Toward a Semantic Representation of the Joconde Database

Jean-Claude Moissinac, François Rouzé, Piyush Wadhera, Bastien Germain

Toward a Semantic Representation of the Joconde Database

Jean-Claude Moissinac
LTCI, Télécom Paris
Institut polytechnique de Paris
19 Place Marguerite Perey,
91120 Palaiseau
France
Email: moissinac@enst.fr

François Rouzé
Independant researcher
France
Email: francois.rouze
@gmail.com

Piyush Wadhera
and Bastien Germain
Réciproque
12 Rue Saint-Maur, 75011 Paris
France
Email: {piyush.wadhera,bastien.germain}
@reciproque.com

***Abstract*—The Joconde database is a French database, which describes about 600,000 works from French art collections. In the *Data&Musée* project, we process data from museums and monuments. We have chosen to model the data using a knowledge graph approach. We enrich the data of the project partners with data from other sources. In this article, we present the semantic representation that we have adopted for the Joconde database and the methods used to obtain this representation. Our semantic representation of the Joconde database is available as Open data as the SemJoconde dataset. We believe that the SemJoconde data can become useful references for work on the use of semantic techniques in the cultural field.**

Keywords–RDF; semantic web; culture; SemJoconde; cultural heritage; LOD.

I. Introduction

This paper introduces a novel dataset named SemJoconde that contains a large number of artworks. This dataset is published as Open Data. This dataset is produced from the Joconde database, of which we have generated an enriched semantic version. We present the dataset and the methods used to obtain this representation. We define a semantic model based on CIDOC-CRM -Conceptual Reference Model- and interlink as many entities as possible to Wikidata [1]. Wikidata is a large semantic dataset about world things, linked to Wikipedia pages. Links with Wikidata are created for creators, domains, places, etc.

This work is part of the *Data&Musée* project [2], in which we process data from museums and monuments. The goal is to reply to questions like: is not a visitor to the Louvre also a visitor to the Eiffel Tower? Better still, a visitor who is satisfied with his Middle Ages journey at the Louvre, isn't he a future visitor to the ramparts and the old town of Carcassonne? So, beside collecting data about the visitors, we are collecting knowledges about the artworks and cultural institutions in France. Building the SemJoconde dataset is part of this process.

This paper follows some works related to semantic representation of data in the cultural heritage domain [3][4], which are generally limited to represent the collection of an unique collection, except Europeana [5]. Our contribution is the dataset itself rather than novel methods, which are mainly simples ways to get entity linking [6][7]. In this article, we present the model, and the process to translate from the JSON version of the database to the semantic interlinked version. We think that it will be useful for communities in the graph technologies domain -graph embedding, reasoning, etc.- and in the cultural heritage domain. Section II presents related works. Section III presents sources used to build SemJoconde. Section V presents the methods used to build SemJoconde and some insight to evaluate the quality of the results. Section VI gives an idea of the technical structure of the dataset. Section II-C presents related datasets. Section VII concludes and suggests future works.

II. Related works

A. Entity Matching and Entity Linking

Several part of our work deal with entity matching: we search for entities in text [6][7]. The problem is composed of entity recognition, entity disambiguation and entity linking. In our case, the searched entity is known by its label (e.g., Claude Monet) and, sometimes, some complementary data (e.g., period of the work) and we expect to produce a link/URI - Uniform Resource Identifier- in some Linked Data dataset. It is a well-known problem with different approaches proposed depending on each context.

In our work, the problem is simplified by the fact that we do not need to recognize the entities and the entities types in a text: we need only to search for identifiers corresponding to labels for which we know the type. We tried different approaches to produce the links and the simple approach presented in this article gives good results.

B. Production and applications of Cultural Heritage datasets

In this section, we will present previous works about the build process of semantically structured datasets in the cultural heritage domain.

