Persons, GLAM institutes and collections: an analysis of entity linking based on the COURAGE registry

Ghazal Faraj*

Eötvös Loránd University,
Pázmány Péter stny. 1/C., 1117,
Budapest, Hungary
Email: Ghazal.faraj@gmail.com
*Corresponding author

András Micsik

Institute for Computer Science and Control (SZTAKI),
Eötvös Loránd Research Network (ELKH),
Lágymányosi u. 11., Budapest, Hungary
Email: micsik@sztaki.hu

Abstract: It is an important task to connect encyclopaedic knowledge graphs by finding and linking the same entity nodes. Various available automated linking solutions cannot be applied in situations where data is sparse, private or a high degree of correctness is expected. Wikidata has grown into a leading linking hub collecting entity identifiers from various registries and repositories. To get a picture of connectability, we analysed the linking methods and results between the COURAGE registry and Wikidata, VIAF, ISNI and ULAN. This paper describes our investigations and solutions while mapping and enriching entities in Wikidata. Each possible mapped pair of entities received a numeric score of reliability. Using this score-based matching method, we tried to minimise the need for human decisions, hence we introduced the term human decision window for the mappings where neither acceptance nor refusal can be made automatically and safely. Furthermore, Wikidata has been enriched with related COURAGE entities and bi-directional links between mapped persons, organisations, collections, and collection items. We also describe the findings on coverage and quality of mapping among the above mentioned authority databases.

Keywords: linked data; cultural heritage; link discovery; entity linking; authority data; metadata quality; Wikidata; VIAF; ISNI; ULAN.

Reference to this paper should be made as follows: Faraj, G. and Micsik, A. (2021) 'Persons, GLAM institutes and collections: an analysis of entity linking based on the COURAGE registry', *Int. J. Metadata, Semantics and Ontologies*, Vol. 15, No. 1, pp.39–49.

Biographical notes: Ghazal Faraj is a PhD student at Eötvös Loránd University, Faculty of Informatics. She received her master's degree in Engineering Information Technology from Budapest University of Technology and Economics in July 2018. Previously, she worked in analysing, designing, implementing, and testing software projects in Syriatel a telecommunication company. In 2017, she received her first master's degree in Software Engineering and Information Systems from Damascus University.

András Micsik is a team leader within the Department of Distributed Systems at SZTAKI (Institute for Computer Science and Control). Recently, he works on the application of Semantic Web and Linked Open Data in various areas such as e-science or web services. He contributed to several international research projects on the topics of software services, metadata vocabularies, semantic interoperability, digital libraries, etc.

This paper is a revised and expanded version of a paper entitled 'Enriching Wikidata with Cultural Heritage Data from the COURAGE Project' presented at the 'MTSR 2019, 13th International Conference on Metadata and Semantics Research', Rome, Italy, 28–31 October 2019.

1 Introduction

It is always an important and interesting task to connect encyclopaedic knowledge graphs by finding and linking nodes representing the same entities. There are many challenges while performing entity linking on several heterogeneous datasets. One problematic area that is dealt with in this paper is the case when the amount and quality of available data are mostly insufficient for automated linking.

Currently, Wikidata gathers cultural heritage (CH) data extensively, also via dedicated campaigns (WikiProject Cultural Heritage, 2020). For example, Europeana data providers are encouraged to use Wikidata as a source for enriching data and to connect their vocabularies to Wikidata (Europeana, 2017). Wikidata is the largest structured data storage connected to Wikipedia, Wikisource, and others (Erxleben et al., 2014). Thereby it creates new ways for managing Wiki* data on a global scale and interlinks datasets with suitable relationships that humans and machines understand (Vrandečić and Krötzsch, 2014). Wikidata has been developed to become a multilingual and global registry that integrates and manages all existing cultural heritage data. Wikidata can be seen as a linking hub connecting its entities to several different external authorities (Baker et al., 2019). In fact, they have aimed to become a focal point for interconnecting heritage collections and linking to other external data sources (Malyshev et al., 2018; Allison-Cassin and Scott, 2018).

Recent statistics in Wikidata Statistics (2020) demonstrated that Wikidata contains more than 86 million entities, approximately one billion statements, and over 800 million labels and descriptions that are available in many languages. Moreover, its entities are connected to more than 1750 different identifiers. However, cultural heritage entities are still quite briefly described within Wikidata.

On the flip side, there is a rich dataset with cultural heritage data, called the registry of the COURAGE (Cultural Opposition: Understanding the CultuRal HeritAGE of Dissent in the Former Socialist Countries) project. The COURAGE project was founded to study strategies for socialist-era cultural resistance during 1950–1990 and to highlight the variety of alternative cultural scenes that flourished in Eastern Europe before 1989 despite rigorous government control (see http://cultural-opposition.eu/). This project has gathered historic people, organisations, groups, collections, and featured items in an online RDF registry. The registry has been used to create virtual and real exhibitions, scientific publications, and learning material. It is also planned to serve as a basis for further narratives and digital humanities (DH) research (Apor et al., 2018).

The COURAGE Ontology underlying the registry contains approximately 100 classes, 220 object properties, and 170 data properties (Micsik, 2019). The main entities of the COURAGE dataset are collections, people, groups, and organisations. Also, some major events in their history and featured items are provided for each collection.

COURAGE has a scope limited in both time and region, but the data was 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 2–3 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.

Our study aimed to use data of the COURAGE registry to enrich Wikidata after mapping its entities with the COURAGE dataset and also to cross-check specific authority identifiers. In our earlier study (Faraj and Micsik, 2019), we successfully found all matching entities between Wikidata and COURAGE regarding people and organisations. We found that entities represented in both knowledge graphs are mostly of the types of person, group, and organisation. In order to make the enrichment process more comprehensive, our investigations were expanded to collections and their featured items (or one could say highlighted pieces). Additionally, the quality and coverage of authority data in significant international databases (VIAF, ISNI, and ULAN) have also been analysed based on curated knowledge in COURAGE and the authority identifiers stored in Wikidata.

