Semantic Web 14 (2023) 55–100
DOI 10.3233/SW-223026
IOS Press

Instance level analysis on linked open data connectivity for cultural heritage entity linking and data integration

Go Sugimoto [a,b,c]

[a] *Austrian Centre for Digital Humanities and Cultural Heritage, Sonnenfelsgasse 19, 1010, Vienna, Austria*
[b] *Donau University Krems, Dr.-Karl-Dorrek-Straße 30, 3500 Krems, Austria*
[c] *Vrije Universiteit Amsterdam, De Boelelaan 1081, NU Building, 1081 HV, Amsterdam, The Netherlands*
E-mail: go_savethequeen@hotmail.com

Editors: Jose Emilio Labra Gayo, University of Oviedo, Spain; Anastasia Dimou, IDLab, Ghent University, Belgium; Katherine Thornton, Yale University Library, USA; Anisa Rula, University of Milano-Bicocca, Italy and University of Bonn, Germany
Solicited reviews: Miel Vander Sande, Ghent University–imec–IDLab, Belgium; Herminio Garcia-Gonzalez, Kazerne Dossin, Belgium; Joe Raad, Vrije Universiteit Amsterdam, Netherlands; Efstratios Kontopoulos, Centre for Research & Technology – Hellas, Greece; two anonymous reviewers

Abstract. In cultural heritage, many projects execute Named Entity Linking (NEL) through global Linked Open Data (LOD) references in order to identify and disambiguate entities in their local datasets. It allows users to obtain extra information and contextualise the data with it. Thus, the aggregation and integration of heterogeneous LOD are expected. However, such development is still limited partly due to data quality issues. In addition, analysis on the LOD quality has not sufficiently been conducted for cultural heritage. Moreover, most research on data quality concentrates on ontology and corpus level observations. This paper examines the quality of the eleven major LOD sources used for NEL in cultural heritage with an emphasis on instance-level connectivity and graph traversals. Standardised linking properties are inspected for 100 instances/entities in order to create traversal route maps. Other properties are also assessed for quantity and quality. The outcomes suggest that the LOD is not fully interconnected and centrally condensed; the quantity and quality are unbalanced. Therefore, they cast doubt on the possibility of automatically identifying, accessing, and integrating known and unknown datasets. This implies the need for LOD improvement, as well as the NEL strategies to maximise the data integration.

Keywords: Linked Open Data quality, graph traversals, connectivity, RDF, Named Entity Linking, cultural heritage contextualisation, data integration, R, network analysis

1. Introduction

In recent years, Linked Open Data (LOD) has been widely acknowledged and data rich institutions have generated a large volume of LOD. As of May 2020, the LOD Cloud website reports 1,301 datasets with 16,283 links.[1] The real power of LOD originates from a very simple philosophy of the Web inventor. Berners-Lee [5] states *"include links to other URIs. so that they can discover more things"*, hence the name "Linked" (Open) Data. LOD transforms

[1] https://lod-cloud.net/#about, last accessed 2022-01-22.

distributed data in Resource Description Framework (RDF) into a connected global knowledge graph and allows us to find and formulate new information and knowledge [6]. This vision seems to be particularly suited for research activities. However, it seems that this scenario is not happening as quickly as we expected. It is still unclear whether we have discovered something significant in this manner. One of the reasons for this problem is the gap between the LOD producers and consumers, which is heavily attributed to data quality. Zaveri et al. [45] state that there is less focus on how to use good quality data than to how to publish it.

In this paper, we explore the problems of LOD quality from the user's point of view. In particular, we analyse the linking quality of LOD from a research perspective in the field of cultural heritage and Digital Humanities (DH). Our study on this fundamental aspect of LOD should be able to provide a better understanding of a bottleneck of LOD practices. Although we concentrate on these domains, we believe that our analysis is equally valuable in other domains, because the analysed data is highly generic.

In cultural heritage and DH, many projects create and use a wide range of LOD for research purposes. In the course of populating and improving LOD, they often execute curatorial tasks such as Named Entity Recognition (NER), entity extraction, entity/coreference resolution, and Named Entity Linking (NEL) [7,10,16,41,46]. These are the tasks to identify, disambiguate, and extract entities/concepts from data, and to reconcile and make references to entities in another data. Thus, we can find more information on the web. In this article, we use NEL as a catch-all term for all these tasks.

For example, Europeana executes NEL in a large number of cultural heritage datasets and creates links to widely known LOD sources including GeoNames, DBpedia, and Wikidata that this paper discusses [35,37]. Jaffri et al. [28] echo this view, stating that many datasets are linked with DBpedia entities through the `owl:sameAs` property. In practice, this means that information about the same entity (e.g., place, person, event etc.) is stored in different LOD datasets on different servers. As Tomasuzuk and Hayland-Wood [39] indicate, RDF enables us to join data stored at disparate sources and provide the user with an integrated perspective of this data. This is called data integration. For instance, if one dataset only supplies partial information about an entity, NEL allows us to retrieve more information from all linked datasets, by "merging" data through links. In this regard, NEL serves as a building block of LOD, fostering connection, compilation, aggregation, and contextualisation of (distributed) information.

What is not investigated in cultural heritage and DH is, what impact NEL and subsequent data integration have for future research? Currently, there is a tendency for entity linking to become a purpose by itself, without examining the consequences of the linking. Due to the relative infancy of LOD in the field, perhaps most effort has been put into the aspect of data discoverability on the web, which NEL also facilitates. This function of LOD may not require extensive use cases after NEL is performed. In any case, data producers are often not fully aware of the next steps for research using LOD, as well as the needs of the data users. Although not limited to these domains, Data on the Web Best Practices[2] observes: "the openness and flexibility of the web create new challenges for data publishers and data consumers, such as how to represent, describe and make data available in a way that it will be easy to find and to understand".

Currently, the benefit of data integration using NEL is often restricted to the data sources within a single institution or domain. For instance, an advanced semantic search is developed for the historical newspapers in the Netherlands [42]. In fact, the investigation of the aggregation and integration of heterogeneous LOD from different data providers is rather rare [16], or done with relatively small multiple sources. A few exceptional cases are found in museums and institutions in France [1] and Spain [29]. Still, the formation of new knowledge based on complex queries across distributed LOD resources is not easily implemented. As such, the full potential of LOD has been neither fully explored nor verified. The practice of LOD-based research using distributed data still faces many challenges.

In terms of data linking quality, computer science communities have intensively worked on this issue in the past years. Critical quality issues of linking have been frequently raised and discussed in the studies of LOD [3,4,9,11–15,23,31–34,43,45]. We discuss this in Section 2 in more detail. However, one specific aspect helps here to explain our motivation. Most previous research regards `owl:sameAs` as a central property for LOD linkages, because it is a W3C recommended standard and serves as a bridging link between identical entities. We also think that it plays an important role to automate data processing using federated SPARQL queries in dispersed datasets, because we

[2] https://www.w3.org/TR/dwbp/ last accessed 2021-01-26.

know the property beforehand without knowing heterogeneous and complicated ontologies of individual datasets. At least there is no doubt that LOD information can be automatically traversed and aggregated by simply following the links through this property. Therefore, we are interested to understand the future prospect of LOD automation by examining commonly used properties.

Taking this background into account, this article aims to evaluate the quality of widely known (referential) LOD as the target resources of NEL. In particular, the linking quality and connectivity is analysed in detail in order to provide an overview of the current "state of NEL ecosystem". To this end, we examine LOD entities/instances through lookups. With a special emphasis on multi-level traversability in the LOD cloud, we can estimate the impact of NEL for end-users. In other words, our research questions are as follows.

- **RQ1**: When a local dataset links to a global LOD, what level of information can we find?
- **RQ2**: How can we follow links *"to discover more things"*?
- **RQ3**: How are the entities in (the core part of) the LOD cloud connected to each other and can be navigated?
- **RQ4**: What kind of information can be obtained by automatic graph traversals through standardised properties like `owl:sameAs`?
- **RQ5**: What are the linking and content patterns for different types of entities?

As LOD potentially enables us to undertake machine-assisted research with the help of more automated data integration and processing, this project serves as a reality check for the current practices of LOD in the field.

The structure of this paper is as follows. Section 2 explores the related research. Section 3 describes objectives, scopes, and methodology. Section 4 presents the analysis of 100 entities in five categories relevant to cultural heritage data integration and contextualisation. The final section summarises the discussions and outlines ideas for future work.

2. Related work

Over the last years quantitative research has been carried out intensively for the LOD quality. The landscape of previous studies is examined in an in-depth survey by Zaberi et al. [45]. They analyse 30 academic articles on data quality frameworks and report 18 quality dimensions and 69 metrics, as well as 20 tools. Many studies investigate the linking quality, but some aim to assess broader aspects of LOD quality. For instance, Färber et al. compare DBpedia, Freebase, OpenCyc, Wikidata, and YAGO with 34 quality criteria [20]. They span from accuracy, trustworthiness, and consistency to interoperability, accessibility, and licences. Schmachtenberg et al. [34] update the 2011 report on LOD, using the Linked Data crawler, analysing the change of LOD (8 million resources) over the years. Debattista et al. [13] provide insights into the quality of 130 datasets (3.7 billion quads), using 27 metrics. However, the linking on which this paper would like to focus is a small part of the metrics. Mountantonakis and Tzitzikas [31] have developed a method for LOD connectivity analysis, reporting the results of connectivity measurements for over 2 billion triples and 400 LOD Cloud datasets. A rather unusual project has been conducted by Guéret et al. [22]. They concentrate on the creation of a framework for the assessment of LOD mappings using network metrics. They specifically look into the quality of automatically created links in the LOD enrichment scenario.

In parallel, a number of valuable contributions have been made to scrutinise `owl:sameAs` and "problem of co-reference" [28]. Firstly, there are critical discussions about the proliferation of `owl:sameAs` semantics [23]. Secondly, several large scale statistical analyses uncover the status of `owl:sameAs` networks to detect errors for 558 million links [32], verify the proliferation [14,15] (4352 and 8.7 million links respectively), and propose solutions. Most projects concentrate on macro studies and statistical observations of the comprehensive cross-domain LOD cloud, applying metrics to measure the data quality through dumps and SPARQL endpoints. Their methodologies help us to gain a holistic view of the development of the LOD cloud in terms of linking quality.

There are also a few examples of "semi-micro" research, using domain specific datasets. Ahlers [3] analyses the linkages of GeoNames (11.5 million names). He reveals some cross-dataset and cross-lingual issues and distribution biases. Debattista et al. [11] inspect the Ordnance Survey Ireland (50 million spatial objects) in order to identify errors in the data mapping for the LOD publishing and check the conformance to best practices. Although the

datasets pass the majority of 19 quality metrics in the Luzzu framework [12], the low number of external links (only DBpedia) is clearly our concern.

The studies for the cultural heritage domain are relatively new. Candela et al. state that there has been so far no quantitative evaluation of the LOD published by digital libraries [8]. They systematically analyse the quality of bibliographic records from four libraries with 35 criteria covering 11 dimensions to provide a benchmark for the library community. The research on the LOD quality for a broader cultural heritage including museums and archives is scarce.

Apart from Mountantonakis and Tzitzikas, macro research projects oftentimes treat data sources (or corpora) as a whole, when investigating `owl:sameAs` link connectivity. In other words, the data connectivity is examined regardless of the user mobility at an instance level. For example, their research does not reveal if the connection for a specific instance such as Mozart is available between data source A and B, even if they detect many links between the instances in the two sources. This is because the domain coverage may be different: A originates from a Polish library and B from a Greek museum. Mozart could be found in both, but could be in neither. To this end, it is necessary to observe trees (Mozart as an instance) not forests (the data source A and B as a collection of instances).

In addition, most macro analyses are not designed for multiple graph traversals. One of the exceptions is Idrissou et al. [26] who indeed claim that gold standards for entity resolution do not go beyond two datasets. Interestingly, they develop hybrid-metrics that combine structure and link confidence score to estimate the quality of links between entities for six datasets from the social science domain. Although we agree that accurate automated evaluation of links is much needed, our study aims to gain deeper understanding of smaller sampling entities.

Going back to our analogy, we currently cannot know how much and what kind of data we can find by following a link from Mozart in data source A to an entity in source B, which provides links to an entity in source C. Therefore, a close observation of instances is needed. The instance level maneuverability indicates whether and how users can navigate themselves in the knowledge graphs and can obtain related information from various data sources, and potentially integrate them.

3. Objectives and methodology

We explain the process of defining objectives and methodology in four sub-sections. The first section describes the scope of the linking quality evaluation. The second section discusses the nature of research in cultural heritage and DH in relation to conceptual models and ontologies, in order to specify the object of analysis. The third section details the data sampling. The fourth section deals with the technical methods of a wide range of analyses.

3.1. Scope of analysis and graph traversals

This paper will not repeat the comprehensive statistical analyses on the LOD quality according to the existing or newly created comprehensive metrics. In contrast to previous research, we deploy a micro analysis. Our research deals with a small ecosystem of LOD in the cultural heritage NEL, based on an empirical qualitative and quantitative method. In particular, it focuses on user maneuverability for arbitrary LOD entities. We analyse multi-level graph traversability using standardised properties, especially bearing the automatic data traversals and integration in mind.

The primary goal is to create "traversal maps" of major LOD data sources at an instance level. "Traversal maps" are maps illustrating all possible routes of graph traversals in the LOD cloud (RQ3). We specialise in the route of standardised properties including `owl:sameAs` (RQ4). Naturally, the collections of instances covering the same topic (i.e. categories in Section 3.2) are of vital importance for the analysis (RQ5). Subsequently, it is expected to provide a better understanding of which referential resources are accessible in what way between multiple sources (RQ1 and RQ2). This scope enables us to deliver an observation more from the data user's perspective than the producer's. The traversal maps should be helpful for the end-users to orient themselves in the LOD cloud and formulate strategies for data navigation and integration to capitalise NEL.

The use case for the LOD traversals in this article is the following: we/user manually look up a LOD entity/resource identified. Then, they follow available links in the entity to reach identical and/or the most related LOD

resources. For example, one may traverse an RDF graph from a resource in DBpedia to a resource in Wikidata via `owl:sameAs`:

```
dbr:1969 owl:sameAs wd:Q2485 .
```

Hyperlinks are documented and counted to generate traversal maps. To support the link quality analysis, information about other content is also documented and counted (RQ5). It includes the amount of `rdfs:label`, `rdf:type`, `skos:prefLabel` and `skos:altLabel` as well as `rdf:resource`, and `rdf:about` (see Section 3.4). The traversal continues as long as it is within the specified datasets boundaries (see Section 3.3). The reason to evaluate lookups instead of data dumps and SPARQL queries is that they play a vital role to publicly and openly raise awareness of the data existence that NEL essentially needs. To our knowledge, none of the previous studies works on lookups.

Regarding the link types, the W3C recommended properties, `owl:sameAs`, `rdfs:seeAlso`, and `skos:exactMatch` are used.[3] It is a common practice that information providers set `owl:sameAs` links to URI aliases [4,6]. In addition, `schema:sameAs` is included, due to its popularity. One of the advantages of those standards is that the properties are widely known (see Section 2), implying no prior knowledge is required to access and process data. As Hartig [24,25] observes, it is highly important that the end users can obtain data from initially unknown data sources. In other words, they should be able to discover new LOD sources at run-time by following RDF links [6].

Since `rdfs:seeAlso` may be asymmetric, our analysis is not limited to LOD and symmetric graphs. This means that the sources and destinations of incoming and outgoing links are not 100% synchronised as identical LOD entities. For example, "Italy" in Getty TGN contains `rdfs:seeAlso` for an HTML representation (http://www.getty.edu/vow/TGNFullDisplay?find=&place=&nation=&subjectid=1000080). This is allowed in the specification.[4] Another reason to avoid strict co-references is that it is hard to find and evaluate the same identity only by URIs. For instance, a VIAF record provides a link to Getty ULAN in the following syntax: http://vocab.getty.edu/ulan/500240971-agent. This resolves to http://vocab.getty.edu/ulan/500240971. In general, redirects introduce technical complexity for the analysis. As a consequence, the links to the same domain name in the URIs (e.g. getty.edu is same as vocab.getty.edu) are regarded as the same destination, regardless the identity and format of the entity. In this way, our analysis attempts to bypass complicated discussions over the accurate semantics of properties such as `owl:sameAs` [23].

When assessing the quality of LOD, proprietary properties cannot be ignored. They often contain interesting and specialised information. However, we put less emphasis on them. Compared to standardised properties, these properties may not be frequently used as a means to connect the data sources within the core part of the LOD cloud. Another reason is extensively explained in Section 3.4 in the context of difficulties in the data quality comparison, and our compromised approach is described.

Documentation on an instance is recorded in separate tabs in a spreadsheet for each source. VBA scripts are created to aggregate and/or facet datasets. Subsequently, various types of tables and charts are generated. In order to increase the research transparency and reproducibility, our datasets and documentation are fully archived in the Zenodo Open Access repository (https://doi.org/10.5281/zenodo.5913136).

3.2. Core questions and contextualisation in cultural heritage ontology

In order to narrow the scope of the LOD evaluation, this article focuses on addressing typical and generic core questions for cultural heritage and DH alike. For instance, one of the largest cultural heritage data platforms is Europeana. It has created the Europeana Data Model (EDM)[5] in order to capture heterogeneous cultural heritage information. Its Primer[6] notes that "EDM will let users browse Europeana in revealing new ways. It answers the

[3] One property per ontology is selected.

[4] https://www.w3.org/TR/rdf-schema/, last accessed 2021-01-26.

[5] https://pro.europeana.eu/resources/standardization-tools/edm-documentation, last accessed 2021-01-26.

[6] https://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Primer_130714.pdf, last accessed 2021-01-26.

'Who?', 'What?', 'When?', 'Where?' questions, and makes connections between the networks of stories that will animate Europeana's content". EDM features five classes (agent, event, place, time-span, concept) for this purpose, which are called contextual entities, because they enrich and "contextualise" cultural heritage objects. Although these 4'W' questions are common sense for scientific research in general, they manifest the essence of cultural heritage research: without them, researchers are hardly able to solve any other research questions in their disciplines. Thus, they provide the contextualisation or foundation of research.

The importance of the four core questions is also reflected in other cultural heritage ontologies. CIDOC-CRM "provides the "semantic glue" needed to mediate between different sources of cultural heritage information, such as that published by museums, libraries and archives".[7] It centres "Event" as a core entity, connecting "Agent", "Time-Span", "Objects", and "Place". In the library sector, DCMI Metadata Terms[8] also defines almost identical entities: "Agent", "PeriodOfTime", "PhysicalResource", and "Location" among others. In addition, FRBR[9] is a conceptual reference model for libraries which introduces hierarchical concepts of cultural works (i.e. work, manifestation, expression, and item). The Group 1 entities (the products of intellectual and artistic endeavor) are relevant to the What question, whereas the Group 2 entities (person and corporate body) are related to Who. Group 3 (the subjects of intellectual or artistic endeavor) is associated with other W-questions.

Therefore, the evaluation of LOD in this article concentrates on these four questions and use them as categories of our investigation. We employ the following terminology to be more specific: agents (for Who), events (for What), objects and concepts (for What), dates (for When), and places (for Where). Due to the genericness of the categories, investigating the five categories not only helps us to answer our research questions, but also makes our analysis valuable for research outside the cultural heritage field.

3.3. Data sources

Our study introduces two basic strategies for the selection of datasets/data sources. It examines LOD in (1) RDF/XML with (2) unrestricted look-up access (i.e. no API keys). Although there are other RDF serialisation formats, RDF/XML is the only commonly available one for all the data sources described below.[10] On top of the technical setup, we consider popularity (through literature [8,16,36,46]), data volume, coverage, and actual linkages for the selection. The aforementioned LOD cloud is also taken into account, as one of the comprehensive visualisations of LOD networks. Consequently, the following nine data sources which include significant content for the cultural heritage and DH are chosen for examination: (1) Getty vocabularies (ULAN (Union List of Artist Names), AAT (Art & Architecture Thesaurus), and TGN (Thesaurus of Geographic Names)), (2) GeoNames, (3) VIAF, (4) WorldCat FAST, (5) DBpedia, (6) Wikidata, (7) the Library of Congress, (8) BabelNet, and (9) YAGO.

There are two exceptions for the selection criteria. Wikipedia delivers its articles in HTML, but it may be studied as an indicator, because it has a unique position as a global reference on the web inside and outside the LOD context [2,3,30,41]. Indeed, the data in DBpedia and YAGO are derived from Wikipedia.[11] Wikidata has a close tie with Wikipedia project. The other case is Europeana. It provides an alpha version API with a public API key.[12] However, it is one of the most valuable LOD resources in the cultural heritage sector, and therefore, it is included.

As this study deploys a qualitative analysis, a manageable level of data sampling is considered. It selects twenty representative instances/entities from five categories defined in the Section 3.2 (Table 1), resulting in 100 entities in total.[13] In order to objectively and systematically select the most relevant entities, we consulted the "Wikipedia

[7] http://www.cidoc-crm.org/, last accessed 2021-01-26.

[8] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/, last accessed 2021-01-26.

[9] https://www.ifla.org/publications/functional-requirements-for-bibliographic-records, last accessed 2021-01-26.

[10] This is mainly due to GeoNames that only provides RDF/XML, KLM, and HTML representation for lookups. This is already a discovery of LOD quality in terms of standardisation.

[11] YAGO2 is used for our study.

[12] https://pro.europeana.eu/page/entity, last accessed 2021-01-26. Only those who read the documentation can find the API key and the URI syntax to access the lookup service.

[13] For practical reasons, it concentrates on the English version as the primary resource of an entity when multiple language versions exist (e.g. DBpedia). Nonetheless, other language versions are documented as a reference.

