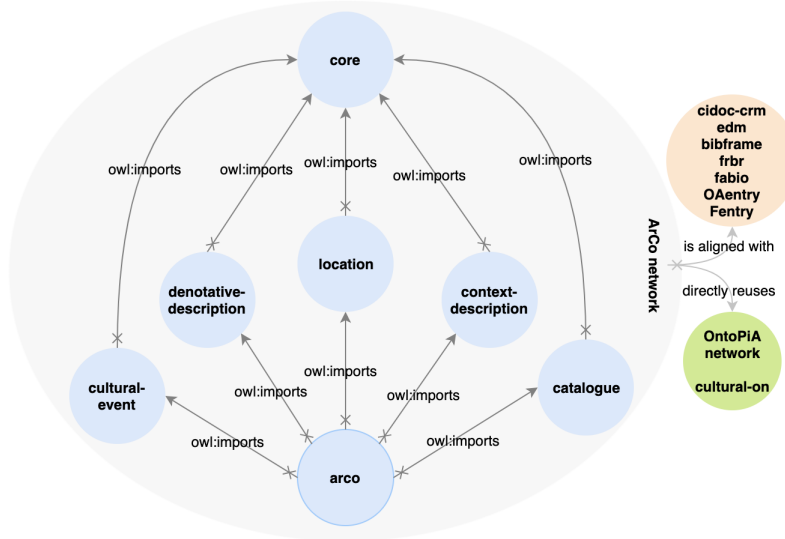


Figure 3: ArCo ontology network. Blue circles depict ArCo modules, the green circle indicates directly reused ontologies (they are embedded into ArCo), the orange circle indicates indirectly reused ontologies (some of their patterns are reused as templates, and alignment axioms are provided).

An important requirement that impacted the design of ArCo consists in expressing a same concept both with *n*-ary relation patterns that enable high-level modeling needs such as temporal indexing, state changes, model evolution, meta-classes, etc., and *shortcut* binary relations, which support lightweight modeling and intuitive navigation. With reference to Figures 4a and 4b, in order to represent the *material* of a cultural property, ArCo has: (1) a class of (reified) *n*-ary relations (`a-dd:CulturalEntityTechnicalStatus`) that include the possible `a-dd:Materials`, e.g. `data:carta` (paper), of a `:CulturalProperty`, and (2) an object property `:hasMaterial` that directly links a `:CulturalProperty` to a `:Material`, and which is defined as a property chain `[a-dd:hasTechnicalStatus 0 a-dd:includesTechnicalCharacteristic]` that makes it a shortcut of the *n*-ary relation.

In the remainder of this Section we provide details about each of the ArCo modules, with their main concepts, the reused ODPs, and the resulting Description Logic (DL) expressivity.

**The arco module** (prefix `:` and DL expressivity  $SOIQ(\mathcal{D})$ ) is the root of the network: it imports all the other modules. It formally represents top-level distinctions from the CH domain, following the definitions given in [ICCD cataloguing standards](#). The top-level class is `:CulturalProperty`, which is modeled as a *partition* of two classes: `:TangibleCulturalProperty`, e.g. a photograph, and `:IntangibleCulturalProperty` e.g. a traditional dance.

`:TangibleCulturalProperty` is further specialized in `:MovableCulturalProperty`, e.g. a painting, and `:ImmovableCulturalProperty`, e.g. an archaeological site. Additional, more specific types are defined down the hierarchy<sup>4</sup>: `:DemoEthnoAnthropologicalHeritage`, `:ArchaeologicalProperty`, `:ArchitecturalOrLandscapeHeritage`, `:HistoricOrArtisticProperty`, `:MusicHeritage`, `:NaturalHeritage`, `:NumismaticProperty`, `:PhotographicHeritage`, `:ScientificOrTechnologicalHeritage`, `:HistoricOrArtisticProperty`. Other distinctions are implemented, as in between `:ComplexCulturalProperty`, e.g. a carnival costume, consisting of an aggregate of more than one `:CulturalPropertyComponent`, e.g. hat, trousers, etc., and `:CulturalPropertyResidual`, i.e. the only residual of a cultural property, such as the handle of an amphora.