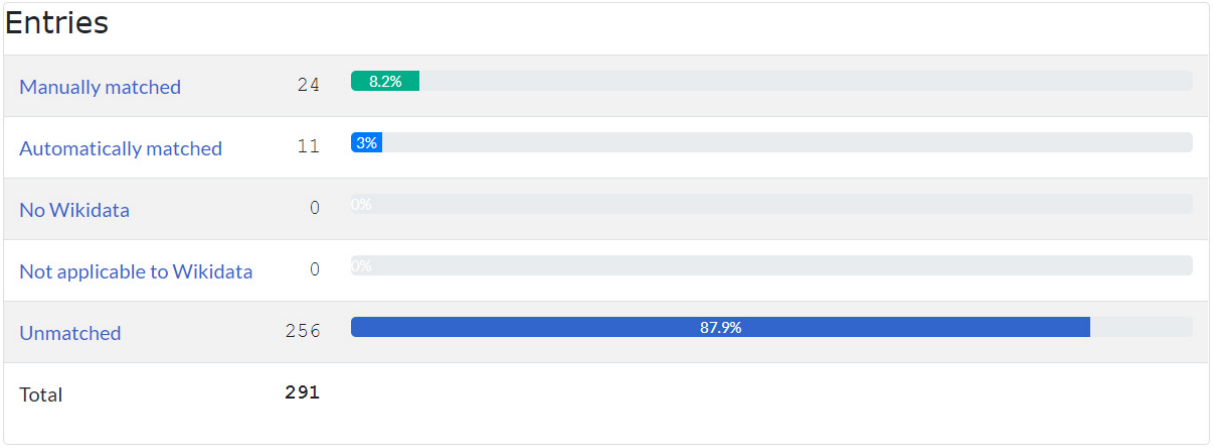
The remainder of the paper is organised as follows: Section 2 surveys the link discovery tools and entity resolution approaches that are related to our research. Section 3 describes preliminary statistics, the requirements for the matching approach, and how the matching process was carried out. Section 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. In Section 5, we analyse the corresponding connection links to other big authority databases (VIAF, ISNI, and ULAN). Finally, the summary of our contributions and the conclusion are in Section 6.

2 Related work

One of the main ideas about the web of data besides representing data to be understandable by a machine is to set mutual 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 by Nentwig et al. (2017), but most of them seem abandoned for three 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 (Nentwig et al., 2017). Silk and LIMES support more link types than other tools that just determine owl:sameAs and they provide a GUI for an interactive use (Isele et al., 2011; Ngomo and Auer, 2011). KNOFUSS just supports the owl:sameAs link type and string similarity approach (Nikolov et al., 2007). 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 farther 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 them often results in a 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.

Figure 1 Organisation 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 (see https://meta.wikimedia.org/wiki/Mix%27n%27match/Manual). We tried to use the tool with organisation entities but unfortunately, the outcomes were not useful (see Figure 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.

MusicWeb is a web-based application that integrates various linked open datasets in the topic of music. The authors' work relied on the SameAs.org service only for finding co-references between datasets (Mora-Mcginity et al., 2016).

The study by Hajra and Tochtermann (2017) aimed to enrich the scientific publications of Digital Libraries (DL) from other repositories. The authors increased the interlinking among different DLs considering all existing metadata such as title, authors, abstract, and keywords. In order to measure the relatedness of the retrieved publications from different repositories, TF-IDF and Cosine Similarity were used and compared with the Deep Learning approach through Word2Vec implementation of Word Embeddings.

Hickey and Toves (2014) manage ambiguity in VIAF by clustering similar authorities and analysing 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 by Larson and Janakiraman (2011), but only the names are used for matching.

Norway's Historical Population Registry (HPR) attempts to cover the population of the country and all resident places between 1800 and 1964, using the details found in censuses and church documents (Thorvaldsen, 2015). This project will be open for all relevant studies which develop the project content. In general, the HPR has quite precise and detailed data about persons compared to Wikidata.

Wikidata was used as a linking hub by Neubert (2017) to connect two economic-related authorities. The author linked Integrated Authority Files (GND) and Research Papers for Economics (RePEc) author identifiers with their corresponding ones in Wikidata. He applied a semi-automatic approach and matched them using Wikidata's Mix'n'match tool. Then, he used the existing VIAF identifiers to map additional entities. The QuickStatements tool was used also to add persons who did not already exist in Wikidata.

The LINKing System for historical family reconstruction (LINKS) project aimed to recreate the Dutch families of the nineteenth and early twentieth centuries. They used the GENLIAS project, which is a digitised index of all civil certificates from this period (see https://iisg.amsterdam/en/hsn/projects/links). During their work, they suffered from miss-spelling and ambiguity of the first and last names, but they solved it using dynamic parsing.

The study by Koho et al. (2020) describes the reconciliation process of person instances in several person registries. The authors applied a matching algorithm based on weights and linked entities to match people in the set of pre-existing person instances. All three registers were part of WarSampo (Finnish World War 2 on the Semantic Web).

Again, the projects listed above could rely on more detailed and curated data than what was available in Wikidata for our case.

3 Entity linking

Some basic properties in Wikidata such as birthplace, birthdate or profession are 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. On the other hand, COURAGE data were created by historians with thorough quality control.

As a first step, we carried out a quick analysis by matching entities based on a few properties. This analysis was to investigate and examine the currently available data about people, organisations, collections, and featured items in Wikidata and COURAGE datasets.

