Micro Analysis of Linked Open Data Quality and Graph Traversals for Cultural Heritage Research

Go Sugimoto

Austrian Centre for Digital Humanities, Sonnenfelsgasse 19, 1010 Vienna, Austria

Abstract. In cultural heritage, many projects have generated a large amount of Linked Open Data (LOD) in the hope that it transforms scattered data into connected global graphs, which are supposed to advance our research with machine-assisted intelligent tools. However, the investigation of aggregation and integration of heterogeneous LOD is rather limited partly due to data quality issues. To this end, the author examines end-user’s “researchability” of LOD, especially in terms of graph connectivity and traversability. Three W3C recommended properties (owl:sameAs, rdfs:seeAlso, and skos:exactMatch) as well as schema:sameAs are inspected for 80 instances/entities for ten widely known data sources in order to create traversal maps. In addition, data content (literals, rdf:about, rdf:resource, rdf:type, skos:prefLabel, skos:altLabel) is assessed to capture the overview of the data quantity and quality. The empirical micro study with network analyses reveals that the major LOD provides relatively low number of outbound links, proprietary RDF properties, and few reciprocal vectors. These quality issues suggest that the LOD may not be fully interconnected and centrally condensed, confirming the outcomes of previous studies. Thus, their homogeneousness casts a doubt on the possibility of automatically identifying and accessing unknown datasets, which implies the needs of traversing strategies to maximise research potentials.

Keywords: Linked Open Data quality, deep graph traversals, RDF analyses, cultural heritage researchability, network analysis
1. Background – Linked Open Data Quality

Many years have passed since RDF became a W3C recommendation. During the following decade(s), Linked Open Data (LOD) has been widely acknowledged and data rich institutions have generated a large portion of LOD. The real power of LOD originates from a very simple philosophy of the inventor of the web. Berners-Lee [1] states “include links to other URIs, so that they can discover more things”, hence the name “Linked” (Open) Data. As of March 2019, the LOD Cloud website reports 1,239 datasets with 16,147 links[1]. It is envisaged that the use of LOD would advance our research, transforming a wide range of distributed data into connected global graphs [2] which facilitate highly intelligent data processing thanks to the development of logical framework in the future. Consequently, LOD enables us to find and formulate new information or knowledge in the name of data-driven research [2].

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. In cultural heritage and humanities, many projects create and use a wide range of LOD for research purposes. In particular, they often execute curatorial tasks such as (named) entity linking and data enrichment [3]. For example, a place name found in a local database is linked to a corresponding entity in GeoNames that offers a global identifier. In other cases, information found in GeoNames is appended to local databases. Jaffri et al. [4] echo this view, stating that many datasets are linked with DBpedia entities through owl:sameAs. In contrast, the investigation of the aggregation and integration of heterogeneous LOD is rather rare [3]; let alone the formation of new knowledge based on complex SPARQL queries across distributed LOD resources. As such, the real value of LOD has been neither fully explored nor verified. The embodiment of LOD-based research still faces many challenges.

One of the problems is the quantity and quality of data. According to Voltz et al. [5], the value of the Web of Data rises and falls with the quantity and quality of links between data sources. A white paper reports that only 10% of cultural heritage materials have been digitised [6], therefore, the availability of cultural heritage information on the web is still extremely partial and limited. The quantity of LOD related to such resources would be even smaller. In terms of data quality, there have been active computer science communities in recent years. Critical quality issues have been frequently raised and discussed in the studies of LOD [5,7–20], alongside the progress over the years. They share the challenges of LOD quality for the coming years.

Although the first problem needs to be addressed eventually, it is rather an economic and political issue; therefore, this paper investigates the data quality issues that are discussed to a lesser extent in cultural heritage community. It focuses on a micro study of LOD “researchability” for the domain. More specifically, the multi-level traversability and quality of LOD entity resources are examined in detail. As LOD potentially empowers the research communities for the goal of machine-assisted research in the future, 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 deals with methodology, related research, and scope. It also briefly discusses about the nature of research in cultural heritage and humanities in relation to conceptual models and schemas/ontologies of LOD. Section 3 presents the results of the analyses of LOD. From Section 3.1 to 3.5 “traversal maps” are illustrated for four categories of cultural heritage entities (agents, events, dates, and places). Section 3.6 and 3.7 describe the statistical examinations of connectivity and data content. The last section summaries the discussions and outlines visions for future work.

2. Methodology

2.1. Related Work

It is crucial to review previous literature in the area of LOD quality. Over the last years quantitative research has been carried out intensively. Schmachtenberg et al. [7] update the 2011 report on LOD, using Linked Data crawler, analysing the change of LOD (8,038,396 resources) over the years. Debattista et al. [8] provide insights into the quality conformance of 130 datasets (3.7 billion quads), using 27 Linked Data quality metrics. Mountanotakis and Tzitzikas [9] have developed a method for LOD connectivity analysis, reporting the results of connectivity measurements for over 2 billion triples and 400 LOD Cloud datasets. In addition, a number of valuable contributions have been made to scrutinise “problem of co-reference” [4]. First, there are critical discussions about the proliferation of owl:sameAs semantics [18]. Second, several large scale statistical

---

1 https://lod-cloud.net/#about, last accessed 2019-09-25
analyses uncover the status of owl:sameAs networks to detect errors for 558 million links [13], verify the proliferation [10,11] (4352 and 8.7 million links respectively), and propose solutions. Most projects concentrate on macro studies and statistical observations of the comprehensive LOD cloud, often using data through dumps and SPARQL endpoints. There are also a few pieces of “semi-micro” research, using domain specific datasets. Ahlers [19] analyses the linkages of GeoNames (11.5 million names), while Debattista et al. [20] inspect the Ordnance Survey Ireland. However, the studies for the cultural heritage domain that this paper covers are not found.