**The core module** (prefix `core:` and DL expressivity  $SHI(\mathcal{D})$ ) represents general-purpose concepts orthogonal to the whole network, which are imported by all other ontology modules. This module reuses a number of patterns, for example the [Part-of](#), the [Classification](#) and the [Situation](#) patterns.

**The catalogue module** (prefix `a-cat:` and DL expressivity  $SOIF(\mathcal{D})$ ) provides means to represent catalogue records, and link them to the cultural properties they are a record of. Different types of `a-cat:CatalogueRecord` are defined, based on the typology of cultural property they describe. `a-cat:CatalogueRecords` have `a-cat:-CatalogueRecordVersions`, which are modeled by implementing the [Sequence](#) pattern.

**The location module** (prefix `a-loc:` and DL expressivity  $SHIF(\mathcal{D})$ ) addresses spatial and geometrical information. A cultural property may have multiple locations, motivated by different perspectives: history, storage, finding, etc. Sometimes they coincide, sometimes they do not. Those perspectives are represented by the class `a-loc:-LocationType`. A certain location type of a cultural property holds during a time interval. This concept is modeled by `a-loc:TimeIndexedTypedLocation`, which implements and specialises the [Time-Indexed Situation](#) pattern. This module also defines the concept of `a-loc:CadastralIdentity` of a cultural property, e.g. the cadastral unit, in which the cultural property is located.

**The denotative description module** (prefix `a-dd:` and DL expressivity  $SOIQ(\mathcal{D})$ ) encodes the characteristics of a cultural property, as detectable and/or detected during the cataloguing process and measurable according to a reference system. Examples include measurements e.g. length, constituting materials e.g. clay, employed techniques e.g. melting, conservation status e.g. good, decent, bad.

To represent those characteristics we reuse the [Description and Situation](#) and the [Classification](#) patterns. Figure 4 shows how we model the `a-dd:CulturalEntityTechnicalStatus`, intended as a situation in which a cultural entity (e.g. a cultural property) has some of these characteristics (e.g. is square-shaped). Each characteristic is *classified* by a `a-dd:TechnicalConcept`, e.g. the `a-dd:Shape`<sup>5</sup>. These concepts are used in the `a-dd:CulturalEntityTechnicalDescription`, that is the conceptualization of the relevant technical characteristics of a cultural entity (see also the beginning of Section 4 for more details).

**The context description module** (prefix `a-cd:` and DL expressivity  $SOIQ(\mathcal{D})$ ) represents attributes that do not result from a measurement of features in a cultural property, but are associated with it. Examples include: information about authors, collectors, copyright holders; relations to other objects such as inventories, bibliography, protective measures, collections; activities such as surveys, conservation interventions; involvement in situations, e.g. commission, coin issuance, estimate, legal proceedings. In order to represent an `a-cd:ArchivalRecordSet`, i.e. fonds, series, subseries, which a cultural property is a member of, we reuse the [Born Digital Archives](#) pattern.

**The cultural event module** (prefix `a-ce:` and DL expressivity  $SOIQ(\mathcal{D})$ ) models cultural events, i.e. events involving cultural properties. It extends, with few classes and properties (e.g. `a-ce:Exhibition`), the [Cultural-ON](#)

<sup>4</sup>Cf. the [diagram](#) on Github.

<sup>5</sup>For the most common values we provide a controlled vocabulary.

