AGDLI: ArCo, GVP and DBpedia Linking Initiative

Stefano Faralli[1][0000−0003−3684−8815], Andrea Lenzi[2][0000−0002−8997−9862], and Paola Velardi[2][0000−0003−0884−1499]

[1] University of Rome UnitelmaSapienza, Italy
stefano.faralli@unitelmasapienza.it
[2] Sapienza University of Rome, Italy
{lenzi,velardi}@di.uniroma1.it

Abstract. We present the *ArCo*, *GVP and DBpedia Linking Initiative* (AGDLI), a research activity within the project *SMARTOUR: intelligent platforms for tourism*, funded by the Italian Ministry of University and Research. Our initiative is aimed at linking *ArCo*'s cultural entities to the well known *Getty Vocabulary Program* and DBpedia ontologies, with the main goal of providing a semantically rich representation of the Italian cultural heritage for tourism-related knowledge-based applications. In this paper we provide a detailed description of the initiative and describe the current research developments and outcomes.

Keywords: ArCo · Getty Vocabularies · DBpedia · knowledge-based applications

1 Introduction

Nowadays, we are observing an increasing number of novel semantically-enabled and knowledge-based applications. Hence, Linked Open Data are more and more gaining the attention from public administrations and industries all over the world. In this paper[3], we describe the *ArCo, GVP and DBpedia Linking Initiative* (*AGDLI*). Our initiative is a research activity part of the *SMARTOUR: intelligent platform for tourism* project (see Section Acknowledgements). The main goal of the initiative is to study semi-supervised methodologies to generate semantically rich definitions of Italian cultural heritage entities, to be used in different knowledge-based tourism related applications, such as recommender systems [6] and semantically-enriched augmented reality tools for point of interests discovery [5]. To this end, we decided to link the entities defined in *ArCo*[4] [2] with the concepts defined in the *Getty Vocabulary Program*[5] (*GVP*) [4] and

[4] http://wit.istc.cnr.it/arco/?lang=en.

[5] https://www.getty.edu/research/tools/vocabularies/.

DBpedia [1][6] ontologies. *ArCo* is a state-of-the art knowledge graph of the Italian cultural heritage, which defines 169 million triples describing 820 thousand cultural entities. In *ArCo* (see Figure 1), important properties - such as the type (*dc:type*) - are valued with literals or not linked with existing ontologies e.g. authorship attributions (*I0:Agent*). The GVP is a top ontology on which the *Art & Architecture Thesaurus®* (*AAT*), the *Getty Thesaurus of Geographic Names®* (*TGN*), and the *Union List of Artist Names®* (*ULAN*) vocabularies are based on. *AAT*, *TGN* and *ULAN* vocabularies provide semantic definitions for concepts useful for cataloging, documenting and retrieving information related to art, architecture, and other material culture. By targeting both the *GVP* and *DBpedia* ontologies, we can generate, with high coverage, links for *ArCo* entities and their properties. This may considerably enrich the *ArCo* ontology, which currently defines only 14 high-level classes with a depth of 4, while e.g. the *AAT* ontology provides more than 55K domain specific concepts divided in 8 facet taxonomies with an average height of 13 levels. In this paper, we provide a description of the current research outcomes and future work of the *AGDLI* initiative.

2 The linking initiative

In Figure 1, we depict an excerpt of the *ArCo* schema. In this diagram we highlight some of the properties of *ArCo CulturalProperty* entities that we link to the *GVP* ontology. Specifically, we are investigating semi-supervised methodologies to automatically: i) mine and link the *dc:type* and *rdfs:label* properties of *CulturalProperty* instances to the *AAT*; ii) link the cities of the addresses of *CulturalProperty* instances to the *TGN*; iii) link the agents of authorship attributions of *CulturalProperty* instances to the *ULAN*; iv) normalize the date intervals of *CulturalProperty* instances into a machine readable format[7], such as the *Open Date Range Format*[8]. We note that *ArCo* is in Italian, while the *GVP* is mainly in English, which represents an additional challenge of our linking initiative. In Figure 2, we depict an example of mining concepts from *ArCo* entities' textual descriptions and linking them to corresponding concepts in the *AAT*. In this task:

1. we automatically translated from English to Italian the AAT terms. To this end, we used the Google Translate API[9]. Note that, we preserved the original Italian terminology when already provided by the AAT;

[6] https://www.dbpedia.org/.