Apart from Mountantonakis and Tzitzikas, macro research projects oftentimes investigate data sources as a whole especially in the sense of owl:sameAs. The connectivity is examined regardless the mobility at an instance level. For example, their research does not reveal if the connection for Mozart is available between data source A and B, even if there are links. This is because the domain coverage may be different: A originates from an encyclopedia and B from biological records. To this end, it is necessary to observe trees (Mozart) not forests (data source A and B as a whole). The instance level maneuverability is the key for the users to navigate themselves in the network.

2.2. Scope of Traversing

In contrast to previous research, this paper deploys a micro analysis. It deals with the small ecosystem of LOD in the cultural heritage domain, based on an empirical qualitative and quantitative method. In particular, it focuses on “researchability and explorability” on arbitrary LOD entities in the sense of graph traversals in order to assess how researchers are able to find useful information in the LOD cloud automatically. The primary goal is to create “traversal maps” of major LOD data sources at an instance level. Naturally, the collections of instances covering the same topic (i.e. categories in Section 2.3) are of vital importance. Subsequently, it is expected to have a better understanding of what LOD resources are accessible in what way between multiple sources, rather than to repeat comprehensive statistical analyses of LOD quality.

The use case of the LOD traversals in this paper is the following: the author/user manually looks up a LOD entity/resource identified by a URI that uses the http:// scheme. Then, s/he follows available links in the entity to reach identical and/or the most related LOD resources. Hyperlinks are documented to generate a traversal map during this process (Figure 3, 4, 6, 8, 10). The traversal continues as long as it is within the specified datasets boundaries (see Section 2.4).

For example, one may traverse an RDF graph from a resource in DBpedia to a resource in Wikidata, which may supply additional information about the resource of DBpedia:

```xml
<rdf:Description rdf:about="http://dbpedia.org/resource/1969">
  <owl:sameAs rdf:resource="http://dbpedia.org/resource/1969">
  <owl:sameAs rdf:resource="http://www.wikidata.org/entity/Q2485"/>
</rdf:Description>
```

For this purpose, the W3C recommended properties, owl:sameAs, rdfs:seeAlso, and skos:exactMatch are used. It is a common practice that information providers set owl:sameAs links to URI aliases [2,12]. In addition, schema:sameAs is also included, due to its popularity. One of the advantages of the standard is that the properties are widely known as long as RDF/XML is used (see Section 2.4), implying no prior knowledge is required to process data. This enables the users’ automation for LOD traversals. As Hartig [21,22] observes, it is of high importance that the end users need to 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 [2].

Since rdfs:seeAlso may be asymmetric, the analysis is not limited to LOD and symmetric graphs. In other words, 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=&amp;place=&amp;nation=&amp;subjectid=1000080). This is allowed in the specification². Another reason to avoid strict co-references is that it is hard to chase 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

---

² https://www.w3.org/TR/rdf-schema/, last accessed 2019-09-25
In order to narrow the scope of the LOD evaluation, this article focuses on answering typical and generic research questions of cultural heritage. For instance, one of the biggest cultural heritage data platforms is Europeana. It has created the Europeana Data Model (EDM) in order to capture heterogeneous cultural heritage information. Its Primer notes that “EDM will let users browse Europeana in revealing new ways. It answers the ‘Who?’, ‘What?’, ‘When?’, ‘Where?’ questions, and makes connections between the networks of stories that will animate Europeana’s content”. Although these 4W questions are a common sense for scientific research in general, they manifest the essence of humanities research: without them, researchers are hardly able to solve any research questions in their disciplines.

The importance of the four core questions is also reflected in other ontologies of cultural heritage. 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”. It centers “‘Event’ as a core entity, connecting “‘Agent”, “Time-Span”, “Objects”, and “Place”. In the library sector, DCMI Metadata Terms also defines almost identical entities: “‘Agent”, “PeriodOfTime”, “PhysicalResource”, and “Location” among others. In addition, FRBR is a conceptual reference model for libraries. The Group 1 primary entities are relevant to the What question, whereas the Group 2 entities are related to Who. The Group 3 is associated with other W-questions.

Therefore, the evaluation of LOD in this article concentrates on the four categories and employs the following terminology to be more specific: agents (for Who), events (for What3), dates (for When), and places (for Where).

2.4. Data Sources

This project sets two basic requirements for the selection of datasets/data sources. It inspects 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. On top of the minimum requirements, the author considers popularity, data volume, coverage, and linking for the selection. The aforementioned LOD cloud is also taken into account, as one of the comprehensive visualisations of LOD. Consequently, the following 9 data sources which include significant content for the cultural heritage and humanities are chosen for inspection: 1)Getty vocabularies (ULAN , AAT , TGN ), 2)GeoNames, 3)VIAF , 4)WorldCat FAST, 5)DBpedia, 6)Wikidata, 7)Librarian of Congress, 8)BabelNet, and 9)YAGO. The overview report of LOD projects in cultural heritage [3] indicates that some big players use few of the 9 data sources.

There is one exception for the selection criteria. Wikipedia articles are encoded in HTML, but Wikipedia may be studied as an indicator, because it has a unique position as a global reference on the web. Indeed, some major LOD data sources link to it. Thus, only incoming links are examined. It should be also noted that LOD datasets derived from Wikipedia (i.e. Wikidata and DBpedia) are included. On the other hand, the above-mentioned Europeana is valuable for cultural heritage digital collections. However, three issue are seen as a drawback. Firstly, the volume of its entity lookup is relatively small. Secondly, it is described as a demo alpha version with a public API key. Thirdly, only JSON-LD has been available until recently. It is decided to exclude it from the analysis.