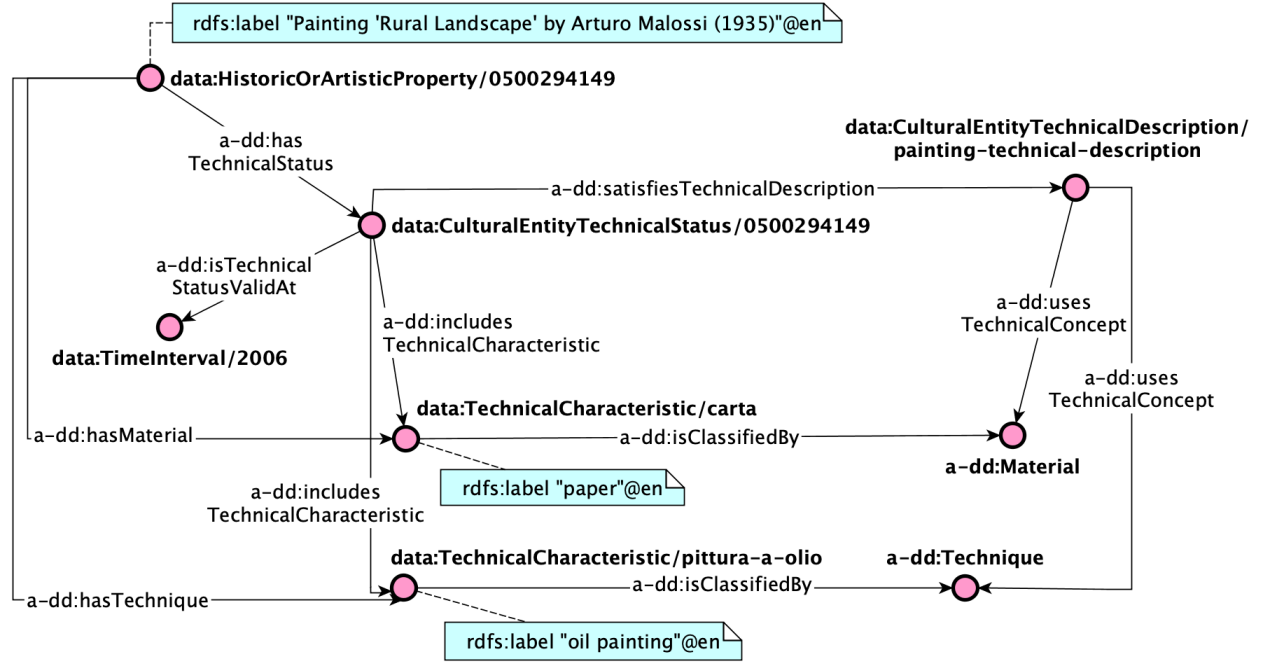
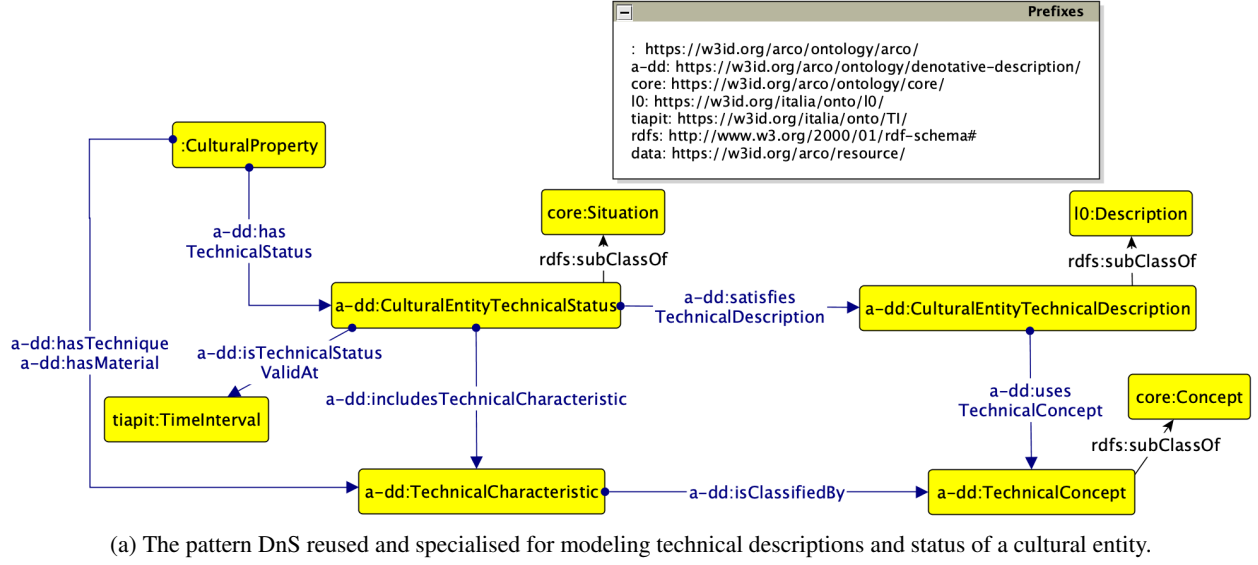