The Getty Foundation has a knowledge graph about its collections. The Foundation described in detail the choices about vocabularies and ontologies used by the Knowledge graph and the process of building it [3]. The Getty Foundation proposes a list of vocabularies and entities using these vocabularies [3]: specifically Art Architecture Thesaurus (AAT) and Union List of Artist Names (ULAN). Both AAT and ULAN are thesauri containing structured terminology for art, architecture, decorative arts, archival materials, visual surrogates, conservation, and bibliographic materials. A very interesting document explains how the foundation build the vocabularies (see below).

The foundation uses the Open Data Commons Attribution License. These vocabularies comply with thesaurus construction standards (NISO- National Information Standards Organization and ISO-International Organization for Standardization), and are developed through contributions from the user community, complied and disseminated by the Getty Vocabulary Program and Getty Digital, and released finally in XML, JSON,RDF N-Triples and via a Sparql endpoint.

In 2012, the Amsterdam Museum published a work on the use of linked data. They start from XML data and present the process of converting the data to linked data with "man in the middle" [8].

In [9], the authors present their approach to build a service for converting legacy data into linked data. They focus on the problems resulting from heterogeneity of the sources, which is not a problem for SemJoconde: we have only one source.

Rijks Museum is one of the first major museums to publish data about its collections according to the principles of Linked Open Data [4]. The Rijks Museum has published a Linked Open Data -LOD- dataset with more than 350000 objects in the version of March 2016. In [4], the authors explain their approach, which is the result of several successive projects. So, the result benefits from a progressive consolidation.

JocondeLab [10] is a French project, which worked on a semantic model to get semantic representation of the Joconde database. To our knowledge, the representation obtained by JocondeLab is not available in Open Data, nor based on CIDOC-CRM.

Globally, we observe that more and more cultural institutions are considering Linked Open Data as a value for the future of their collections and their visitors.

C. Other related datasets

In this section, we present some significant datasets for our projects. As SemJoconde, several are based on CIDOC-CRM. Others are sources of inspiration or candidates for useful links, in the spirit of Linked Open Data.

The Europeana project [5] produces an aggregation of different sources of European cultural content - librairies, archives; audiovisual collections, theme-based content, as well as regional and national aggregators. Europeana follows the rules of Linked Open Data (LOD). As for SemJoconde, the schema is largely layered over the CIDOC-CRM model and includes concepts from ORE and Dublin core as well. The EDM -Europeana Data Model- is a flexible data model that combines object-centric, contextual and event-centric approaches to data representation. It uses URIs for addressing accessible resources.

The British Museum dataset [11] is organised using the CIDOC-CRM model, with the objective of harmonising with other international cultural heritage data. Although based on a linked data service, dataset licensing combines Creative Commons, and BM Licensing for 3D and HD content. Linked data is available in RDF and via a SPARQL Endpoint.

Paris Musées Collections [12] is a dataset of artworks curated by the members of the Paris Musées consortium. The dataset enrichment was an OpenData project executed in 2019-2020, but no Linked Data enrichment is available. Most data is open access (Creative Commons CC0), there is licensed HD and 3D content, as well as some specific licensed content. Data is available through an API, and dissemination on Wikimedia commons and Europeana is in the process. This dataset, as Joconde, is a source for our project Data&Musée. We have modeled part of these works with the CIDOC-CRM model.

DataTourisme [13] is a French LOD project regarding touristic offer and points of interest. The ontology supports Schema and Dublin Core vocabularies amongst others. It is a source of useful links, mainly for practical data about museums and monuments, but also about point of interest around them.

Geonames [14] proposes a massive list of geographical entities with their coordinates using the WGS84 latitude longitude system (World Geodetic System, 1984) and some other data about these entities: administrative links, country, etc. The dataset is collaborative and allows contributions using a wiki interface. It is available under a Creative commons licence. The data is accessible in a zip file and through webservices.