Table 1
100 entities in five categories selected for analysis

ID	Agents[1]	Events[2]	Dates	Places[3]	Objects and Concepts
1	Carl Linnaeus	World War II	1987	United States	Book of Kells
2	Jesus	World War I	1986	United Kingdom	Vasa (ship)
3	Aristotle	American Civil War	1985	France	The Garden of Earthly Delights (paining)
4	Napoleon	FA Cup	1983	England	Rosetta Stone
5	Adolf Hitler	Vietnam War	1980	Germany	Palazzo Pitti (building)
6	Julius Caesar	Academy Awards	1984	Canada	Boeing 747
7	Plato	Cold War	1982	Australia	Sgt. Pepper's Lonely Hearts Club Band (album)
8	William Shakespeare	Korean War	1968	Japan	Tosca (opera)
9	Albert Einstein	American Revolutionary War	1979	Italy	Blade Runner (film)
10	Elizabeth II	UEFA Champions League	1969	Poland	Uncle Tom's Cabin (novel)
11	Michael Jackson	UEFA Europa League	1978	India	Ming Dynasty
12	Madonna (entertainer)	Olympic Games	1967	Spain	Ukiyo-e (art)
13	Ludwig van Beethoven	Stanley Cup	1981	London	Angkor Wat (building)
14	Wolfgang Amadeus Mozart	Super Bowl	1977	Russia	Toraja (ethnic group)
15	Pope Benedict XVI	Iraq War	1976	New York City	Byzantine Empire
16	Alexander the Great	War of 1812	1975	Brazil	Mars (planet)
17	Charles Darwin	Gulf War	1964	California	Tamil language
18	Barack Obama	Spanish Civil War	1966	New York	Influenza (disease)
19	Mary (mother of Jesus)	World Series	1965	The Netherlands	The King and I (musical)
20	Queen Victoria	EFL Cup	1960	Sweden	Like a Rolling Stone (song)

[1]The priority is given in the following order: page rank, 2Drank (24 languages), and page rank (female).
[2]International events are prioritised, thus a couple of specific events such as US censuses are removed.
[3]Top 20 places are extracted from the general list.

most referenced articles"[14] (2011) for the top 20 places and dates, whereas a scientific article about the interaction of top people in Wikipedia is used for the 20 agents [17]. In addition, the top 20 events are retrieved by a SPARQL query from the EventKG endpoint[15] as follows:

```
PREFIX eventKG-s: <http://eventKG.l3s.uni-hannover.de/schema/>
PREFIX eventKG-g: <http://eventKG.l3s.uni-hannover.de/graph/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX sem: <http://semanticweb.cs.vu.nl/2009/11/sem/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX dbr: <http://dbpedia.org/resource/>
SELECT ?dbp ?links {
 ?event rdf:type sem:Event .
 GRAPH eventKG-g:dbpedia_en { ?event owl:sameAs ?dbp . } .
  {
   SELECT ?event (SUM(?link_count) AS ?links) WHERE {
   ?relation rdf:type eventKG-s:Relation .
   ?relation rdf:object ?event .
   GRAPH eventKG-g:wikipedia_en { ?relation eventKG-s:links ?link_count . }.
  } GROUP BY ?event
 }
```

[14]https://en.wikipedia.org/wiki/Wikipedia:Most-referenced_articles, last accessed 2019-09-25.
[15]http://eventkginterface.l3s.uni-hannover.de/sparql (last accessed 2019-09-25).

```
} ORDER BY DESC(?links)
LIMIT 30
```

It is not trivial to nominate 20 objects and concepts, because cultural heritage and DH cover an extremely broad field. In fact, there are countless numbers of material entities such as museum objects and buildings. Moreover, millions of archaeological objects are even unnamed. Indeed, many object entities are not globally and uniquely identifiable, because they have not (yet) been created in the global references. As such, it is much more challenging to implement entity linking for those entities. Nevertheless, we manually selected 20 entities from the featured articles of Wikipedia.[16] They aim to represent a wide range of chronological, geographical, and thematic diversity.[17]

The actual number of entities analysed is 836 (859 occurrences), since some sources do not have the entities the others have. Full details of the entity coverage per data source are provided in Appendix A. Statistically speaking, in case of missing entities, they are included in the calculation and the data values are counted as null.[18] In addition, there are double identity/occurrence (or a kind of "duplicate"[19]) in some sources. The double identities are consolidated as one identity.[20] When an entity lookup is not accessible for technical reasons, the data is included in the statistics as a zero value.[21]

In practice, it is not feasible to fully automate the analysis process. In order to properly document the data quality, it is required to search, identify, and verify the same entity across 11 data sources. The quality of each entity needs to be manually double-checked. The main problem of our analysis is semantic disambiguation. It is even not always possible to accurately find an entity. For instance, the challenges of disambiguation and entity matching across multiple LOD sources are presented by Farag [19]. In our case, three reasons are worth mentioning: (a) the lack of cross linking between data sources makes it hard to find all available entities, (b) the entities are confusingly organised and hidden from the mainstream contents, especially in aggregated LOD, and (c) the search functionalities on the website of the data sources may have limited capacity and have not been optimised. In these cases, lookups are executed on a best-effort basis.[22] Another justification of our manual evaluation is the lack of gold standard. In fact, the research on the LOD quality in digital libraries requires manual reviews for several metrics [8].

3.4. Analysis methodologies

In this study, we conduct both qualitative and quantitative analysis. As for the qualitative approach, this paper presents some examples that are found during the manual inspection of LOD instances. As for the quantitative

[16] https://en.wikipedia.org/wiki/Wikipedia:Featured_articles, last accessed 2020-03-10.

[17] This research investigates tens of thousands of global entities that are reasonably well known and one could look up and refer to as sources for NEL, rather than millions or billions of instances of cultural heritage objects that could be hard to refer to globally. On one hand, encyclopedia-based and authority-file based LOD sources such as Wikidata and VIAF deal with the former and generate LOD by a top-down approach. On the other hand, Europeana takes a bottom-up data aggregation approach to build LOD for over 50 million digital objects from the records held by thousands of cultural heritage organisations. Most of them are unique and not well known. Next to their instance-level LOD, Europeana offers a highly limited amount of entity lookups relevant to their LOD that our study evaluates.

[18] For example, WorldCat does not seem to have entities 1976, 1979, and Europa League.

[19] This article only tries to identify the data about the same entity without judging if the data contents are duplicated or not. It seems that the double identity is a leftover of merging entities during data aggregation. Such examples include Aristotle in VIAF (https://viaf.org/viaf/268271999/ and https://viaf.org/viaf/7524651/) and California in YAGO (http://lod.openlinksw.com/describe/?url=http%3A%2F%2Fyago-knowledge.org%2Fresource%2FCalifornia)(last accessed 2021-01-26).

[20] During the entity identification process, we already recognise interesting patterns in the coverage of entities across the data sources. A typical case is the mosaic of availability for the objects and concepts. In the Getty Vocabulary, Ukiyo-e would be included as an artistic style, not an individual artwork, whereas Book of Kells, Garden of Earthly Delights, Sgt. Papers, Blade Runner, Uncle Tom's Cabin, and the King and I are not, because they are unique. Symbolically the latter group is all included in WorldCat, the Library of Congress, and VIAF as well as BabelNet, DBpedia, and YAGO. It seems to make sense to consider this pattern as the coverage difference between record-orientated library authority files and concept-orientated museum vocabularies.

[21] For example, unfortunately Italy in BebelNet has constantly returned HTTP 500 error during our analysis (http://babelnet.org/rdf/page/s00047705n).

[22] In addition, it is noted that this study does not guarantee technical feasibility of traversing via lookup services in reality. The project only documents and analyses the availability of links, not the validity of links. For example, it is the responsibility of LOD providers to adequately implement and maintain content negotiation and HTTP redirect.

approach, we generate chord diagrams in R[23] to examine the basic flow of incoming and outgoing links within the 11 data sources. We deploy Data to Viz, based on the circlize package.[24] For the creation of traversal maps, we import matrix data from spreadsheets to R and generate network diagrams with igraph[25] packages. In addition, we calculate the amount and percentage of links and provide different views on the quality. Moreover, a basic network analysis is also conducted with R to objectively evaluate the characteristics of the small LOD network. It turns out that this approach is useful, because Guéret et al. [22] subsequently proposed a linking quality method with some of the network metrics we use in the R analysis.

Furthermore, this paper also analyses other data content (such as literals) in addition to the links. This is important, as we cannot obtain a full picture of link quality without studying the content of the link destination. In an RDF graph, there can be three types of nodes: IRIs,[26] literals, and blank nodes.[27] As the blank nodes are not heavily used in our target datasets and add extra complexity, we limit ourselves to literals. For this purpose, first we simply extend our calculation to check the use of four W3C standardised properties, mainly for literals. The amount and percentage of `rdfs:label`, `rdf:type`, `skos:prefLabel`, and `skos:altLabel` are calculated. In addition, the total amount of content associated with `rdf:resource` and `rdf:about` is assessed. These two properties are at the centre of RDF/XML and are used to describe and connect resources. Although there are other important properties than the six properties described above, they are the most fundamental and frequently used properties to describe entities. These statistics allow us to obtain basic holistic views on the data content. However, they are not sufficient to draw conclusions.

The challenge is how to objectively compare and evaluate the content quality of different LOD sources. The major problems are: (a) there is no standard theory about what is regarded as high quality, and (b) it is hard to evaluate the quality of semantics. In terms of (a), for example, the number of links (edges) or labels/literals (strings) alone would not be able to indicate the data quality. In terms of (b), the same hyperlinks and labels can be found in different context. For example, the link "http://www.example.com" can be found in `skos:exactMatch` or `dcterms:isPartOf`, while the string "Book of Kells" can be in `skos:prefLabel` or `rdfs:label`. Both of these cases carry the same information, but there is no easy way to assess the quality of semantics of the properties. This is especially the case when proprietary properties are used. It is practically impossible to judge the quality, due to the nature of freedom in LOD. Moreover, we cannot give any preference to a hyperlink or literal as the object of a property.

To minimise the impact of a biased evaluation, Python scripts[28] are developed to supplement our analyses. They compare the overlap of data content in each LOD source without any interpretations/assumptions. Technically this means that the scripts analyse the objects of the main entity with string matching, and calculate the amount of unique content. The objects include both edges and literals, where URIs are considered as string values to be compared. In other words, the semantics of the properties are not evaluated. Although this method may not be the most accurate way to measure the content quality, it allows us to perform systematic and automatic measurements. It provides us with a sense of the amount of information and the coverage or diversity of data contents.

It is anticipated that a broad mix of above-mentioned methods can provide new insights into the linking quality at different levels.

[23] https://www.r-project.org/, last accessed 2021-01-26.

[24] https://www.data-to-viz.com/graph/chord.html, last accessed 2021-01-26.

[25] https://igraph.org/r/ last accessed 2021-01-26.

[26] Internationalised Resource Identifier is the generalisation of URI that supports Unicode characters. For our convenience, URI is used as a synonym of IRI in this paper.

[27] https://www.w3.org/TR/rdf11-concepts/ last accessed 2022-01-20.

[28] Available at https://github.com/GO5IT/LOD_analysis and https://doi.org/10.5281/zenodo.5913595 including the data generated.

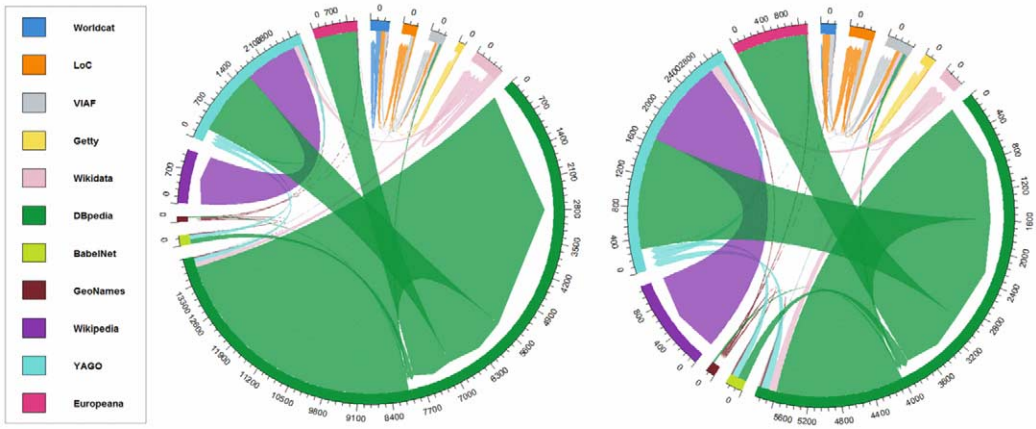


Fig. 1. Chord diagram illustrating the amount of link flows between 11 data sources (left) and after removing inverse links (right).

Table 2
The total and average number of outgoing links (to the 11 data sources) held by the data sources

ID	A	B	C	D	E	F	G	H	I	J	K	Total
Source	YAGO	WorldCat	Wikidata	VIAF	Library of Congress	Getty	GeoNames	Europeana	DBpedia	BabelNet	Wikipedia	
Total	2713	259	192	171	102	69	23	903	5832	210	0	10474
Average	27.4	2.7	1.9	3.1	1.1	1.6	1.0	36.1	58.3	2.1	0.0	12.5

4. Linked open data analysis

4.1. Overall traversal map

The first analysis starts with chord diagrams. Figure 1 primarily focuses on the number of links and their origins and destinations within the 11 data sources. The source data which produce Fig. 1 is found in Appendix B.

The total number of links amounts to 10474. The dominance of DBpedia is obvious, occupying over 66.2% of the entire linkages (Fig. 1 left). It is also noticeable that self-links significantly contribute to the volume of the links. YAGO supplies a substantial amount of links to DBpedia and Wikipedia. This results in the influential position of Wikipedia (5.2%), although it is not LOD. Surprisingly, Europeana comes fourth, despite the significantly limited amount of available entities (Appendix A). WorldCat, the Library of Congress, and VIAF somewhat share similar numbers of links. The outgoing and incoming links are unbalanced for Europeana.

From these numbers we can derive the following: the average number of links in all sources is 952.2, whereas the medians are 2.1 and 149 for both outgoing and incoming links. In fact, the amount of outgoing hyperlinks found in each source is moderate, given the entire size of those datasets (i.e. millions of triples); on average it is mostly under four links per entity (Table 2). These small figures are alarming, because this survey focuses on well-known sources often used for NEL for the cultural heritage datasets. It is clear that there is a great deal of room for improvement. Nevertheless, DBpedia, Europeana, and YAGO stand out, showing more promising quality for LOD with high number of links per entity.

When inverse traversals are removed from the statistics, the situation looks largely different (Fig. 1 right). The sum of the links decreases to 6166. DBpedia loses an ample number of links (47.3%), whereas YAGO gains most (24.2%). Such a dramatic shift is an evidence of abundance of inverse properties described in DBpedia. If we scrutinise the data closely, we notice that this is mostly due to the inverse use of `rdfs:seeAlso` in DBpedia. For instance, the entity of Sweden contains:

```
dbr:Lund rdfs:seeAlso dbr:Sweden .
```

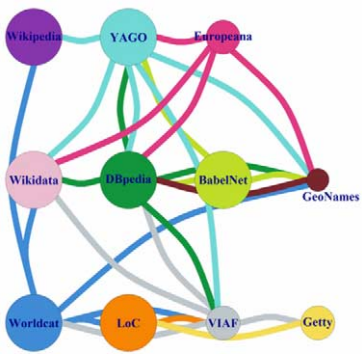


Fig. 2. The overall "traversal map" shows available links/paths through four standardised properties between the 100 entities in 11 data sources (after self-links to the same domain is removed).[29]

Figure 2 is the simplified overall "traversal map" for all data sources. It is a network diagram, illustrating all possible paths between the 11 data sources. However, since we observe a very high volume of links in DBpedia, YAGO, and Europeana, volumes and self-links/loops (i.e. links pointing to the same data source/domain) are not included in this figure. Thus, the diagram concentrates on the routes of traversals (i.e., the users' mobility and traversability).

It is clear that the traversing routes are not equally available across the data sources, and thus, it may be hard to navigate the LOD network. It is found that YAGO delivers four connections as well as one to Wikipedia. The next contenders are Europeana and DBpedia with four outgoing connections. In contrast, Wikidata has no outgoing connections.[30] Whilst GeoNames only links to DBpedia, the Library of Congress and Getty have one channel. With regard to incoming connections, GeoNames is an attractive destination to which five sources refer. Wikidata and DBpedia are also a centre of gravity, inviting five connections. On the other hand, Europeana and BabelNet receive no links. Whereas the lack of incoming links to BabelNet may be surprising, in Europeana's case it is not, because it is not equipped with a truly public lookup. This would mean that the generation of LOD dump and/or SPARQL endpoint may not be sufficient. It is best to publicly declare entities that are resolvable via lookups without access restrictions. WorldCat and Getty are both only reached by VIAF.

It is particularly remarkable that reciprocal links are quite rare. There are several nodes/vertices which can be reached via only particular edge(s)/path(s). This implies that network is not desirably populated by the standard properties, and that the users would not be able to efficiently obtain information through these properties. They need to follow the best paths to retrieve the identical or closely matching information. It is possible for data publishers to use other RDF properties, but it would be an irregular practice.

Idrissou et al. [26] stress that a full mesh (fully connected network) has the highest quality in their link quality metrics. When they compare different structures (e.g. ring, line, star, mesh, tree), the more a network resembles a fully connected graph, the higher the quality of the links in the network for all metrics (bridge, diameter, closure). One might argue that a full mesh is not necessarily a prerequisite of high data quality. This may be true for much LOD, however, let us remember that we focus on the most well-known data sources that many other LOD tend to link to. Therefore, it helps the connectivity of LOD on the web as a whole. Guéret et al. [22] use clustering coefficient and `owl:sameAs` chains as their criteria for high quality.

Figure 3 depicts traversal maps faceted by four link types. From now on, inverse properties are included but loops are excluded for the traversal map visualisation. Thus, the distortion of the "route diagram" that we avoided in Fig. 2 is minimal. However, the rest of the statistics (matrix data and in the texts) include both inverse properties and loops, so that they reflect the actual situation.

[29] In traversal maps (Fig. 3, 4, 5, 6, 7, 8), the sizes of the vertices correspond to their volumes of the available entities. Colours are assigned by the origin of the edges. The widths of edges represent their weights (except Fig. 2).

[30] 192 self-links are omitted.

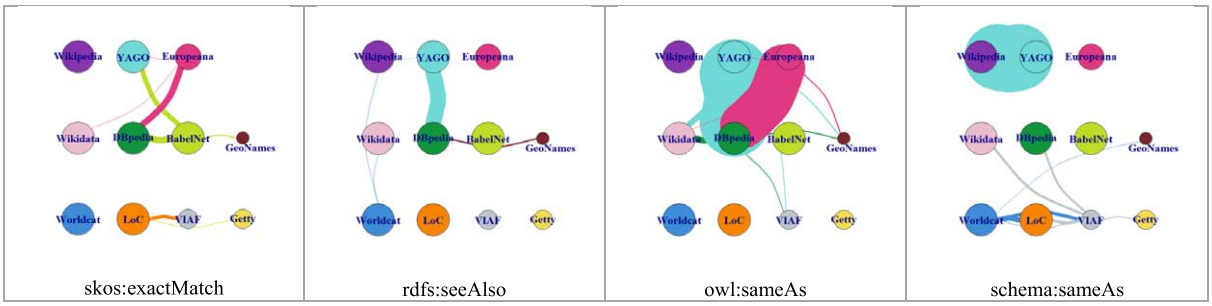


Fig. 3. The overall traversal map by each standardised property (after removing self-links to the same domain).

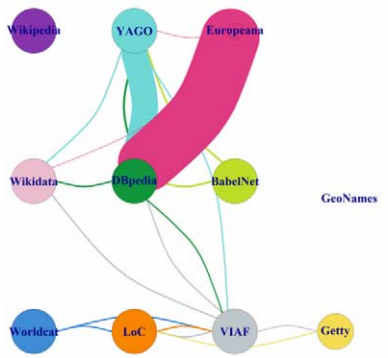


Fig. 4. The overall traversal map for agent entities.

Although we decided to avoid discussions on interpretations of link semantics, there is at least a clear difference between `owl:sameAs` (as well as `skos:exactMatch` and `schema:sameAs`) and `rdfs:seeAlso`. It can be clearly seen that Europeana, the Library of Congress, and BabelNet are the only data publishers using `skos:exactMatch`. `rdf:seeAlso` is used mostly by YAGO, while GeoNames and WorldCat are also visible. However, the proportions of `owl:sameAs` and `schema:sameAs` are higher. In particular, Europeana and YAGO provide a large amount of connections to either DBpedia or Wikipedia. We also realise that WorldCat and VIAF opt more for `schema:sameAs`. In general, Fig. 3 suggests that the data creators made different ontological decisions on the choice of standardised properties. We will explore this further in the following sections.

4.2. Agent traversal map

Figure 4 depicts the traversal map for agents. Appendix B includes the source matrix data and the traversal maps for all four properties. In general, agents have much less influence from loops than from other categories, because 72.4% of links are still present after removing recursive links, compared to the overall 42.0%. The most eye-catching result is Europeana. Especially, it uses `owl:sameAs` to link to DBpedia. In cultural heritage, VIAF plays a valuable role for agents as an aggregation of authority files of national libraries. For instance, it is the only source which offers four outgoing paths. This category has only three sets of nodes that have bilateral links. Therefore, segmentation is visible in the network and truly standardised LOD connectivity is limited.

In Table B3 in Appendix B, the role of DBpedia is expectedly prominent for incoming links, attracting 1555 links (80%). Unlike the outgoing links, Wikidata captures 121 referrals, making it the second highest source. Manual examination found that VIAF had only 72 incoming links, however, it contains more links which connect its entity to data sources outside the 11 sources, than any of the other sources. For instance, only four links with `schema:sameAs` are recorded for Beethoven. However, the destinations of a further eight links include the national libraries of France, Germany, Japan, Spain, and Sweden.

Table 3
The amount of outgoing links that the 11 data sources hold in each agent entity (* means duplicate consolidation)

		Linneaes	Jesus	Aristotle*	Napoleon	Hitler	Caesar*	Plato	Shakespeare	Einstein	Elizabeth II	M Jackson	Madonna	Beethoven	Mozart	Benedict XVI	Alex Great*	Darwin	Obama	Mary	Q Victoria	SUM
A	YAGO	23	38	24	28	28	24	24	24	26	27	37	34	25	31	29	23	28	31	0	26	530
B	WorldCat	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	60
C	Wikidata	5	0	5	3	6	4	2	2	4	0	1	0	3	1	1	1	4	4	2	7	55
D	VIAF	12	3	11	12	12	20	10	12	12	10	13	11	12	12	11	13	12	11	6	12	227
E	LoC	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	20
F	Getty	3	0	1	3	3	3	3	3	3	3	1	0	3	3	0	1	3	1	0	3	40
G	GeoNames	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
H	Europeana	0	0	122	0	119	0	117	119	0	0	0	88	114	119	0	0	0	0	0	0	798
I	DBpedia	24	59	25	31	29	26	26	25	27	28	39	36	26	32	30	24	29	34	19	27	596
J	BabelNet	4	0	4	7	6	7	4	6	4	4	4	4	4	4	3	5	4	3	5	4	86
K	Wikipedia	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	SUM	75	104	196	88	207	88	190	195	80	76	99	177	191	206	78	71	84	88	36	83	2412

The amount of outgoing links held by 11 data sources in each entity is visualised in Table 3. When comparing the total amount in this table and in Table B3 in Appendix B, we notice that 1945 incoming links are received within the 11 data sources, out of 2412 outgoing links (80%).[31] Whereas Europeana has 798 outbound links (33%), DBpedia and YAGO follow at 596 and 530 respectively. There is a considerable gap between the highest number of outgoing links across 11 sources (Hitler, 207) and the lowest (Mary, 36). The highest cluster are from Europeana, however, the outgoing links in Europeana are unevenly distributed. Only Aristotle, Hitler, Plato, Shakespeare, Madonna, Beethoven and Mozart are present. This would offer evidence that art and cultural figures are more important for the cultural heritage objects that Europeana deals with than politicians and scientists. DBpedia and YAGO show a similar pattern, mainly due to the tight connections between them. In there, we observe relative popularity for Jesus, Michael Jackson, and Madonna.