[7] This initiative's aim is intended to provide a ready to use resource for time-based tourism applications.

[8] *Dublin Core Collection Description: Open Date Range Format* http://www.ukoln.ac.uk/metadata/dcmi/date-dccd-odrf/.

[9] https://cloud.google.com/translate/.

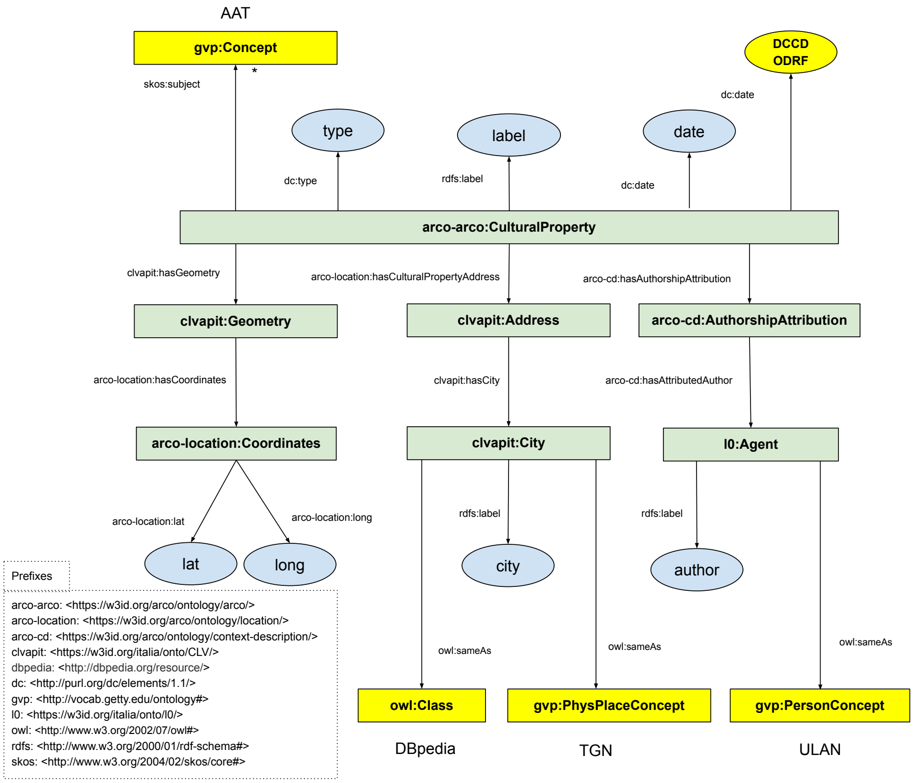
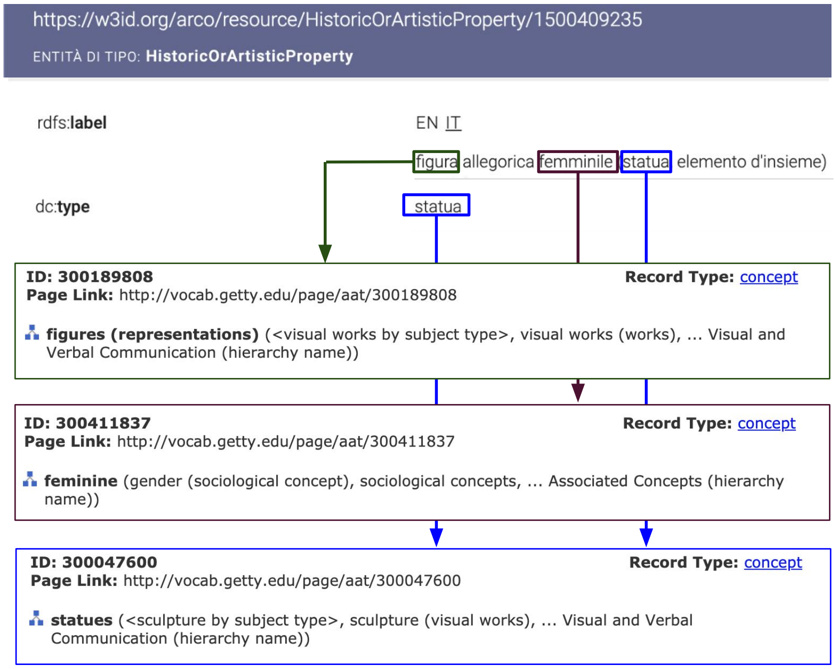