DBPedia [15] is a large dataset of entities based on Wikipedia data. It is a community effort to extract structured information from Wikipedia and to make this information available on the Web. DBpedia allows you to ask sophisticated queries. DBPedia is interlinked with a lot of other datasets. A RDF dump is available, and queries can be send to a SPARQL endpoint. DBpedia-Fr is similar and build from the french Wikipedia.

Wikidata [1] is a large dataset of world things linked to Wikipedia pages. Wikidata is a project of the Wikimedia foundation. As Wikidata offered the best coverage of the museums and monuments partners in the *Data&Musée* project, we privilege links with Wikidata. Wikidata allows you to ask sophisticated queries and to link other datasets on the Web and to Wikipedia. Similar to Wikipedia (creative commons) RDF, SPARQL as well as semantic web sitemaps are available to obtain the data. The RDF data is structured in N-Triples.

Yago [16] is a semantic knowledge base derived from Wikipedia, WordNet and GeoNames. Its specificity comes from the accuracy scores that have been manually attached to the data. The data and resources are available in many formats including RDF and TSV.

III. Sources

In this section, we describe the data sources that allowed us to build SemJoconde.

A. Joconde database

The Joconde database describes 589,278 works of art from French collections. It is established by the French Ministry of Culture. An extraction was made available in Open Data via the Open platform for French public data [17]. It is available in several formats including JSON. It is the extraction in this format that we used. An open license allowing free reuse is associated with this data.

Each Joconde database record has 14 fields

- 'STAT': status of the work: owner, place, etc.; for example: "propriété de la commune ; achat ; Château-Thierry ; musée Jean de La Fontaine",
- 'EPOQ': eras associated with the work; for example: "Paléolithique" or "Qing (1644-1911)",

- 'DOMN': fields associated with the work; for example: "dinanderie" or "Néolithique" or "photographie",
- 'INV' : an inventory number,
- 'TECH': techniques used by the work; for example: "matière plastique (moulé, imprimé)" (plastic (molded, printed)),
- 'DIMS': dimensions of the work; for example: "H. 27 ; l. 6.1 ; P. 4.2",
- 'LOCA': place of conservation and / or exhibition of the work; for example: "Grenoble ; musée Stendhal",
- 'DENO': object types; for example: "silex" (flint) or "tombeau" (tomb),
- 'TITR': title associated with the work (a simple string),
- 'AUTR': creators of the work; for example: "RODIN Auguste"
- 'DECV': elements concerning the discovery of the work; not used in this article,
- 'COPY': always '© Direction des musées de France',
- 'REF' : a unique identifier for the work; for example: 'AE037477',
- 'PERI': periods associated with the work; for example: 2e quart 20e siècle

B. Analysis

For some fields, we analyze the dataset. The goal is to find the values used for these fields and the count of works associated with each value of each field. Table I shows in the 'Dataset' column the count of values for the fields: AUTR, DOMN, DENO, LOCA, EPOQ, PERI.

C. Wikidata alignment and ground truth

In this work, we favor a mapping between Joconde vocabulary and Wikidata.

Thanks to the project WikiProject Vocabulaires Joconde [18], we have a ground truth. In this project, volunteers try to link manually the Joconde vocabulary with Wikidata. They use some tools to help humans to produce and validate such links. Links are notably available for creators, domains, places, epochs, periods, techniques.

TABLE I. GROUND TRUTH (14/7/2020).

Category (field)	*Validated*	*Dataset*	*%*
Creators (AUTR)	2560	37828	6.7
Domains (DOMN)	168	168	100.
Object types (DENO)	77	5766	1.3
Places (LOCA)	35	3593	0.9
Epochs (EPOQ)	500	831	60.1
Periods (PERI)	60	346	17.3

Table I shows the state of the ground truth at 14/7/2020. Corresponding files are available on github (and other files related to this article) [19].