WorldCat holds exactly three links per entity.[32] One is caused by the description of a new WorldCat identifier via the inverse property of `rdf:seeAlso`. The other two are `schema:sameAs` which links to the Library of Congress and VIAF. Similarly, the Library of Congress has exactly one link per entity (`skos:exactMatch` to VIAF).[33] These two cases suggest evenly distributed and highly normalised RDF content, probably due to systematically generated links between the library sources.

Whilst most data sources cover all 20 agents, Jesus Madonna, Benedict XVI, and Mary are totally missing in Getty vocabularies. Similarly, the number of VIAF links is sharply reduced for Jesus and Mary. This is understandable since Getty ULAN and VIAF are typically orientated toward artists and authors in the context of libraries and museums, and religious figures are harder to be recognised as agent entities. Indeed, Jesus has the lowest number of links for five data sources (Mary for four data sources). As such, it is remarkable that Jesus is relatively high in DBpedia (59 links). It is also interesting that non-artists figures such as Einstein, Elizabeth II, and Obama are found in ULAN.

4.3. *Events traversal map*

Figure 5 clearly illustrates the lack of links. Bilateral links are extremely rare: only between YAGO and DBpedia. As a result, it is not possible, for example, to move from the Library of Congress to Wikidata. This implies that the entry point to a network determines the movement within it. DBpedia contains far more links than other sources. Although Europeana has only one entity in this category (i.e. World War I), it manages to draw a thick line (`skos:exactMatch`) in the figure (111 links).

In general, events were not found in VIAF during the manual data exploitation, however, it turns out that WorldCat and the Library of Congress refer to it seven times each. For example, the former links to the World Series in French

[31] In the coming sections, we will compare outgoing links (the tables in the text) with incoming links (the overall tables in Appendix B).

[32] There is the forth link (`rdfs:seeAlso`). It is provided by not `rdf:resource`, but the anyURI typed literal, therefore, it is excluded from the analysis.

[33] `skos:closeMatch` is excluded from the analysis.

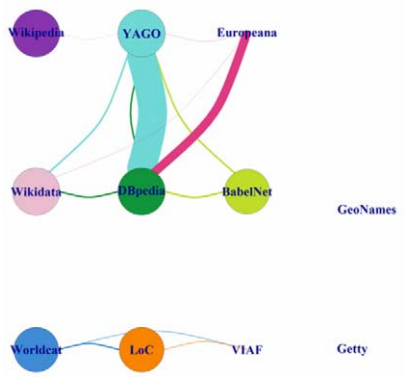


Fig. 5. The overall traversal map for event entities.

Table 4
The amount of outgoing links that the 11 data sources hold in each event entity (* means duplicate consolidation)

		WW II	WW I	World Series	War of 1812	Vietnam War	Super Bowl	Stanley Cup	Spanish Civil War	Olympic Games*	Korean War	Iraq War	Gulf War	FACup	Europa League	EFLCup	Cold War	Champions League	American Rev War	American Civil War	Academy Awards	SUM
A	YAGO	21	24	21	19	33	23	18	22	25	22	34	23	26	21	16	21	23	20	20	22	454
B	WorldCat	2	2	3	2	2	3	3	2	3	2	2	2	3	0	3	1	3	2	2	2	44
C	Wikidata	3	4	0	0	3	2	1	1	2	2	3	1	2	2	0	3	1	1	3	2	36
D	VIAF	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E	LoC	2	2	1	2	2	1	1	2	1	2	2	2	1	0	1	2	1	2	2	2	31
F	Getty	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	GeoNames	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
H	Europeana	0	113	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	113
I	DBpedia	122	61	23	28	39	24	19	26	23	25	41	29	27	22	17	29	24	23	28	24	654
J	BabelNet	4	7	3	4	4	3	3	0	9	4	3	4	3	3	3	4	3	7	7	4	82
K	Wikipedia	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	SUM	154	213	51	55	83	56	45	53	63	57	85	61	62	48	40	60	55	55	62	56	1414

(`skos:prefLabel` is Séries mondiales (Base-ball) and `skos:altLabel` is World Series (Base-ball)). Another 13 cases are all sporting events and awkwardly labelled as corporate entity in VIAF. Although those entities may be exceptional cases, they also reveal interesting cataloguing practices (or perhaps errors) by libraries in data modelling or mapping. Whatever the reasons are, we may face challenges in the future to tackle errors and inconsistency for semantic reasoning.

In terms of each entity (Table 4), the most appealing entity is World War II, followed by World War I and the Iraq War. Europeana's contribution to World Wart I is considerable. Although the EFL Cup is the lowest, the gaps between entities are relatively subtle except the top three (i.e., median 49.5, average 57.5).

The principal reason for the prominence of DBpedia for the World War II is `rdfs:seeAlso` inverse links which include the DBpedia entities of agents (e.g. Winston Churchill), places (e.g. Leipzig), ships (e.g. USS Hornet), and the lists and articles derived from Wikipedia (e.g. tanks in the German Army, history of propaganda). In this case it is advantageous for the users to discover and access detailed information about the war. However, as RDF representation is not guaranteed for `rdfs:seeAlso`, this situation would hamper predicting the source of link destination and decreasing the possibility of efficient and/or automatic data processing.

4.4. Dates traversal map

It is striking that the volume of links is very low (Fig. 6). Out of 881 outgoing links, 863 links are consumed within the 11 data sources, implying a high level of closure in the network. In addition, only three sources are referenced: DBpedia, Wikidata, and the Library of Congress. Although YAGO provides many links to DBpedia and Wikidata via `owl:sameAs`, it does not receive any incoming links. Since bilateral links do not exist, the movement in the network is highly restricted. There are only three possible paths. Consequently, the fluctuation of linking patterns is also minor (Table 5).

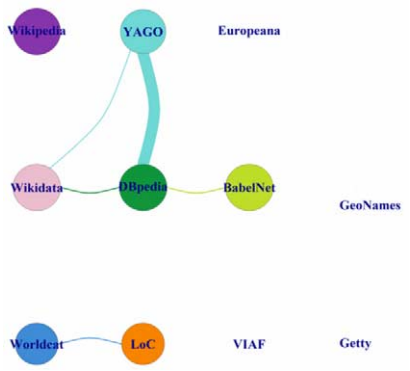


Fig. 6. The overall traversal map for date entities.

Table 5
The amount of outgoing links that the 11 data sources hold in each date entity

		1987	1986	1985	1983	1980	1984	1982	1968	1979	1969	1978	1967	1981	1977	1976	1975	1964	1966	1965	1960	SUM
A	YAGO	13	13	13	13	13	13	13	13	13	13	0	13	13	13	13	13	13	13	13	13	247
B	WorldCat	2	2	2	2	2	2	2	2	0	2	2	2	2	2	0	2	2	2	2	2	36
C	Wikidata	4	3	3	4	3	4	4	1	2	2	2	3	3	3	3	2	4	4	3	3	60
D	VIAF	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E	LoC	2	2	2	2	2	2	2	2	0	2	2	2	2	2	0	2	2	2	2	2	36
F	Getty	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	GeoNames	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
H	Europeana	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
I	DBpedia	26	24	25	25	26	27	26	28	23	21	23	20	25	26	26	25	24	23	20	19	482
J	BabelNet	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	20
K	Wikipedia	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	SUM	48	45	46	47	47	49	48	47	39	41	30	41	46	47	43	45	46	45	41	40	881

The economy of the creation of date entities may show serious issues. 1978, 1979, and 1976 do not seem to exist in YAGO, the Library of Congress, and WorldCat, while other consecutive years in the 1970's are available (see Appendix A). Such inconsistency would become problematic, when queries are constructed to look for answers to research questions on years and periods. In semantic queries, erroneous links and data omissions require careful presentation to LOD users in the future, in order to avoid misinterpretation and misjudgment.

One reason for this phenomenon is the lack of recognition and/or needs for numeric date instant entities, in comparison with other date representations, including textual dates (e.g. "End of the 17th century"), numeric durations (e.g. "1880–1898"), and periods and eras (e.g. "Bronze Age" and "Roman Republic"). For example, a quick search indicates the entity for "Neolithic" exists in all our data sources except GeoNames, VIAF, and Europeana.

In cultural heritage, numeric dates are often stored in a database as string/literal data type, when encoded in XML or RDF. They can be typed as date in the XML Schema (e.g. `xsd:date`). Thus, they are not designed for NEL, although it would have many advantages, especially for data linking and integration. What is clear is that users have currently a very limited possibility to execute NEL for numeric dates. To fill this gap, we have recently started a project to create LOD for the numeric date entities [38].

4.5. Places traversal map

Traversability for places is better than in other categories. YAGO dominates the scene for outgoing links (Fig. 7). Interestingly VIAF comes third despite its focus on agent entities. The Library of Congress, Getty TGN, and GeoNames contain an almost consistent number of links, each typically pointing to DBpedia. Users need to be careful regarding Europeana, because it does not provide the entities for the USA at all (USA, California, New York, and New York City). This type of inconsistency may be problematic for NEL implementers. They should scrutinise the occurrences of their place entities in their local datasets before selecting the right NEL targets. Strangely, no outgoing links are found for Australia, Canada, France, Germany, Italy, Japan, and Russia in Wikidata.

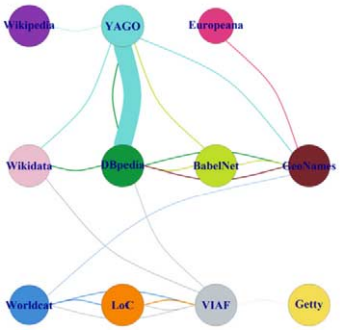


Fig. 7. The overall traversal map for place entities.

Table 6
The amount of outgoing links that the 11 data sources hold in each place entity

		USA	UK	France	England	Germany	Canada	Australia	Japan	Italy	Poland	India	Spain	London	Russia	New York City	Brazil	California	New York	Netherlands	Sweden	SUM
A	YAGO	43	37	40	34	35	31	34	33	35	34	35	44	32	35	31	39	35	31	45	46	729
B	WorldCat	4	4	4	3	3	4	4	4	3	3	4	3	4	3	0	3	3	3	3	3	65
C	Wikidata	4	1	0	5	0	0	0	0	0	4	2	4	2	0	2	3	2	5	5	2	41
D	VIAF	11	7	7	5	8	9	11	10	10	7	11	10	11	8	11	6	10	10	9	4	175
E	LoC	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	21
F	Getty	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	20
G	GeoNames	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	21
H	Europeana	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	0	1	1	17
I	DBpedia	52	49	469	131	481	253	266	187	314	417	238	164	63	275	78	133	208	231	45	121	4175
J	BabelNet	1	7	6	6	6	6	7	6	0	6	6	6	9	6	6	6	9	7	6	6	118
K	Wikipedia	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	SUM	119	109	531	189	537	307	326	244	366	475	300	235	125	331	131	194	270	290	117	186	5382

The presence of GeoNames, in particular, facilitates more fluid movements in the network. Although Ahlers [3] claims that it is the largest contributor to geospatial LOD and is intensely cross-linked with DBpedia, it is a disadvantage that it only connects to DBpedia. This makes the overall mobility less ideal. Apart from a link to VIAF, Getty TGN only contains 20 self-links mostly in the form of `rdfs:seeAlso` for a HTML representation. RDF/XML for New York City (tgn:7007567) holds:

```
tgn:7007567 rdfs:seeAlso <http://www.getty.edu/
vow/TGNFullDisplay?
find=&place=&nation=&subjectid=7007567> .
```

Therefore, it is a dead end in terms of network traversals, of which the users need to be aware during their traversing. Europeana is disappointing including only 17 outgoing links only to GeoNames.

If loops are included, DBpedia holds 86% of all outgoing links. This is caused by a vast number of inverse links. For example, in case of Australia, 255 out of 266 outgoing links in DBpedia are those inverse `rdfs:seeAlso` links to DBpedia itself. It is possible to find both important and less important links:

```
dbr: Health_care_in_Australia rdfs:seeAlso dbr:Australia .
```

On one hand, the DBpedia loops may be confusing, especially due to the use of ambiguous `rdfs:seeAlso` links and the flexibility of information provided. On the other hand, they allow users to find unexpected related information that other LOD sources do not provide, leading to the serendipity that LOD is good at.

In Table 6, the lowest entities are surprisingly: the Netherlands, United Kingdom, and United States. This is chiefly attributed to fewer numbers of DBpedia links. However, the reason for this is unclear. On the contrary, the top entities receive a large quantity of links, which include Germany, France, and Poland.

The outgoing links are the lowest for United Kingdom, followed by the Netherlands, and United States. In contrast, Poland, Germany and France are the top three. The cause is obvious: the numbers are affected by the uneven

pattern of links in DBpedia. The amount of links in other sources are instead more or less evenly spread across different entities. It would be intriguing to investigate the reasons by inspecting the corresponding entities in Wikipedia articles and the linking mechanism behind the DBpedia transformation. It would reveal pros and cons of a crowdsourcing approach to LOD, as opposed to authority approach such as the Library of Congress, VIAF, and Getty from libraries and museums.

4.6. Objects and concepts traversal map

Objects and concepts are the subject matter in which cultural heritage researchers would be most interested. To a large degree, they are the target entities of contextualisation which is substantiated through data integration and inferences, thus, the contextualised entities are out of our scope. Rather we analyse them as the entities supporting contextualisation (Fig. 8). 1844 outgoing links are recorded of which 91% are bounded for the 11 data sources. Network closure also persists in this category. 81.3% of all incoming links concentrate on Wikipedia (1085), with DBpedia (100) and Wikidata (43) lagging far behind. The same can be said for outgoing links: YAGO (1212) and the rest. This happens, because YAGO provides a considerable number of links to Wikipedia.

Although Europeana produces LOD out of digital cultural heritage objects, its entity API is merely an experimental reference point, thus, no contribution is observed in our traversal scenarios. Interestingly, VIAF plays an authoritative role for this category. It serves a small number of links to five sources. Although the number of outgoing links from BabelNet is not high, it performs better in this category.

During the process of identifying and collecting the entities, some data quality issues are recognised. The significant concepts of cultural objects in FRBR, namely Work, Manifestation, Expression, and Item, are not easily conceptualised and encoded in the LOD observed. For example, taking a book as an example, we consider a single physical copy of a book as Item. Then, all published copies of the book which share the same ISBN are defined as Expression. Manifestation is considered as a book in a specific language by a specific author, whereas Work is a higher level of abstraction to cover the idea or the fundamental creation of the book by an author. Therefore, for instance, VIAF holds records on The King and I as Expression (motion picture) and Work (the original artwork). However, partly due to the technical mechanism of VIAF, Work may not be easily created. Similarly, Wikipedia has a disambiguation page for the King and I to distinguish the original musical from films and music products associated to the musical. This implies some difficulties in terms of co-reference resolution during NEL, as well as graph traversing.

As this category is deliberately broad and vague in principle, it is not possible to see clear-cut results. For example, GeoNames has entities for Palazzo Pitti and Angkor Wat, which could be classified as places and object simultaneously. Nevertheless, it reminds us that the data modelling for cultural heritage entities is intentionally complex. There could be entities that have multi-types. Depending on the perspective, the data modelers and users would need to find a common view on both practical use and theoretical truth and/or fuzziness of datasets. For instance, Palazzo Pitti could be a geographical place, as well as a building structure, concept, or organisation. However, complicated roles may introduce unnecessary complexity for real usage, confusing end users.

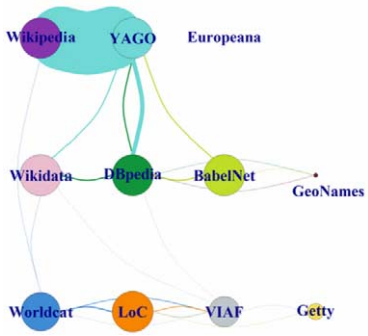


Fig. 8. The overall traversal map for objects and concepts entities.

Table 7

The amount of outgoing links that the 11 data sources hold in each object and concept entity

		Book of Kells	Vasa	Garden of E Delights	Rosetta Stone	Palazzo Pitti	Boeing 747	Sgt. Pepper's	Tosca	Blade Runner	Uncle Tom's Cabin	Ming Dynasty	Ukiyo-e	Angkor Wat	Toraja	Byzantine Empire	Mars	Tamil language	Influenza	King and I	Like a Rolling Stone	SUM
A	YAGO	52	50	44	88	55	77	18	53	65	8	113	59	112	31	153	23	161	0	10	40	1212
B	WorldCat	3	3	3	3	3	2	3	3	3	3	0	4	2	2	3	2	2	2	3	3	52
C	Wikidata	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	VIAF	5	2	4	3	8	0	4	2	5	5	1	0	4	0	3	2	0	0	6	2	56
E	LoC	1	1	1	1	1	2	1	2	1	1	2	2	2	2	1	2	6	3	2	1	35
F	Getty	0	0	0	0	2	0	0	0	0	0	1	1	1	0	1	1	1	1	0	0	9
G	GeoNames	0	0	0	0	1	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	3
H	Europeana	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
I	DBpedia	17	16	16	19	18	20	21	21	21	19	23	19	20	12	39	25	25	24	11	20	406
J	BabelNet	3	3	3	4	4	3	3	3	3	3	5	3	4	3	5	5	5	4	2	3	71
K	Wikipedia	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	SUM	81	75	71	118	92	104	50	84	98	39	145	88	147	50	205	60	200	34	34	69	1844

Table 8

Network analysis measurements by category

Measurement	Overall	Agents	Events	Dates	Places	Objects & Concepts
Reciprocity	0.345	0.316	0.154	0.000	0.381	0.381
Transitivity	0.505	0.600	0.692	0.600	0.420	0.447
Mean Distance	1.919	1.791	1.235	1.167	1.826	1.878
Diameter	4	4	2	2	4	4
Edge Density	0.264	0.173	0.118	0.045	0.191	0.191

Another interesting finding is that Mars appears in TGN of the Getty vocabulary. It is normally considered that the vocabulary contains place names on earth, as one expects from GeoNames. There could be some surprise for LOD users in terms of how data is conceptualised and modelled, and from where data is obtained, especially when automatic data collection and integration are implemented in the future.

Regarding the individual entities (Table 7), Byzantine Empire and Tamil language in YAGO display a distinct pattern. The cause of this pattern seems to be clear; it includes links to language orientated resources such as language codes, maybe suggesting an important role of language resources in the LOD scenario. For other entities in YAGO it is hard to find exact causes and correlations between the entities with more links (Rosetta Stone, Ming Dynasty, Angkor Wat) and the ones with fewer links (Uncle Tom's Cabin, Influenza, King and I). The results from Getty imply the exclusion of specific objects.

4.7. Network analysis

We deploy a network analysis using R to supplement the so far relatively subjective impressions and interpretations of the traversal maps (Table 8). Although the work of Idrissou et al. [26] is highly relevant here, unfortunately we are unable to use their metrics, because they are based on undirected weighted graph with link strength (confidence scores). As seen in the traversal maps, reciprocity is generally low. The unavailability of bilateral links are obvious for dates and events. Mean distance is short, mostly under 2.0. Diameter is the length of the longest geodesic. We have rather short diameters, implying connections are limited within a small circle. Edge density is the ratio of the number of edges and the number of possible edges. Here we observe low density.

In addition, centrality is calculated, using three methods: Closeness (in and out), Eigen Vector, and Betweenness (Fig. 9). The Closeness statistically suggests the LOD hubs of outgoing and incoming links. The overall Closeness is similar across 11 sources. However, the contrast between Wikidata and Wikipedia as an incoming source and BabelNet and Europeana as an outgoing source can be observed. It is rather unexpected that there are no big differences between the sources for the centrality by Eigen Vector. Thus, the dominance of DBpedia (and to a less extent YAGO) is not clearly visible in the chart. VIAF and DBpedia seem to sit in-between position, mediating the linking flows. Moreover, a radar chart (Fig. 10) shows the indicator by R for the roles of vertex. The vertex is called a "hub" if it functions as a node to hold many outgoing edges, while it is called "authority" if it serves as a node

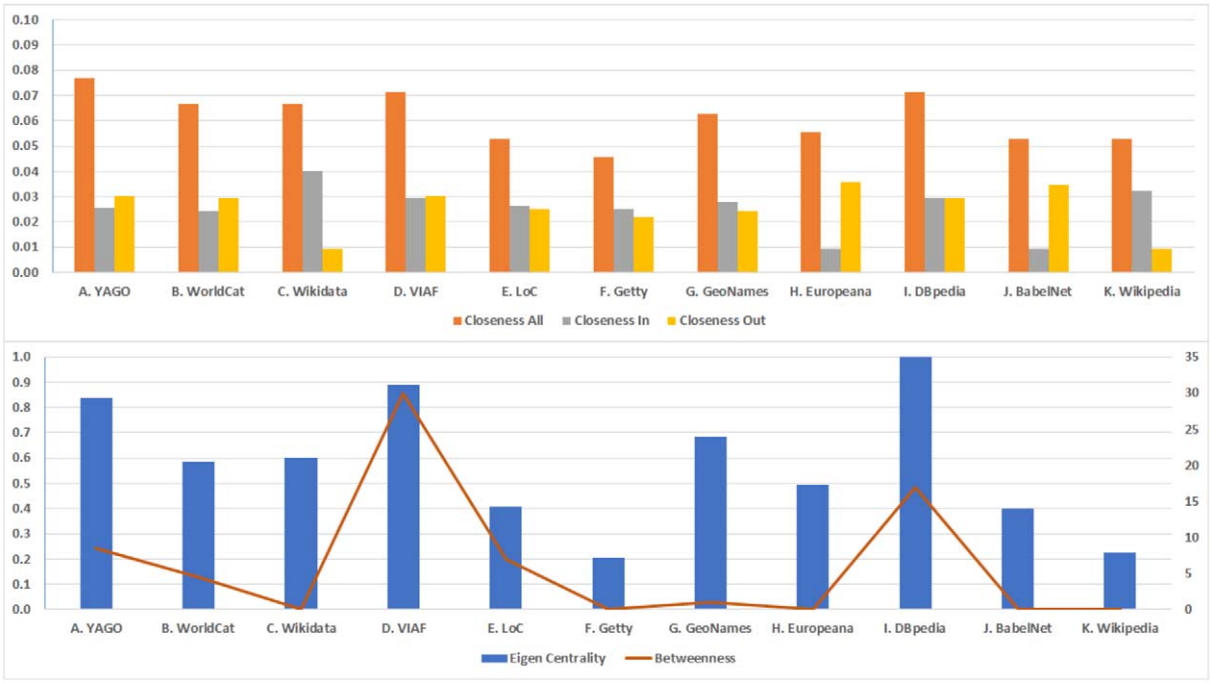


Fig. 9. Closeness (above) and centrality (below, left Y axe) and betweenness (below, right Y axe) for 11 data sources.

