Enriching Wikidata with cultural heritage data from the COURAGE project

Ghazal Faraj[1][0000-0002-6705-7568] and András Micsik[2][0000-0001-9859-9186]

[1] Eötvös Loránd University, Pázmány Péter stny. 1/C., 1117, Budapest, Hungary
[2] MTA SZTAKI DSD, Lágymányosi u. 11., Budapest, Hungary
ghazal.faraj@gmail.com,micsik@sztaki.hu

Abstract.
Creating links manually between large datasets becomes an extremely tedious task. Although the linked data production is growing massively, the interconnecting needs improvement. This paper presents our work regarding detecting and extending links between Wikidata and COURAGE entities with respect to cultural heritage data. The COURAGE project explored the methods for cultural opposition in the socialist era (cc. 1950-1990), highlighting the variety of alternative cultural scenes that flourished in Eastern Europe before 1989. We describe our methods and results in discovering common entities in the two datasets, and our solution for automating this task. Furthermore, it is shown how it was possible to enrich the data in Wikidata and to establish new, bi-directional connections between COURAGE and Wikidata. Hence, the audience of both databases will have a more complete view of the matched entities.

Keywords: Linked Data, Cultural Heritage, Wikidata, Link Discovery, Link Disambiguation.

1 Introduction

The COURAGE (Cultural Opposition: Understanding the CultuRal HeritAGE of Dissent in the Former Socialist Countries) project explored the methods for cultural opposition in the socialist era (cc. 1950-1990) [1]. One of the project goals was to highlight the variety of alternative cultural scenes that flourished in Eastern Europe before 1989 in spite of rigorous government control. The project has compiled a registry of historic collections, people, groups, events and sample collection items stored in an RDF triple store. The registry is available online and has been used to create virtual and real exhibitions and learning material. It is also planned to serve as a basis for further narratives and digital humanities (DH) research [2]. The main entities of the COURAGE dataset are:

- Collections, the main focus of the research;
- Interviews with key persons of collections;

- People, groups, and organizations playing an important role in the history of the collection, for example, owners, founders, operators, collectors;
- Some major events in the history of collections;
- Featured items from each collection.

The registry schema is called the COURAGE Ontology, which contains cca. 100 classes, 220 object properties, and 170 data properties [3].

Wikidata is the main storage for structured data which is related to Wikipedia, Wikisource, and others [4] thus it creates new ways for managing Wiki* data on a global scale [5]. This data is freely available online, regularly updated by volunteers worldwide and is extremely correlated and connected to other datasets. The most important advantage of using Wikidata is linking datasets with appropriate relationships that can be understandable by humans and machines.

According to the recent statistics, Wikidata contains more than 57 million entities. They have approximately 718 million statements, and over 800 million labels and descriptions which are available in 350 languages or more [6].

The production of Linked Data is growing massively these days, but the linking between these datasets needs to be improved. Typically, the following anomalies exist in the linked data world: different entities describe the same individual in different datasets, or similar statements are described differently in different datasets. The closer we get in the elimination of these anomalies, the more complete knowledge we can serve to users.

Currently, both Europeana and Wikidata collect cultural heritage (CH) data extensively. Wikidata had a campaign dedicated to collecting cultural heritage data [7]. Europeana is about digital cultural heritage in general, including metadata, illustrations, narratives, and many other aspects. Europeana data providers are encouraged to use Wikidata as a source for enriching data and to connect their vocabularies to Wikidata [8].

Following this guideline, the current paper aims to connect Wikidata and COURAGE datasets. We found that the overlapping set of resources is mostly of the types: person, group and organization, so our investigations were based on these entity classes. The research questions we address include:

- How safely can we identify matching entities in Wikidata and COURAGE?
- How can we extend Wikidata and COURAGE so that the audience of both databases gets more facts about matched entities?

The remainder of the paper is organized as follows: Section 2 surveys the link discovery tools and entity resolution approaches which are related to our research. Section 3 describes preliminary statistics, the requirements for the matching approach and how the matching process was carried out. Section 3 also discusses the results generated by the matching algorithm. Extending Wikidata after determining the injected properties and generating the triples file are presented in Section 4. Finally, we conclude the paper in Section 5.