As this project deploys not only a quantitative but also qualitative analysis, a manageable level of data sampling is considered. It takes twenty representative instances/entities for each category defined in the Section 2.3 (Table 1). For practical reasons, it concentrates on the English version as the primary resource of an entity, when multiple language versions

---

3 This project focuses on events for the What question, although there are other important cultural heritage entities such as objects and buildings.
exist. Nonetheless, other language versions are documented as a reference. In order to objectively and systematically select the most relevant entities, the author consults the “Wikipedia most referenced articles” (2011) for top 20 places and dates, whereas a scientific article about the interaction of top people in Wikipedia is used for 20 agents [23]. In addition, top 20 events are retrieved by a SPARQL query from the EventKG endpoint. The actual number of entities analysed is 660 (Table 2), since some sources do not have the entities the others have. For instance, VIAF holds only agent and place entities. The entities are manually looked up and examined with a help of a self-developed automation tool called RDF Beams.

Statistically speaking, in case of missing entities, they are included in the calculation and the data values are counted as null. For example, WorldCat does not seem to have 1976, 1979, and Europa League. In addition, there are dual identity or “duplicate” in some sources. In particular, they appear in aggregated sources such as YAGO and VIAF. It seems that the dual identity is a leftover of merging entities during data aggregation. Such examples include Aristotle in VIAF and California in YAGO. The dual entities are consolidated for the statistical purpose.

Not only the links in the entities, but also other important data contents are scrutinised. W3C’s standard properties including rdfs:label, rdf:type, skos:prefLabel, and skos:altLabel are recorded. Moreover, the number of literals, rdf:resource, and rdf:about supports the rough understanding of the entire links and content volume.

Although this approach is quite time consuming, it is not feasible to fully automate this process, because of co-reference. It is required to search, identify, and verify the same entity across 10 data sources. The quality of each entity needs to be manually double-checked and documented. In particular, accurately finding an entity is not always possible for several reasons: the lack of cross linking between data sources makes it hard to find all available entities. The entities are confusingly organised and hidden from the mainstream contents, especially in aggregated LOD. The search functionalities at the website of the data sources may have limited capacity and are not optimised. On those grounds, lookups are executed on a best-effort basis.

After the complicated procedure, entity data are collected in separate tabs in an EXCEL file for each source. Individual data are also aggregated and/or faceted, using VBA. Subsequently, a various type of charts are generated. In addition, basic network analyses are conducted by R to capture and visualise the characteristics of the small network of the LOD.

3. Researchability of Linked Open Data

3.1. Overall Traversal Map

The fist analysis starts with a chord diagram which is populated in R with Data to Viz, based on the circlize package (Figure 1). The number of links and their origins and destinations within 10 data sources is demonstrated. The total number of links amounts to 7883. The dominance of DBpedia is obvious, occupying over 77.6% of the entire linkages. It is also noticeable that self-links significantly contribute to the volume of the links.

YAGO is the second biggest pie (11.1%) supplying a substantial amount of links to DBpedia as well as other destinations. Wikidata, VIAF, Library of Congress, and WorldCat somewhat share similar numbers of links. The next cluster would be BabelNet, GeoNames and Getty vocabularies, while Wikipedia is almost invisible.

Regarding the interactions of the data sources, the connections are relatively limited to neighboring sources. DBpedia holds 6754 incoming and 5475 outgoing links, while YAGO does 218 and 1528 links respectively, and the two data sources are tightly connected. Those figures are disproportionally high, compared to those of other sources. While Wikidata has 392 incoming and 192 outgoing links, Getty vocabularies provide 64 and 60 links. Indeed, the average number of links is 788.3, whereas the medians are 111 and 161.5 for outgoing and incoming.

When inverse traversals are removed from Figure 1, the situation looks largely different (Figure 2). DBpedia loses an ample number of links (57.8%). On the contrary, YAGO gains most, aggressively chasing DBpedia to cement its second position (23.1%).

8. See endnotes for the SPARQL query
https://viaf.org/viaf/7524651/, last accessed 2019-09-25
Such dramatic shift is an evidence of abundance of inverse properties described in DBpedia. In fact, this is mostly due to the inverse use of rdfs:seeAlso in DBpedia. For instance, the entity of Sweden contains:

```
<rdf:Description rdf:about="http://dbpedia.org/resource/Lund">
  <rdfs:seeAlso rdf:resource="http://dbpedia.org/resource/Sweden"/>
</rdf:Description>
```

Table 1. 80 selected entities in 4 categories

<table>
<thead>
<tr>
<th>Agents</th>
<th>Events</th>
<th>Dates</th>
<th>Places</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carl Linnaeus</td>
<td>World War II</td>
<td>1987</td>
<td>United States</td>
</tr>
<tr>
<td>Jesus</td>
<td>World War I</td>
<td>1986</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>Aristotle</td>
<td>American Civil War</td>
<td>1985</td>
<td>France</td>
</tr>
<tr>
<td>Napoleon</td>
<td>FA Cup</td>
<td>1983</td>
<td>England</td>
</tr>
<tr>
<td>Adolf Hitler</td>
<td>Vietnam War</td>
<td>1980</td>
<td>Germany</td>
</tr>
<tr>
<td>Julius Caesar</td>
<td>Academy Awards</td>
<td>1984</td>
<td>Canada</td>
</tr>
<tr>
<td>Plato</td>
<td>Cold War</td>
<td>1982</td>
<td>Australia</td>
</tr>
<tr>
<td>William Shakespeare</td>
<td>Korean War</td>
<td>1968</td>
<td>Japan</td>
</tr>
<tr>
<td>Albert Einstein</td>
<td>American Revolutionary War</td>
<td>1979</td>
<td>Italy</td>
</tr>
<tr>
<td>Elizabeth II</td>
<td>UEFA Champions League</td>
<td>1969</td>
<td>Poland</td>
</tr>
<tr>
<td>Michael Jackson</td>
<td>UEFA Europa League</td>
<td>1978</td>
<td>India</td>
</tr>
<tr>
<td>Madonna (entertainer)</td>
<td>Olympic Games</td>
<td>1967</td>
<td>Spain</td>
</tr>
<tr>
<td>Ludwig van Beethoven</td>
<td>Stanley Cup</td>
<td>1981</td>
<td>London</td>
</tr>
<tr>
<td>Wolfgang Amadeus Mozart</td>
<td>Super Bowl</td>
<td>1977</td>
<td>Russia</td>
</tr>
<tr>
<td>Pope Benedict XVI</td>
<td>Iraq War</td>
<td>1976</td>
<td>New York City</td>
</tr>
<tr>
<td>Alexander the Great</td>
<td>War of 1812</td>
<td>1975</td>
<td>Brazil</td>
</tr>
<tr>
<td>Charles Darwin</td>
<td>Gulf War</td>
<td>1964</td>
<td>California</td>
</tr>
<tr>
<td>Barack Obama</td>
<td>Spanish Civil War</td>
<td>1966</td>
<td>New York</td>
</tr>
<tr>
<td>Mary (mother of Jesus)</td>
<td>World Series</td>
<td>1965</td>
<td>The Netherlands</td>
</tr>
<tr>
<td>Queen Victoria</td>
<td>EFL Cup</td>
<td>1960</td>
<td>Sweden</td>
</tr>
</tbody>
</table>