Figure 4: An example (in Graafoo notation) of pattern implementation in ArCo. The *Situation*, *Classification* and *Description* ODPs are reused as templates for modeling technical characteristics of cultural properties.

ontology. This module provides an implementation of the `a-ce:RecurrentEvent` ODP, which we have defined based on ArCo's requirements<sup>6</sup>.

<sup>6</sup>A thorough description of this new ODP is beyond the scope of this paper. It will be described in a dedicated publication, and shared on the [ODP portal](#).



### 4.3 ArCo dataset

ArCo dataset currently counts 169,151,644 triples. Table 1 gives an overview of the dataset, indicating, for the most prominent concepts defined in ArCo, the size of the corresponding extension.

ArCo dataset provides 24,008 owl:sameAs axioms linking to 18,746 distinct entities in other datasets. Link discovery is limited to authors (8,884 links) and places (9,862 links). Targets of ArCo links are: [DBpedia](#) (12,622 linked entities), [Deutsche National Bibliothek](#) (152 linked entities) [Zeri&LODE](#) (847 linked entities), [YAGO](#) (860 linked entities) [Europeana](#) (30 linked entities), [Linked ISPR](#) (598 linked entities) [Wikidata](#) (2,091 linked entities), [Geonames](#) (1,466 linked entities), and from Getty vocabularies: [ULAN](#) (13 linked entities) and [TGN](#) (67 linked entities). Entity linking is performed with [LIMES](#), configured to use a Jaccard distance computed on the `rdfs:label` literals associated with the entities. We use an extremely selective threshold (0.9 on a [0-1] range), below which candidate links are cut off. We tested lower threshold values with manual inspection on 10% of the produced links: 0.9 is the minimum to approximate 100% reliability of results. The LIMES configuration files used in the linking process are available on [Zenodo](#).

Table 1: Dataset statistics.

Metric	Result	Metric	Result
# instances of CulturalEntity	822,452	# of triples of CulturalProperty hasAuthorshipAttribution AuthorshipAttribution	1,428,018
# instances of CulturalProperty	781,902	# of instances of agents having role Author	54,204
# instances of TangibleCulturalProperty	781,900	# instances of TimeIndexedTypedLocation	1,085,521
# instances of IntangibleCulturalProperty	2	# instances of LocationType	24
# instances of MovableCulturalProperty	775,148	avg # of TimeIndexedTypedLocation per CulturalEntity	1.39
# instances of ImmovableCulturalProperty	6,752	# instances of TechnicalConcept	22
# instances of HistoricOrArtisticProperty	511,733	# instances of TechnicalCharacteristic	22,719
# instances of PhotographicHeritage	20,360	# of triples of CulturalEntity hasTechnicalStatus CulturalEntityTechnicalStatus	1,084,548
# instances of ArchaeologicalProperty	149,091	# instances of CatalogueRecord	781,902
# instances of NaturalHeritage	43,964	# instances of CatalogueRecordsVersion	1,767,376
# instances of NumismaticProperty	17,986	# of triples of CulturalProperty hasCadastralIdentity CadastralIdentity	14,683
# instances of ArchitecturalOrLandscapeHeritage	6,505	# of instances of CulturalEvent	40,331
# instances of ScientificOrTechnologicalHeritage	2,687	# instances of Organization	580
# instances of DemoEthnoAnthropologicalHeritage	29,576	avg # of Organization per CulturalProperty	5.2

## 5 Evaluation and Impact of ArCo

### 5.1 Evaluation

ArCo KG has been evaluated by following the approach used in [3], along the following dimensions: usability, logical consistency and requirements coverage.