An entity in Wikidata is addressed by an opaque item identifier which starts with "Q" and a number. This entity is also presented on a page which consists of the following main parts: label, description, a set of aliases, a set of

statements, and a set of external links (Malyshev et al., 2018). The set of statements usually includes type, name, location, and date of birth or creation.

In COURAGE, entities also have a unique identifier, labels, and short descriptions in multiple languages, type, location, webpage, and a list of type-specific statements. Most of the entities have source references and lengthy documentation in at least two languages.

3.1 Preparatory investigations

For our investigations, we collected 556 collection entities and 855 featured item entities from COURAGE. Subsequently, we attempted to match them with Wikidata entities, based on a comparison of some properties: creator, operator (for collections), type, location, and year of creation. To our surprise, we found "The Book of Laughter and Forgetting" (Q2723517) as the only matching featured item entity and no matches for collections. Although Wikidata contains a large number of museums, the collections maintained by these museums were not in the focus of documentation in Wikidata. Regarding the featured items, some of them are art pieces, books, or movies. The reason for not being represented in Wikidata may be their contemporariness and their alternative, non-mainstream nature. We thought that including these would widen the cultural landscape offered by Wikidata, so we decided to inject them into Wikidata using the QuickStatements tool, as it is described later in Section 4.

Regarding person and organisation entities, we retrieved 1218 person entities with 3 properties: name, type, and birthdate from COURAGE. We performed a simple search based on these properties to find all possible Wikidata entities. After this, we classified matched pairs into groups on account of the clarity level of their matching decision. We found that 63.21% of matched person entities have the minimum requirements in both datasets (see Figure 2) to make the matching decision unambiguously (first group). The second group "Ambiguous matching decision" has 36.79% of entities falling into three sub-groups. In the first sub-group, automatic matching was hard to implement, consequently, it needs a human decision. In the second sub-group, there were deficient entities that made a human decision impossible to make due to the lack of data. The last sub-group is about false-positive cases. They were considered as matching entities, because of using two properties only. Therefore, we took more properties into our approach, examples are given later.

As for organisation entities in Wikidata and COURAGE, we used 4 properties for matching: name, type, country, and geocoordinates. The statistics, which were calculated for 457 organisations in COURAGE (see Figure 3), state that 58.84% of organisation 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 because they required a human decision, or they were considered false-positive cases.

Figure 2 Person result classification based on matching decision

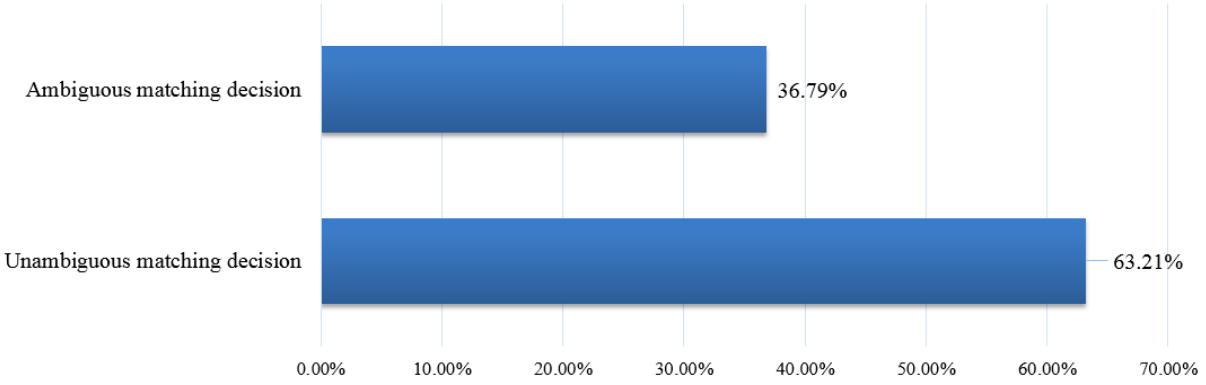
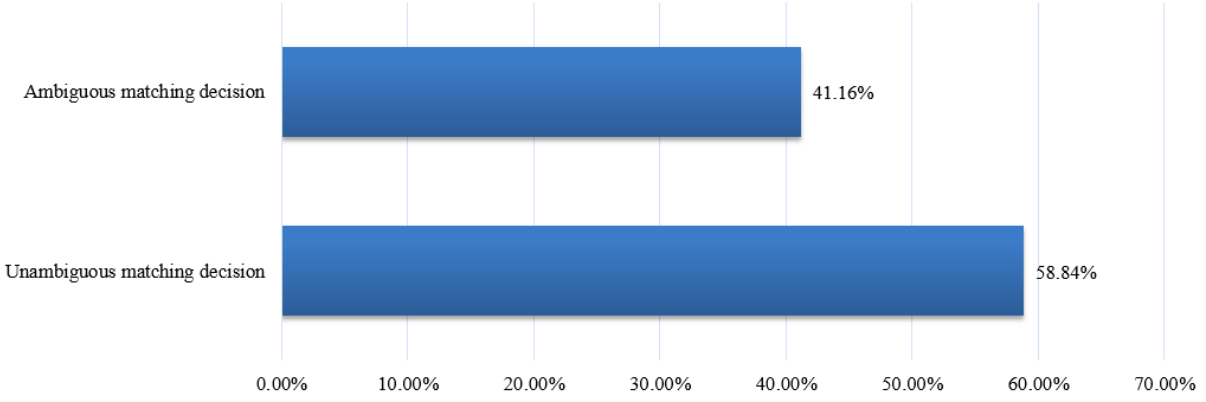


Figure 3 Organisation result classification based on matching decision



For example, "Gerhard Ortinau" (Q101211) is a person entity that 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 was 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 false positive matches. This entity was matched based on the same name and birthdate, but it turned out that it was not the same person as the birthplace was different. This is the reason why we increased the number of properties involved in the matching process.