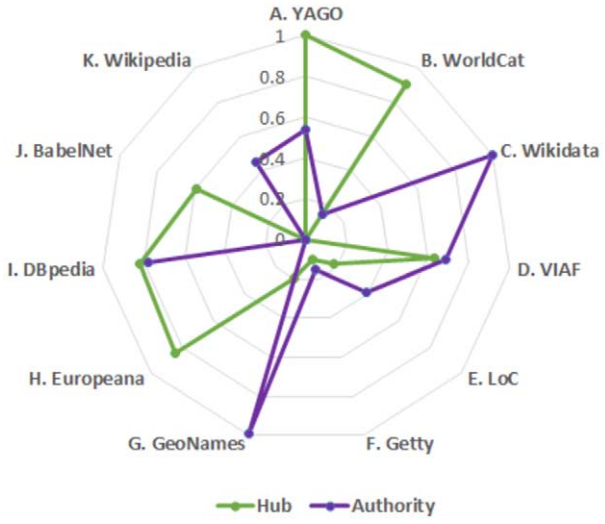


Fig. 10. Indicator by R if a data source is authority or hub.

to attract many incoming edges. Whereas YAGO, WorldCat, and Europeana are hubs, Wikidata and GeoNames are authorities. DBpedia has both characteristics, and is, therefore, a strong influencer for the analysed LOD sources.

Generally speaking, the overall situation shows a mosaic of segmentation even in a small LOD cloud. It is far from a full mesh network, if not data silos, which LOD is supposed to resolve. Our result simultaneously indicates a couple of tightly connected LOD clusters at best. Thus, it is currently hard to implement automatic traversals among the datasets without studying non-standardised properties (i.e. ontologies) and traversal maps.

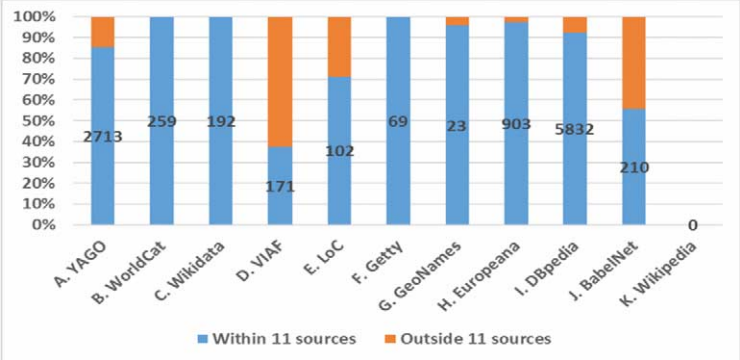


Fig. 11. The ratio of the four standardised property links going within and outside 11 data sources.

4.8. Connectivity and link types in detail

In order to better understand the overall connectivity of LOD datasets, we additionally generated more segmentation and detailed statistics.

Figure 11 illustrates how close the 11 data sources are connected to each other through four standardised properly links. It displays the ratio of the hyperlinks bounding for the domains of the 11 datasets. Thus, it should represent the openness or closure of this small network. A high level of exclusivity for our data sources is observed. On average, 87.8% of links are within the 11 dataset boundary. Except Wikipedia, VIAF remains the lowest source in terms of links to the other datasets, but still holds over 37.3%. The statistics clearly indicate the closed and close connections of the 11 data sources in terms of standardised traversability.

When combing with analysis in the previous sections, this closure and the homogeneity and centrality of the 11 datasets are a worrying sign in the sense that the users of 11 datasets are not able to identify and explore new and unknown datasets beyond those giants of LOD, hampering serendipity for users' research. This phenomenon would also decrease the diversity of the LOD cloud. Our analysis indicates that the identical entities in local cultural heritage datasets cannot be effectively connected to each other through NEL via the 11 global LOD sources. Data integration and/or contextualisation would only be possible if the users know the connectivity of datasets in advance and conduct a federated SPARQL query at known endpoints.

In fact, Ding et al. [15] note that the typical size of `sameAs` networks either remains a small constant or increases slowly, and that single central resources are connected to a number of peripheral resources. This condensed view of LOD is adequately depicted in their cluster analysis and visualisation, where a few LOD data sources investigated in this paper are clearly seen as in-degree or out-degree hub nodes such as DBpedia, GeoNames, Wikipedia, and WorldCat. Correndo et al. [9] also report a power-based LOD network. Moreover, recent research discovers two high-centrality nodes (DBpedia and Freebase) and domain specific naming authorities/hubs such as GeoNames among others [4]. The added value of our study is to reveal the extent of this phenomenon for four different properties at an instance level.

Now, let us take a close look at link types. Figure 12 presents the percentage of the four standard properties used within `rdf:resource`. In RDF/XML, `rdf:resource` is the property to indicate the URI of the object node in a graph.[34] In this sense, it should normally contain all the outgoing links. By dividing the ratio of the four properties, we can highlight the balance between them and other properties including proprietary ones.

The overall percentage is, unsurprisingly, low because the four properties are normally a small part of RDF content. Nevertheless, the range varies from 30% to close to 0%. An exception is Europeana. 90.6% of links use them, demonstrating a high conformity to the standardised RDF properties and highly limited use of proprietary properties. The result suggests relatively high importance of the four properties in the WorldCat and BabelNet datasets. In contrast, Getty vocabularies and Wikidata use other properties almost exclusively. Indeed, a query on

[34] https://www.w3.org/TR/rdf-syntax-grammar/, last accessed 2021-01-26.

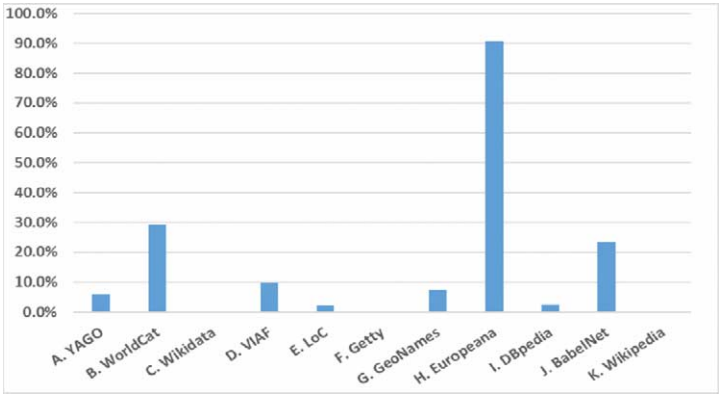


Fig. 12. The percentage of four srandardised properties used for the purpose of rdf:resource linking in 11 data sources.

WDProp[35] lists 8732 unique properties in Wikidata as of 26 January 2021. A manual examination of Wikidata entities further justifies the outcome: the properties are organised by its proprietary `wdt:` with P prefix, while `wdtn:` is the entities with Q prefix.[36]

For example, the entity of France contains 9500 `rdf:resource`, while `wdt:` is used 294 times with `rdf:resource`. 7292 `rdf:type` are included in combination with `rdf:resource`. The W3C properties of our concern are not available at all. `owl:sameAs` only appears occasionally to provide inverse relations for obsolete (mostly duplicate) properties that offer redirects. Erxleben et al. [18] explain that Wikidata is keen to faithfully represent the original data using the language of RDF and linked data properly. In particular, they claim that `owl:sameAs` would often not be justified to relate external URIs to Wikidata. This leads to their hesitation to use this property as well as to include links to many external data.

On the one hand, proprietary properties in Wikidata enable the users to refine the semantics of outbound links. It is useful in some cases where one needs to identify a particular link among tens of `owl:sameAs` links. On the other hand, they make it more difficult to automate graph traversals, when used with other LOD. In addition, there is a question of manageability and usability. As the outgoing link properties can be suggested by the users, the number of the properties could grow sharply. Then, the complication of selecting them will be amplified.

Another issue is that the Wikidata entities do not use human "guessable" URIs, even if they are not absolutely opaque URIs such as hash. For instance, the syntax of the entity URI for Cold War is https://www.wikidata.org/entity/Q8683. They are agnostic about their semantics and are language independent, which prevents human users from guessing the meaning of properties and/or hacking the URIs[37] without examining the ontology behind. We should recognise that self-describing URIs are rated high for the quality metrics of Candela et al. [8].

When we manually examined France in Getty, we found that there were 1783 `rdf:resource`. 1349 SKOS properties are used among which 10 `skos:prefLabel`, 18 `skos:altLabel`, and 1246 `skos:narrower` are present. Whereas 251 Dublin Core Metadata Terms (`dct:`)[38] and 202 Getty Ontology (`gvp:`)[39] are in use, 60 PROV (`prov:`)[40] and 56 SKOS-XL (`skosxl:`)[41] are also found. Although not all properties use `rdf:resource`, the figures provide us a clue about the relation between linking and property usage.

Figure 13 illustrates the ratio of each property among the four properties. Despite the wide spread of research concerning `owl:sameAs`, its use for outgoing links is less than the majority for all outgoing links (42.2%). While

[35] https://rawgit.com/johnsamuelwrites/wdprop/master/index.html, last accessed 2021-01-26.

[36] https://www.wikidata.org/wiki/Wikidata:SPARQL_tutorial last accessed 2021-01-26.

[37] In this context, hacking means the manipulation of URIs to access another data, for example, by changing prefix or suffix. See also http://www.jenitennison.com/2009/07/25/opaque-uris-unreadable-uris.html, last accessed 2021-01-26.

[38] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/, last accessed 2021-01-26.

[39] http://vocab.getty.edu/ontology, last accessed 2021-01-26.

[40] https://www.w3.org/TR/prov-o/, last accessed 2021-01-26.

[41] https://www.w3.org/TR/skos-reference/skos-xl.html, last accessed 2021-01-26.

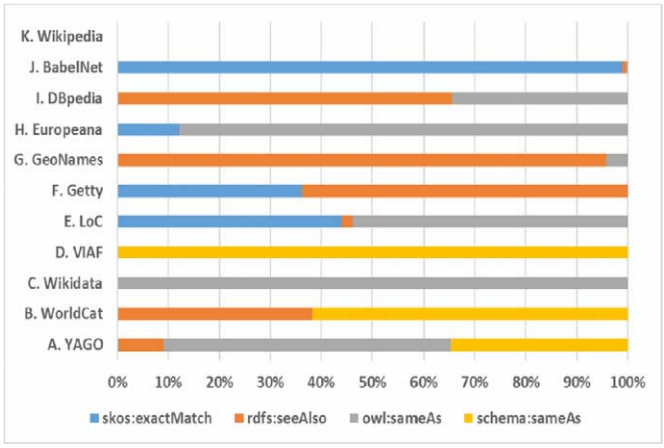


Fig. 13. The ratio of each property among the four standardised properties used in 11 data sources.

38.4% use `rdfs:seeAlso`, `schema:sameAs` and `skos:exactMatch` are in the minority. As GeoNames provides the link to DBpedia with `rdfs:seeAlso`, the equivalent identity cannot be inferred. `skos:exactMatch` is present in BabelNet, Europeana, Getty vocabularies, and the Library of Congress. VIAF exclusively uses `schema:sameAs`, whilst more than half of WorldCat entities are described with it. YAGO also uses it for more than one third of its entities. However, its use is debatable, since the schema.org ontology is not a W3C recommendation.[42] Moreover, Beek et al. [4] point out that it is semantically different from `owl:sameAs`.

From Fig. 12 and 13, it becomes clear that some data providers set different strategies to design their ontologies in spite of the W3C recommendations. The results indicate that it is not feasible to traverse LOD and collect information, if the users specify only one type of property. As seen throughout Section 4, the need of traversing strategies is also verified from this perspective.

4.9. Literals

This section examines the quality of other data content to supplement the analysis of link quality. The content-related four W3C standard properties are analysed, namely, `rdfs:label`, `rdf:type`, `skos:prefLabel`, and `skos:altLabel`. Figure 14 shows the ratio of each property among the four properties used in the 11 data sources.

Here one can also observe the characteristics of data sources. The contrast between `rdfs:label` and SKOS vocabularies is one focal point. Interestingly BabelNet prefers to use the former this time, in place of the latter. It is noted that GeoNames only uses `rdf:type`, primarily because it employs proprietary properties for the name of places (`gn:`):

```
<https://sws.geonames.org/6251999/>
gn:name "Canada";
gn:alternateName "Canada"@nn, "Kanuadu"@olo, "Ca-na-đa"@vi , "Kanada"@nds,
"Kanada"@mt, "ภาษาถา"@lo;
```

The library sector (VIAF, the Library of Congress, and WorldCat) uses `skos:altLabel` extensively. Generally speaking, it is evident that the use of properties is diverse and not standardised. Therefore, automatic retrieval of basic information such as entity labels would require good understanding of each data source before data processing begins.

We further investigate the core constructs of RDF/XML. The use of `rdf:resource` and `rdf:about` is analysed. The average amount of `rdf:resource`, `rdf:about`, and literals is shown in Table 9. In general, contrast is clearly visible between the data providers with a high volume of content (Wikidata, YAGO, DBpedia) and the

[42] https://schema.org/docs/howwework.html, last accessed 2021-01-26.

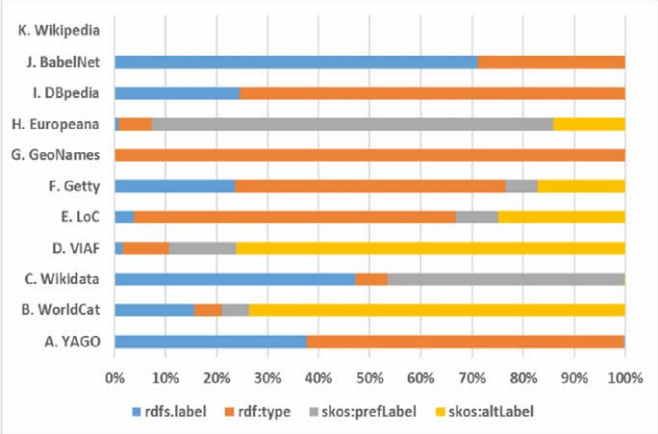


Fig. 14. The ratio of each content-related property among the four content-related properties used in 11 data sources.

Table 9
The average number (per entity) of rdf:resource, rdf:about, and literals for each data source

ID	A	B	C	D	E	F	G	H	I	J	K	Total
Source	YAGO	WorldCat	Wikidata	VIAF	Library of Congress	Getty	GeoNames	Europeana	DBpedia	BabelNet	Wikipedia	
rdf:resource	530.6	9.3	4696.1	84.1	62.2	595.4	14.3	41.0	2546.5	16.3	0.0	8595.9
rdf:about	1.1	8.1	2164.8	27.5	93.5	73.6	2.0	1.3	2285.8	5.9	0.0	4663.7
Literals	105.2	43.7	50723.9	448.1	230.5	207.2	176.1	138.1	82.6	2.9	0.0	52158.3

rest. Somehow Getty has competitive numbers. We are also curious about the low average of 1.1 for `rdf:about` in YAGO. When we had a close look at the dataset, we discovered that it used a single instance of `rdf:about` for the entity itself, for example, as follows:

```
<rdf:Description rdf:about="http://dbpedia.org/resource/World_War_II">
</rdf:Description>
```

Similarly, each entity in GeoNames contains it exactly twice (2.0 for `rdf:about`):

```
<gn:Feature rdf:about="http://sws.geonames.org/2077456/"></gn:Feature>
<foaf:Document
rdf:about="http://sws.geonames.org/2077456/about.rdf"></foaf:Document>
```

The second `rdf:about` preserves the technical metadata about the entity such as a Creative Commons license and creation date.

Moreover, we investigate the amount of literals. However, they have to be treated carefully, as they may include less relevant information about the entity. Despite the caveats, the figures do provide a rough idea of how much content is described in each LOD instance. Manual inspection indicates that the number of literals in some LOD is extremely high. This is not only due to an enormous amount of technical metadata, but also to repetitions (e.g. literals expressed in several schemas) and language variations in them. For example, there are in total over 4.5 million literals and, on average, more than 50 thousand for the 100 entities in Wikidata.

4.10. Content coverage

This section presents our attempt to further enhance the results of Section 4.9. Our Python scripts compare the content differences of the 100 instances across the 11 data sources (see Fig. C.1 in Appendix C). The amount of unique content of a single entity and the ratio are automatically calculated, and the aggregated view for the 11 data

Table 10
The number of unique data content per data source in each category (values in parentheses indicate coverage in percentage)

ID	Data Source	Overall	Agents	Events	Dates	Places	Objects & Concepts
A	YAGO	251293 (56.2)	19201 (34.2)	24227 (55.0)	418 (0.9)	202215 (72.5)	5232 (24.2)
B	WorldCat	3667 (0.8)	886 (1.6)	346 (0.8)	276 (0.6)	1876 (0.7)	287 (1.3)
C	Wikidata	69183 (15.5)	18944 (33.8)	6708 (15.2)	4074 (8.8)	34068 (12.2)	5389 (24.9)
D	VIAF	11207 (2.5)	5695 (10.2)	0 (0.0)	0 (0.0)	4702 (1.7)	810 (3.7)
E	LoC	11980 (2.7)	2587 (4.6)	2253 (5.1)	774 (1.7)	4997 (1.8)	1369 (6.3)
F	Getty	23894 (5.3)	1605 (2.9)	0 (0.0)	0 (0.0)	21783 (7.8)	506 (2.3)
G	GeoNames	3284 (0.7)	0 (0.0)	0 (0.0)	0 (0.0)	3200 (1.1)	84 (0.4)
H	Europeana	3746 (0.8)	1375 (2.5)	256 (0.6)	0 (0.0)	2115 (0.8)	0 (0.0)
I	DBpedia	128307 (28.7)	20212 (36.0)	20003 (45.4)	40951 (88.1)	35469 (12.7)	11672 (53.9)
J	BabelNet	1866 (0.4)	359 (0.6)	345 (0.8)	358 (0.8)	497 (0.2)	307 (1.4)
K	Wikipedia	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
–	**Full Coverage**	**447065 (100.0)**	**56068 (100.0)**	**44044 (100.0)**	**46489 (100.0)**	**278807 (100.0)**	**21657 (100.0)**

sources is shown in Table 10. In theory, they should represent the coverage and diversity of content (for a data source). The table is grouped by categories (i.e. all entities within are aggregated), because the instances tend to show similar patterns within the same category. "Full coverage" indicates the total amount of the unique content that 11 data sources hold as a whole (thus 100% coverage). It means that overlapping content is calculated once. The percentage of a data source indicates the ratio of the unique content against the full coverage.

In the overall column of Table 10, YAGO holds the largest amount of unique content (56.2%), which also implies that it is the data source with the most diverse content. It is nearly double the size of DBpedia. It may be also surprising that Wikidata contains just over a half of the DBpedia data. When we look at this from a cross-domain LOD perspective, the Library of Congress and WorldCat are considered as small-scale datasets, while the number of BabelNet content is even smaller. Obviously, data sources containing fewer entities provide less content.

Regarding the agents category, DBpedia exceeds YAGO and Wikidata. As expected VIAF is also prominent. However, the number is rather disappointing, compared to these three sources.

With regard to events, the reasons why the Library of Congress has relatively high number of contents is mostly due to `bflc:subjectOf` link. DBpedia provides a large number of seemingly Wikipedia derived content, ranging from links (related persons, places, events, and digital resources) to literal descriptions in different languages.

In the dates category, DBpedia has substantial advantage (88.1%). Other sources are unlikely to offer highly informative content. We also conducted manual inspection on our data sources. We discovered that the high volume of DBpedia in general was most likely due to a large number of links (derived from Wikipedia article `dbo:wikiPageExternalLink` (i.e., external links, further reading in Wikipedia) and `dbo:wikiPageWikiLink` (i.e., many useful links in Wikipedia). Wikidata is the second highest source (as it contains labels in many languages), but it is hard to understand the target resource with opaque entity names (`wd:Qxxxx`). The Library of Congress has useful links to their library resources related to the date (`bflc:subjectOf`). The Library of Congress and WorldCat use SKOS to connect to broader concepts of decade. It is noticeable that the library-based LOD sources (WorldCat, the Library of Congress, VIAF) have many overlapping content. BabelNet also uses `skos:broader`, but it seems the links are generated programmatically and it uses proprietary IDs (like Wikidata). Thus, it is hard for machine (and humans) to understand the meaning of the links. In addition, for some reason, the RDF representation of an entity has a significantly lower number of links compared to the HTML representation, therefore, some useful information may be lost.

YAGO shows strength in the places category, given that the ratios are more evenly distributed across all sources due to the availability of the entities in this popular category. Interestingly, Getty Vocabularies (TGN) performs relatively well, whereas GeoNames is not as good as we expected. New and diverse information may not be found in the latter.

As for objects and concepts category, the strength of DBpedia persists. It seems that it extracted a great deal of data from Wikipedia. Understandably, Wikipedia articles would be more exciting for human users than a collection of factual data in LOD.

In general, this analysis suggests: (a) the concentration of (diverse) content in DBpedia, YAGO, and Wikidata, and (b) data richness in specific proprietary properties. A critical question is how the 11 LOD producers facilitate users to find them among hundreds of properties, in order to access rich information, especially if they are unfamiliar with their ontologies. The hurdle could be higher for the data integration by federated queries in multiple LOD sources.

Table 11 illustrates the amount of data overlaps per category. While the one-source column indicates the number of non-overlapping content for the source (i.e., unique content), other columns indicate the number of overlapping content (i.e, two to ten sources hold identical string). Interestingly, the content covering all data sources does not exist at all. This implies that even the most standard English label cannot be found in every source. Over 75% of content is unique. However, overlaps in two sources are relatively high for agents, events, and objects and concepts. The numbers drop sharply for the overlap in more than two sources. However, very high coverage is also seen for agents, places, and objects and concepts. One reason for these phenomena would be the contrasting volume of data sources. As we have seen earlier, the disproportionately high volume of DBpedia, YAGO, and Wikidata makes the rest of the sources look insignificant. Therefore, although there are some highly overlapping content, the percentages remain very low.

Our assumption is two-fold: (1) the higher the coverage, the more accessible the data, yet the more redundancy in the LOD cloud, and (2) the lower the coverage, the more serendipity with unique content, yet redundant traversals. From this perspective, it is too early for us to judge how much users benefit from a large amount of unique content, and/or how much they suffer from redundant information in multiple sources, because we do not have gold standard for data quality.