**Usability.** The numerousness of axioms and annotations, and the use of naming conventions, gives an indication of the easiness to use an ontology and understand its commitment [11, 3]. Every ArCo ontology entity has a camel-case ID, at least one label, and one comment, both in English and Italian, and is accompanied by a detailed documentation. Many classes are also annotated with examples of usage in Turtle. The ontology contains 395 restrictions, 130 disjointness axioms, 37 alignments with 7 external ontologies. 59 classes and properties are directly reused from other ontologies.

**Semantic consistency and Requirements coverage.** We refer to Section 3 for a description of the testing phase, which allows us to assess semantic consistency and requirements coverage of ArCo. We have performed: 18 tests for inference verification, which raised 3 issues; 29 tests for provoking errors, which detected 14 cases of missing axioms. 53 CQs could be converted into SPARQL queries and provide the expected results. All issues have been fixed. In addition, we received 35 issues on GitHub, solved by the design team.

## 5.2 Potential impact of ArCo

ArCo first release is dated January 2018. Since then there is evidence of an emerging and growing community around it. A first (of a series of) webinar, attended by 10 participants, has been recently held. ArCo’s [mailing-list](#) counts 27 subscriptions and 37 threads, so far. Between beginning of January 2019 and end of March 2019, [ArCo release site](#) has been accessed 496 times by 170 distinct users. Between March 25th, 2019 and April 4th, 2019, [ArCo SPARQL endpoint](#) has been queried 1262 times and the [GitHub page](#) has had 29 unique visitors and two clones. In the last 12 months, [ArCo’s official webpage](#) had 1084 unique visitors. We are aware of at least five organisations already using ArCo in their (independent) projects. [Google Arts & Culture](#), in agreement with ICCD, is digitalising its collection of historical photographs (500.000). LOD about these pictures are modelled with ArCo and are ingested by Google Arts & Culture from ICCD SPARQL endpoint. [Regesta.exe](#)<sup>7</sup> uses ArCo for publishing LOD about artworks owned by [IBC-ER](#). [Synapta](#)<sup>8</sup> reuses ArCo ontologies for representing musical instruments belonging to [Sound Archives & Musical Instruments Collection](#) (SAMIC); [OnData](#) works on linking data about areas of Italy hit by the earthquakes in 2016 to ArCo’s data in the context of the project [Ricostruzione Trasparente](#). [InnovaPuglia](#) is extending its ontologies with, and linking its [LOD](#) to, ArCo KG.

## 6 ArCo: availability, sustainability, and licensing

**Availability.** ArCo namespaces are introduced in Section 4. We create permanent URIs with the [W3C Permanent Identifier Community Group](#). ArCo KG is available through [MiBAC’s official portal](#) and [SPARQL endpoint](#), and on [GitHub](#) (cf. Subsection 4.1). Its ontology modules are indexed by, and can be retrieved from, [Linked Open Vocabulary](#) (LOV). Additionally, ArCo is published on [Zenodo](#), which provides its DOI [10.5281/zenodo.2630447](#). ArCo’s [community channel](#) on Zenodo aggregates all its material (data, experiment configurations, results, etc.).

**ArCo sustainability** is guaranteed by MiBAC’s commitment to maintain and evolve ArCo, by following the XD methodology. In addition, CNR is committed to support and collaborate with MiBAC, based on their long-term and established collaboration as well as their shared objectives on this matter. ICCD’s experts received guidelines and training for maintaining ArCo KG and for using the software for producing LOD. The docker on GitHub and its running instance online will be also maintained. In addition to the institutional commitment, there is an active community growing around ArCo (additional information in Section 5), which contributes with both new requirements, model extensions, alignments, etc.

**Versioning and licensing.** ArCo is under version control on a public [GitHub repository](#). ArCo KG license is [Attribution-ShareAlike 4.0 International](#) (CC BY-SA 4.0).