IV. SEMANTIC MODEL

We have chosen to rely on the CIDOC-CRM model for our different representations. The CIDOC Conceptual Reference Model (CRM) [20] is a theoretical and practical tool for information integration in the field of cultural heritage. This model is massively used in the cultural heritage domain [21]. For example, Europeana (see Section II-C) uses CIDOC-CRM as a base for its Europeana Data Model (EDM).

Figure 1 shows the model used to represent the works. Properties starting with P and concepts starting with E followed by a number and text, such as P65_is_shown_by and E65_Creation, are properties or concepts defined by CIDOC-CRM. Properties starting with DMP -for Data Musée Property- followed by a number and text, like DMP2_has_description, are defined in our vocabulary. Entities starting with "dmgs:" have a defined URI in our domain where dmgs: is a prefix whose expanded value is "http://datamusee.givingsense.eu/".

As shown in Figure 1 and Section V-C, we need several linked entities to represent a work. An entity A represents the act of creating the work; this entity A is linked by the property P108_has_produced to the physical object P result of the act of creation; entity A is also linked to a conceptual object C by the property P94_has_created. The object P is linked by the property P43_has_dimension to an entity describing the dimensions of the physical object.

V. SEMANTIC TRANSLATION METHOD

Each field of the original data requires interpretation to enter the proposed semantic model. In this section, we present the process used to obtain a semantic representation from these fields.

A. General approach

As each field contains one or more labels for a specific type of data, we have no need for entity recognition, but just parsing each field to split the values for the field. Then, we need to undertake entity linking with a level of disambiguation. The main method for disambiguation is based on prior knowledge: we know that the field LOCA contains a place and the place is in France, the field AUTR contains one or several persons or organizations, etc.

Our strategy is the same for each field:

- we analyze the Joconde dataset to produce a list of possible values (strings) for each field,
- we count the number of works associated with each value (some works have several values),
- for some field, we need to parse the value to produce more useful data (see below in each field)
- we can use any algorithm to match a value against an entity of Wikidata; a simple algorithm is presented in Section V-B,
- humans check the link for the most used values (values covering the most works); in this way, we are able to guarantee good links for the most used values,
- when available, we check the obtained links against a ground truth; so, we have an idea of the quality of our data beyond the human checked links.

We will now see how this strategy is applied for some fields.