2 Related work

Wikidata was established to become a multilingual and global database which contains the entire cultural heritage data for data integration and data management. Moreover, they also aimed to become a focal point for interconnecting heritage collections and providing links to other external data sources [9,10].

One of the main ideas about the web of data besides representing data to be understandable by a machine is to set relationships between entities across knowledge bases. These relationships may be determined automatically using link discovery tools.

There are quite a few link discovery tools mentioned in [11], but most of them seem abandoned for 3 or more years. Silk was the first link discovery tool for finding links between entities and it provides a language to specify the link types which should be discovered between datasets [12]. Silk and LIMES support more link types than other tools which just determine owl:sameAs and they provide a GUI for an interactive use [13]. KNOFUSS just supports owl:sameAs link type and string similarity approach [14]. SERIMI takes input only from SPARQL endpoints as it does not support RDF input. It is restricted to one property for matching and the thresholds must be manually determined. We tried to use some of these tools for our link discovery task, but without any success. We got farthest with LIMES, but still, it was not able to find any links applying either acceptance conditions or unsupervised learning. We think the reason for this was that Wikidata has millions of entities and querying these often results in time-out. Moreover, using the previously mentioned tools usually requires an acceptance threshold for matching, and finding the optimal threshold value requires an iterative method similar to ours.

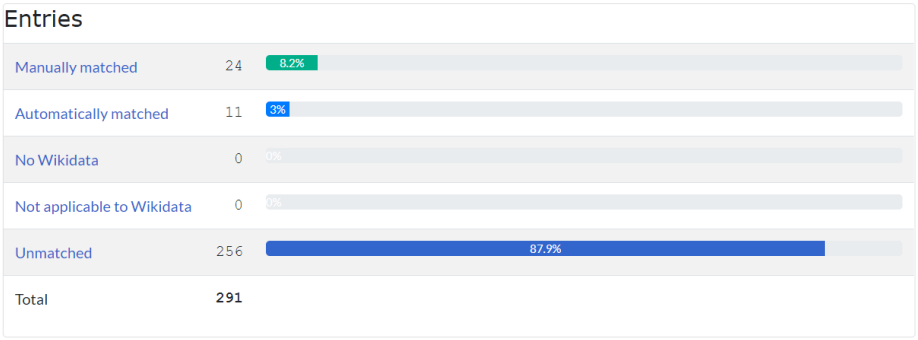

Fig. 1 Organization matching results using Mix'n'match tool

Mix'n'match is a tool developed by Magnus Manske to let the user match entities with Wikidata ones [15]. We tried to use the tool with organization entities but unfortunately, the outcomes were not really useful (see Fig. 1). 3% of the entities were automatically matched with many false positive cases and 87.9% of the entities were unmatched. This happened partly because the sought entity did not exist in Wikidata, and partly because the search method of the tool did not find an unambiguous match.

In [16] authors manage ambiguity in VIAF by clustering similar authorities and analyzing these clusters (or subgraphs). On the other hand, COURAGE and Wikidata have a very low number of duplicates, and we had to select a single best matching entity as a result. Another similar name disambiguation problem is handled in [17], but only the names are used for matching.

3 Matching individuals

COURAGE has a scope limited in both time and region, but the entity data were created by historians with thorough quality control. The entity descriptions are available in at least two languages and they may be quite lengthy. On the contrary, Wikidata entity descriptions are typically 1-2 lines of length, while Wikipedia pages may be 1-3 times longer than COURAGE pages about the same entity.

Wikidata lacks the contribution types and roles of people in various cultural groups and collections. Basic properties such as birthplace, gender, profession, etc. are sometimes more precise in one entity than in the other. This creates a delicate situation both when matching individuals and when trying to complement the data in one dataset based on the other.

A Person entity in Wikidata is addressed by an opaque item identifier which starts with "Q" and a number. This entity is also presented in a page which consists of these main parts: label, description, a set of aliases, a set of statements and a set of external links [9]. The set of statements usually includes instance of, image, given name, family name, birthdate and birthplace properties.

In COURAGE, the Person entity also has a unique identifier and a list of statements. Person properties include given name, family name, year of birth, birthplace, profession and some other personal data.