## 7 Cultural Heritage and Knowledge Graphs: related work

Semantic technologies, in particular ontologies and LOD, are widely adopted within the CH domain for facilitating researchers, practitioners and generic users to study and consume cultural objects. Notable examples include: the [EDM](#) and its [datasets](#), the [CIDOC-CRM](#), the [Rijksmuseum collection](#) [9], the [Zeri Photo Archive](#) [8], the [Getty Vocabularies](#), the [IBC-ER](#), the [Smithsonian Art Museum](#), the [LODLAM](#), the [OpenGLAM](#), the [Google Arts & Culture](#). ArCo substantially enriches the existing LOD CH cloud with invaluable data on the Italian CH and a network of ontologies addressing overlooked modelling issues.

Relevant related research discusses good practices for developing ontologies and LOD for CH [6, 19, 1, 14, 9]. ArCo draws from these lessons learnt, as well as from good practices in pattern-based ontology engineering [12].

There are commonalities between CIDOC-CRM, EDM and ArCo. Nevertheless, CIDOC-CRM and EDM resulted insufficient against the requirements that ArCo needs to address. For example, modeling the diagnosis of a paleopathology in anthropological material, the coin issuance, the Hornbostel-Sachs classification of musical instruments, etc., all motivate the need for developing extended ontologies for representing CH properties. To further support this claim we compute the terminological coverage of EDM and CIDOC-CRM against ArCo CQs (cf. Section 3), and compare it with ArCo’s. We model this task as an ontology matching task between an RDF vocabulary representing ArCo’s CQs, and the ontologies being tested. The coverage measure is computed as the percentage of matched entities. We use Sketch Engine [16] to extract a reference vocabulary of keywords from ArCo’s CQs. The result is an [RDF vocabulary of 66 terms](#). Ontology matching is performed with Silk [20] by using the *substring* metric with 0.5 as

<sup>7</sup>Their blog post on ArCo: [Regesta’s blog](#).

<sup>8</sup>Their blog post on ArCo: [Synapta’s blog](#).

threshold. The result shows the following coverage values (0: no coverage, 1: coverage): ArCo 0.68, CIDOC-CRM 0.29, EDM 0.12. However, ArCo ontologies are aligned to (i.e. indirectly reuse) CIDOC-CRM and EDM [15], as well as to BIBFRAME, FRBR and FaBiO (for bibliographic data), and to FEntry and OAEntry (dedicated to photographs and artworks). Directly reused ontologies include: Cultural-ON (Cultural ONtology), which models Italian cultural institutes, sites, and cultural events [17], and is maintained by MiBAC; OntoPiA, a network of ontologies and controlled vocabularies, based on DUL patterns, which model top-level information crossing different domains (e.g. People, Organisation, Location) and recommended as a standard by MiBAC<sup>9</sup>.

## 8 Conclusion and ongoing work

This paper presented ArCo: the knowledge graph of Italian Cultural Heritage (CH), encoded and published as linked open data (LOD). ArCo is a robust Semantic Web resource, nevertheless it is an evolving creature. As such, it can be further improved and enriched in many ways.

Concerning identity, our URI production strategy may produce possible ambiguous identifiers, i.e. same URI for different entities of the world. Because involved entities mainly represent Italian authors and organisations (cultural properties are uniquely identified, places have robust keys), we expect a small number of ambiguous cases as compared to the dimension of the dataset. Nevertheless, we are currently working on spotting them by applying a set of heuristics and validating them with the help of experts. We are also experimenting different techniques for improving external linking, including LIMES’ machine learning modules, key- and linkkey-based interlinking methods [2], and crowdsourcing (involving experts) for both validation and enrichment.

There are aspects that are yet to be modelled: for example some specific characteristics of naturalistic heritage, e.g. slides and phials associated to an *herbarium*, or the optical properties of a stone, to name a few. ArCo development includes associating pictures to each cultural entity, which are available, but not in the dataset yet. Additional effort must be put to complete the translation of the data to other languages. At the moment, the data are expressed in Italian as from the Catalogue, 17,906,639 entities have an English label. A first step is to complete the English translation. Although this is a task to be performed by experts, we are considering supporting them with automatic translation. In order to facilitate reuse of ArCo ontologies, we plan to develop additional tool support for CH data owners, and to address requirements coming from the library and archive domains.