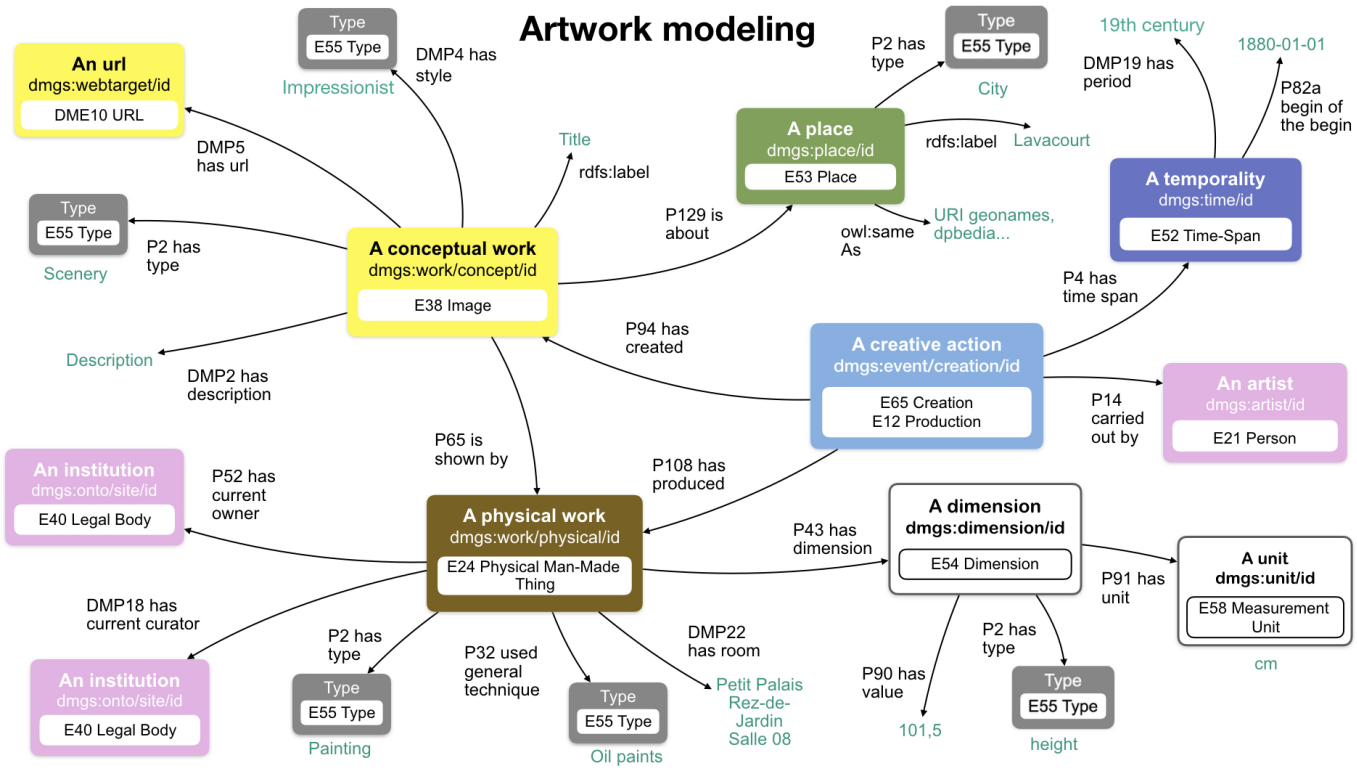

Figure 1. Artwork modeling.

B. *Entity matching and simple algorithm*

Several algorithms have been tried, like using DBpedia Spotlight [22] to get links with DBpedia or Aïda to get links with Yago [23]. The results presented here are obtained with a very simple algorithm based on the search service of Wikidata to get links with Wikidata. The search service gives us some entities corresponding to a label and some variants:

Algorithm:

- produce variants of the label: the label, the label in lowercase, the label in uppercase, the label in title case (each word with the first char in uppercase), and finally, if the label has several words, we attempt to move the first word to the last position,
- check the Wikidata search service for each variant,
- filter the results by some types,
- if only one entity is found, we keep that one; if several entities are found, we keep only the one which matches the label in lower case or none (a better disambiguation must be used in a future release)

For example, the Wikidata query template used to get the creators is in the github repository, file wikidataQueryTemplateForWord2UrisCreators.rq. The search service of Wikidata is combined with the knowledge that we search for some types of creators:

- painter "http://www.wikidata.org/entity/Q1028181"
- sculptor "http://www.wikidata.org/entity/Q1281618"
- drawer "http://www.wikidata.org/entity/Q15296811"
- artist "http://www.wikidata.org/entity/Q483501"
- visualartist "http://www.wikidata.org/entity/Q3391743"
- photographer "http://www.wikidata.org/entity/Q33231"
- engraver "http://www.wikidata.org/entity/Q329439"
- ceramicist "http://www.wikidata.org/entity/Q7541856"

See Section /refcreators for results.

Similar strategies are used for the other fields.

C. *URIs and REF field*

The original data presents a unique identifier for each work. We will use this identifier to build several URIs needed for our model. Each work gives rise to the creation of at least 4 entities: the creative act, at least one physical object, a conceptual object, several URIs for the dimensions of the physical object.