Person and organization preliminary statistics. For our investigations, we collected 1218 person entities with 3 properties: name, type, and birthyear from COURAGE. We performed a simple search based on these properties to find all possible Wikidata entities. After this, we classified the matched pairs into two groups based on the possibility of a clear matching decision (Fig. 2). We found that for 63.21% of matched person entities the matching decision can be made unambiguously (first group), meaning that type, name and birthyear were the same. The second group "Ambiguous matching decision" has 36.79% of person entities which could be further divided into three subgroups. In the first sub-group entities have enough data so that an expert can make a decision. In the second sub-group there are many missing properties which make a human decision impossible. The last sub-group is about false-positive cases, when a false match was made in spite of equal person properties. Therefore, we had to take more properties into our approach.

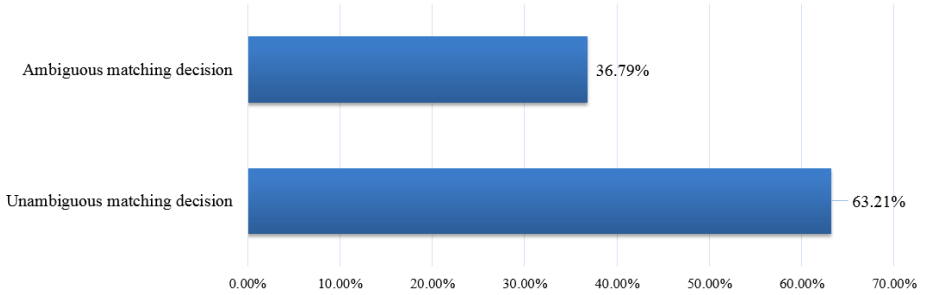

Fig. 2 Person results classification based on matching decision

For example, "Gerhard Ortinau" (Q101211) is a person entity which belongs to the first group where all specified properties exist and that made the matching decision clear. "Dragoș Petrescu" (Q18545324) belongs to the first subgroup in the second group where the birthdate property value was missing, still an expert may be able to make the matching decision based on other properties. The "Ion Dumitru" (Q23309144) entity is also in the second group since the human matching decision was ambiguous due to missing critical data in the coupled Wikidata and COURAGE entities. "Patti Smith" (Q27582022) is an example of the false positive matches. This entity was matched based on the same name and birthyear, but it turned out that it was not the same person as the birthplace was different.

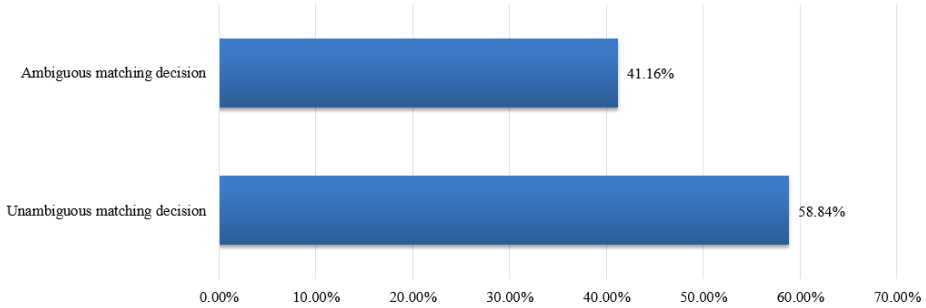

Fig. 3 Organizations results classification based on matching decision

As for organization entities in Wikidata and COURAGE, we used 4 properties for matching: name, type, country, and GPS. The statistics, which were calculated for 457 organizations in COURAGE, state that 58.84% of organization entity pairs belong to the first group where a matching decision could be made unambiguously. Consequently, 41.16% of the entities were in the second group.

3.1 Metrics for similarity

After analyzing the data regarding people and organizations, we set up many suitable characteristics to identify them, such as name, type, location, birthdate, founding year, etc. Unfortunately, not all of these properties exist in both Wikidata and COURAGE

datasets. Therefore, we considered two basic sets of criteria. The first set contains 5 properties that identify organizations: name, city, country, GPS and year of founding. The second set contains 3 keys that identify people: name, birthplace, and birthdate. We assumed that the 'type' property is always correct in both COURAGE and Wikidata. Therefore, we used it for data filtering without considering it as a key in the latter formulas.