3.2 Metrics for similarity

After checking the preliminary statistics of people and organisations, we set up a set of characteristics to identify organisation entities: name, city, country, geocoordinates, and year of founding. Besides, a second set to identify person entities: 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.

Based on the simple statistics, we determined how to find correct matching decisions. A scoring system was introduced to provide points for each candidate entity based on the matching status as below (see Figure 4).

Figure 4 The algorithm of calculating matching scores for persons

```
Matching Algorithm
    person= selected person entity in COURAGE
    points = 0
    Get Wikidata candidates wList based on p data
    For each person entity w in wList
        If person.name Exact match w.name then
            points += 4
        else If p.name contains w.name then
            points += 2
        else If Levenshtein distance(p.name, w.name) <=1 then
            points += 1
        else
            // no points are added
        If p.birthplace Exact match w.birthplace then
            points += 2
        else If missing values in p. birthplace or w. birthplace then
            points += 1
        else
            // no points are added
        If p.birthdate Exact match w.birthdate then
            points += 2
        else If missing values in p.birthdate or w.birthdate then
            points += 1
        else
            // no points are added
    End for
    Return points
```

The established metrics for matched person entities between COURAGE and Wikidata are:

- Name: The results of the comparison were checked after removing the diacritics: 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 the Wikidata entity is exactly equal to the COURAGE entity birthplace, it gets 2 points. But if one of the values is missing, the comparison value gets 1 point. Otherwise, it gets 0 points.
- Birthdate: similarly to persons' birthplace.

The metrics for organisations are per property:

- Name: similarly to persons' names.
- City and Country: if the 2 city properties and 2 country properties are exactly equal, it gets 4 points. if just the 2 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.
- Geocoordinates: 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: if the year of the Wikidata entity is exactly equal to the COURAGE entity foundation year it gets 2 points, or 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 (≤ 1.6 km), may get the most points.

3.3 Matching algorithm

We aimed to apply a reliable matching process on the person and organisation entities in Wikidata and COURAGE. The process goal was to minimise the cases where the human decision is needed. This approach was described in detail in our published paper (Faraj and Micsik, 2019).

An algorithm was developed in C# for matching person and organisation entities. First, organisation data was downloaded using the COURAGE SPARQL endpoint, cleaned, and imported into a local database. Then, all Wikidata candidates which were usually between 1 and 6 were retrieved based on the name containment. For each similarity with the characteristic identifications, we provided points based on the similarity status. Second, the total score was calculated based on the provided points and weights which were determined for each property. Likewise, the total score for person entities was calculated.

In order to determine the best weight combination for the taken sample, we generated all possible weight combinations between [0, 2] with a step increment of 0.1. The lower

threshold for *totalScore Tlo* is the largest threshold below which only non-matching pairs will be seen in this sample. The upper threshold *Tup* is the smallest threshold above which only matching pairs will be seen. Between *Tlo* and *Tup*, one finds the ambiguous pairs, which we called the *human decision window*. Our goal was to minimise the number of items in the human decision window.

As a result, we found that the foundation year of organisations can be discarded from the matching process, as it has no impact on the size of the human decision window.

Overall, the thresholds and weights related to the least items in the human decision window were applied to the entire sets of person and organisation entities respectively. After this, a random manual checking was performed without facing any error. The result of the statistics demonstrated that 78.64% of person entities and 80.5% of organisation entities could be safely matched automatically with Wikidata entities.

4 The enrichment process

After mapping COURAGE entities to Wikidata entities, we set up a list of transferable properties. A table of the corresponding properties in COURAGE and Wikidata was built. These properties were classified as properties used for matching and new properties.

As a next step, we collected the common properties between people and organisations as shown in Table 1. Regarding other properties that were also used for matching and enriching, they are displayed in the tables (Table 2, Table 3).

Table 1 General properties for both persons and organisations

Courage	*Wikidata*	
mainImage	P18/P154	Image/logo image
website	P856	official website
place	P276	location
Item Courage URI	P973	Described at URL

Table 2 Properties used for matching and enriching person data

Courage	*Wikidata*	
hasGivenName	P735	given name
hasFamilyName	P734	family name
birthDate	P569	date of birth
birthPlace	P19	place of birth
deathDate	P570	date of death
hasNickName	P1449	nickname
hasSex	P21	sex or gender
memberOf	P463	member of
ownerOf	P1830	owner of
hasCreatorRole	P6379	has works in the collection(s)
creatorOf	P170	inverse of creator

Table 3 Properties used for matching and enriching organisation data

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

Statements were generated in the format of the QuickStatements tool, which supports adding and removing data in Wikidata (Thorvaldsen et al., 2015). The QuickStatements file was generated from the matching database using a custom script. The file had 1765 statements for person and organisation entities. For person entities, we enriched 385 Wikidata entities successfully (Table 4).

Table 4 Sample of person properties in the generated file

Item	*Property*	*Value*	*Source*
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

In total 143 organisation entities were enriched (Table 5). We also generated another file with different syntax to create new entities based on the predefined list of properties (Table 8).

Table 5 Sample of organisations properties in the generated file

Item	*Property*	*Value*	*Source*
Q11179076	P276	Q1085 (Prague)	
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

After adding person and organisation entities, we aimed to place these agents into context, and show their roles and activities in recent history. We found that the creator and 'has works in collection' properties accept only Wikidata

entities as an object. Therefore, we planned to find the existing entities of featured items and collections in COURAGE and inject the new ones.

First, we created a list of transferable properties for collection and featured item entities as in Table 6 and Table 7 respectively.