Table 2. The actual number of entities available for the analyses

<table>
<thead>
<tr>
<th>Data sources</th>
<th>Entities</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>WorldCat</td>
<td>80 (77)</td>
<td>1976, 1979, Europa League and New York City do not seem to exist. New York City in Wikidata confusingly links to New York in WorldCat, while New York in Wikidata does not link to WorldCat. Both World War I and II are an entity, but World War I and II also exist as separate entities, whose types are called periodisation. The latter two are not included in this analysis.</td>
</tr>
<tr>
<td>Library of Congress</td>
<td>80 (77)</td>
<td>1976, 1979, and Europa League do not seem to exist</td>
</tr>
<tr>
<td>VIAF</td>
<td>40</td>
<td>Only agents and places exist. However, Jesus, Madonna, Mary, and Pope Benedict XVI do not seem to exist</td>
</tr>
<tr>
<td>Getty vocabularies</td>
<td>40 (36)</td>
<td>Only agents and places exist. However, Jesus, Madonna, Mary, and Pope Benedict XVI do not seem to exist</td>
</tr>
<tr>
<td>Wikidata</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>DBpedia</td>
<td>80</td>
<td>Spanish Civil War exists only in other languages than English</td>
</tr>
<tr>
<td>GeoNames</td>
<td>20</td>
<td>Only place entities exist</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>80</td>
<td>Only incoming links are analysed</td>
</tr>
<tr>
<td>YAGO</td>
<td>80 (79)</td>
<td>1978 does not seem to exist. 10 dual identify data are consolidated (Brazil, California, India, London, The Netherlands, New York, Olympic Games, Russia, Spain, and Sweden)</td>
</tr>
<tr>
<td>Total</td>
<td>660 (648)</td>
<td></td>
</tr>
</tbody>
</table>

13 The priority is given in the following order: page rank, 2Drank (24 languages), and page rank (female).
14 International events are prioritised, thus a couple of specific events such as US censuses are removed.
15 Top 20 places are extracted from the general list.
Fig. 1. Chord diagram for the links between 10 data sources

Fig. 2. Chord diagram after removing inverse links

Table 3. The number and average number (per entity) of links held by 10 data sources

<table>
<thead>
<tr>
<th>Links</th>
<th>WorldCat</th>
<th>Library of Congress</th>
<th>VIAF</th>
<th>Getty</th>
<th>Wikidata</th>
<th>DBpedia</th>
<th>BabelNet</th>
<th>GeoNames</th>
<th>Wikipedia</th>
<th>YAGO</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>207</td>
<td>78</td>
<td>153</td>
<td>60</td>
<td>192</td>
<td>5475</td>
<td>170</td>
<td>20</td>
<td>0</td>
<td>1528</td>
<td>7883</td>
</tr>
<tr>
<td>Average</td>
<td>2.6</td>
<td>1.0</td>
<td>3.8</td>
<td>1.5</td>
<td>2.4</td>
<td>68.4</td>
<td>2.1</td>
<td>1.0</td>
<td>0.0</td>
<td>19.1</td>
<td>11.9</td>
</tr>
</tbody>
</table>
The sum of the links decreases to 3722. Meanwhile, the average and medians slump to 372.2, 110.5, and 113.5 respectively. The amount of outgoing hyperlinks found in each source is moderate, given the entire size of those datasets (e.g. millions of triples); on average it is mostly under four links per entity, up to 68.4 for DBpedia (Table 3). There is a great deal of room for “linked” data improvement.

The chord diagrams are useful to understand the overall tendency of the data sources, and, in particular, one-to-one relations between two data sources. However, it is best to deploy network diagrams to scrutinise the multiple traversals among the 10 sources.

Figure 3 is created by igraph packages and illustrates the overview of the traversability for all entities across all data sources. Volumes and self-links/loops (i.e. pointing to the same data source/domain) are not visualised in order to hide the heavy weights between DBpedia and YAGO. In this way, Figure 3 focuses on the routes of traversals (i.e. the users’ mobility) rather than their quantities of the links.

It is clear that the traversing routes are not equally available. There are several nodes/vertices which can be reached via only particular edge(s)/path(s). This implies that the users cannot automatically obtain information as they enter the LOD network. They need to formulate strategies in order to follow the best paths to retrieve the information in which they are interested.