In Wikidata, 52% of the possible organization matches missed geocoordinates and 31% missed city location. Furthermore, the year of foundation was unknown for 39%.

Since matching methodology improvement is a continuous process, first we made simple statistics to determine how to find correct matching decisions. Based on several experimental studies a scoring system was introduced to provide points for each candidate entity based on the matching status as below.
The metrics established for matching person entities were:

- Name: we removed the diacritics and checked the results of the comparison: if the name of Wikidata entity is exactly equal to the COURAGE entity name, it gets 4 points, containing the name it gets 2 points, and if the Levenshtein distance was at most 1 it gets 1 point. Otherwise, the comparison of the two names gets 0 points.
- Birthplace: if the birthplace of Wikidata entity exactly equals to the COURAGE entity birthplace, it gets 2 points. If one of the values is missing, the score is 1 point. Otherwise, it gets 0 points.
- Birthdate: if the year in birthdate for Wikidata entity exactly equals to the birthyear of COURAGE entity it gets 2 points, if the difference is 1 year between values it gets 1 point. Otherwise, it gets 0 points.

The metrics for matching organizations were:

- Name: similarly to persons' names.
- City and Country: if the city properties and country properties exactly equal, it gets 4 points. if just the city properties are exactly equal, it gets 2 points. If one of the values is missing, the comparison gets 1 point. Otherwise, it gets 0 points.
- GPS: if the distance between the resource locations is less than 1.6 km, it gets 3 points. If it is missing, the comparison gets 1 point. Otherwise, it gets 0 points.
- Year of foundation: if the year of Wikidata entity exactly equals to the COURAGE entity foundation year it gets 2 points, if the difference is 1 year between values it gets 1 point. Otherwise, it gets 0 points.

Regarding the scores approach, the exact equality status and the distance between resource locations may get the most points.

3.2 Matching algorithm

The aim of our work was to develop and implement a relatively reliable matching process on person and organization entities. There was no human capacity for research of matching individuals one-by-one, and a fully automatic matching also proved to be

unfeasible. Therefore, we aimed at detecting the cases where a human decision was needed but at the same time also minimizing the number of such cases.

Approach. A matching algorithm was developed in C# for the previous purpose. Firstly, we executed a SPARQL query via the COURAGE SPARQL endpoint and downloaded organization data keys which are: name, city, country, GPS and founding year. After cleaning this data, we imported it to our database. Next, we ran a C# script for each item to get all possibly related entities from Wikidata based on its type and name containment. In the beginning, we compared the name, type, city, country, GPS and founding year at once. But in order to enhance the performance, we followed sequential steps, by comparing the name and the type as a first step. After which, we moved to compare city and GPS then the country and founding year. For each exact or partial similarity with the 5 keys (name, city, country, GPS and founding year) we provided points in all conditions according to the previous rules. The number of Wikidata candidates for COURAGE entities was between 1 and 6. The previous five key points with their weights produced a total score for the match:

$$wo1*namePoints + wo2*cityPoints + wo3*countryPoints + wo4*GPSPoints + wo5*foundingYearPoints = totalScore \quad (1)$$

Secondly, we downloaded person data keys which are: name, birthplace, and birthyear from COURAGE dataset. Similarly to organizations, we executed a SPARQL query and applied the same methodology on this data. The total score was calculated as:

$$wp1* namePoints + wp2* birthPlacePoints + wp3* birthdatePoints = totalScore \quad (2)$$

During the matching algorithm, points were assigned to each metric in each matched pair and thus a matrix of matching points was built. Based on this matrix, weighted matching scores (*totalScore)* were calculated for each matched pair in the sample.