AAT
gvp:Concept
*
skos:subject
type
dc:type
label
rdfs:label
date
dc:date
DCCD
ODRF
dc:date
arco-arco:CulturalProperty
clvapit:hasGeometry
arco-location:hasCulturalPropertyAddress
arco-cd:hasAuthorshipAttribution
clvapit:Geometry
clvapit:Address
arco-cd:AuthorshipAttribution
arco-location:hasCoordinates
clvapit:hasCity
arco-cd:hasAttributedAuthor
arco-location:Coordinates
clvapit:City
l0:Agent
arco-location:lat
arco-location:long
lat
long
rdfs:label
city
rdfs:label
author
owl:sameAs
owl:sameAs
owl:sameAs
owl:Class
gvp:PhysPlaceConcept
gvp:PersonConcept
DBpedia
TGN
ULAN
Prefixes
arco-arco: <https://w3id.org/arco/ontology/arco/>
arco-location: <https://w3id.org/arco/ontology/location/>
arco-cd: <https://w3id.org/arco/ontology/context-description/>
clvapit: <https://w3id.org/italia/onto/CLV/>
dbpedia: <http://dbpedia.org/resource/>
dc: <http://purl.org/dc/elements/1.1/>
gvp: <http://vocab.getty.edu/ontology#>
l0: <https://w3id.org/italia/onto/l0/>
owl: <http://www.w3.org/2002/07/owl#>
rdfs: <http://www.w3.org/2000/01/rdf-schema#>
skos: <http://www.w3.org/2004/02/skos/core#>

Fig. 1. A diagram representing an excerpt of *ArCo* ontology (green boxes) and the links to external classes and properties (yellow boxes) our initiative is aimed to generate.

2. we applied standard text pre-processing techniques (e.g., tokenization, lowercasing) to *ArCo* entities' textual descriptions. To this end, we adopted the Stanford NLP API[10];
3. we automatically collected all the occurrences of Italian *AAT* terms in *ArCo* entities' *rdfs:label* and *dc:type* resulting preprocessed textual properties. In this step, for each ArCo's entity, we obtained a collection of links with AAT concepts. As a result, we obtained a collection of ambiguous links to all the AAT concepts having the same *skosxl:literalForm*[11] (see the example of as described in Figure 2,).
4. since these tasks are error-prone, we performed a manual refinement of the translated *AAT*'s terms, fixing translation errors and adding synonyms, singular, plural and hypernymous forms for terms occurring in the the textual properties of missing linked ArCo's entities;
5. we repeated steps 3 and 4 until an adequate coverage was reached.

[10] https://nlp.stanford.edu/software/.
[11] https://www.w3.org/TR/skos-reference/skos-xl.html#literalForm.



https://w3id.org/arco/resource/HistoricOrArtisticProperty/1500409235

ENTITÀ DI TIPO: **HistoricOrArtisticProperty**

rdfs:**label** EN IT

figura allegorica femminile (statua elemento d'insieme)

dc:**type** statua

ID: 300189808 **Record Type:** concept
Page Link: http://vocab.getty.edu/page/aat/300189808

figures (representations) (<visual works by subject type>, visual works (works), ... Visual and Verbal Communication (hierarchy name))

ID: 300411837 **Record Type:** concept
Page Link: http://vocab.getty.edu/page/aat/300411837

feminine (gender (sociological concept), sociological concepts, ... Associated Concepts (hierarchy name))

ID: 300047600 **Record Type:** concept
Page Link: http://vocab.getty.edu/page/aat/300047600

statues (<sculpture by subject type>, sculpture (visual works), ... Visual and Verbal Communication (hierarchy name))

Fig. 2. An example of automatically mined and linked concepts form the *rdfs:label* and the *dc:type* properties of the *arco-arco:CulturalProperty* described at https://w3id.org/arco/resource/HistoricOrArtisticProperty/1500409235. Note that both the singular form Italian words "figura" (*figure*) and "statua" (*statue*) were correctly linked to the corresponding English plural forms "figures" and "statues".