Here are the rules to build each URI, where {REF} must be replaced by the value of the REF field in the source:

- URI for the creative act: http://datamusee.givingsense.eu/event/creation/{REF}
- URI for the physical object: http://datamusee.givingsense.eu/work/physical/{REF}
- URI for the conceptual object: http://datamusee.givingsense.eu/work/concept/{REF}
- URI for the dimensions of the physical object: http://datamusee.givingsense.eu/dimension/{REF}_X, where X is a number generated for each dimension

D. *Domains: field DOMN*

As this field is completely covered by the ground truth, we use directly the proposed links.

E. Object types: field TECH

9697 terms are used for the 'TECH' field. The hundred most used cover more than 88% of the works. Many values used for this field are artistic techniques - drawing, painting, mosaic, etc- in particular in the most used values. We searched for corresponding entities in Wikidata. A useful class is "http://www.wikidata.org/entity/Q11177771", with the label "artsistic technique". So, with the property P31 (instance of) or P279 (subclass of), we were able to find all the artistic techniques known by Wikidata. We found 306 of them (result obtained on July 13, 2020). Then, we search for corresponding techniques values in Joconde. We did the same with the instances and subclass of "http://www.wikidata.org/entity/Q3300034", with the label "painting material". We found 116 of them. Then, we found 45 exact match in the Joconde data for one class or the other; we checked all of them. These 45 techniques covers 254630 works (43% of the works). Note: the SPARQL queries used to do it on Wikidata Query Service are available on the github repository referenced above.

In the ground truth, there is no association for the TECH field. So, some more work must be done to complete and to assert the quality of our results for this field.

F. Object types: field AUTR

The field AUTR gives a string naming the creator of a work. Some works have no known creator (99194 works; 16.83%). Many (63199; 10.72%) have the creator named 'anonymous'. But for the others (426885), we will try to find a matching entity in Wikidata. Creators are persons or organizations.

We are particularly interested in the most productive creators. We have chosen a threshold of 10 or more works per selected creator. There are 5217 creators in this category. They produced 98.07% of the works attributed to a creator.

We have benefited in particular from the work carried out by the Wikidata-Joconde project [18]. This project associates terms used in the Joconde database with Wikidata entities, with a human validation process. As of 5/30/2020, 2560 associations were validated for creators. 1325 are in our target of productive creators. They cover 25.04% of the attributed works.

Our algorithm V-B allows to find 1173 Wikidata entities associated with the designation of the creator by the AUTR field in Joconde, of which 1168 correspond to the entities validated by the Wikidata-Joconde project.

To evaluate our results, we use precision, recall and F1 measures.

Ncw = number of creators validated by the Wikidata Joconde project and targets of the evaluation

Nct = number of creators for which our algorithm finds a Wikidata entity

Nce = number of exact links found

Pc = precision relative to creators = Nce/Nct

Rc = recall relative to creators = Nce/Ncw

$F1c$ = 2*Pc*Rc/(Pc+Rc)

We also considered the 100 creators with the greatest number of works except 'anonymous'. Of these 100 creators, a Wikidata link was found for 59 of them. Of these 59, 28 were among the links already validated by the ground truth. We proceeded to a human validation of the other 31 links: all of them were exact. On these 59 links, an accuracy of 100% was therefore obtained. The recall cannot be evaluated, since for creators not found, we do not have a method to tell if the creator is not in Wikidata or if our algorithm failed to find it. Assuming that all creators are listed in Wikidata, we get a lower bound of the recall: 59%; and a lower bound for the F1 measure: 74.21.

TABLE II. RESULTS FOR CREATORS

Measure	*Value*
Ncw	1315
Nct	1180
Nce	1177
Pc	99.74
Rc	88.83
F1c	93.97

We have manually checked 10 links for creators among the most productive, covering 47029 works (11.01% of attributed works). The list is: RODIN Auguste (13231 works), MOREAU Gustave (6816), CHASSERIAU Théodore (5010), DELACROIX Eugène (4136), COROT Jean-Baptiste Camille (4114), INGRES Jean Auguste Dominique (3202), STEINLEN Théophile Alexandre (2916), LE BRUN Charles (2882), PICASSO Pablo (2496), HEBERT Ernest (2226). Ten correct links are found by our algorithm for these 10 creators.