We additionally created intriguing views of the amount of unique content per entity for each category. Figure 15 provides a view for the events category. In this case, content diversity is clearly visible, ranging from the rich volume of Byzantine Empire and Mars, to poor volume of Traja and Like a Rolling Stone. The details of other categories and short comments are found in Appendix C.

5. Conclusions

5.1. Challenges for cultural heritage datasets

This research strives to uncover gaps between the data producers and consumers. Indeed, our evaluation of 11 LOD providers reveals a clear sign of data quality issues from a user perspective, which have neither been examined in this detail nor on an instance level by other studies. While it verifies some results of the previous research, it also pinpoints additional issues, in particular, issues specific to the cultural heritage domain, as well as the different types of link properties and literals.

Our analysis confirms the observations of Ahlers and Debattista et al. [3,11] that a limited number of links are found for major LOD datasets, with the exception of the relatively ample amount for DBpedia (RQ1). A large proportion of LOD sources may not be fully connected and unevenly interlinked for the representative entities (RQ2, 3). This result also reflects previous LOD studies on the overall quality and `owl:sameAs` networks [9,15]. In particular, power-law-based networks and closures have been found for the LOD cloud. Moreover, centrality can be observed for not only linkages, but also for data content.

"High-volume and high-quality" datasets are biased toward a couple of data sources, especially generic knowledge bases (RQ3). Consequently, it is uncertain if users and researchers would be able to find new information, let alone to answer more specialised questions that they are interested in. As Zaveri et al. pointed out [45], assuring data quality is particularly a challenge in LOD as the underlying data stems from multiple autonomous and evolving data sources.

Some valuable information about the same entity is not easily reachable due to the lack of links, and/or redundantly long traversing (RQ2). For example, it is not possible for a user looking at Beethoven in Getty ULAN to obtain relevant artists and songs in BabelNet. Generally speaking, due to the heterogeneity of LOD quality and linking patterns, it seems that the automation of graph traversals and the subsequent data integration currently require more human effort than necessary (RQ4).

Table 11
The amount of ovelaping content per category[1]

Category	1 Source	2 Sources	3 Sources	4 Sources	5 Sources	6 Sources	7 Sources	8 Sources	9 Sources	10 Sources	SUM
Agents	42871 (76.3)	12373 (22.0)	627 (1.1)	231 (0.4)	58 (0.1)	20 (0.0)	5 (0.0)	14 (0.0)	0 (0.0)	–	56199 (100.0)
Events	34308 (77.9)	9437 (21.4)	262 (0.6)	29 (0.1)	8 (0.0)	0 (0.0)	0 (0.0)	–	–	–	44044 (100.0)
Dates	46163 (99.3)	290 (0.6)	36 (0.1)	0 (0.0)	0 (0.0)	0 (0.0)	–	–	–	–	46489 (100.0)
Places	255184 (91.5)	20051 (7.2)	1500 (0.5)	823 (0.3)	582 (0.2)	173 (0.1)	234 (0.1)	100 (0.0)	160 (0.1)	0 (0.0)	278807 (100.0)
Objects & Concepts	16880 (84.2)	2694 (13.4)	386 (1.9)	65 (0.3)	19 (0.1)	10 (0.0)	2 (0.0)	2 (0.0)	0 (0.0)	–	20058 (100.0)
Overall	**393028 (88.9)**	**43934 (9.9)**	**2764 (0.6)**	**1131 (0.3)**	**664 (0.2)**	**201 (0.0)**	**241 (0.1))**	**116 (0.0)**	**160 (0.0)**	**0 (0.0)**	**442239 (100.0)**

[1]"1 source" column indicates the number of content without any overlap (unique content). From the "2 sources " column to "10 sources", the number of overlapping content is seen. Values in parentheses indicate percentages within each category. Due to the lack of entities in data sources, some celles are blank.

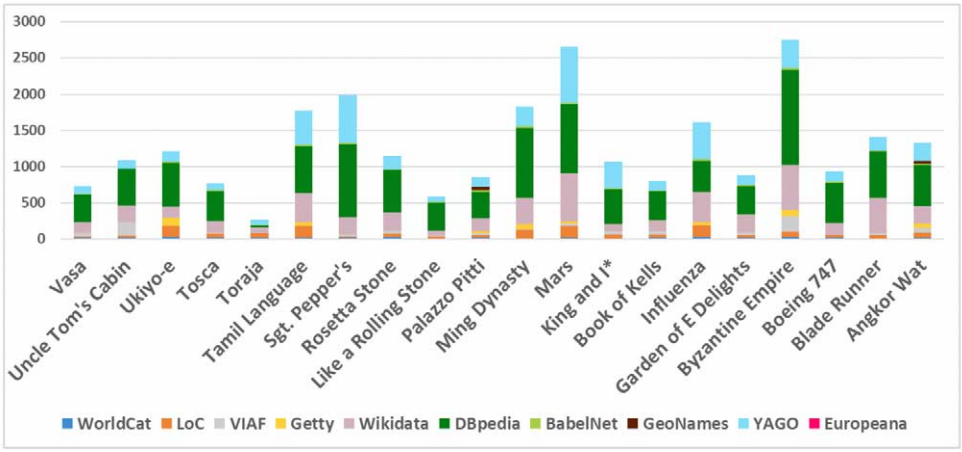


Fig. 15. The amount of content for events entities per data source.

Those are serious shortcomings for our research scenarios. In other words, the quality of a hundred representative entities from major LOD providers has not yet met the basic needs of researchers.

From a user's perspective, our analyses also provide an insight into LOD that previous research has not been able to deliver. For example, it became clear that some objects and concepts may introduce complication, because links between LOD resources may be missing and/or confusingly created (RQ3, 5). There seem to be a different number of corresponding records, depending on the type of concepts in FRBR (work, manifestation, expression, and item). Unlike skilled librarians, average users on the web would not be able to distinguish four types of FRBR resources and solve co-references on their own. However, this is not a technical problem of LOD, but an issue about the different perceptions and/or understanding of users about the conceptualisation of entities. This "semantic gap" between the data consumers and data producers has the potential to cause problems for research in the future.

As we have seen, an obstacle for interoperability and data processing automation is proprietary properties. LOD is not as powerful as it can be, as long as human users analyse related data every time when traversing data, because they are not initially aware of data sources and their ontologies in their query time [40]. This is particularly true for a large amount of data for which manual analysis is unrealistic. According to Bizer et al. [6], it is a good practice to reuse terms from well-known RDF vocabularies wherever possible, and only if they do not provide the required terms should data publishers define new, data source-specific terminology. In the interoperability metric of Candela et al. [8], the use of external vocabularies is also favoured for the LOD quality assessment. At the same time, we found that rich information tended to be "hidden" in proprietary properties among many other properties (RQ1, 2). Without close manual examination of ontology and data itself, it would not be easy to automate data processing (RQ4).

5.2. *Limitations of our analysis*

Admittedly, this article has some limitations. It focuses on the analysis of LOD entities which provide a context for cultural heritage research. For example, as mentioned earlier, Europeana has enriched its digital object datasets with named entities. One may find cultural heritage objects with `owl:sameAs` links to GeoNames or DBpedia. However, the entity collection of Europeana analysed in this paper has been created separately from the object datasets. Europeana offers (a) LOD instances (i.e. digital cultural heritage objects via OAI-PMH and SPARQL endpoint), (b) their related entities (i.e. contextual entity via REST APIs that we analysed), and (c) the ontology (i.e. Europeana Data Model). Therefore, co-reference resolution should occur in situations such as SPARQL queries, so that the related instances could actually "meet" via an identical entity in the same repository. Thus, it is usually not possible to see such data integration in the lookup scenario we used in our research (RQ2).

In addition, in case of external entity linking, federated queries are required to investigate the data integration across different LOD sources, which is slightly out of the scope of this paper.[43] For the same reason, we could not apply such sophisticated network metrics as developed by Idrissou et al. [26], because they cannot be easily evaluated in the lookup scenario. Moreover, due to different characteristics of graphs (i.e. weights and directions), it is necessary to heavily customise the metrics. We take those issues for the upcoming research.

Furthermore, largely due to the manual-based methodology, the sample size remains the bare minimum. However, LOD is oftentimes populated programmatically, although crowdsourced LOD such as Wikidata would have more manual curation by human users. In fact, we show that much LOD content is relatively standardised or normalised; the number of links at a data source is relatively similar and consistent across entities in the same category (RQ5). It is therefore doubtful that if a large-scale sampling would make our results considerably different.

Nevertheless, our research should aim for the fusion of manual and automatic evaluation in the future. As Idrissou et al. [26] stress, we agree that the links must often be human validated, since entity resolution algorithms are far from being perfect. We also consent to computer support that can accurately estimate the quality of LOD, because the manual analysis is both a costly and an error-prone process.

It is also worth mentioning that there are some technical challenges concerning the automatic analysis of LOD. We encountered many small problems to collect and analyse the data. For example, data is sometimes not consistent (RQ1, 2, 4, 5). YAGO has an issue with special characters in the data. We observed this for Uncle Tom's Cabin and Sgt Peppers Lonely Heart Club Band. In case of the former, YAGO's URI is different from that of the DBpedia URI, while all other URIs are identical for the two sources. Thus, error handling was required for those exceptional entities in Python scripts. In addition, the stability of URIs is extremely important, but not always guaranteed. If we look at a broader range of LOD resources, we know that, for example, there was certain impact, when the GND, the German integrated authority records, changed their entity URIs from HTTP to HTTPS in 2019.[44]

5.3. Recommendations for data consumers and producers

Despite those caveats for limitations, the investigation in this paper clearly indicates that NEL in local databases may not be as sufficient as one may think (RQ1). Our study observes an iceberg of a large variation in data quality on the web [45]. Thus, it would be wrong to expect that NEL automatically generates synergies for LOD data integration. Indeed, successful projects applying such data integration are highly limited so far in our field. Careful strategies are required to identify efficient traversals and obtain data such as multilingual labels and links to global and/or local databases, and integrate heterogeneous datasets in a useful fashion (RQ2, 3). One recommendation for the NEL strategy would be to refer to hubs such as YAGO, DBpedia, and WorldCat as much as possible, from where the W3C standardised links to other major LOD resources are available. At the same time, one should be aware that YAGO and WorldCat would be the best choice to find information in Wikipedia. While WorldCat is not connected to DBpedia, it has links to the Library of Congress, which DBpedia does not. Contrary to many practices of NEL in cultural heritage, links to Wikidata would be recommended if the users have a good understanding of its proprietary properties to access other data sources. In addition, our traversal maps can be used as an orientation guide for the NEL implementers.

It is ironic that although Wikidata generally receives high numbers of incoming links from other sources and holds a substantial amount of information, it does not offer the standardised way of providing outgoing links at all. This could be a controversial issue for the efficiency and/or "democratisation" of LOD. A limited amount of new data could be obtained from WorldCat, BabelNet, and GeoNames. It is therefore not promising to carry out serious research with such data as it seems that some datasets tend to serve merely as global identifiers, rather than new sources of information (RQ1).

Simultaneously, the use of opaque URIs and a large number of proprietary properties in Wikidata should be more intensively discussed by the LOD publishers and consumers, especially by the NEL implementers, because Wikidata is becoming a NEL standard in cultural heritage [36].

[43] There is also a serious technical problem with scalability for federated SPARQL queries on the web, which makes it hard to conduct analysis of our kind.

[44] https://wiki.dnb.de/display/DINIAGKIM/HTTP+vs.+HTTPS+in+resource+identification, last accessed 2022-01-18.

In any case, providing multiple links during NEL will increase interoperability, because it may avoid redundant traversals and give us more flexibility (RQ2). At the same time, we can also advise the maintainers of 11 LOD sources to fully link to each other, as well as to provide more links to other local datasets as much as possible. The reciprocal links will allow users to integrate truly interdisciplinary and heterogeneous datasets. In a way, our study identifies the myth of NEL and verifies the obstacles of LOD (RQ1). NEL is a step necessary to the use of multiple datasets in LOD [26]. However, linking is the means, not the goal.

5.4. Discussions on local datasets

The connection between local datasets and globally known reference resources that this paper deals with has been largely uninvestigated (RQ2). This entails that the local-to-local (L2L) connections via global sources are not well known, although LOD and NEL are designed to perform this task. One exception is demonstrated by Waagmeester et al. [44], describing four cases with federated SPARQL queries to connect Wikidata with local datasets. Yet, our research clarifies that the 11 global LOD sources do not easily enable us to integrate local datasets due to the lack of links to them (RQ1). In addition, if two local datasets point to different global sources, they need to traverse more than one graph in order to link each other. This means that the destination of NEL determines the usability of L2L data integration. In any case, a feasibility study on the L2L data integration would be one of the next tasks for our research. We could extend it further by exploring what innovative research we could actually do after NEL and federated queries. Pilot use cases are needed to simulate and evaluate data aggregation, contextualisation and integration as the outcomes of NEL in the cultural heritage field, followed by semantic reasoning and creation of new knowledge. Otherwise there is a risk that LOD would end up with an idealistic vision without concrete impact on our society.

Related to this, there are also problems with local datasets. It is known that some LOD in cultural heritage is not adequately and sufficiently published. For instance, Francorum Online[45] has technical problems. Pleiades[46] provides RDF/XML, but does not offer links to major LOD that are available in JSON. Other LOD projects (LOCAH[47] and PCDHN[48]) have other problems such as sustainable funding. From a quantity perspective, it is hoped that more local LOD will be published and connected to improve the overall "researchability" for the domain.

5.5. Further research and development in semantic web

To enhance the analysis carried out in this article, it would be interesting to investigate the LOD traversability in comparison with all the LOD properties actually used. For instance, Linked Open Vocabularies[49] is a good starting point to analyse the acceptance of a broad range of properties for LOD and the implications of standardisation and proliferation of vocabularies. In addition, the automated graph traversals and data integration can be examined, using SPARQL queries. Although our research concentrates on lookup because of the NEL setting, analysis on federated queries can uncover the real research scenarios of the end users.

As Berners-Lee states [5] that "statements which relate things in the two documents must be repeated in each" and further, "a set of completely browsable data with links in both directions has to be completely consistent, and that takes coordination, especially if different authors or different programs are involved." As such, reciprocal links and lookups need to be added with care. For the next step, it seems necessary for the web community to help major LOD dataset maintainers to identify incoming LOD as much as possible, and enrich the datasets to create reciprocal links. Even if a full mesh network is not an aim for many LOD data sources, it would be critical for the LOD creators to be aware of and interconnect with other LOD data sources in order to provide a way to find as much new information as possible (RQ1, 2, 3).

[45] http://francia.ahlfeldt.se/index.php, last accessed 2021-01-26.
[46] https://pleiades.stoa.org/, last accessed 2021-01-26.
[47] http://data.archiveshub.ac.uk/, last accessed 2021-01-26.
[48] https://dataverse.library.ualberta.ca/dataset.xhtml?persistentId=doi:10.7939/DVN/URXSGC, last accessed 2021-01-26.
[49] https://lov.linkeddata.es/dataset/lov/, last accessed 2021-01-26.

Python analysis let us remember that data overlaps across data sources are duplicate information (RQ1, 5). On the positive side, fewer traversals are needed to find the same information. On the negative side, data is redundant. As the size of the LOD cloud grows, it may confuse users in the vast amount of information like a needle in a haystack. Use cases by researchers would help to evaluate the pros and cons of the LOD's distributed data approach. In this regard, we also need to find a way to adequately manage and use aggregation services of LOD.

One example which enables the users to compare LOD sources is SILK [43]. Although it is limited to two data sources, it provides support to create and maintain interlinks. Their update notification service is also particularly valuable. It is also possible and realistic that third-party services would be developed for the integration of LOD data sources [21,27]. However, there are limited numbers of web applications capable of crawling the web and detecting incoming links of LOD. Some projects offer data dumps containing such information. Yet, they often do not provide an interactive interface. Furthermore, research on LOD search engines is advancing somewhat slowly. Although there are some projects including Swoogle, Sindice, and LODatio [12], many are experimental, out-of-date, or un-user friendly. It is hoped that next generation of search engines for LOD will be developed.

This paper highlights the reality of a reasonable set of LOD datasets in cultural heritage, but the discussion is applicable for other domains. By removing the obstacles found in this article, LOD traversing and date integration become more feasible for end-users with help of automatised tools.

Acknowledgements

This work was partially supported by the EU Horizon 2020 project InTaVia: In/Tangible European Heritage – Visual Analysis, Curation and Communication (http://intavia.eu) under grant agreement No 101004825.

Appendix A. Entity coverage per data source

Table A1

The occurrences of 100 entities in 11 data sources (A to K) (zero indicates absence. More than one means duplicate entities)

		A	B	C	D	E	F	G	H	I	J	K	SUM	Occurence SUM
		YAGO	Worldcat	Wikidata	VIAF	LoC	Getty	GeoNames	Europeana	DBpedia	BabelNet	Wikipedia		
1	Carl	1	1	1	1	1	1	0	0	1	1	1	9	9
2	Jesus	1	1	1	1	1	0	0	0	1	1	1	8	8
3	Aristotle	1	1	1	2	1	1	0	1	1	1	1	11	10
4	Napoleon	1	1	1	1	1	1	0	0	1	1	1	9	9
5	Adolf	1	1	1	1	1	1	0	1	1	1	1	10	10
6	Julius	1	1	1	2	1	1	0	0	1	1	1	10	9
7	Plato	1	1	1	1	1	1	0	1	1	1	1	10	10
8	William	1	1	1	1	1	1	0	1	1	1	1	10	10
9	Albert	1	1	1	1	1	1	0	0	1	1	1	9	9
10	Elizabeth II	1	1	1	1	1	1	0	0	1	1	1	9	9
11	Michael	1	1	1	1	1	1	0	0	1	1	1	9	9
12	Madonna	1	1	1	1	1	0	0	1	1	1	1	9	9
13	Ludwig van	1	1	1	1	1	1	0	1	1	1	1	10	10
14	Wolfgang	1	1	1	1	1	1	0	1	1	1	1	10	10
15	Pope	1	1	1	1	1	0	0	0	1	1	1	8	8
16	Alexander	1	1	1	2	1	1	0	0	1	1	1	10	9
17	Charles	1	1	1	1	1	1	0	0	1	1	1	9	9
18	Barack	1	1	1	1	1	1	0	0	1	1	1	9	9
19	Mary	1	1	1	1	1	0	0	0	1	1	1	8	8
20	Queen	1	1	1	1	1	1	0	0	1	1	1	9	9
1	World War	1	1	1	0	1	0	0	0	1	1	1	7	7
2	World War	1	1	1	0	1	0	0	1	1	1	1	8	8
3	American	1	1	1	0	1	0	0	0	1	1	1	7	7
4	FA Cup	1	1	1	0	1	0	0	0	1	1	1	7	7
5	Vietnam	1	1	1	0	1	0	0	0	1	1	1	7	7
6	Academy	1	1	1	0	1	0	0	0	1	1	1	7	7
7	Cold War	1	1	1	0	1	0	0	0	1	1	1	7	7
8	Korean	1	1	1	0	1	0	0	0	1	1	1	7	7
9	American Revolution	1	1	1		1	0	0	0	1	1	1	7	7
10	UEFA	1	1	1	0	1	0	0	0	1	1	1	7	7
11	UEFA	1	0	1	0	0	0	0	0	1	1	1	5	5
12	Olympic	2	1	1	0	1	0	0	0	1	1	1	8	7
13	Stanley	1	1	1	0	1	0	0	0	1	1	1	7	7
14	Super	1	1	1	0	1	0	0	0	1	1	1	7	7
15	Iraq War	1	1	1	0	1	0	0	0	1	1	1	7	7
16	War of	1	1	1	0	1	0	0	0	1	1	1	7	7
17	Gulf War	1	1	1	0	1	0	0	0	1	1	1	7	7
18	Spanish	1	1	1	0	1	0	0	0	1	0	1	6	6
19	World	1	1	1	0	1	0	0	0	1	1	1	7	7
20	EFL Cup	1	1	1	0	1	0	0	0	1	1	1	7	7
1	1987	1	1	1	0	1	0	0	0	1	1	1	7	7
2	1986	1	1	1	0	1	0	0	0	1	1	1	7	7
3	1985	1	1	1	0	1	0	0	0	1	1	1	7	7
4	1984	1	1	1	0	1	0	0	0	1	1	1	7	7
5	1983	1	1	1	0	1	0	0	0	1	1	1	7	7
6	1982	1	1	1	0	1	0	0	0	1	1	1	7	7
7	1981	1	1	1	0	1	0	0	0	1	1	1	7	7
8	1980	1	1	1	0	1	0	0	0	1	1	1	7	7
9	1979	1	0	1	0	0	0	0	0	1	1	1	5	5
10	1978	0	1	1	0	1	0	0	0	1	1	1	6	6
11	1977	1	1	1	0	1	0	0	0	1	1	1	7	7
12	1976	1	0	1	0	0	0	0	0	1	1	1	5	5
13	1975	1	1	1	0	1	0	0	0	1	1	1	7	7
14	1969	1	1	1	0	1	0	0	0	1	1	1	7	7
15	1968	1	1	1	0	1	0	0	0	1	1	1	7	7
16	1967	1	1	1	0	1	0	0	0	1	1	1	7	7
17	1966	1	1	1	0	1	0	0	0	1	1	1	7	7
18	1965	1	1	1	0	1	0	0	0	1	1	1	7	7
19	1964	1	1	1	0	1	0	0	0	1	1	1	7	7
20	1960	1	1	1	0	1	0	0	0	1	1	1	7	7
1	United	1	1	1	1	1	1	1	1	1	1	1	11	11
2	United	1	1	1	1	1	1	1	1	1	1	1	11	11
3	France	1	1	1	1	1	1	1	1	1	1	1	11	11
4	England	1	1	1	1	1	1	1	1	1	1	1	11	11
5	Germany	1	1	1	1	1	1	1	1	1	1	1	11	11
6	Canada	1	1	1	1	1	1	1	1	1	1	1	11	11
7	Australia	1	1	1	2	1	1	1	1	1	1	1	12	11
8	Japan	1	1	1	1	1	1	1	1	1	1	1	11	11
9	Italy	1	1	1	1	1	1	1	1	1	1	1	11	11
10	Poland	1	1	1	1	1	1	1	1	1	1	1	11	11
11	India	2	1	1	1	1	1	1	1	1	1	1	12	11
12	Spain	2	1	1	1	1	1	1	1	1	1	1	12	11
13	London	2	1	1	1	1	1	1	1	1	1	1	12	11
14	Russia	2	1	1	1	1	1	1	1	1	1	1	12	11
15	New York	1	0	1	1	1	1	1	0	1	1	1	9	9
16	Brazil	2	1	1	1	1	1	1	0	1	1	1	11	10
17	California	2	1	1	1	1	1	1	0	1	1	1	11	10
18	New York	2	1	1	1	1	1	1	1	1	1	1	12	11
19	The	2	1	1	1	1	1	1	1	1	1	1	12	11
20	Sweden	2	1	1	1	1	1	1	1	1	1	1	12	11
1	Book of	1	1	1	1	1	0	0	0	1	1	1	8	8
2	Vasa	1	1	1	1	1	0	0	0	1	1	1	8	8
3	The Garden of	1	1	1	1	1	0	0	0	1	1	1	8	8
4	Rosetta	1	1	1	1	1	0	0	0	1	1	1	8	8
5	Palazzo	1	1	1	2	1	1	0	1	1	1	1	11	10
6	Boeing 747	1	1	1	0	1	0	0	0	1	1	1	7	7
7	Sgt.	1	1	1	2	1	0	0	0	1	1	1	9	8
8	Tosca	1	1	1	1	1	0	0	0	1	1	1	8	8
9	Blade	1	1	1	1	1	0	0	0	1	1	1	8	8
10	Uncle	1	1	1	1	1	0	0	0	1	1	1	8	8
11	Ming	1	0	1	1	1	1	0	0	1	1	1	8	8
12	Ukiyo-e	1	1	1	0	1	1	0	0	1	1	1	8	8
13	Angkor	1	1	1	1	1	1	1	0	1	1	1	10	10
14	Toraja	1	1	1	0	1	0	0	0	1	1	1	7	7
15	Byzantine	1	1	1	2	1	1	0	0	1	1	1	10	9
16	Mars	1	1	1	1	1	1	0	0	1	1	1	9	9
17	Tamil	2	1	1	0	3	1	0	0	1	1	1	11	8
18	Influenza	1	1	1	0	1	0	0	0	1	1	1	8	8
19	The King	1	1	1	3	2	0	0	0	1	1	1	11	8
20	Like a	1	1	1	1	1	0	0	0	1	1	1	8	8
	SUM	110	95	100	64	100	44	22	25	100	99	100	859	836
	Occurence SUM	99	95	100	55	97	44	22	25	100	99	100	836	