To determine the best weights two random sample sets were created with 300 matched pairs for persons and 50 matched pairs for organizations. Each pair was manually checked as matching or non-matching. Next, the scores were calculated in the sample sets for all possible weights between [0, 2] with a step increment of 0.1. After this, various indicator values for the goodness of the weights were calculated: the lower threshold *Tlo* is the largest *totalScore* value below which only non-matching pairs will be seen in this sample. The upper threshold *Tup* is the smallest *totalScore* value above which only matching pairs will be seen. Between *Tlo* and *Tup*, one finds the ambiguous pairs, which we called the human decision window. The least number of items in the window (*windowSize*) is the best. The *minError* count is generated for each threshold in the sample based on how many cases are below this threshold but they are matched, and above the threshold but are not matched. Finally, we calculated the minimum threshold *Tmin* at which the number of error cases (*minError*) is the lowest.

Findings and results. The results of all the prior calculations indicated that the foundation year of organizations is not an important property, because whatever the weights

were, we got the same size for the human decision window. Consequently, we could eliminate it from the properties list before applying the matching process on all the data.

Overall, we took the person and organization weights related to the least items in the human decision window and applied these weights and thresholds on the entire person and organization entities respectively. After which, we checked 50 random entities from the matched cases and also 50 random entities from not matched cases without facing any incorrect decision. We also checked manually the cases inside the window. The statistics of the result showed that 78.64% of person entities and 80.5% of organization entities could be safely matched automatically with Wikidata entities.

The person outcomes state that the human decision window has more than one value for the *Tlo* and the *Tup*. However, the *windowSize* inside this window is 121 (Table 1).

Table 1 Threshold calculations for matching persons

Tlo	Tup	windowSize	wName	wPlace	wYear
4.4	6.1	121	0.8	1.3	1.4
5	6.9	121	0.9	1.5	1.6
5.5	7.5	121	1	1.6	1.7
5.5	7.7	121	1	1.6	1.8
5.6	7.7	121	1	1.7	1.8

On the other hand, the organization matching result (Table 2) shows that the *Tlo* of human decision window is 9.2 and the *Tup* is 13.1. The *windowSize* is 51 (for the whole set). Consequently, the corresponding weights *wCity*, *wCountry*, *wGPS*, and *wName* values are the best weights among all weight sets.

Table 2 Threshold calculation for matching organizations

Tlo	Tup	windowSize	wCity	wCountry	wGPS	wName
9.2	13.1	51	1.9	1.8	2	1.8

4 Establishing connections

As a next step, a list of transferable properties has been set up and triples to extend Wikidata have been compiled. We created a table of matching properties in COURAGE and Wikidata. These properties can be grouped into two categories for each entity type: properties used for matching and new properties.

First, we gathered the common properties between people and organizations to avoid duplication as shown in Table 3. Regarding other properties, they are displayed in the tables (Table 4, Table 5) below.

Table 3 General properties for both persons and organizations

Courage	Wikidata	
public#mainImage	P18/P154	Image/logo image
courage.owl#website	P856	official website
courage.owl#place	P276	location
Item Courage URI	P973	Described at URL

Table 4 Properties matched for person data

Courage	Wikidata	
courage.owl#hasGivenName	P735	Given name
courage.owl#hasFamilyName	P734	Family name
courage.owl#birthDate	P569	date of birth
courage.owl#birthPlace	P19	Birth Place
courage.owl#deathDate	P570	date of death
courage:hasNickName	P1449	nickname
courage:hasSex	P21	sex or gender
courage:memberOf	P463	member of
courage:ownerOf	P1830	owner of
courage:hasCreatorRole	P6379	has works in the collection(s)
courage:creatorOf	P170	inverse of creator

Table 5 Properties matched for organization data

Courage	Wikidata	
courage.owl#yearOfFunding	P571	inception
courage.owl#country	P17	country
courage.owl#city	P131/ P159	located in the administrative territorial entity / headquarters location
courage.owl#lat, courage.owl#long	P625	coordinate location
courage.owl#instType	P31	instance of
courage.owl#ownerRoleOf	P1830	owner of
courage.owl#leader	P488/ P1037	chairperson / director or manager
courage.owl#operatorRoleOf	P126	maintained by

Based on the final transferable properties list, we generated triples in the format of the QuickStatements tool, which allows the bulk addition of Wikidata items [18]. For the implementation, an algorithm was established to generate a file which contained the needed triples to do this extension. The file has 1765 statements for person and organization entities. For person entities, we enriched 385 Wikidata entities successfully (Table 6).