Table 6 Properties used for enriching collection data

Courage	*Wikidata*	
hasTopic	P921	main subject
contentLanguage	P407	language of work or name
country	P17	country
place	P6375	street address
website	P856	official website
hasCreationDate	P571	inception
operator	P137	operator
collector	P6241	collection creator
Item Courage URI	P973	described at URL

Table 7 Properties used for enriching featured item data

Courage	*Wikidata*	
hasItemTopic	P921	main subject
contentLanguage	P407	language of work or name
collection	P195	collection
website	P856	official website
hasCreationDate	P571	inception
creator	P170	creator
Item Courage URI	P973	described at URL

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

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

As a next step, we generated triples in a similar format like the person and organisation triples to inject them using the QuickStatements tool. The file for collections has 9375 statements. After execution, 566 Wikidata collection entities were added successfully (Table 9). Subsequently, the 'has works in collection' property was added using a different file to all person entities which have a creator role in their corresponding collection in COURAGE.

Concerning the featured items file, it has 8110 statements as 852 Wikidata featured item entities were added successfully (Table 10).

Table 9 Sample for creating a new collection entity

Statements				
CREATE				
LAST	Len	"Invisible Society of Soviet-era Lithuania"		
LAST	Llt	"Nematoma sovietmečio visuomenė"		
LAST	P31	Q2668072		
LAST	P407	Q9083	S248	Q64784883
LAST	P17	Q37	S248	Q64784883
LAST	P6375	lt: "01130 Vilnius Vokiečių gatvė 10, Lithuania"	S248	Q64784883
LAST	P856	"http://www.visuomenesovietmeciu.tspmi.vu.lt/"	S248	Q64784883
LAST	P137	Q7931198	S248	Q64784883
LAST	P571	+2015-01-01T00:00:00Z/9	S248	Q64784883
LAST	P921	Q152416	S248	Q64784883
LAST	P921	Q832237	S248	Q64784883

Table 10 Sample for creating a new featured item entity

Statements				
CREATE				
LAST	Len	"Painting "Zodiako dvyniai" (Zodiac twins)"		
LAST	Llt	"Paveikslas "Zodiako dvyniai""		
LAST	P31	Q3305213		
LAST	P973	"http://courage.btk.mta.hu/courage/individual/n12291"		
LAST	P7037	"12291"		
LAST	P195	Q93272447	S248	Q64784883
LAST	P407	Q9083	S248	Q64784883
LAST	P170	Q12677671	S248	Q64784883

5 Investigation of authority identifiers

We already knew the matching ratio of authorities between COURAGE and Wikidata, but we were also interested in this aspect of other big authority databases. Fortunately, Wikidata contains the corresponding identifiers in other databases. VIAF, ISNI, and ULAN were selected for deeper investigation.

Linking and mapping COURAGE entities to their corresponding ones in Wikidata provides access to several authorities, in particular in the case of person and organisation entities. This advantage allows us to validate each pair in COURAGE and Wikidata with their respective entities in other registries. One of these authority IDs is VIAF (Virtual International Authority File) which is a joint project of many national libraries (see https://www.oclc.org/en/viaf.html). We chose VIAF because most of the person and organisation entities in Wikidata have a VIAF ID. Besides, it is well maintained for human names, organisations, and other bibliographic data by the participating national libraries. ISNI (International Standard Name Identifier) is our second choice of authority. It is an ISO standard for uniquely identifying the contributors to media content including artists, researchers,

publishers, and more (see http://www.isni.org/). The last one is ULAN (Union List of Artist Names) which is a free online database created by the Getty Research Institute and now it is maintained by the Getty Vocabulary Program. At the time of writing, it had over 300,000 artists and 720,000 names (see https://www.getty.edu/research/tools/vocabularies/ulan/).

To check the interlinking quality between COURAGE entities and their VIAF, ISNI, and ULAN identifier IDs via Wikidata, the following steps were applied:

Data collection. We retrieved person and organisation entity data including VIAF, ISNI, and ULAN IDs via the Wikidata SPARQL endpoint. Only entities mapped from COURAGE were selected.

Comparison of data. For person entities, name, birthdate, and deathdate properties were compared. For organisation entities, the comparison was based on name and country properties.

Comparison methodology. A C# program was written to check authority data. First, the VIAF, ISNI, and ULAN external links were retrieved for each entity in our database. To check the VIAF data, we read the entity XML file, and four properties (givenName, familyName, birthdate, and deathDate) were compared with their corresponding properties in the COURAGE dataset. The ISNI entity was treated in a similar manner after reading its XML file. Then, the properties: forename, surname, and marcDate were compared with their corresponding ones in COURAGE. The same procedure was performed for the organisation entity by examining mainName, and nameOfLocation.

In the case of ULAN, the JSON file of the entity was retrieved, and label, estStart, and estEnd properties of person entity were checked, and label and location properties of organisation entity were compared to their corresponding properties in COURAGE. Afterwards, the results of matching were saved in our database. Finally, manual random checking was performed to evaluate the results.

Data evaluation. During our investigation we faced problems that can be classified into three categories: 1) inappropriate links; when different entities are linked together. For instance, the entity "Šuhevič, Ûrìj-Bogdan" (315536053) in VIAF is related to "Yuriy Shukhevych" (Q4528122) in Wikipedia with a different name and country. We also found two invalid ISNI links and one incorrect ISNI ID. 2) Lack of data: five ambiguous Wikidata and VIAF ID pairs were found. The cause was missing significant data in all cases. Six uncertain pairs of Wikidata and ISNI IDs were also detected. 3) Duplicate links were found to both VIAF and ISNI. Altogether 21 items have at least two different links to ISNI and 17 items have duplicate links to VIAF from Wikidata. For example, the entity "National Széchényi Library" in Wikidata is linked to six similar VIAF entities. For ULAN IDs, all IDs were correct. Altogether, the number of errors found in these datasets may be seen as negligible.