When in and out degrees of the network are compared, it is found that YAGO functions as a hub, delivering 5 connections to Wikidata, DBpedia, and GeoNames, VIAF, as well as Wikipedia. YAGO also has 5 outgoing connections, but is interestingly more connected to other sources such as Getty vocabularies and WorldCat. Similarly, DBpedia plays a central hub role to link to Wikidata, YAGO, GeoNames, and VIAF. In contrast, Wikidata has no outgoing connections (192 self-links are omitted). Whilst GeoNames only links to DBpedia, Library of Congress and Getty have one channel. With regard to incoming connections, GeoNames is an authority to which YAGO, DBpedia, BabelNet, and WorldCat refer. VIAF is also a centre of gravity, attracting 4 connections. DBpedia is also a strong contender for an authority hub. Wikidata is connected by 3 sources. On the other hand, BabelNet receives no connection. WorldCat and Getty are both reached only by VIAF. As such, Figure 3 gives an impression of graph patterns. One is the centrality of a few data sources. At the same time, fragmentation is also visible in the outskirts of the network including Getty and Library of Congress. A dead end and an isolated island (Wikidata and BabelNet respectively) are also observed, where it is hard to navigate the LOD network.

3.2. Agent Traversal Map

Figure 4 exhibits the traversal map of agents, by faceting agent entities out of the overall statistics. From now on, inverse properties are included but loops are excluded for the traversal map visualisation. Thus, the distortion of the “route diagram” especially regarding DBpedia is minimal. However, the rest of the statistics (bar charts and in the texts) include both inverse properties and loops, so that they reflect the actual situation.

As GeoNames does not hold agent entities, it is isolated. Wikipedia is not referred in case of agent entities. BabelNet is a data source without incoming links. This category has only three sets of nodes that have bilateral links, namely between DBpedia and VIAF, YAGO and DBpedia, VIAF and the Library of Congress. Therefore, segmentation is visible in the network and truly standardised LOD connectivity is rather limited.

The role of DBpedia is prominent, attracting 740 links. It is now more evident that YAGO feeds a large amount of links to DBpedia, compared to other lines of traversals. When looking into the data of the individual entities, interestingly, WorldCat is referred exactly one time per entity (20 links) as self-link. This is caused by the description of a new identifier of WorldCat. For example, the instance of Shakespeare is described as follows, where “29048” is a new identifier replacing “fst00029048”:

```
<rdf:Description rdf:about="fst00029048">
  <rdfs:seeAlso rdf:resource="29048"/>
</rdf:Description>
```

YAGO and Wikidata almost level at 76 and 75 for incoming links. VIAF captures 72 referrals, overshadowing the Library of Congress at 50. In cultural heritage, VIAF plays a valuable role for agents as an aggregation of authority files of national libraries. For instance, although only 4 links with schema:sameAs are recorded for Beethoven in VIAF for this analysis, 8 other links are found which bound

---

16  https://igraph.org/r/ last accessed 2019-09-25
for outside the 10 data sources. The destinations include the national library of Spain, Sweden, France, Japan, and Germany.

The situation of link destinations found in each entity is visualised in Figure 5. A spike for DBpedia is Jesus, which is the overall winner delivering 93 links. Michael Jackson deserves the second place (72 links), followed by the group of Barack Obama (68), Queen Victoria (65), and Adolf Hitler (64). However, the median (56.5) and average (58.8) of the whole entities are not extremely far from the numbers of the top groups. Michael Jackson of YAGO is somehow highly referred (11 links), compared to other entities (average 3.75 links). Mary obtains a considerably low number of links for DBpedia (19 links). Whilst most data sources cover all 20 agents, Mary, Madonna, Jesus, Benedict XVI, are missing in Getty vocabularies. The number of VIAF links is also sunken for Mary, Madonna, and Jesus. This is understandable that Getty ULAN and VIAF are typically oriented toward artists and authors in the context of libraries and museums. In contrast, Jesus is the highest targeted entity for DBpedia (84 links) proving its importance.

### 3.3. Events Traversal Map

As for events (Figure 6), GeoNames and Getty vocabularies do not have entities. Apart from DBpedia (947 links), Wikidata (76) and YAGO (62) are the most frequently referred. In contrast, WorldCat (19) and Library of Congress (30) are not widely recognised, while no sources links to BabelNet. A little deviation is VIAF. In general events were not found in VIAF during the manual data exploitation, however, it turns out that two resources (WorldCat and Library of Congress) refer to it 7 times each. For example, WorldCat links to World Series in French (skos:prefLabel is Séries mondiales (Base-ball) and skos:altLabel is World Series (Base-ball)). Other 13 cases are all sport events and labelled as corporate entity in VIAF.

Compare to Figure 3 and 4, Figure 6 clearly exposes the lack of links. Even if there are links, the connections are minimally established. Unlike agents, Wikipedia is connected by YAGO. Bilateral links are extremely rare: only between YAGO and DBpedia. As a result, the obstacles of the LOD orientation are more serious than the overall situation. It is not possible, for example, to move from Library of Congress to Wikidata. This implies that the entry point to the network determines the movement in it.

In terms of each entity (Figure 7), the most appealing entity would be World War II (143 links). It is followed by World War I (90) and Iraq War (75). Although EFL Cup is the lowest (39), the gaps between entities are relatively subtle except the top 3 (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 detail information about the war, although the semantic of the rdfs:seeAlso is weak. As seen in Section 2.2, even the RDF representation is not guaranteed, which would hamper LOD data processing.

A clear pattern is also recognised for what kind of links constitute an entity. A typical case of an entity would have: one WorldCat link, a couple of Library of Congress links, four Wikidata links, three YAGO links, and sometimes one VIAF link. As such, the links in the 20 entities within the sample LOD data sources are highly standardised or normalised. This result suggests that content pattern is quite homogeneous and it is likely that the majority of entities to answer cultural heritage W-questions have similar tendencies with only few irregularities, therefore, the number of sampling is reasonable, representing the data quality of 10 data sources relatively well.