Table 6 Sample of person properties in the generated file

Item	Property	Value	Source property	
Q112688	P734	Q2168571	S248	Q64784883
Q112688	P973	"http://courage.btk.mta.hu/courage/individual/n13144"		
Q112688	P1830	"http://courage.btk.mta.hu/courage/individual/n25127"	S248	Q64784883

While for organization entities we enriched 143 Wikidata entities (Table 7).

Table 7 Sample of organizations properties in the generated file

Item	Property	Value	Source property	
Q11179076	P276	Q1085		
Q11179076	P973	"http://courage.btk.mta.hu/courage/individual/n100194"	S248	Q64784883
Q11179076	P571	+1949-01-01T00:00:00Z/9	S248	Q64784883
Q11179076	P625	@50.0755381/14.4378005	S248	Q64784883

We also generated another file with different syntax to create new entities (Table 8).

Table 8 Sample of creating a new entity in the generated file

Statements				
CREATE				
LAST	Len	“Gardzienice Theatre”		
LAST	Lpl	“Teatr Gardzienice”		
LAST	P31	Q43229		
LAST	P973	http://courage.btk.mta.hu/courage/individual/n45835		
LAST	P571	+1977-01-01T00:00:00Z/9	S248	Q64784883
LAST	P131	Q5522662	S248	Q64784883
LAST	P625	@51.110556/22.8586111	S248	Q64784883
LAST	P856	http://gardzienice.org	S248	Q64784883

Our contribution was enriching and linking the person and organization entities as the dashed lines show in Fig. 4. Person and organization Wikidata entities are mapped to COURAGE entities via property P973 (Described at URL). Following this, we also created new person and organization Wikidata entities for non-matched COURAGE entities. Later, when scholars have time, they can create the Wikipedia pages for these new entities.

We found it hard to establish links other than ‘has id’ between Wikidata and COURAGE entities. For example, the creator and 'has works in collection' properties

accept only Wikidata entities as an object. Thus, it was impossible to direct Wikidata readers' attention to artifacts authored by a person. Similarly, we could not refer to roles (owner, operator, supporter, etc.) taken by persons or groups at collections in the COURAGE registry. However, we could create member and leader links between persons and groups as they were all in Wikidata after our data injection.

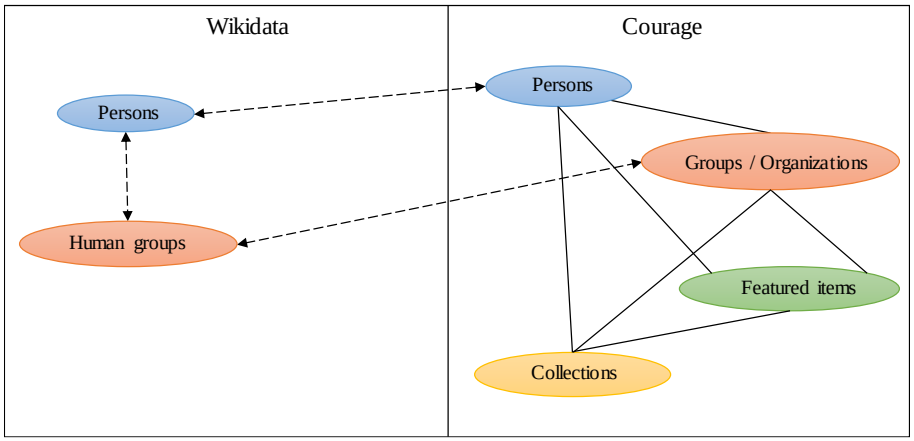

Fig. 4 Main connections inside COURAGE and with Wikidata

5 Conclusion

Our aim was to connect the linked data registry of COURAGE into a broader linked data context, for which Wikidata seemed to be the ideal candidate. The COURAGE project made an extensive research on the cultural heritage of former European socialist countries, resulting in high quality linked data about available collections on the subject and their surrounding personal networks. The common point of integration was found to be persons, groups and organizations. To match these entities in the two datasets, a score-based method has been shaped, and automated link discovery has been performed successfully on 78% of person entities and 80% of group/organization entities. For the remaining matching candidates, a human decision was needed in order to maintain the good quality of links between the datasets.