Appendix B. Source matrix data

Table B1

Matrix data which generated the chord diagrams (Fig. 1)

1.1 Matrix data with inverse (Fig. 1 left)

	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	98	0	20	0	0	0	0	0	0	0	0	118
LoC	93	43	52	13	0	0	0	0	0	0	0	201
VIAF	58	59	0	0	0	16	0	0	0	16	0	149
Getty	0	0	18	56	0	0	0	0	0	0	0	74
Wikidata	1	0	43	0	192	100	0	0	0	98	8	442
DBpedia	0	0	38	0	0	5599	108	23	0	1397	870	8035
BabelNet	0	0	0	0	0	0	0	0	0	0	0	0
GeoNames	8	0	0	0	0	22	20	0	0	20	17	87
Wikipedia	1	0	0	0	0	0	0	0	0	1094	0	1095
YAGO	0	0	0	0	0	95	82	0	0	88	8	273
Europeana	0	0	0	0	0	0	0	0	0	0	0	0
SUM	259	102	171	69	192	5832	210	23	0	2713	903	10474

1.2 Matrix data without inverse (Fig. 1 right)

	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	1	0	20	0	0	0	0	0	0	0	0	21
LoC	93	43	52	13	0	0	0	0	0	0	0	201
VIAF	58	59	0	0	0	16	0	0	0	16	0	149
Getty	0	0	18	56	0	0	0	0	0	0	0	74
Wikidata	1	0	43	0	0	100	0	0	0	98	8	250
DBpedia	0	0	38	0	0	1581	108	23	0	1397	870	4017
BabelNet	0	0	0	0	0	0	0	0	0	0	0	0
GeoNames	8	0	0	0	0	22	20	0	0	20	17	87
Wikipedia	1	0	0	0	0	0	0	0	0	1094	0	1095
YAGO	0	0	0	0	0	94	82	0	0	88	8	272
Europeana	0	0	0	0	0	0	0	0	0	0	0	0
SUM	162	102	171	69	0	1813	210	23	0	2713	903	6166

Table B2

Matrix data which generated the traversal map per W3C standard property (Fig. 3)

2.1 skos:exactMatch traversal map

	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	0	0	0	0	0	0	0	0	0	0	0	0
LoC	0	3	0	13	0	0	0	0	0	0	0	16
VIAF	0	59	0	0	0	0	0	0	0	0	0	59
Getty	0	0	0	12	0	0	0	0	0	0	0	12
Wikidata	0	0	0	0	0	0	0	0	0	0	1	1
DBpedia	0	0	0	0	0	0	108	0	0	0	109	217
BabelNet	0	0	0	0	0	0	0	0	0	0	0	0
GeoNames	0	0	0	0	0	0	20	0	0	0	0	20
Wikipedia	0	0	0	0	0	0	0	0	0	0	0	0
YAGO	0	0	0	0	0	0	82	0	0	0	1	83
Europeana	0	0	0	0	0	0	0	0	0	0	1	1
SUM	0	62	0	25	0	0	210	0	0	0	112	409

2.2 rdfs:seeAlso traversal map

	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	97	0	0	0	0	0	0	0	0	0	0	97
LoC	0	1	0	0	0	0	0	0	0	0	0	1
VIAF	0	0	0	0	0	0	0	0	0	0	0	0
Getty	0	0	0	44	0	0	0	0	0	0	0	44
Wikidata	1	0	0	0	0	0	0	0	0	0	0	1
DBpedia	0	0	0	0	0	4100	0	23	0	261	0	4384
BabelNet	0	0	0	0	0	0	0	0	0	0	0	0
GeoNames	0	0	0	0	0	0	0	0	0	0	0	0
Wikipedia	1	0	0	0	0	0	0	0	0	10	0	11
YAGO	0	0	0	0	0	0	0	0	0	0	0	0
Europeana	0	0	0	0	0	0	0	0	0	0	0	0
SUM	99	1	0	44	0	4100	0	23	0	271	0	4538

2.3 owl:sameAs traversal map

	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	0	0	0	0	0	0	0	0	0	0	0	0
LoC	0	39	0	0	0	0	0	0	0	0	0	39
VIAF	0	0	0	0	0	16	0	0	0	16	0	32
Getty	0	0	0	0	0	0	0	0	0	0	0	0
Wikidata	0	0	0	0	192	100	0	0	0	98	7	397
DBpedia	0	0	0	0	0	1469	0	0	0	1132	761	3362
BabelNet	0	0	0	0	0	0	0	0	0	0	0	0
GeoNames	0	0	0	0	0	22	0	0	0	20	17	59
Wikipedia	0	0	0	0	0	0	0	0	0	0	0	0
YAGO	0	0	0	0	0	95	0	0	0	87	7	189
Europeana	0	0	0	0	0	0	0	0	0	0	0	0
SUM	0	39	0	0	192	1702	0	0	0	1353	792	4078

2.4 schema:sameAs traversal map

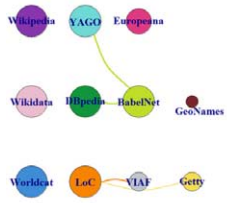
	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	1	0	20	0	0	0	0	0	0	0	0	21
LoC	93	0	52	0	0	0	0	0	0	0	0	145
VIAF	58	0	0	0	0	0	0	0	0	0	0	58
Getty	0	0	18	0	0	0	0	0	0	0	0	18
Wikidata	0	0	43	0	0	0	0	0	0	0	0	43
DBpedia	0	0	38	0	0	0	0	0	0	0	0	38
BabelNet	0	0	0	0	0	0	0	0	0	0	0	0
GeoNames	8	0	0	0	0	0	0	0	0	0	0	8
Wikipedia	0	0	0	0	0	0	0	0	0	1084	0	1084
YAGO	0	0	0	0	0	0	0	0	0	1	0	1
Europeana	0	0	0	0	0	0	0	0	0	0	0	0
SUM	160	0	171	0	0	0	0	0	0	1085	0	1416

Table B3

Matrix data which generated the traversal map for agents (Fig. 4)

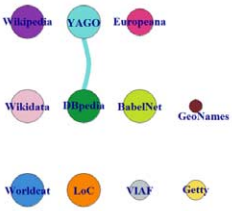
	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	20	0	0	0	0	0	0	0	0	0	0	20
LoC	20	0	20	12	0	0	0	0	0	0	0	52
VIAF	20	20	0	0	0	16	0	0	0	16	0	72
Getty	0	0	15	28	0	0	0	0	0	0	0	43
Wikidata	0	0	20	0	55	20	0	0	0	19	7	121
DBpedia	0	0	19	0	0	408	24	0	0	343	761	1555
BabelNet	0	0	0	0	0	0	0	0	0	0	0	0
GeoNames	0	0	0	0	0	0	0	0	0	0	0	0
Wikipedia	0	0	0	0	0	0	0	0	0	0	0	0
YAGO	0	0	0	0	0	26	23	0	0	26	7	82
Europeana	0	0	0	0	0	0	0	0	0	0	0	0
SUM	60	20	74	40	55	470	47	0	0	404	775	1945

3.1 skos:exactMatch traversal map for agents



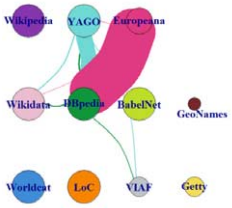
	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	0	0	0	0	0	0	0	0	0	0	0	0
LoC	0	0	0	12	0	0	0	0	0	0	0	12
VIAF	0	20	0	0	0	0	0	0	0	0	0	20
Getty	0	0	0	0	0	0	0	0	0	0	0	12
Wikidata	0	0	0	0	0	0	0	0	0	0	0	0
DBpedia	0	0	0	0	0	0	24	0	0	0	0	24
BabelNet	0	0	0	0	0	0	0	0	0	0	0	0
GeoNames	0	0	0	0	0	0	0	0	0	0	0	0
Wikipedia	0	0	0	0	0	0	0	0	0	0	0	0
YAGO	0	0	0	0	0	0	23	0	0	0	0	23
Europeana	0	0	0	0	0	0	0	0	0	0	0	0
SUM	0	20	0	24	0	0	47	0	0	0	0	91

3.2 rdfs:seeAlso traversal map for agents



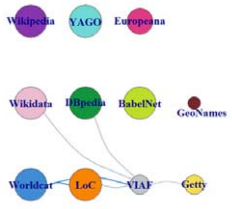
	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	20	0	0	0	0	0	0	0	0	0	0	20
LoC	0	0	0	0	0	0	0	0	0	0	0	0
VIAF	0	0	0	0	0	0	0	0	0	0	0	0
Getty	0	0	0	16	0	0	0	0	0	0	0	16
Wikidata	0	0	0	0	0	0	0	0	0	0	0	0
DBpedia	0	0	0	0	0	104	0	0	0	77	0	181
BabelNet	0	0	0	0	0	0	0	0	0	0	0	0
GeoNames	0	0	0	0	0	0	0	0	0	0	0	0
Wikipedia	0	0	0	0	0	0	0	0	0	0	0	0
YAGO	0	0	0	0	0	0	0	0	0	0	0	0
Europeana	0	0	0	0	0	0	0	0	0	0	0	0
SUM	20	0	0	16	0	104	0	0	0	77	0	217

3.3 owl:sameAs traversal map for agents



	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	0	0	0	0	0	0	0	0	0	0	0	0
LoC	0	0	0	0	0	0	0	0	0	0	0	0
VIAF	0	0	0	0	0	16	0	0	0	16	0	32
Getty	0	0	0	0	0	0	0	0	0	0	0	0
Wikidata	0	0	0	0	55	20	0	0	0	19	7	101
DBpedia	0	0	0	0	0	304	0	0	0	266	761	1327
BabelNet	0	0	0	0	0	0	0	0	0	0	0	0
GeoNames	0	0	0	0	0	0	0	0	0	0	0	0
Wikipedia	0	0	0	0	0	0	0	0	0	0	0	0
YAGO	0	0	0	0	0	26	0	0	0	26	7	59
Europeana	0	0	0	0	0	0	0	0	0	0	0	0
SUM	0	0	0	0	55	362	0	0	0	327	775	1519

3.4 schema:sameAs traversal map for agents



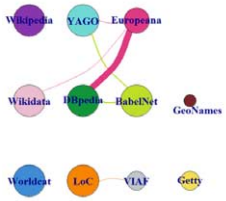
	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	0	0	0	0	0	0	0	0	0	0	0	0
LoC	20	0	20	0	0	0	0	0	0	0	0	40
VIAF	20	0	0	0	0	0	0	0	0	0	0	20
Getty	0	0	15	0	0	0	0	0	0	0	0	15
Wikidata	0	0	20	0	0	0	0	0	0	0	0	20
DBpedia	0	0	19	0	0	0	0	0	0	0	0	19
BabelNet	0	0	0	0	0	0	0	0	0	0	0	0
GeoNames	0	0	0	0	0	0	0	0	0	0	0	0
Wikipedia	0	0	0	0	0	0	0	0	0	0	0	0
YAGO	0	0	0	0	0	0	0	0	0	0	0	0
Europeana	0	0	0	0	0	0	0	0	0	0	0	0
SUM	40	0	74	0	0	0	0	0	0	0	0	114

Table B4

Matrix data which generated the traversal map for events (Fig. 6)

	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	19	0	0	0	0	0	0	0	0	0	0	19
LoC	18	12	0	0	0	0	0	0	0	0	0	30
VIAF	7	7	0	0	0	0	0	0	0	0	0	14
Getty	0	0	0	0	0	0	0	0	0	0	0	0
Wikidata	0	0	0	0	36	20	0	0	0	20	1	77
DBpedia	0	0	0	0	0	563	23	0	0	361	109	1056
BabelNet	0	0	0	0	0	0	0	0	0	0	0	0
GeoNames	0	0	0	0	0	0	0	0	0	0	0	0
Wikipedia	0	0	0	0	0	0	0	0	0	1	0	1
YAGO	0	0	0	0	0	20	21	0	0	21	1	63
Europeana	0	0	0	0	0	0	0	0	0	0	0	0
SUM	44	19	0	0	36	603	44	0	0	403	111	1260

4.1 skos:exactMatch traversal map for events

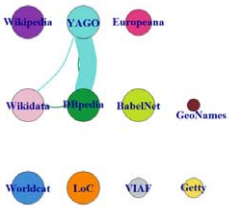


	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	0	0	0	0	0	0	0	0	0	0	0	0
LoC	0	0	0	0	0	0	0	0	0	0	0	0
VIAF	0	7	0	0	0	0	0	0	0	0	0	7
Getty	0	0	0	0	0	0	0	0	0	0	0	0
Wikidata	0	0	0	0	0	0	0	0	0	0	1	1
DBpedia	0	0	0	0	0	0	23	0	0	0	109	132
BabelNet	0	0	0	0	0	0	0	0	0	0	0	0
GeoNames	0	0	0	0	0	0	0	0	0	0	0	0
Wikipedia	0	0	0	0	0	0	0	0	0	0	0	0
YAGO	0	0	0	0	0	0	21	0	0	0	1	22
Europeana	0	0	0	0	0	0	0	0	0	0	1	1
SUM	0	7	0	0	0	0	44	0	0	0	112	163

4.2 rdfs:seeAlso traversal map for events

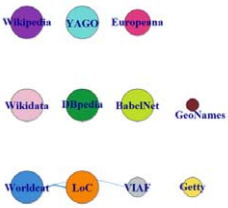
	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	19	0	0	0	0	0	0	0	0	0	0	19
LoC	0	0	0	0	0	0	0	0	0	0	0	0
VIAF	0	0	0	0	0	0	0	0	0	0	0	0
Getty	0	0	0	0	0	0	0	0	0	0	0	0
Wikidata	0	0	0	0	0	0	0	0	0	0	0	0
DBpedia	0	0	0	0	0	263	0	0	0	82	0	345
BabelNet	0	0	0	0	0	0	0	0	0	0	0	0
GeoNames	0	0	0	0	0	0	0	0	0	0	0	0
Wikipedia	0	0	0	0	0	0	0	0	0	1	0	1
YAGO	0	0	0	0	0	0	0	0	0	0	0	0
Europeana	0	0	0	0	0	0	0	0	0	0	0	0
SUM	19	0	0	0	0	263	0	0	0	83	0	365

4.3 owl:sameAs traversal map for events



	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	0	0	0	0	0	0	0	0	0	0	0	0
LoC	0	12	0	0	0	0	0	0	0	0	0	12
VIAF	0	0	0	0	0	0	0	0	0	0	0	0
Getty	0	0	0	0	0	0	0	0	0	0	0	0
Wikidata	0	0	0	0	36	20	0	0	0	20	0	76
DBpedia	0	0	0	0	0	298	0	0	0	279	0	577
BabelNet	0	0	0	0	0	0	0	0	0	0	0	0
GeoNames	0	0	0	0	0	0	0	0	0	0	0	0
Wikipedia	0	0	0	0	0	0	0	0	0	0	0	0
YAGO	0	0	0	0	0	20	0	0	0	21	0	41
Europeana	0	0	0	0	0	0	0	0	0	0	0	0
SUM	0	12	0	0	36	338	0	0	0	320	0	706

4.4 schema:sameAs traversal map for events



	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	0	0	0	0	0	0	0	0	0	0	0	0
LoC	18	0	0	0	0	0	0	0	0	0	0	18
VIAF	7	0	0	0	0	0	0	0	0	0	0	7
Getty	0	0	0	0	0	0	0	0	0	0	0	0
Wikidata	0	0	0	0	0	0	0	0	0	0	0	0
DBpedia	0	0	0	0	0	0	0	0	0	0	0	0
BabelNet	0	0	0	0	0	0	0	0	0	0	0	0
GeoNames	0	0	0	0	0	0	0	0	0	0	0	0
Wikipedia	0	0	0	0	0	0	0	0	0	0	0	0
YAGO	0	0	0	0	0	0	0	0	0	0	0	0
Europeana	0	0	0	0	0	0	0	0	0	0	0	0
SUM	25	0	0	0	0	0	0	0	0	0	0	25

Table B5

Matrix data which generated the traversal map for dates (Fig. 8)

	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	19	0	0	0	0	0	0	0	0	0	0	19
LoC	17	18	0	0	0	0	0	0	0	0	0	35
VIAF	0	0	0	0	0	0	0	0	0	0	0	0
Getty	0	0	0	0	0	0	0	0	0	0	0	0
Wikidata	0	0	0	0	45	20	0	0	0	19	0	99
DBpedia	0	0	0	0	0	462	20	0	0	228	0	710
BabelNet	0	0	0	0	0	0	0	0	0	0	0	0
GeoNames	0	0	0	0	0	0	0	0	0	0	0	0
Wikipedia	0	0	0	0	0	0	0	0	0	0	0	0
YAGO	0	0	0	0	0	0	0	0	0	0	0	0
Europeana	0	0	0	0	0	0	0	0	0	0	0	0
SUM	36	18	0	0	60	482	20	0	0	247	0	863

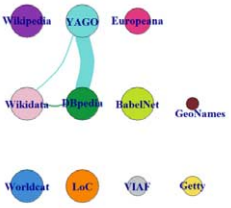
5.1 skos:exactMatch traversal map for dates

	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	0	0	0	0	0	0	0	0	0	0	0	0
LoC	0	0	0	0	0	0	0	0	0	0	0	0
VIAF	0	0	0	0	0	0	0	0	0	0	0	0
Getty	0	0	0	0	0	0	0	0	0	0	0	0
Wikidata	0	0	0	0	0	0	0	0	0	0	0	0
DBpedia	0	0	0	0	0	0	20	0	0	0	0	20
BabelNet	0	0	0	0	0	0	0	0	0	0	0	0
GeoNames	0	0	0	0	0	0	0	0	0	0	0	0
Wikipedia	0	0	0	0	0	0	0	0	0	0	0	0
YAGO	0	0	0	0	0	0	0	0	0	0	0	0
Europeana	0	0	0	0	0	0	0	0	0	0	0	0
SUM	0	0	0	0	0	0	20	0	0	0	0	20

5.2 rdfs:seeAlso traversal map for dates

	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	18	0	0	0	0	0	0	0	0	0	0	18
LoC	0	0	0	0	0	0	0	0	0	0	0	0
VIAF	0	0	0	0	0	0	0	0	0	0	0	0
Getty	0	0	0	0	0	0	0	0	0	0	0	0
Wikidata	0	0	0	0	0	0	0	0	0	0	0	0
DBpedia	0	0	0	0	0	162	0	0	0	0	0	162
BabelNet	0	0	0	0	0	0	0	0	0	0	0	0
GeoNames	0	0	0	0	0	0	0	0	0	0	0	0
Wikipedia	0	0	0	0	0	0	0	0	0	0	0	0
YAGO	0	0	0	0	0	0	0	0	0	0	0	0
Europeana	0	0	0	0	0	0	0	0	0	0	0	0
SUM	18	0	0	0	0	162	0	0	0	0	0	180

5.3 owl:sameAs traversal map for dates



	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	0	0	0	0	0	0	0	0	0	0	0	0
LoC	0	18	0	0	0	0	0	0	0	0	0	18
VIAF	0	0	0	0	0	0	0	0	0	0	0	0
Getty	0	0	0	0	0	0	0	0	0	0	0	0
Wikidata	0	0	0	0	60	20	0	0	0	19	0	99
DBpedia	0	0	0	0	0	300	0	0	0	228	0	528
BabelNet	0	0	0	0	0	0	0	0	0	0	0	0
GeoNames	0	0	0	0	0	0	0	0	0	0	0	0
Wikipedia	0	0	0	0	0	0	0	0	0	0	0	0
YAGO	0	0	0	0	0	0	0	0	0	0	0	0
Europeana	0	0	0	0	0	0	0	0	0	0	0	0
SUM	0	18	0	0	60	320	0	0	0	247	0	645

5.4 schema:sameAs traversal map for dates

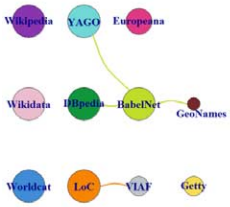
	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	1	0	0	0	0	0	0	0	0	0	0	1
LoC	17	0	0	0	0	0	0	0	0	0	0	17
VIAF	0	0	0	0	0	0	0	0	0	0	0	0
Getty	0	0	0	0	0	0	0	0	0	0	0	0
Wikidata	0	0	0	0	0	0	0	0	0	0	0	0
DBpedia	0	0	0	0	0	0	0	0	0	0	0	0
BabelNet	0	0	0	0	0	0	0	0	0	0	0	0
GeoNames	0	0	0	0	0	0	0	0	0	0	0	0
Wikipedia	0	0	0	0	0	0	0	0	0	0	0	0
YAGO	0	0	0	0	0	0	0	0	0	0	0	0
Europeana	0	0	0	0	0	0	0	0	0	0	0	0
SUM	18	0	0	0	0	0	0	0	0	0	0	18