### 3.4. Dates Traversal Map

Dates cannot be missed when it comes to cultural heritage. As anticipated, dates are rarely set as a LOD entity, thus, only four sources are referenced (Figure 8): DBpedia (568 links), Wikidata (59), Library of Congress (36), and WorldCat (18). Although YAGO provides many links to DBpedia and Wikidata, it does not receive any links this time. Since bilateral links do not exist, the movement in the network is highly limited. The only three possible paths are as follows: One can navigate from YAGO to Wikidata through or without DBpedia. The BebelNet users can reach Wikidata via DBpedia. Users can depart from WorldCat to arrive at Library of Congress. Consequently, the fluctuation of linking patterns is also minor (Figure 9).

The economy of the creation of date entities may show serious issues. 1978, 1979, and 1976 do not seem to exist in YAGO, while other consecutive
years in the 1970’s are available. Such inconsistency would become problematic, when queries are constructed to look for answers to research questions. Erroneous links and data omissions from the solutions/answers based on the semantic queries would require careful presentation to the LOD users in the future.

3.5. Places Traversal Map

Places are another important element of cultural heritage research, exemplifying the diversity of culture in different geographical environments. Figure 10 draws a relatively similar picture to Figure 3, probably symbolising the role of geographical information in the LOD environment. Consequently, traversability is better than other categories.

Without doubt, DBpedia dominates the scene (4302 links), yet even more idiosyncratic than other categories. Wikidata (102) and YAGO (81) are the usual suspects for the second group. However, GeoNames is a new face of reference with 66 link, followed by neck-and-neck WorldCat (37), Library of Congress (39), and VIAF (39). The presence of GeoNames, in particular, facilitates more fluid movements in the network. Although [19] 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. As usual, BebelNet does not seem to be widely considered as a LOD reference source. Getty TGN only contains self-links mostly in the form of rdfs:seeAlso for a HTML representation. RDF/XML for New York City holds:

```
<gvp:Subject rdf:about="http://vocab.getty.edu/tgn/7007567">
  <rdfs:seeAlso rdf:resource="http://www.getty.edu/vow/TGNFullDisplay?find=&amp;place=&amp;nation=&amp;subjectid=7007567"/>
</gvp:Subject>
```

Therefore, it is a dead end in terms of network traversals, of which the users need to be aware during their traversing.

In Figure 11, the lowest positions are surprisingly the Netherlands (66 links), United Kingdom (73), and United States (78). This is chiefly attributed to fewer numbers of DBpedia links. However, the reason for the lack of links are unclear. On the contrary, the top entities receive a large quantity of links, which include Germany (501), France (490), and Poland (433).

The normalised link pattern still persists with two links for WorldCat, Library of Congress, and VIAF. In addition, it is typical to have one Getty TGN link, three or four GeoNames, and slightly more links for Wikidata and YAGO. Occasionally, Wikipedia links are present. Nevertheless, the diversity of DBpedia link patterns is more distinct than other categories.

3.6. Network Analysis

R computes network analyses by using libraries equipped with statistical package (Table 4). In our case, it could be used to objectively verify or reject so far relatively subjective impressions and interpretations of the traversal maps.

The reciprocity of the network is generally low, not exceeding 50%. In particular, date entities have no reciprocal links. It is alarming in terms of traversability of cultural heritage information, as the absence of reciprocal links limit the area of data exploration. On the other hand, the transitivity is slightly higher in general. The mean distances of the network are rather short, implying that there are two traversals at most. Related to this, diameter is the number of nodes for the longest path of the network, taking into account the direction of edges. One example would be in the sequence of Getty, Library of Congress, VIAF, DBpedia, YAGO, and Wikipedia. Long traversals may be needed to access from one data source to another, depending on the entry point of the network.

```
Table 4. Network analysis statistics

<table>
<thead>
<tr>
<th>Measurement</th>
<th>Overall</th>
<th>Agents</th>
<th>Events</th>
<th>Dates</th>
<th>Places</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reciprocity</td>
<td>0.435</td>
<td>0.375</td>
<td>0.200</td>
<td>0.000</td>
<td>0.400</td>
</tr>
<tr>
<td>Transitivity</td>
<td>0.508</td>
<td>0.683</td>
<td>0.643</td>
<td>0.600</td>
<td>0.457</td>
</tr>
<tr>
<td>Mean Distance</td>
<td>2.0</td>
<td>1.8</td>
<td>1.2</td>
<td>1.2</td>
<td>1.7</td>
</tr>
<tr>
<td>Diameter</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>
```

In addition, centrality is calculated, using three methods: Eigen Vector, Betweenness, and Closeness (in, and out) (Figure 12). The centrality of DBpedia (and to a less extent YAGO) is clearer in the chart. The closeness (in and out) also statistically suggests the LOD hubs of outgoing and incoming links. The contrast between Wikidata as an incoming source and BebelNet as an outgoing source is observable. VIAF,
GeoNames, and DBpedia seem to sit in the in-between position, mediating the linking flows.

Generally speaking, the whole situation shows a mosaic of segmentation, if not data silos, which LOD is supposed to resolve, or a couple of tightly connected LOD clusters at best. It is currently hard to implement automatic traversals among the datasets without studying traversal maps.