To link *ArCo*'s cities to *TGN* and *DBpedia* we performed string matching[12] with the corresponding terms and entities. At the time of writing, we are investigating on effective linking methodologies of ArCo's *I0:Agents* with *ULAN* entities, and on *dc:date* normalization.

3 Current Outcomes and Conclusions

In this paper, we introduced the *AGDLI* initiative. As a result, we obtained[13]:

- the automatic translation in Italian of the 55K *AAT* terms;
- a total of 5.6 M triples (*skos:relatedMatch* and *skos:related*) linking the 98.2% (by *dc:type*) and the 99.9% (by *rdfs:label*) of *arco-arco:CulturalProperty* entities to candidate *AAT* concepts;
- a total of 6.6 K triples (skos:relatedMatch) linking the 86.3% of *clvapit:City* instances to candidate *TGN* entities; iv) 4.7 K novel *owl:sameAs* relations, now linking the 100% *clvapit:City* to DBpedia.

[12] We applied different similarity measures e.g., string edit distance-based similarity.

[13] Resources are available under Creative Commons Attribution 4.0 International (CC BY 4.0) at https://sites.google.com/unitelmasapienza.it/agdli/.

As already introduced in Section 2, the next planned activities are aimed at both linking *ArCo*'s authorship attributions to *ULAN* entities and normalizing the *CulturalProperty*'s *dc:date*.

Moreover, we are planning to apply semi-supervised methodologies for the disambiguation of the generated candidate links. For instance, generated links to AAT concepts can be refined with semi-supervised word sense disambiguation approaches, while the generated matches with *TGN* candidates can be disambiguated based on the distance between the geographical coordinates of *ArCo* and *TGN* entities.

Further plans of the *AGDLI* initiative include, among others, the application and investigation of knowledge graph completion methodologies [3] to link isolated (unmatched) entities of the resulting graph, and the adoption of best practices for continuous resource maintenance and deployment.

Acknowledgements

This work was carried out within the research project ”SMARTOUR: intelligent platform for tourism” funded by the Italian Ministry of University and Research with the Regional Development Fund of European Union (PON Research and Competitiveness 2007-2013).

References

1. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: Dbpedia: A nucleus for a web of open data. In: The Semantic Web. pp. 722–735. Springer Berlin Heidelberg, Berlin, Heidelberg (2007). https://doi.org/10.1007/978-3-540-76298-0_52
2. Carriero, V.A., Gangemi, A., Mancinelli, M.L., Marinucci, L., Nuzzolese, A.G., Presutti, V., Veninata, C.: Arco: The italian cultural heritage knowledge graph. In: The Semantic Web - ISWC 2019 - 18th International Semantic Web Conference, Auckland, New Zealand, October 26-30, 2019, Proceedings, Part II. Lecture Notes in Computer Science, vol. 11779, pp. 36–52. Springer (2019). https://doi.org/10.1007/978-3-030-30796-7_3
3. Chen, Z., Wang, Y., Zhao, B., Cheng, J., Zhao, X., Duan, Z.: Knowledge graph completion: A review. IEEE Access **8**, 192435–192456 (2020). https://doi.org/10.1109/ACCESS.2020.3030076
4. Harpring, P.: Development of the getty vocabularies: Aat, tgn, ulan, and cona. Art Documentation: Journal of the Art Libraries Society of North America **29**(1), 67–72 (2010), http://www.jstor.org/stable/27949541
5. Ruta, M., Scioscia, F., De Filippis, D., Ieva, S., Binetti, M., Di Sciascio, E.: A semantic-enhanced augmented reality tool for openstreetmap poi discovery. Transportation Research Procedia **3**, 479–488 (2014). https://doi.org/https://doi.org/10.1016/j.trpro.2014.10.029, https://www.sciencedirect.com/science/article/pii/S2352146514001926, 17th Meeting of the EURO Working Group on Transportation, EWGT2014, 2-4 July 2014, Sevilla, Spain
6. Zhang, Q., Lu, J., Jin, Y.: Artificial intelligence in recommender systems. Complex & Intelligent Systems **7**(1), 439–457 (Feb 2021)