Table B6

Matrix data which generated the traversal map for places (Fig. 10)

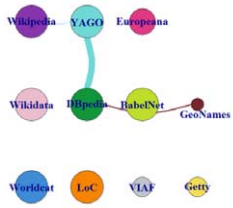
	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	19	0	18	0	0	0	0	0	0	0	0	37
LoC	19		19	0	0	0	0	0	0	0	0	39
VIAF	19	20		0	0	0	0	0	0	0	0	39
Getty	0	0	1	20	0	0	0	0	0	0	0	21
Wikidata	0	0	21	0	41	20	0	0	0	20	0	102
DBpedia	0	0	17	0	0	3854	21	20	0	390	0	4302
BabelNet	0	0	0	0	0	0		0	0	0	0	0
GeoNames	8	0	0	0	0	20	18		0	20	17	83
Wikipedia	0	0	0	0	0	0	0	0		9	0	9
YAGO	0	0	0	0	0	26	20	0	0	35	0	81
Europeana	0	0	0	0	0	0	0	0	0	0		0
SUM	65	21	76	20	41	3920	59	20	0	474	17	4713

6.1 skos:exactMatch traversal map for places



	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	0	0	0	0	0	0	0	0	0	0	0	0
LoC	0	0	0	0	0	0	0	0	0	0	0	0
VIAF	0	20	0	0	0	0	0	0	0	0	0	20
Getty	0	0	0	0	0	0	0	0	0	0	0	0
Wikidata	0	0	0	0	0	0	0	0	0	0	0	0
DBpedia	0	0	0	0	0	0	21	0	0	0	0	21
BabelNet	0	0	0	0	0	0	0	0	0	0	0	0
GeoNames	0	0	0	0	0	0	18	0	0	0	0	18
Wikipedia	0	0	0	0	0	0	0	0	0	0	0	0
YAGO	0	0	0	0	0	0	20	0	0	0	0	20
Europeana	0	0	0	0	0	0	0	0	0	0	0	0
SUM	0	20	0	0	0	0	59	0	0	0	0	79

6.2 rdfs:seeAlso traversal map for places

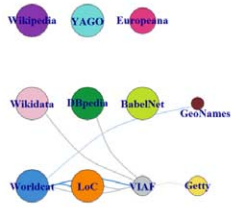


	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	19	0	0	0	0	0	0	0	0	0	0	19
LoC	0	1	0	0	0	0	0	0	0	0	0	1
VIAF	0	0	0	0	0	0	0	0	0	0	0	0
Getty	0	0	0	20	0	0	0	0	0	0	0	20
Wikidata	0	0	0	0	0	0	0	0	0	0	0	0
DBpedia	0	0	0	0	0	3537	0	20	0	101	0	3658
BabelNet	0	0	0	0	0	0	0	0	0	0	0	0
GeoNames	0	0	0	0	0	0	0	0	0	0	0	0
Wikipedia	0	0	0	0	0	0	0	0	0	7	0	7
YAGO	0	0	0	0	0	0	0	0	0	0	0	0
Europeana	0	0	0	0	0	0	0	0	0	0	0	0
SUM	19	1	0	20	0	3537	0	20	0	108	0	3705

6.3 owl:sameAs traversal map for places

	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	0	0	0	0	0	0	0	0	0	0	0	0
LoC	0	0	0	0	0	0	0	0	0	0	0	0
VIAF	0	0	0	0	0	0	0	0	0	0	0	0
Getty	0	0	0	0	0	0	0	0	0	0	0	0
Wikidata	0	0	0	0	41	20	0	0	0	20	0	81
DBpedia	0	0	0	0	0	300	0	0	0	287	0	587
BabelNet	0	0	0	0	0	0	0	0	0	0	0	0
GeoNames	0	0	0	0	0	20	0	0	0	20	16	56
Wikipedia	0	0	0	0	0	0	0	0	0	0	0	0
YAGO	0	0	0	0	0	26	0	0	0	33	0	59
Europeana	0	0	0	0	0	0	0	0	0	0	0	0
SUM	0	0	0	0	41	366	0	0	0	360	16	783

6.4 schema:sameAs traversal map for places



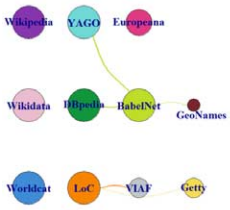
	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	0	0	18	0	0	0	0	0	0	0	0	18
LoC	19	0	19	0	0	0	0	0	0	0	0	38
VIAF	19	0	0	0	0	0	0	0	0	0	0	19
Getty	0	0	1	0	0	0	0	0	0	0	0	1
Wikidata	0	0	21	0	0	0	0	0	0	0	0	21
DBpedia	0	0	17	0	0	0	0	0	0	0	0	17
BabelNet	0	0	0	0	0	0	0	0	0	0	0	0
GeoNames	8	0	0	0	0	0	0	0	0	0	0	8
Wikipedia	0	0	0	0	0	0	0	0	0	0	0	0
YAGO	0	0	0	0	0	0	0	0	0	0	0	0
Europeana	0	0	0	0	0	0	0	0	0	0	0	0
SUM	46	0	76	0	0	0	0	0	0	0	0	122

Table B7

Matrix data which generated the traversal map for objects and concepts (Fig. 12)

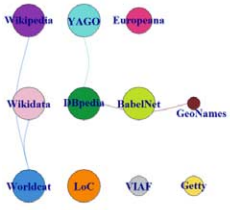
	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	1	0	2	0	0	0	0	0	0	0	0	21
LoC	19	12	13	1	0	0	0	0	0	0	0	45
VIAF	12	12	0	0	0	0	0	0	0	0	0	24
Getty	0	0	2	3	0	0	0	0	0	0	0	10
Wikidata	1	0	2	0	1	20	0	0	0	20	0	43
DBpedia	0	0	2	0	0	31	20	3	0	75	0	412
BabelNet	0	0	0	0	0	0	1	0	0	0	0	0
GeoNames	0	0	0	0	0	2	2	1	0	0	0	4
Wikipedia	1	0	0	0	0	0	0	0	0	1084	0	1085
YAGO	0	0	0	0	0	23	18	0	0	5	0	47
Europeana	0	0	0	0	0	0	0	0	0	0	0	0
SUM	52	24	21	9	0	357	40	3	0	1185	0	1691

7.1 skos:exactMatch traversal map for objects and concepts



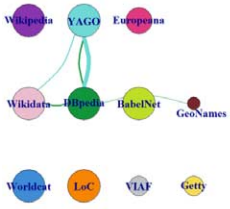
	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	0	0	0	0	0	0	0	0	0	0	0	0
LoC	0	3	0	1	0	0	0	0	0	0	0	4
VIAF	0	12	0	0	0	0	0	0	0	0	0	12
Getty	0	0	0	0	0	0	0	0	0	0	0	0
Wikidata	0	0	0	0	0	0	0	0	0	0	0	0
DBpedia	0	0	0	0	0	0	20	0	0	0	0	20
BabelNet	0	0	0	0	0	0	0	0	0	0	0	0
GeoNames	0	0	0	0	0	0	2	0	0	0	0	2
Wikipedia	0	0	0	0	0	0	0	0	0	0	0	0
YAGO	0	0	0	0	0	0	18	0	0	0	0	18
Europeana	0	0	0	0	0	0	0	0	0	0	0	0
SUM	0	15	0	1	0	0	40	0	0	0	0	56

7.2 rdfs:seeAlso traversal map for objects and concepts



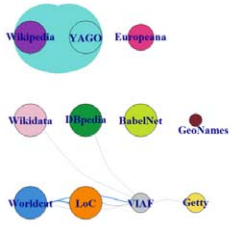
	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	19	0	0	0	0	0	0	0	0	0	0	19
LoC	0	0	0	0	0	0	0	0	0	0	0	0
VIAF	0	0	0	0	0	0	0	0	0	0	0	0
Getty	0	0	0	8	0	0	0	0	0	0	0	8
Wikidata	1	0	0	0	0	0	0	0	0	0	0	1
DBpedia	0	0	0	0	0	34	0	3	0	1	0	38
BabelNet	0	0	0	0	0	0	0	0	0	0	0	0
GeoNames	0	0	0	0	0	0	0	0	0	0	0	0
Wikipedia	1	0	0	0	0	0	0	0	0	0	0	1
YAGO	0	0	0	0	0	0	0	0	0	0	0	0
Europeana	0	0	0	0	0	0	0	0	0	0	0	0
SUM	21	0	0	8	0	34	0	3	0	1	0	67

7.3 owl:sameAs traversal map for objects and concepts



	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	0	0	0	0	0	0	0	0	0	0	0	0
LoC	0	9	0	0	0	0	0	0	0	0	0	9
VIAF	0	0	0	0	0	0	0	0	0	0	0	0
Getty	0	0	0	0	0	0	0	0	0	0	0	0
Wikidata	0	0	0	0	0	20	0	0	0	20	0	40
DBpedia	0	0	0	0	0	271	0	0	0	70	0	341
BabelNet	0	0	0	0	0	0	0	0	0	0	0	0
GeoNames	0	0	0	0	0	2	0	0	0	0	0	2
Wikipedia	0	0	0	0	0	0	0	0	0	0	0	0
YAGO	0	0	0	0	0	23	0	0	0	5	0	28
Europeana	0	0	0	0	0	0	0	0	0	0	0	0
SUM	0	9	0	0	0	316	0	0	0	95	0	420

7.4 schema:sameAs traversal map for objects and concepts



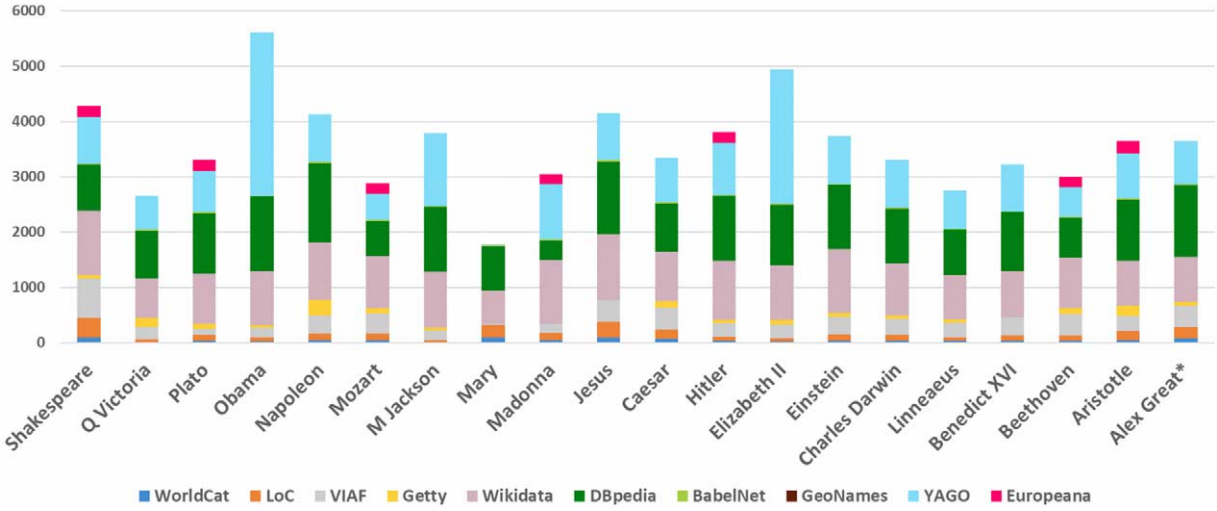
	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	0	0	2	0	0	0	0	0	0	0	0	2
LoC	19	0	13	0	0	0	0	0	0	0	0	32
VIAF	12	0	0	0	0	0	0	0	0	0	0	12
Getty	0	0	2	0	0	0	0	0	0	0	0	2
Wikidata	0	0	2	0	0	0	0	0	0	0	0	2
DBpedia	0	0	2	0	0	0	0	0	0	0	0	2
BabelNet	0	0	0	0	0	0	0	0	0	0	0	0
GeoNames	0	0	0	0	0	0	0	0	0	0	0	0
Wikipedia	0	0	0	0	0	0	0	0	0	1084	0	1084
YAGO	0	0	0	0	0	0	0	0	0	1	0	1
Europeana	0	0	0	0	0	0	0	0	0	0	0	0
SUM	31	0	21	0	0	0	0	0	0	1085	0	1137

Appendix C. Python analysis details

	WorldCat	Full Coverage	Library of Congress	VIAF	Getty ULAN	Wikidata	Dbpedia	BabelNet	GeoNames	YAGO	Europeana
0	29136	29136				29136					
1	Daerwen, 1809-1882	Daerwen, 1809-1882	Daerwen, 1809-1882								
2	Daerwen, 1809-1882	Daerwen, 1809-1882	Daerwen, 1809-1882								
3	Darvin, Charlʹz, 1809-1882	Darvin, Charlʹz, 1809-1882	Darvin, Charlʹz, 1809-1882								
4	Darvin, Charlʹz, 1809-1882	Darvin, Charlʹz, 1809-1882	Darvin, Charlʹz, 1809-1882								
5	Darvin, Tsharlz, 1809-1882	Darvin, Tsharlz, 1809-1882	Darvin, Tsharlz, 1809-1882								
6	Darvin, Tsharlz, 1809-1882	Darvin, Tsharlz, 1809-1882	Darvin, Tsharlz, 1809-1882								
7	Darvin, Čarls, 1809-1882	Darvin, Čarls, 1809-1882	Darvin, Čarls, 1809-1882								
8	Darvin, Čarls, 1809-1882	Darvin, Čarls, 1809-1882	Darvin, Čarls, 1809-1882								
9	Darwin, Carlos R., 1809-1882	Darwin, Carlos R., 1809-1882	Darwin, Carlos R., 1809-1882								
10	Darwin, Carlos R., 1809-1882	Darwin, Carlos R., 1809-1882	Darwin, Carlos R., 1809-1882								
11	Darwin, Charles Robert, 1809-1882	Darwin, Charles Robert, 1809-1882	Darwin, Charles Robert, 1809-1882								
12	Darwin, Charles Robert, 1809-1882	Darwin, Charles Robert, 1809-1882	Darwin, Charles Robert, 1809-1882								
13	Darwin, Charles, 1809-1882	Darwin, Charles, 1809-1882	Darwin, Charles, 1809-1882			Darwin, Charles, 1809-1882					
14	Darwin, Charles, 1809-1882	Darwin, Charles, 1809-1882	Darwin, Charles, 1809-1882			Darwin, Charles, 1809-1882					
15	Darwin, Karol, 1809-1882	Darwin, Karol, 1809-1882	Darwin, Karol, 1809-1882								
16	Darwin, Karol, 1809-1882	Darwin, Karol, 1809-1882	Darwin, Karol, 1809-1882								
17	Dăwin, 1809-1882	Dăwin, 1809-1882	Dăwin, 1809-1882								
18	Dăwin, 1809-1882	Dăwin, 1809-1882	Dăwin, 1809-1882								
19	Sdar-win, 1809-1882	Sdar-win, 1809-1882	Sdar-win, 1809-1882								
20	Sdar-win, 1809-1882	Sdar-win, 1809-1882	Sdar-win, 1809-1882								
21	Sdar-win, Char-le-si Ro-sbe-thi, 18(	Sdar-win, Char-le-si Ro-sbe-thi, 18(	Sdar-win, Char-le-si Ro-sbe-thi, 1809-1882								
22	Sdar-win, Char-le-si Ro-sbe-thi, 18(	Sdar-win, Char-le-si Ro-sbe-thi, 18(	Sdar-win, Char-le-si Ro-sbe-thi, 1809-1882								
23	Tărvin, 1809-1882	Tărvin, 1809-1882	Tărvin, 1809-1882								
24	Tărvin, 1809-1882	Tărvin, 1809-1882	Tărvin, 1809-1882								
25	Tărvin, Cărlas, 1809-1882	Tărvin, Cărlas, 1809-1882	Tărvin, Cărlas, 1809-1882								
26	Tărvin, Cărlas, 1809-1882	Tărvin, Cărlas, 1809-1882	Tărvin, Cărlas, 1809-1882								
27	http://id.loc.gov/authorities/name	http://id.loc.gov/authorities/name	http://id.loc.gov/authorities/name	http://id.loc.go	http://id.loc.gov/authorities/names/n78095637						
28	https://viaf.org/viaf/27063124	https://viaf.org/viaf/27063124									
29	schema:Person	schema:Person		schema:Person			schema:Person				
30	http://en.wikipedia.org/wiki/Charl	http://en.wikipedia.org/wiki/Charles_Darwin					http://en.wikipedia.org/wiki/Ch			http://en.wikipedia.or	
31	http://id.worldcat.org/fast/ontolo	http://id.worldcat.org/fast/ontology/1.0/#facet-Personal									
32	http://id.worldcat.org/fast/ontolo	http://id.worldcat.org/fast/ontology/1.0/#fast									
33	http://id.worldcat.org/fast/29136	http://id.worldcat.org/fast/29136	http://id.worldcat.org/fast/29136			http://id.worldcat.org/fast/29136					
34		madsrdf:Authority	madsrdf:Authority								
35		madsrdf:PersonalName	madsrdf:PersonalName								
36		skos:Concept	skos:Concept		skos:Concept			skos:Concept			

Fig. C.1. Python scripts generate EXCEL files to show the content overlaps across 11 dataset. Content in the same row is overlap (example of Charles Darwin).

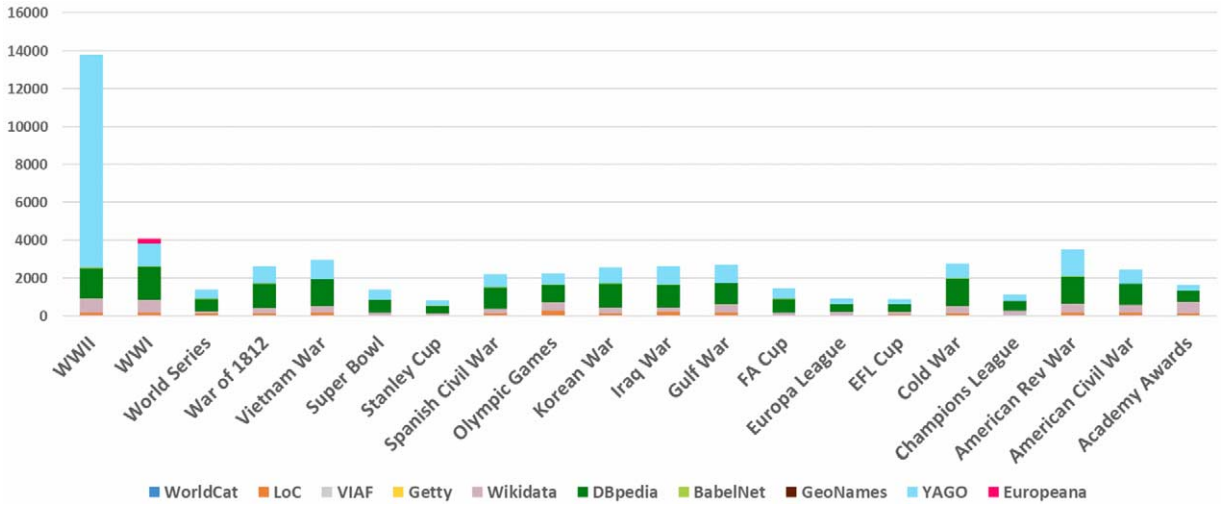
Figure C.2/Table C1
The number of content in agents entities per data source.[1]



	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	YAGO	Europeana	Full Coverage
Shakespeare	92	353	718	57	1167	830	19	0	846	199	3532
Q Victoria	14	40	230	165	718	858	18	0	608	0	2026
Plato	30	104	113	96	901	1097	17	0	742	208	2546
Obama	24	67	178	44	984	1350	10	0	2954	0	4764
Napoleon	39	118	340	281	1025	1444	22	0	852	0	3263
Mozart	40	117	366	101	939	639	18	0	466	198	2268
M Jackson	14	29	173	46	1022	1170	15	0	1317	0	2965
Mary	89	229	6	0	609	808	20	0	8	0	1643
Madonna	40	133	161	0	1152	363	19	0	1002	169	2615
Jesus	97	274	386	0	1209	1307	26	0	853	0	3152
Caesar	64	163	404	122	893	870	20	0	801	0	2570
Hitler	29	74	245	64	1069	1172	17	0	938	195	2872
Elizabeth II	25	60	240	84	983	1105	19	0	2418	0	4078
Einstein	36	109	321	72	1154	1161	14	0	867	0	2965
Charles Darwin	34	100	297	65	936	994	13	0	867	0	2579
Linneaeus	28	63	269	56	802	828	15	0	690	0	2084
Benedict XVI	37	84	339	0	827	1077	10	0	850	0	2331
Beethoven	32	99	387	99	921	723	16	0	535	188	2423
Aristotle	48	164	263	191	811	1113	21	0	814	218	2791
Alex Great*	74	207	390	62	822	1287	30	0	773	0	2732
SUM	886	2587	5826	1605	18944	20196	359	0	19201	1375	56199

[1] Spikes for Obama and Elizabeth II are mostly due to YAGO's high input.