Fig. 3. Network diagram among 10 data sources (after removing volumes and self-links to the same domain)\textsuperscript{17}

Fig. 4. Network diagram for agent entities

\textsuperscript{17}In traversal maps (Figure 3,4,6,8,10), 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 Figure 3).
Fig. 5. The amount of outgoing links to 10 data sources found in 20 agent entities (* means duplicate consolidation)

Fig. 6. Network diagram for event entities
Fig. 7. The amount of outgoing links to 10 data sources found in 20 event entities (* means duplicate consolidation)

Fig. 8. Network diagram for date entities
Fig. 9. The amount of links to 10 data sources found in 20 date entities

Fig. 10. Network diagram for place entities

Fig. 11. The amount of links to 10 data sources found in 20 place entities (* means duplicate consolidation)
3.7. Statistics for Connectivity and Data Content

In order to better understand the researchability of LOD datasets, the author additionally generates more segmentation and detailed statistics. Figure 13 displays the percentage of hyperlinks bounding for the domains of 10 datasets, out of all hyperlinks detected for the four properties. The average 85.9% of links are within the 10 data boundary. The lowest source of VIAF still holds over 37.2%. The statistics indicate the closed and close connections of 10 data sources in terms of standardised traversability.

Generally speaking, these homogeneousness and centrality of the 10 datasets as well as the hubs within are a worrying sign in the sense that the users of 10 datasets are not able to identify and explore new datasets beyond those giants of LOD. [11] note that typical size of sameAs networks is either a small constant or growing slowly, and single central resources are connected to a number of peripheral resources. This condensed picture of LOD is adequately depicted in their cluster analysis and visualisation, where a few LOD data sources observed in this paper are clearly seen as in-degree or out-degree hub nodes such as DBpedia, GeoNames, Wikipedia, and WorldCat. Correndo et al. [14] also report a power-based LOD network. Moreover, recent research [12] discovers two high-centrality nodes (DBpedia and Freebase) and domain specific naming authorities/hubs such as GeoNames among others.

Figure 14 presents the percentage of the four standard properties against the total number of rdf:resource. In RDF/XML, rdf:resource is the property to indicate the URI of the object node in a graph[18]. In this sense, it should normally contain all the outgoing links. By dividing the ratio of the four properties, it should be able to highlight the balance between them and other properties including proprietary ones.

Overall percentage is expectedly low, because the four properties are normally a small part of RDF content. Nevertheless, the range varies from 29% to near 0%. The result suggests the high importance of the four properties in the WorldCat and BabelNet datasets. In contrast, Getty vocabularies and Wikidata use other properties almost exclusively.

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 appears only occasionally to provide inverse relations for obsolete (mostly duplicate) properties that offer redirects.

On one hand, proprietary properties in Wikidata enable the users to refine the semantics of outbound

---

18 https://www.w3.org/TR/rdf-syntax-grammar/, last accessed 2019-09-25
links. It is useful in some cases where one needs to identify a particular link among tens of outgoing links. On the other hand, they make it more difficult to automate graph traversals.

Another issue is that the Wikidata properties do not use human “guessable” URIs, if not absolutely opaque URIs such as machine-encoded ones. For instance, the entity URI of Cold War is https://www.wikidata.org/wiki/Special:EntityData/Q8683.rdf, using the Q-prefix identifier in the suffix. They are agnostic about their semantics, which prevents human users from guessing or knowing the meaning of properties and/or hacking the URIs without examining the ontology behind.

In the case of France in Getty TGN, there are 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:)22 and 202 Getty Ontology (gvp:)23 are in use, 60 PROV (prov:)21 and 56 SKOS-XL (skosxl:)24 are also found. Although not all properties use rdf:resource, the figures demonstrate a rough idea about the relation between linking and property usage. It becomes clear that some data providers set different strategies to design their ontologies in spite of the W3C recommendations.

Figure 15 illustrates another view of the four properties. It is found that there are different approaches to the links. It indicates that it is not feasible to traverse LOD and collect information, if the users specify only one type of property. As seen in Section 3.1, the need of traversing strategies is also verified from this perspective. Despite the wide spread of research concerning owl:sameAs (Section 2.1), its use exceeds the majority only marginally, so as rdfs:seeAlso. As GeoNames provides the link to DBpedia with the latter, the equivalent identity cannot be inferred. skos:exactMatch is slightly unique in the sense that it is not exactly in the RDF/XML family but SKOS vocabularies. Nonetheless, BabelNet, Getty vocabularies, and Library of Congress take it as a major description method.

schema:sameAs is analysed in this paper due to the prospect of its use by LOD data providers. In fact, VIAF exclusively uses it, whilst more than half of WorldCat entities are described with it. However, its use is debatable, since schema.org ontology is not a W3C recommendation. Moreover, Beek et al. [12] point out that it is semantically different from owl:sameAs.

In addition, this paper analyses the data content other than the links to estimate the researchability of the datasets. Figure 16 focuses on other W3C properties for content, namely rdfs:label, rdf:type, skos:prefLabel, and skos:altLabel.

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 also visible that GeoNames uses only rdf:type, primarily because it employs proprietary properties for the name of places:

```xml
<gn:name>Canada</gn:name>
<gn:alternateName
xml:lang="am">ካናዳ</gn:alternateName
>

The library sector (VIAF, Library of Congress, and WorldCat) states skos:altLabel extensively. In general, it is evident that the use of properties is diverse and not standardised.

The amount of rdf:resource, rdf:about, and literals is shown in Table 5. Those two properties are the centre of RDF/XML to describe and connect resources. Therefore, it is suitable to measure the data content. YAGO always uses rdf:about only once for the entity itself, for example, as follows:

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

Similarly, each entity in GeoNames contains it exactly twice:

```xml
https://schema.org/docs/howwework.html, last accessed 2019-09-25
```
The second `rdf:about` preserves the technical metadata about the entity such as a Creative Commons license and creation date.

Moreover, literals have to be carefully treated, as they may include less relevant information about the entity. Despite the caveats, the figures give a rough idea of how much content is described in each LOD. Manual inspection indicates that the number of literals in some LOD are extremely high due to not only an enormous amount of technical metadata, but also repetitions (e.g. literals expressed in several schemas) and language variations in them. For example, there are the total of over 4.5 million literals and the average more than 57 thousands in Wikidata.

The limited amount of new data could be obtained from WorldCat, BebelNet, and GeoNames; it is not promising to carry out serious research with such data. It seems that some datasets tend to serve merely as global references/identifiers, rather than new source of information.
Fig. 13. The percentage of links going inside and outside 10 data sources

Fig. 14. The percentage of 4 properties against rdf:resource
Fig. 15. The percentage of the use of 4 linking properties in data sources

Fig. 16. The percentage of the use of 4 content properties in data sources

Table 5. The average number (per entity) of rdf:resource, rdf:about, and literals for each data source

<table>
<thead>
<tr>
<th>Content</th>
<th>WorldCat</th>
<th>Library of Congress</th>
<th>VIAF</th>
<th>Getty</th>
<th>Wikidata</th>
<th>DBpedia</th>
<th>BabelNet</th>
<th>GeoNames</th>
<th>Wikipedia</th>
<th>YAGO</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>rdf:resource</td>
<td>8.9</td>
<td>66.2</td>
<td>100.0</td>
<td>616.3</td>
<td>5314.9</td>
<td>3064.3</td>
<td>16.9</td>
<td>14.2</td>
<td>0.0</td>
<td>628.6</td>
<td>756916</td>
</tr>
<tr>
<td>rdf:about</td>
<td>7.7</td>
<td>98.7</td>
<td>33.8</td>
<td>74.2</td>
<td>2443.8</td>
<td>2760.1</td>
<td>6.2</td>
<td>2.0</td>
<td>0.0</td>
<td>1.1</td>
<td>429757</td>
</tr>
<tr>
<td>literals</td>
<td>47.1</td>
<td>255.0</td>
<td>598.4</td>
<td>209.8</td>
<td>577150.8</td>
<td>84.9</td>
<td>2.9</td>
<td>189.5</td>
<td>0.0</td>
<td>84.4</td>
<td>4646127</td>
</tr>
</tbody>
</table>
4. Conclusion

4.1. Challenges for Cultural Heritage Datasets

This project strives to uncover gaps between the data producers and consumers. As Data on the Web Best Practices\textsuperscript{26} 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”. The evaluation of ten LOD reveals a sign of data quality issues. It confirms the observations of [19,20] that limited number of links are found for major LOD datasets, apart from the relatively ample amount for DBpedia. A large part of LOD hubs may not be fully connected (and unevenly connected) for the representative entities for humanities research. This result also reflects previous LOD studies on overall quality and owl:sameAs networks \cite{11,14}. In particular, power-law-based networks and closures have been found for the LOD cloud. Moreover, centrality can be observed for not only linkage, but also data content.

It seems that high-quality datasets are hugely biased toward a couple of data sources, especially generic knowledge bases. Consequently, it is uncertain if researchers would be able to find new information, let alone to answer specific questions that they are interested for specific domains. For instance, unless more useful data are connected from LOD hubs, it is hard to study the population and crime rates of London and its neighbouring cities over centuries, the current locations and auction prices of all heritage objects of Mozart, or the comparison of architectural plans of Romanesque churches in France. Such studies require detailed statistics and/or a set of fully digitised multimedia collections other than encyclopedical data.

Some valuable information is not easily reachable simply due to the lack of links, and/or redundantly long traversing. For example, it is currently not possible for a user looking at/accessing to Beethoven in Getty ULAN to obtain relevant artists and songs in BabelNet. This is a serious shortcoming of a research scenario. In summary, it has become clear through this research that the quality of eighty representative entities has not yet met the simple needs of researchers when exploring major LOD.

It is also known that LOD in cultural heritage is not adequately and sufficiently published. For instance, Francorum Online\textsuperscript{27} has technical problems. Pleiades\textsuperscript{28} provides RDF/XML, but does not offer links to major LOD that are available in JSON. Other LOD projects (LOCAH\textsuperscript{29} and PCDH\textsuperscript{30}) have other problems such as funding. From a quantity perspective, it is hoped that more LOD will be published and connected to improve the researchability.

4.2. For the Next Move

An obstacle for data processing automation is proprietary properties. LOD is not as powerful as it can be, as long as human users analyse related data each time when traversing data, because they are not initially aware of data sources and their ontologies in their query time \cite{24}. Although there may be different reasons why they need to be used, the first choice should be given to the standard properties in order to increase interoperability and potentials for automation. According to Bizer et al. \cite{2}, 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.

To enhance the analysis of this article, it would be interesting to investigate the LOD traversability in comparison with all the LOD properties actually used. For instance, Linked Open Vocabulary is a good starting point to analyse the acceptance of a broad range of properties/vocabularies for LOD and the implications of standardisation and proliferation of vocabularies. Due to its constraints, this paper merely manages to shed the first light on the relation between the W3C standards and proprietary properties.

As Berners-Lee states \cite{1} 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 are needed 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

\textsuperscript{26} https://www.w3.org/TR/dwbp/ last accessed 2019-09-25
\textsuperscript{27} http://francia.ahlfeldt.se/index.php, last accessed 2019-09-25
\textsuperscript{28} https://pleiades.stoa.org/, last accessed 2019-09-25
\textsuperscript{29} http://dataarchiveshub.ac.uk/, last accessed 2019-09-25
\textsuperscript{30} http://www.canadiansa.ca/en/pcdn-lnod/, last accessed 2019-09-25
enrich the datasets to create reciprocal vector links. One such example which enables the users to compare data sources is SILK [5]. Although it is limited to two data sources, it supports them 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 [25,26]. 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 interactive interface. Furthermore, research on LOD search engine is somewhat advancing slowly. Although there are some projects including Swowise, Sindice, and LODatio [25], many are experimental, out-of-date, or user-unfriendly. It is hoped that next generation 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 obstacles found in this article, LOD traversing becomes more feasible for the end-users with a help of automated tools.

References