As a result, matched entities have links to the corresponding Wikidata entity in the COURAGE registry, and Wikidata users may choose to navigate to matched COURAGE entities for more information. On the other hand, the link to Wikidata on the COURAGE side provides access to many other authority IDs (e.g. VIAF, IMDB) collected in Wikidata. Furthermore, Wikidata has been enriched with data present in COURAGE registry, including official websites and connections between persons and organizations. In the future, the insertion of collections and artifacts to Wikidata may give further benefits for Wikidata users, even if these entities are for a quite specialized

interest at the moment, as only a minimal number of such items exist currently in Wikidata.

Acknowledgement The project has been supported by the European Union, co-financed by the European Social Fund (EFOP-3.6.3-VEKOP-16-2017-00002).

References

1. COURAGE project homepage. http://cultural-opposition.eu/
2. Apor, B., Apor, P., Horváth, S. eds.: The Handbook of COURAGE. Budapest, (2018), doi:10.24389/handbook
3. Micsik, A.: Courage registry - open dataset 1.1 (July 2019), doi: 10.5281/zenodo.3333540
4. Erxleben, F., Günther, M., Krötzsch, M., Mendez, J., Vrandečić, D.: Introducing Wikidata to the linked data web. In: International Semantic Web Conference, pp. 50-65, Springer, (2014), doi: 10.1007/978-3-319-11964-9_4
5. Vrandečić, D., Krötzsch, M.: Wikidata: A free collaborative knowledgebase. Communications of the ACM, 57(10), (2014), doi: 10.1145/2629489
6. Wikidata Statistics. https://www.wikidata.org/wiki/Wikidata:Statistics
7. WikiProject Cultural heritage. https://www.wikidata.org/wiki/Wikidata:WikiProject_Cultural_heritage
8. Why data partners should link their vocabulary to Wikidata: a new case study. Europeana pro page. https://pro.europeana.eu/post/why-data-partners-should-link-their-vocabulary-to-wikidata-a-new-case-study
9. Malyshev, S., Krötzsch, M., González, L., Gonsior, J., Bielefeldt, A.: Getting the Most out of Wikidata: Semantic Technology Usage in Wikipedia's Knowledge Graph, 17th International Semantic Web Conference, Monterey, Proceedings, Part II, pp. 376-394, CA, USA, October 8–12, 2018.
10. Allison-Cassin, S.; Scott, D.: Wikidata: a platform for your library's linked open data. In: Code4Lib, 40 (2018).
11. Nentwig, M., Hartung, M., Cyrille, A., Ngomo, N., Rahm E.: A Survey of Current Link Discovery Frameworks, Semantic Web Journal 2.224 (2017), doi:10.3233/SW-150210
12. Isele, R., Jentzsch, A., Bizer, C.: Efficient Multidimensional Blocking for Link Discovery without losing Recall. In: 14th International Workshop on the Web and Databases, WebDB, Athens. (2011).
13. Ngomo, A.C.N. and Auer, S.: LIMES - a time-efficient approach for large-scale link discovery on the web of data. In: IJCAI, pp. 2312-2317 (2011), doi: 10.5591/978-1-57735-516-8/IJCAI11-385
14. Nikolov, A., Uren, V., Motta, E.: KnoFuss: A comprehensive architecture for knowledge fusion. In: Proceedings of the 4th international conference on Knowledge capture, pp. 185–186, ACM, (2007)
15. Mix'n'match Manual Wikimedia page, https://meta.wikimedia.org/wiki/Mix%27n%27match/Manual
16. Thomas B. Hickey and Jenny A. Toves: Managing Ambiguity In VIAF. D-Lib Magazine 20(7/8), doi:10.1045/july2014-hickey
17. Larson, R., Janakiraman, K.: Connecting Archival Collections: The Social Networks and Archival Context Project. In: Research and Advanced Technology for Digital Libraries. TPDL 2011. Lecture Notes in Computer Science, vol 6966. Springer, Berlin, Heidelberg
18. QuickStatements help page. https://www.wikidata.org/wiki/Help:QuickStatements