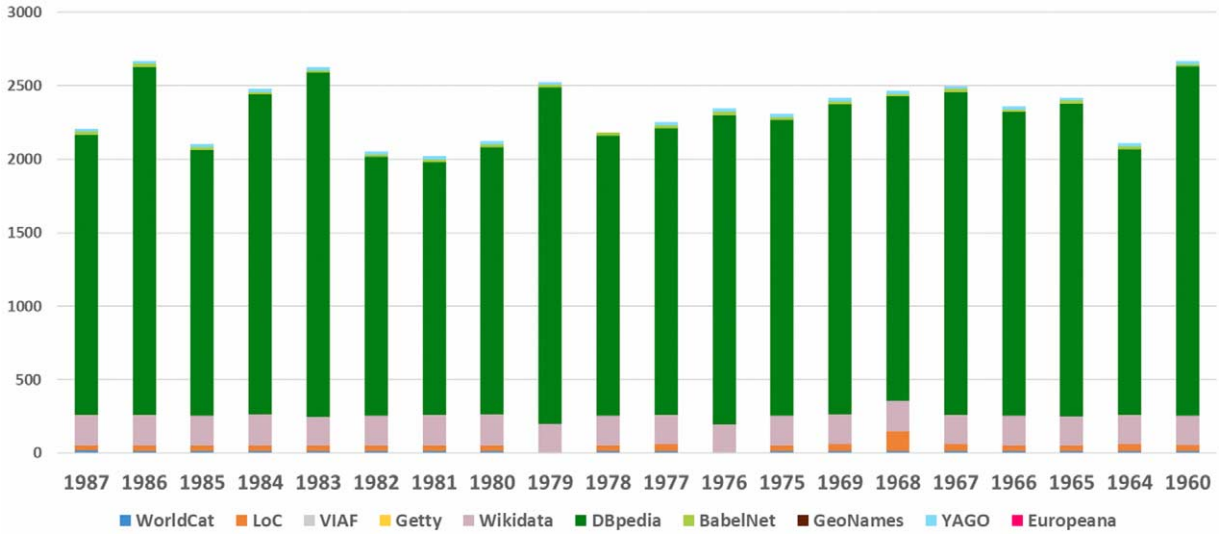
Figure C.3/Table C2
The number of content in events entities per data source.[1]



	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	YAGO	Europeana	Full Coverage
WWII	26	151	0	0	749	1601	15	0	11273	0	12811
WWI	26	155	0	0	677	1740	21	0	1189	256	2939
World Series	39	111	0	0	89	662	16	0	471	0	924
War of 1812	12	128	0	0	266	1308	18	0	897	0	2156
Vietnam War	19	147	0	0	350	1426	15	0	991	0	2288
Super Bowl	11	28	0	0	143	663	18	0	539	0	1016
Stanley Cup	11	30	0	0	109	378	17	0	278	0	571
Spanish Civil War	10	124	0	0	262	1113	13	0	679	0	1656
Olympic Games	10	261	0	0	436	941	28	0	566	0	1749
Korean War	11	125	0	0	326	1245	14	0	817	0	1922
Iraq War	33	176	0	0	255	1186	9	0	965	0	1946
Gulf War	26	143	0	0	447	1115	12	0	949	0	2013
FA Cup	17	41	0	0	130	715	16	0	551	0	1036
Europa League	0	0	0	0	202	421	12	0	290	0	723
EFL Cup	21	45	0	0	148	395	14	0	253	0	665
Cold War	11	132	0	0	364	1484	16	0	740	0	2235
Champions League	14	33	0	0	233	498	18	0	339	0	908
American Rev War	16	146	0	0	498	1421	29	0	1401	0	3140
American Civil War	17	147	0	0	431	1095	27	0	729	0	1953
Academy Awards	16	130	0	0	593	582	17	0	310	0	1393
SUM	**346**	**2253**	**0**	**0**	**6708**	**19989**	**345**	**0**	**24227**	**256**	**44044**

[1] YAGO contains a large amount of content for WWII.

Figure C.4/Table C3
The number of content in dates entities per data source.[1]

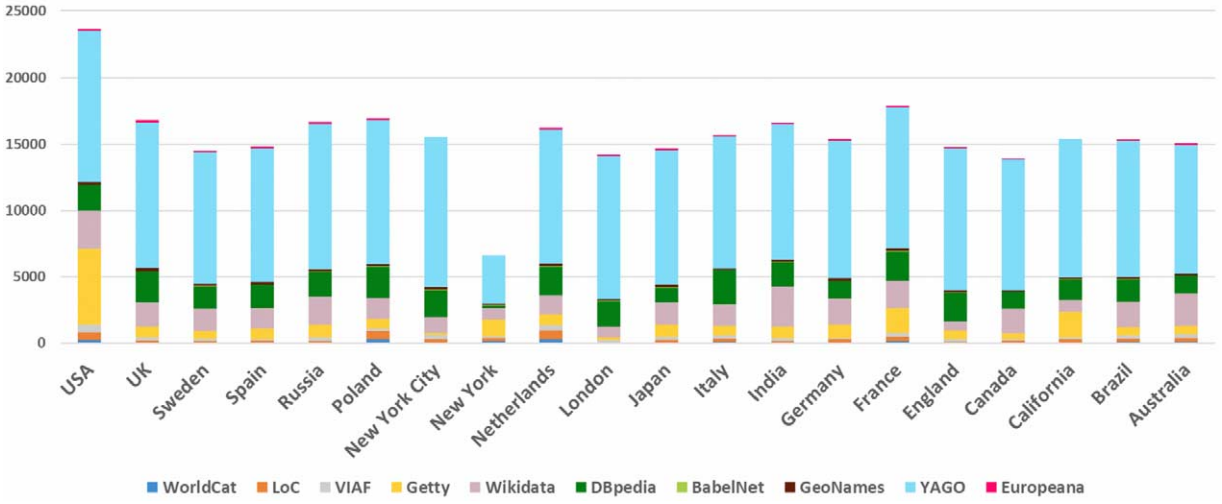


	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	YAGO	Europeana	Full Coverage
1987	18	35	0	0	207	1909	20	0	22	0	2190
1986	16	35	0	0	207	2368	22	0	22	0	2650
1985	16	34	0	0	205	1810	20	0	22	0	2087
1984	15	37	0	0	213	2175	15	0	22	0	2458
1983	15	34	0	0	198	2341	15	0	22	0	2606
1982	15	34	0	0	205	1764	14	0	22	0	2035
1981	15	34	0	0	209	1722	21	0	22	0	2004
1980	15	36	0	0	214	1820	17	0	22	0	2105
1979	0	0	0	0	201	2288	15	0	22	0	2519
1978	15	35	0	0	207	1907	22	0	0	0	2169
1977	16	44	0	0	202	1953	19	0	22	0	2236
1976	0	0	0	0	193	2110	21	0	22	0	2339
1975	15	35	0	0	205	2014	18	0	22	0	2289
1969	15	46	0	0	205	2111	18	0	22	0	2398
1968	15	133	0	0	207	2073	15	0	22	0	2446
1967	15	44	0	0	200	2198	19	0	22	0	2479
1966	15	36	0	0	203	2071	14	0	22	0	2341
1965	15	37	0	0	197	2130	18	0	22	0	2399
1964	15	43	0	0	200	1811	20	0	22	0	2092
1960	15	42	0	0	196	2376	15	0	22	0	2647
SUM	**276**	**774**	**0**	**0**	**4074**	**40951**	**358**	**0**	**418**	**0**	**46489**

[1] A highly normalised/standardised content pattern is seen across date entities.

Figure C.5/Table C4

The number of content in places entities per data source.[1]



	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	YAGO	Europeana	Full Coverage
USA	241	580	579	5725	2847	1979	12	176	11329	150	21351
UK	58	134	227	852	1828	2318	33	237	10942	192	15022
Sweden	34	100	219	585	1647	1698	27	159	9925	120	12829
Spain	56	153	93	828	1512	1758	22	174	10113	140	13283
Russia	30	126	265	996	2106	1874	25	139	10979	114	15208
Poland	273	627	228	687	1584	2371	22	177	10832	139	14840
New York City	0	307	309	151	1191	2026	36	233	11289	0	13958
New York	153	230	128	1266	860	203	34	131	3631	0	6278
Netherlands	285	697	363	804	1458	2177	22	204	10081	155	13845
London	21	48	159	188	832	1928	28	101	10836	71	12932
Japan	63	167	240	937	1668	1131	19	193	10125	140	13334
Italy	84	235	251	744	1636	2555	0	134	9954	105	13914
India	37	110	227	876	3035	1838	24	136	10219	110	15229
Germany	74	202	100	1003	2006	1303	27	188	10351	146	13909
France	127	334	329	1870	2030	2241	26	180	10612	143	15502
England	21	57	194	679	689	2172	30	139	10730	100	12831
Canada	50	147	78	441	1878	1298	24	81	9882	57	12963
California	64	208	175	1934	880	1549	35	110	10446	0	14117
Brazil	92	242	245	635	1904	1708	24	155	10256	118	13777
Australia	113	293	293	582	2477	1324	27	153	9683	115	13685
SUM	**1876**	**4997**	**4702**	**21783**	**34068**	**35451**	**497**	**3200**	**202215**	**2115**	**278807**

[1]New York (state) might be low, due to its less popular concept, compared to countries and big cities. Somehow USA stands out, with a lot of unexpected contribution from Getty TGN.

Table C5

The number of content in objects and concepts entities per data source (see Fig. 15 in the main text)

	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	YAGO	Europeana	Full Coverage
Vasa	11	24	46	0	146	386	11	0	107	0	626
Uncle Tom's Cabin	13	25	193	0	226	516	0	0	113	0	1015
Ukiyo-e	24	152	0	118	151	609	19	0	137	0	1096
Tosca	11	53	21	0	162	413	11	0	97	0	718
Toraja	17	56	0	0	87	23	11	0	71	0	215
Tamil Language	15	162	0	49	403	655	14	0	481	0	1392
Sgt. Pepper's	11	24	22	0	246	1007	14	0	666	0	1479
Rosetta Stone	19	46	44	0	249	592	11	0	188	0	985
Like a Rolling Stone	9	20	11	0	71	383	19	0	75	0	547
Palazzo Pitti	16	37	21	30	178	367	22	47	134	0	703
Ming Dynasty	0	120	6	77	369	963	24	0	269	0	1659
Mars	12	164	34	28	668	957	21	0	767	0	2118
King and I*	9	46	48	0	95	482	18	0	375	0	718
Book of Kells	15	47	43	0	147	406	11	0	134	0	669
Influenza	23	159	0	48	418	433	20	0	512	0	1440
Garden of E Delights	17	36	32	0	253	395	18	0	126	0	776
Byzantine Empire	27	63	221	91	626	1310	17	0	393	0	2282
Boeing 747	12	41	0	0	164	560	18	0	143	0	863
Blade Runner	9	42	29	0	490	639	11	0	191	0	1298
Angkor Wat	13	72	62	65	240	569	17	37	253	0	1100
SUM	283	1389	833	506	5389	11665	307	84	5232	0	21699

References

[1] M. Achichi, P. Lisena, K. Todorov, R. Troncy and J. Delahousse, DOREMUS: A graph of linked musical works, in: *The Semantic Web – ISWC 2018*, D. Vrandečić, K. Bontcheva, M.C. Suárez-Figueroa, V. Presutti, I. Celino, M. Sabou, L.-A. Kaffee and E. Simperl, eds, Springer International Publishing, Cham, 2018, pp. 3–19. doi:10.1007/978-3-030-00668-6_1.

[2] E. Agirre, A. Barrena, O.L. de Lacalle, A. Soroa, S. Fernando and M. Stevenson, Matching cultural heritage items to Wikipedia, in: *LREC*, 2012.

[3] D. Ahlers, Linkage quality analysis of GeoNames in the Semantic Web, in: *Proceedings of the 11th Workshop on Geographic Information Retrieval*, ACM, New York, NY, USA, 2017, pp. 10:1–10:2. doi:10.1145/3155902.3155904.

[4] W. Beek, J. Raad, J. Wielemaker and F. van Harmelen, sameAs.cc: The closure of 500M owl: sameAs statements, in: *The Semantic Web*, A. Gangemi, R. Navigli, M.-E. Vidal, P. Hitzler, R. Troncy, L. Hollink, A. Tordai and M. Alam, eds, Springer International Publishing, 2018, pp. 65–80. doi:10.1007/978-3-319-93417-4_5.

[5] T. Berners-Lee, *Linked Data – Design Issues*, 2009, https://www.w3.org/DesignIssues/LinkedData.html (accessed April 24, 2018).

[6] C. Bizer, T. Heath and T. Berners-Lee, Linked data: The story so far, in: *Semantic Services, Interoperability and Web Applications: Emerging Concepts*, IGI Global, 2011, pp. 205–227. doi:10.4018/978-1-60960-593-3.ch008.

[7] G. Candela, P. Escobar, R.C. Carrasco and M. Marco-Such, A linked open data framework to enhance the discoverability and impact of culture heritage, *Journal of Information Science* **45** (2019), 756–766. doi:10.1177/0165551518812658.

[8] G. Candela, P. Escobar, R.C. Carrasco and M. Marco-Such, Evaluating the quality of linked open data in digital libraries, *Journal of Information Science* **48**(1) (2020), 21–43. doi:10.1177/0165551520930951.

[9] G. Correndo, A. Penta, N. Gibbins and N. Shadbolt, Statistical analysis of the owl: SameAs network for aligning concepts in the linking open data cloud, in: *Database and Expert Systems Applications*, S.W. Liddle, K.-D. Schewe, A.M. Tjoa and X. Zhou, eds, Springer, Berlin, Heidelberg, 2012, pp. 215–230. doi:10.1007/978-3-642-32597-7_20.

[10] M. De Wilde, M.D. Wilde and S. Hengchen, Semantic enrichment of a multilingual archive with linked open data, *DHQ* **011** (2018), 4.

[11] J. Debattista, E. Clinton and R. Brennan, Assessing the quality of geospatial linked data – experiences from ordnance survey Ireland (OSi), in: *CEUR Workshop Proceedings (CEUR-WS. Org) Proceedings of the Posters and Demos Track of the 14th International Conference on Semantic Systems Co-Located with the 14th International Conference on Semantic Systems (SEMANTiCS 2018)*, Vienna, Austria, September 10–13, 2018, Vol. 2198, 2018. http://ceur-ws.org/Vol-2198/paper_94.pdf (accessed September 12, 2018).

[12] J. Debattista, C. Lange and S. Auer, Luzzu – A Framework for Linked Data Quality Assessment, 2014, http://arxiv.org/abs/1412.3750 (accessed October 18, 2018).

[13] J. Debattista, C. Lange, S. Auer and D. Cortis, Evaluating the quality of the LOD cloud: An empirical investigation, *Semantic Web* **9** (2018), 859–901. doi:10.3233/SW-180306.

[14] L. Ding, J. Shinavier, T.W. Finin and D.L. McGuinness, owl: sameAs and Linked Data: An Empirical Study, 2010. doi:10.13016/M2XP6V734.

[15] L. Ding, J. Shinavier, Z. Shangguan and D.L. McGuinness, SameAs networks and beyond: Analyzing deployment status and implications of owl: sameAs in linked data, in: *The Semantic Web – ISWC 2010*, P.F. Patel-Schneider, Y. Pan, P. Hitzler, P. Mika, L. Zhang, J.Z. Pan, I. Horrocks and B. Glimm, eds, Springer, Berlin, Heidelberg, 2010, pp. 145–160. doi:10.1007/978-3-642-17746-0_10.

[16] J. Edelstein, L. Galla, C. Li-Madeo, J. Marden, A. Rhonemus and N. Whysel, Linked Open Data for Cultural Heritage: Evolution of an Information Technology, 2013, http://www.whysel.com/papers/LIS670-Linked-Open-Data-for-Cultural-Heritage.pdf (accessed April 24, 2018).

[17] Y.-H. Eom, P. Aragón, D. Laniado, A. Kaltenbrunner, S. Vigna and D.L. Shepelyansky, Interactions of cultures and top people of Wikipedia from ranking of 24 language editions, *PLoS ONE* **10** (2015), e0114825. doi:10.1371/journal.pone.0114825.

[18] F. Erxleben, M. Günther, M. Krötzsch, J. Mendez and D. Vrandečić, Introducing Wikidata to the Linked Data Web, in: *The Semantic Web – ISWC 2014*, P. Mika, T. Tudorache, A. Bernstein, C. Welty, C. Knoblock, D. Vrandečić, P. Groth, N. Noy, K. Janowicz and C. Goble, eds, Springer International Publishing, Cham, 2014, pp. 50–65. doi:10.1007/978-3-319-11964-9_4.

[19] M. Farag, Entity Matching and Disambiguation Across Multiple Knowledge Graphs, Master's Thesis, University of Waterloo, 2019, https://uwspace.uwaterloo.ca/handle/10012/14750 (accessed January 26, 2021).

[20] M. Färber, F. Bartscherer, C. Menne and A. Rettinger, Linked data quality of DBpedia, freebase, OpenCyc, Wikidata, and YAGO, *SW* **9** (2017), 77–129. doi:10.3233/SW-170275.

[21] T. Gottron, A. Scherp, B. Krayer and A. Peters, Get the Google Feeling: Supporting Users in Finding Relevant Sources of Linked Open Data at Web-Scale, 2012, https://pdfs.semanticscholar.org/43a9/670c57fec2a4d05ff72429886e457a88e59f.pdf.

[22] C. Guéret, P. Groth, C. Stadler and J. Lehmann, Assessing linked data mappings using network measures, in: *The Semantic Web: Research and Applications*, E. Simperl, P. Cimiano, A. Polleres, O. Corcho and V. Presutti, eds, Springer, Berlin, Heidelberg, 2012, pp. 87–102. doi:10.1007/978-3-642-30284-8_13.

[23] H. Halpin, P.J. Hayes, J.P. McCusker, D.L. McGuinness and H.S. Thompson, When owl: sameAs isn't the same: An analysis of identity in linked data, in: *The Semantic Web – ISWC 2010*, P.F. Patel-Schneider, Y. Pan, P. Hitzler, P. Mika, L. Zhang, J.Z. Pan, I. Horrocks and B. Glimm, eds, Springer, Berlin, Heidelberg, 2010, pp. 305–320. doi:10.1007/978-3-642-17746-0_20.

[24] O. Hartig, SQUIN: A traversal based query execution system for the web of linked data, in: *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, ACM, 2013, pp. 1081–1084. doi:10.1145/2463676.2465231.

[25] O. Hartig and M.T. Özsu, Walking without a map: Ranking-based traversal for querying linked data, in: *International Semantic Web Conference*, Springer, 2016, pp. 305–324.

[26] A. Idrissou, F. van Harmelen and P. van den Besselaar, Network metrics for assessing the quality of entity resolution between multiple datasets, *Semantic Web* **12** (2021), 21–40. doi:10.3233/SW-200410.

[27] A. Jaffri, H. Glaser and I. Millard, *Managing URI Synonymity to Enable Consistent Reference on the Semantic Web*, 2008, https://eprints.soton.ac.uk/265614/ (accessed April 9, 2019).

[28] A. Jaffri, H. Glaser and I. Millard, URI Disambiguation in the Context of Linked Data, 2008, https://eprints.soton.ac.uk/265181/ (accessed April 9, 2019).

[29] R. Maturana, M. Ortega, S. López-Sola, M.E. Alvarado and M.J. Ibáñez, Mismuseos.net: Art after technology. Putting cultural data to work in a linked data platform, in: *Veni@OKCon*, 2013.

[30] D. Milne and I. Witten, *Learning to Link with Wikipedia, International Conference on Information and Knowledge Management, Proceedings*, 2008. doi:10.1145/1458082.1458150.

[31] M. Mountantonakis and Y. Tzitzikas, High performance methods for linked open data connectivity analytics, *Information* **9** (2018), 134. doi:10.3390/info9060134.

[32] J. Raad, W. Beek, F. van Harmelen, N. Pernelle and F. Saïs, Detecting erroneous identity links on the web using network metrics, in: *The Semantic Web – ISWC 2018*, D. Vrandečić, K. Bontcheva, M.C. Suárez-Figueroa, V. Presutti, I. Celino, M. Sabou, L.-A. Kaffee and E. Simperl, eds, Springer International Publishing, Cham, 2018, pp. 391–407. doi:10.1007/978-3-030-00671-6_23.

[33] A. Rula, A. Maurino and C. Batini, Data quality issues in linked open data, in: *Data and Information Quality: Dimensions, Principles and Techniques*, C. Batini and M. Scannapieco, eds, Springer International Publishing, Cham, 2016, pp. 87–112. doi:10.1007/978-3-319-24106-7_4.

[34] M. Schmachtenberg, C. Bizer and H. Paulheim, Adoption of the linked data best practices in different topical domains, in: *International Semantic Web Conference*, 2014. doi:10.1007/978-3-319-11964-9_16.

[35] A. Simon, D.V. Suero, E. Hyvönen, E. Guggenheim, L.G. Svensson, N. Freire, R. Simon, R. Bailly, R. Wyns, S. van Hooland, S. Wang, V. Alexiev, J. Stiller, A. Isaac and V. Petras, *EuropeanaTech Task Force on a Multilingual and Semantic Enrichment Strategy: Final Report*, 2014, https://pro.europeana.eu/files/Europeana_Professional/EuropeanaTech/EuropeanaTech_taskforces/MultilingualSemanticEnrichment/Multilingual%20Semantic%20Enrichment%20report.pdf.

[36] K. Smith-Yoshimura, Analysis of 2018 international linked data survey for implementers, *The Code4Lib Journal* **42** (2018), https://journal.code4lib.org/articles/13867 (accessed January 24, 2021).

[37] J. Stiller, V. Petras, M. Gäde and A. Isaac, Automatic enrichments with controlled vocabularies in Europeana: Challenges and consequences, in: *Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection*, M. Ioannides, N. Magnenat-Thalmann, E. Fink, R. Žarnić, A.-Y. Yen and E. Quak, eds, Springer International Publishing, Cham, 2014, pp. 238–247. doi:10.1007/978-3-319-13695-0_23.

[38] G. Sugimoto, Building linked open date entities for historical research, in: *Metadata and Semantic Research MTSR 2020*, Springer International Publishing, Cham, 2021.

[39] D. Tomaszuk and D. Hyland-Wood, RDF 1.1: Knowledge representation and data integration language for the web, *Symmetry* **12** (2020), 84. doi:10.3390/sym12010084.

[40] J. Umbrich, A. Hogan, A. Polleres and S. Decker, Link traversal querying for a diverse web of data, *Semantic Web* **6** (2015), 585–624. doi:10.3233/SW-140164.

[41] S. van Hooland, M. De Wilde, R. Verborgh, T. Steiner and R. Van de Walle, Exploring entity recognition and disambiguation for cultural heritage collections, *Digital Scholarship in the Humanities* **30** (2015), 262–279. doi:10.1093/llc/fqt067.

[42] T. van Veen, J. Lonij and W.J. Faber, Linking named entities in Dutch historical newspapers, in: *Metadata and Semantics Research*, E. Garoufallou, I. Subirats Coll, A. Stellato and J. Greenberg, eds, Springer International Publishing, Cham, 2016, pp. 205–210. doi:10.1007/978-3-319-49157-8_18.

[43] J. Volz, C. Bizer, M. Gaedke and G. Kobilarov, Discovering and maintaining links on the web of data, in: *The Semantic Web – ISWC 2009*, A. Bernstein, D.R. Karger, T. Heath, L. Feigenbaum, D. Maynard, E. Motta and K. Thirunarayan, eds, Springer, Berlin, Heidelberg, 2009, pp. 650–665. doi:10.1007/978-3-642-04930-9_41.

[44] A. Waagmeester, E. Willighagen, N.Q. Rosinach, E. Mitraka, S. Burgstaller-Muehlbacher, T.E. Putman, J. Turner, L.M. Schriml, P. Pavlidis, A.I. Su and B.M. Good, Linking Wikidata to the rest of the Semantic Web, in: *Proceedings of the 9th International Conference Semantic Web Applications and Tools for Life Sciences*, 2017, p. 2.

[45] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann and S. Auer, Quality assessment for Linked Data: A survey: A systematic literature review and conceptual framework, *Semantic Web* **7** (2015), 63–93. doi:10.3233/SW-150175.

[46] M.L. Zeng, Semantic enrichment for enhancing LAM data and supporting digital humanities. Review article, *EPI* **28** (2019). doi:10.3145/epi.2019.ene.03.