## References

- [1] Chris J. van Aart et al. “Mobile Cultural Heritage Guide: Location-Aware Semantic Search”. In: *Proc. of EKAW 2010*. (Lisbon, Portugal). Lecture Notes in Computer Science. Springer, 2010, pp. 257–271.
- [2] Manuel Atencia et al. “Data interlinking through robust linkkey extraction”. In: *Proc. of ECAI 2014*. (Prague, Czech Republic). IOS Press, 2014, pp. 15–20.
- [3] Eva Blomqvist et al. “Experiments on Pattern-Based Ontology Design”. In: *K-CAP 2009*. ACM, 2009.
- [4] Eva Blomqvist et al. “Experimenting with eXtreme Design”. In: *Proc. of EKAW 2010*. (Lisbon, Portugal). Vol. 6317. Lecture Notes in Computer Science. Springer, 2010, pp. 120–134.
- [5] Eva Blomqvist et al. “Ontology Testing - Methodology and Tool”. In: *Proc. of EKAW 2012*. (Galway City, Ireland). Springer, 2012, pp. 216–226.
- [6] Victor de Boer et al. “Amsterdam Museum Linked Open Data”. In: *Semantic Web 4.3* (2013), pp. 237–243.
- [7] Piero Andrea Bonatti et al. “Knowledge Graphs: New Directions for Knowledge Representation on the Semantic Web (Dagstuhl Seminar 18371)”. In: *Dagstuhl Reports* 8.9 (2019), pp. 29–111.
- [8] Marilena Daquino et al. “Enhancing Semantic Expressivity in the Cultural Heritage Domain: Exposing the Zeri Photo Archive as Linked Open Data”. In: *JOCCH* 10.4 (2017), 21:1–21:21.
- [9] Chris Dijkshoorn et al. “Modeling cultural heritage data for online publication”. In: *Applied Ontology* 13.4 (2018), pp. 255–271.
- [10] Martin Doerr. “The CIDOC Conceptual Reference Module: An Ontological Approach to Semantic Interoperability of Metadata”. In: *AI Magazine* 24.3 (2003), pp. 75–92.
- [11] Aldo Gangemi et al. “Modelling Ontology Evaluation and Validation”. In: *Proc. of ESWC 2006*. (Budva, Montenegro). Springer, 2006, pp. 140–154.
- [12] Pascal Hitzler et al., eds. *Ontology Engineering with Ontology Design Patterns - Foundations and Applications*. Vol. 25. Studies on the Semantic Web. IOS Press, 2016.

<sup>9</sup>OntoPiA is a *de facto* standard for open data of the Italian Public Administration.

- [13] Pascal Hitzler et al. “Towards a Simple but Useful Ontology Design Pattern Representation Language”. In: *Proc. of WOP 2017*. (Vienna, Austria). 2017.
- [14] Eero Hyvönen. “Semantic Portals for Cultural Heritage”. In: *Handbook on Ontologies*. Ed. by Steffen Staab et al. International Handbooks on Information Systems. Springer, 2009, pp. 757–778.
- [15] Antoine Isaac et al. “Europeana Linked Open Data - data.europeana.eu”. In: *Semantic Web 4.3* (2013), pp. 291–297.
- [16] Adam Kilgarrieff et al. “Itri-04-08 the sketch engine”. In: *Information Technology* 105 (2004), p. 116.
- [17] Giorgia Lodi et al. “Semantic Web for Cultural Heritage Valorisation”. In: *Data Analytics in Digital Humanities*. Springer, 2017, pp. 3–37.
- [18] Valentina Presutti et al. “The Role of Ontology Design Patterns in Linked Data Projects”. In: *Proc. of ER 2016*. (Gifu, Japan). Springer, 2016, pp. 113–121.
- [19] Pedro A. Szekely et al. “Connecting the Smithsonian American Art Museum to the Linked Data Cloud”. In: *Proc. of ESWC 2013*. (Montpellier, France). Springer, 2013, pp. 593–607.
- [20] Julius Volz et al. “Silk-a link discovery framework for the web of data.” In: *LDOW 2009*. Vol. 538. CEUR-ws, 2009.