For the 5127 productive creators, we found 2199 links by our algorithm. A simple extrapolation from the results obtained on the ground truth, with $Pc = 99.74$, suggests a result of around $2199*Pc/100 = 2193$ correct links, which is 878 new links beyond the ground truth.

G. Localisation: fields LOCA and STAT

The LOCA is generally composed of a city name, followed by an institution or organization name, separated by a semicolon.

We will skip the entity linking of the city, because it is a very classical problem with good results using a lot of available tools. So, our focus will be the organization or institution. Each institution has the same city coupled with her in each occurrence of the institution in a LOCA field value. So, the count of institutions is the count of different values in the LOCA field: 3593. No link to Wikidata is available in the ground truth.

For our algorithm, we selected the following types:

- museum "www.wikidata.org/entity/Q33506"
- glam ".../entity/Q1030034"
- cultural institution ".../entity/Q5193377"
- cultural organization ".../entity/Q29918292"

And we add a filter against the city: the institution found must be in the good city.

We selected institutions with more than 100 works in Joconde, the 'richest' institutions. So, 304 institutions were selected. They are covering 580035 works (98.43%). For these institutions, we found 155 links. We undertook manual check on the first quarter of the list (first 76 museums). We found 42

TABLE III. RESULTS FOR A SELECTION OF LOCALIZATION

Measure	*Value*
Searched museums	76
Found links	42
Exact links	42
Precision	100
Recall	55.26
F1c	71.18

links for these museums; all found links have been checked manually: all are exact. So, on this sample, we have:

The STAT field is similar to the LOCA field in the sense that it contains mainly a city and an organization/institution. So, the STAT field is processed similarly to the LOCA field.

VI. SEMJOCONDE DATASET

In our triple store, Fuseki, we have a dataset named SemJoconde. The main components of this dataset are the following RDF graphs:

- one graph contains the works,
- one graph contains the creators,
- one graph contains the institutions and organizations,
- one graph contains the cities.

These graphs are linked together and are linked with Wikidata. These graphs are available with a Creative Commons licence in the github repository [19]. It is evolving on daily basis and will soon have a description with VOID triples [24].

For entities not found in Wikidata by the previously described methods, we produce our own URIs and, in the future, expect to complete these URIs by owl:sameAs links to other Knowledge Graphs, like Getty, BNF, Europeana, British Museum, Wikidata, DBpedia, Yago (see Section II-C), etc.

In addition, the github repository includes JSON files which list the domains, the authors and the techniques encountered in the database, with their frequency of use. It includes queries to Wikidata Query Service, which contributes to the process of building this dataset.

VII. CONCLUSION AND FUTURE WORKS

In this paper, we introduce a new LOD dataset. With close to 600000 artistic works described by triples. A work to produce links over Wikidata entities is presented. A good coverage of works interlinked with Wikidata by at least one property is our goal and we see some preliminary results as links for 59% of the creators with a precision of more than 99% and similar results for the institutions.

In the future, we expect to improve the coverage and consolidate our results by exploiting the context more intensively. For example, we can use the PERI field (period) to improve the selection of a creator or improve the links with institutions by knowing the creators presented in them.

Also, we intend to use the SemJoconde graph in recommendation projects using graph embedding methods.

ACKNOWLEDGMENT

This work is conducted as part of the *Data&Musée* project selected in the $23rd$ call for projects of the Fonds Unique Interministériel (FUI) and certified by Cap Digital and Imaginove.

REFERENCES

[1] "Wikidata," 2020, URL: https://www.wikidata.org [retrieved: October, 2020].

[2] "Data&Musée," 2020, URL: http://datamusee.fr [retrieved: October, 2020].