We created summary statistics by counting the previously mentioned authority IDs for each person and organisation entity. The statistics for 660 person and 563 organisation entities are described in Table 11. The VIAF row refers to the number of entities that have at least one VIAF ID. The same holds for ISNI and ULAN rows. The 'Only VIAF' row contains the number of entities that have only VIAF ID and neither ISNI nor ULAN IDs (and similarly for only ISNI and only ULAN rows). The 'VIAF & ISNI' row presents the number of entities that have at least one VIAF ID and one ISNI ID. Similarly, the other rows were calculated. The 'None found' row presents the number of entities that contain neither of these IDs (see also Figure 5, Figure 6). It can be seen that only a few persons and organisations have ULAN IDs: 10% and 2% respectively. The cause may be that the ULAN dataset focuses on a specific domain which is digital art history and artist data. It can also be seen that the 'Only ISNI' and 'Only ULAN' categories contain just a couple of items, and thus they can be seen as a subset of the VIAF dataset.

Table 11 Person and organisation entity summary statistics

	People	*Organisations*
VIAF	520	184
Only VIAF	54	74
ISNI	466	113
Only ISNI	1	3
ULAN	58	8
Only ULAN	2	0
VIAF & ISNI	465	110
Only VIAF & ISNI	410	102
VIAF & ULAN	56	8
Only VIAF & ULAN	1	0
ISNI & ULAN	55	8
Only ISNI & ULAN	0	0
VIAF & ISNI & ULAN	55	8
None found	192	190
Total	606	361

ISNI has just 9% smaller person coverage than VIAF, but 20% smaller organisation coverage compared to VIAF in the case of COURAGE. The 'Only VIAF' categories (without ISNI and ULAN) contain 8.9% of person entities and 20.4% of organisation entities. On the other hand, VIAF has more errors and duplications due to the large amount of data that is loaded from many national libraries. Loading this amount of data makes VIAF much bigger than Wikidata as VIAF has over 35 million persons/organisations while Wikidata has just 8 million persons/organisations. The entity de-duplication task has many challenges, including the problem of many languages used and the ambiguity of spelling/transliteration of person names. Finally, still, 32% of persons and 53% of organisations in COURAGE fell into the 'None found' set. This set contains: 1) entities that do not have links from Wikidata to any of these datasets. For instance, the researcher Anna Dąbrowska could be found in VIAF with the id '165508374' but there is no link for this item between COURAGE and VIAF via Wikidata. 2) Entities that could not be found in any of these authority databases. For instance, the architect Ştefan Gane was not found in any of these authorities even using manual search by name.

Figure 5 Person authority IDs summary

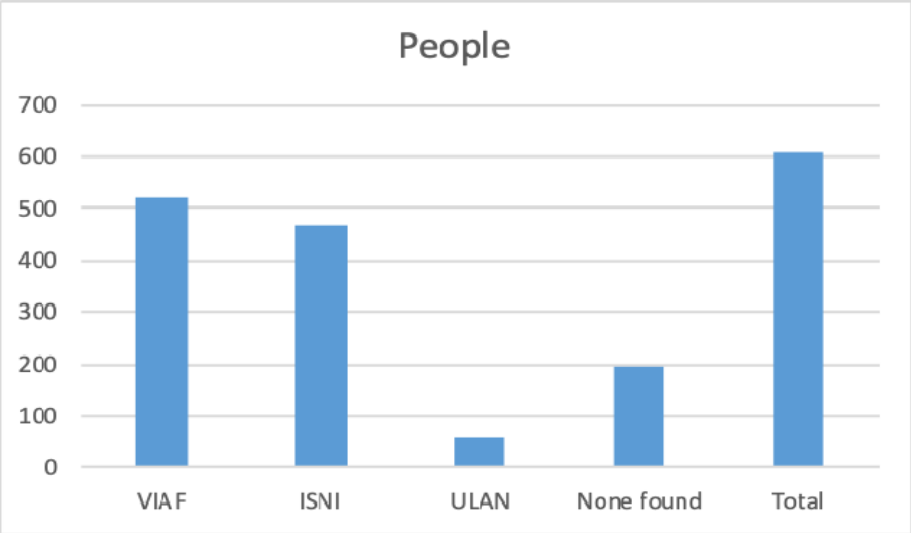
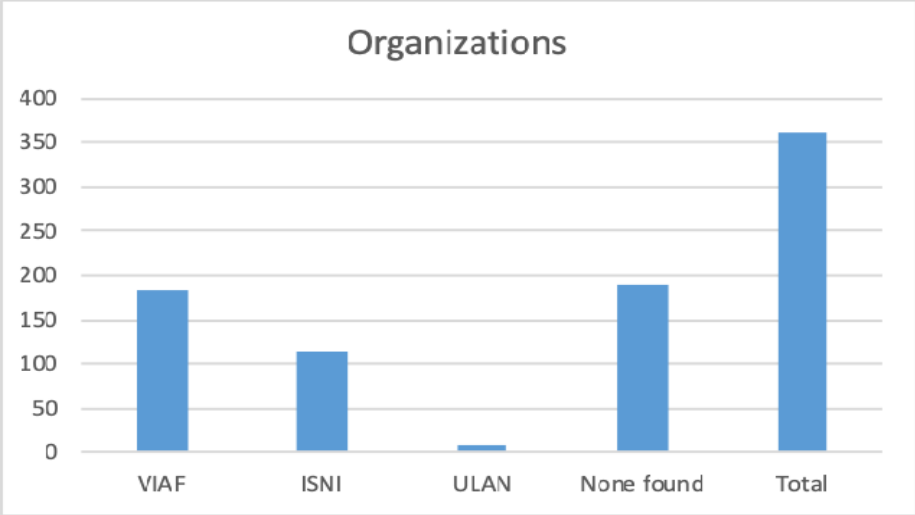


Figure 6 Organisation authority IDs summary



6 Conclusion

Our goal was to enrich and link the pre-existing person, organisation, collection, and featured item entities in Wikidata with the results composed and collected by a hundred experts in the COURAGE project. The dashed lines in Figure 7 summarise the new link types established during our work. On the Wikidata side, the involved entities contain already existing and newly created items as well. As the coverage of COURAGE entities in Wikidata was much smaller than expected, we had to add 1779 new entities in total to Wikidata so that we could represent most of the interconnections among the entities revealed during the research in COURAGE. Furthermore, we inspected the linking quality in more detail regarding person and organisation COURAGE entities. Data from three big authority datasets (VIAF, ISNI, and ULAN) were compared with COURAGE data.

We tested various available solutions for entity linking, but we had to create a new solution in order to reach the desired quality of mappings. In this solution we tried to minimise the need for human decisions, hence we introduced the term human decision window for the mappings where neither acceptance nor refusal can be made automatically and safely. Still, we found that in our case about 20% of the items to be mapped fell into the human decision window.

Our final results as shown in Figure 8: 77% of COURAGE entities were mapped to Wikidata entities. 23% of the entities out of 77% existed in Wikidata. These existing entities were enriched with the available data from COURAGE. The remaining 54% of new entities were injected into Wikidata. Regarding the 23% of unmapped entities, their data was not published for public usage yet in COURAGE and some of them have privacy on the most important properties so adding them has no added value to our work.

Figure 7 Main connections inside COURAGE and between COURAGE and Wikidata

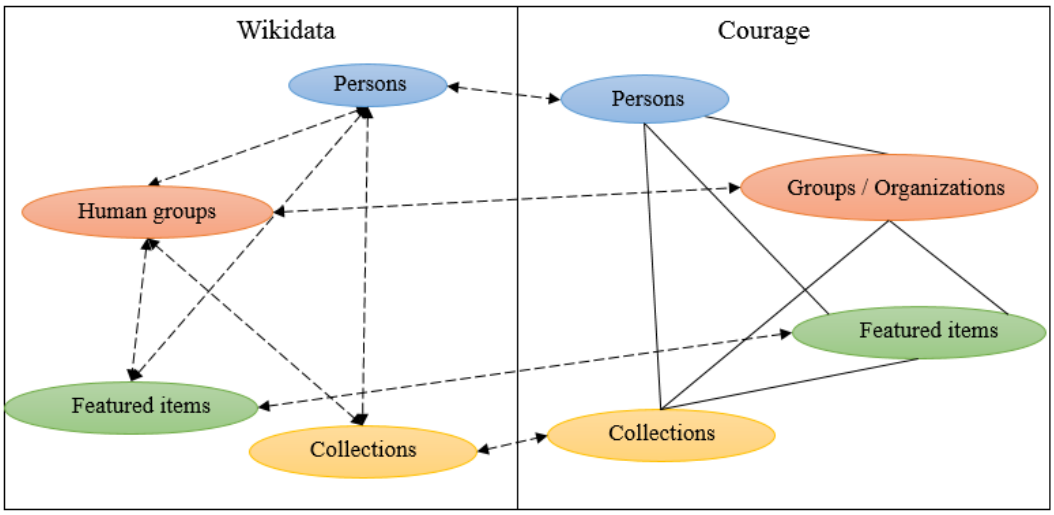
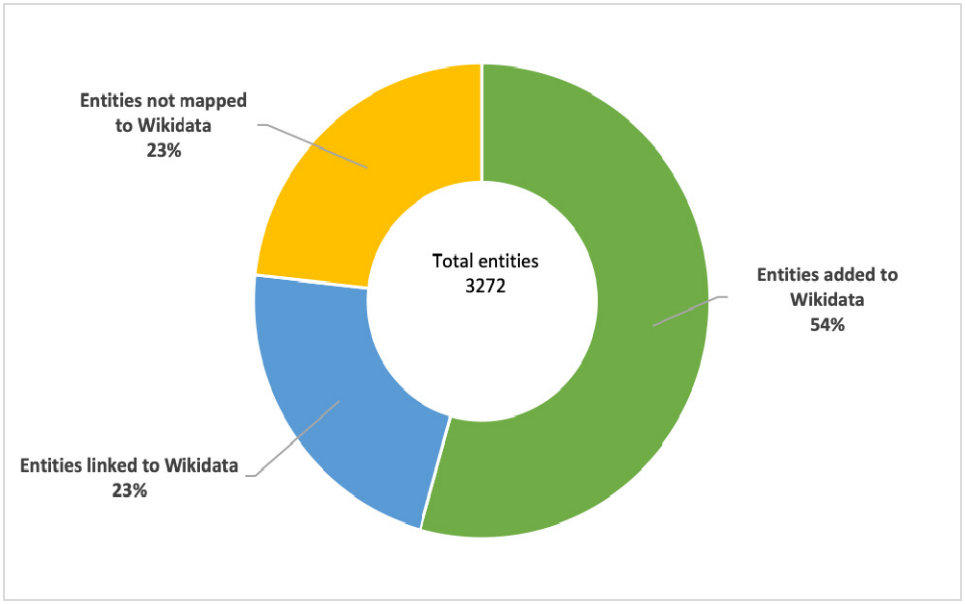


Figure 8 The percentage of linking between COURAGE and Wikidata entities



The COURAGE project has high-quality cultural heritage linked data related to the period 1950-1990, which has been successfully linked to Wikidata entities. We applied a score-based method on matched person and organisation entities and successfully performed an automated link discovery on 78% of person entities and 80% of group/organisation entities. Subsequently, we continued with adding internal linkages between the previously mentioned Wikidata entities using properties: leader, founder, creator, has works in the collection, collector, etc. By mapping COURAGE topics to Wikidata general concepts, the findability of the new resources has also been improved.