[3] "Getty Research: Editorial Guidelines," 2018, URL: http://www.getty.edu/research/tools/vocabularies/guidelines/index.html [retrieved: September, 2020].

[4] C. Dijkshoorn et al., "The rijksmuseum collection as linked data," Semantic Web, vol. 9, no. 2, 1 2018, pp. 221–230.

[5] "Europeana project," 2020, URL: https://www.europeana.eu [retrieved: October, 2020].

[6] J. Hoffart et al., "Robust Disambiguation of Named Entities in Text," in Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, Edinburgh, Scotland, 2011, pp. 782–792.

[7] J. Volz, C. Bizer, M. Gaedke, and G. Kobilarov, "Silk–a link discovery framework for the web of data," Proceedings of the 2nd Linked Data on the Web Workshop, 01 2009.

[8] V. de Boer et al., "Supporting linked data production for cultural heritage institutes: The amsterdam museum case study," in The Semantic Web: Research and Applications. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 733–747.

[9] E. Mäkelä, E. Hyvönen, and T. Ruotsalo, "How to deal with massively heterogeneous cultural heritage data - lessons learned in culturesampo," Semantic Web, vol. 3, 01 2012, pp. 85–109.

[10] "CIDOC-CRM: Conceptual Reference Model," 2020, URL: http://jocondelab.iri-research.org/jocondelab/about/ [french; retrieved: September, 2020].

[11] "British Museum dataset," 2020, URL: https://info.datatourisme.gouv.fr [retrieved: October, 2020].

[12] "Paris Musées Collections," 2020, URL: https://apicollections.parismusees.paris.fr [retrieved: October, 2020].

[13] "DataTourisme," 2020, URL: https://old.datahub.io/dataset/british-museum-collection [retrieved: October, 2020].

[14] "Geonames," 2020, URL: https://www.geonames.org [retrieved: October, 2020].

[15] "DBpedia," 2020, URL: https://dbpedia.org [retrieved: October, 2020].

[16] "Yago," 2020, URL: https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago [retrieved: October, 2020].

[17] "Data Gouv," 2020, URL: https://www.data.gouv.fr/en/datasets/collections-des-musees-de-france-extrait-de-la-base-joconde-en-format-xml/ [retrieved: October, 2020].

[18] "Wikidata:WikiProject Vocabulaires Joconde," 2020, URL: http://www.wikidata.org/wiki/Wikidata:WikiProject_Vocabulaires_Joconde/en [retrieved: September, 2020].

[19] "Repository for SemJoconde," 2020, URL: https://github.com/datamusee/semjoconde [retrieved: September, 2020].

[20] "CIDOC-CRM: Conceptual Reference Model," 2020, URL: http://www.cidoc-crm.org/ [retrieved: September, 2020].

[21] V. Alexiev, V. C. Ivanov, and M. e. Grinberg, Eds., Practical Experiences with CIDOC CRM and its Extensions (CRMEX 2013) Workshop, 17th International Conference on Theory and Practice of Digital Libraries (TPDL 2013), Dec. 2013.

[22] J. Daiber, M. Jakob, C. Hokamp, and P. N. Mendes, "Improving efficiency and accuracy in multilingual entity extraction," in Proceedings of the 9th International Conference on Semantic Systems, ser. I-SEMANTICS '13. New York, NY, USA: Association for Computing Machinery, 2013, p. 121–124. [Online]. Available: https://doi.org/10.1145/2506182.2506198

[23] J. Hoffart, "Discovering and disambiguating named entities in text," in Proceedings of the 2013 SIGMOD/PODS Ph.D. Symposium, ser. SIGMOD'13 PhD Symposium. New York, NY, USA: Association for Computing Machinery, 2013, p. 43–48. [Online]. Available: https://doi.org/10.1145/2483574.2483582

[24] "Describing Linked Datasets with the VoID Vocabulary," 2011, URL: https://www.w3.org/TR/void/ [retrieved September, 2020].