Linking to Wikidata on the COURAGE site provides also access to other authority IDs. Matched pairs of COURAGE and Wikidata entries were validated with their respective entities in VIAF, ISNI, and ULAN registries by comparing the common properties of person and organisation entities. This study gives us a better understanding of the differences in coverage and quality of linking in authority data, at least within the period of the last 70 years.

References

Allison-Cassin, S. and Scott, D. (2018) 'Wikidata: a platform for your library's linked open data', *Code4Lib*, Vol. 40.

Apor, B., Apor, P. and Horváth, S. (Eds) (2018) *The Handbook of COURAGE*, Budapest, doi: 10.24389/handbook.

Baker, T., Neubert, J. and Waagmeester, A. (2019) *Wikidata as a hub for the linked data cloud*, Tutorial at DCMI conference in Seoul. Available online at: https://jneubert.github.io/wd-dcmi2019/#/

Erxleben, F., Günther, M., Krötzsch, M., Mendez, J. and Vrandečić, D., (2014) 'Introducing Wikidata to the linked data web', *International Semantic Web Conference*, Springer, pp.50–65, doi: 10.1007/978-3-319-11964-9_4.

Europeana (2017) *Why data partners should link their vocabulary to Wikidata: a new case study*. Available online at: https://pro.europeana.eu/post/why-data-partners-should-link-their-vocabulary-to-wikidata-a-new-case-study

Faraj, G. and Micsik, A. (2019) 'Enriching Wikidata with cultural heritage data from the COURAGE project', *Metadata and Semantic Research. 13th International Conference, MTSR 2019*, Rome, Italy, 28–31 October, *Communications in Computer and Information Science*, Vol. 1057, Springer, p.460, https://doi.org/10.1007/978-3-030-36599-8_37.

Hajra, A. and Tochtermann, K. (2017) 'Linking science: approaches for linking scientific publications across different LOD repositories', *International Journal of Metadata, Semantics and Ontologies*, Vol. 12, No. 124, doi: 10.1504/IJMSO.2017.090778.

Hickey, T.B. and Toves, J.A. (2014) 'Managing ambiguity in VIAF', *D-Lib Magazine*, Vol. 20, Nos. 7/8, doi: 10.1045/july2014-hickey.

Isele, R., Jentzsch, A. and Bizer, C. (2011) 'Efficient multidimensional blocking for link discovery without losing recall', *14th International Workshop on the Web and Databases*, WebDB, Athens.

Koho, M., Leskinen, P. and Hyvönen, E. (2020) *Integrating Historical Person Registers as Linked Open Data in the WarSampo Knowledge Graph*. Available online at: https://seco.cs.aalto.fi/publications/2020/koho-et-al-integrating-person-registers.pdf

Larson, R. and Janakiraman, K. (2011) 'Connecting archival collections: the social networks and archival context project', *Research and Advanced Technology for Digital Libraries. Lecture Notes in Computer Science*, Vol. 6966, Springer, Berlin, Heidelberg, doi: 10.1007/978-3-642-24469-8.

Malyshev, S., Krötzsch, M., González, L., Gonsior, J. and Bielefeldt, A. (2018) 'Getting the most out of Wikidata: semantic technology usage in Wikipedia's knowledge graph', *17th International Semantic Web Conference, Monterey, Proceedings, Part II*, CA, USA, 8–12 October, pp.376–394, doi: 10.1007/978-3-030-00668-6_23.

Micsik, A. (2019) *Courage registry - open dataset 1.1*, doi: 10.5281/zenodo.3333540.

Mora-Mcginity, M., Allik, A., Fazekas, G. and Sandler, M. (2016) 'MusicWeb: music discovery with open linked semantic metadata', Garoufallou, E., Subirats Coll, I., Stellato, A. and Greenberg, J. (Eds): *Metadata and Semantics Research. MTSR 2016. Communications in Computer and Information Science*, Vol. 672, Springer, Cham, doi: 10.1007/978-3-319-49157-8_25.

Nentwig, M., Hartung, M., Cyrille, A., Ngomo, N. and Rahm, E. (2017) 'A survey of current link discovery frameworks', *Semantic Web Journal*, Vol. 2, No. 224, doi: 10.3233/SW-150210.

Neubert, J. (2017) 'Wikidata as a linking hub for knowledge organization systems? Integrating an authority mapping into Wikidata and learning lessons for KOS mappings', *NKOS@TPDL 2017*, pp.14–25.

Ngomo, A.C.N. and Auer, S. (2011) 'LIMES – a time-efficient approach for large-scale link discovery on the web of data', *IJCAI*, pp.2312–2317, doi: 10.5591/978-1-57735-516-8/IJCAI11-385.

Nikolov, A., Uren, V. and Motta, E. (2007) 'KnoFuss: a comprehensive architecture for knowledge fusion', *Proceedings of the 4th International Conference on Knowledge Capture*, ACM, pp.185–186, doi: 1298406.1298446.

Thorvaldsen, G., Andersen, T. and Sommerseth, H.L. (2015) 'Record linkage in the historical population register for Norway', *Population Reconstruction*, Springer, pp.155–171, doi: 10.1007/978-3-319-19884-2.

Vrandečić, D. and Krötzsch, M. (2014) 'Wikidata: a free collaborative knowledgebase', *Communications of the ACM*, Vol. 57, No. 10, doi: 10.1145/2629489.

Wikidata Statistics (2020) Available online at: https://www.wikidata.org/wiki/Wikidata:Statistics

WikiProject Cultural Heritage (2020) Available online at: https://www.wikidata.org/wiki/Wikidata:WikiProject_Cultural_heritage