University of Economics, Prague

Faculty of Informatics and Statistics


# DBPEDIA LINKAGE ANALYSIS LEVERAGING ON ENTITY SEMANTICS

MASTER THESIS

Study programme: Applied Informatics

Field of study: Knowledge and Web Technologies

Author: Bc. David Fuchs

Supervisor: prof. Ing. Vojtěch Svátek, Dr.

Prague, May 2019

## Declaration

I hereby declare that I am the sole author of the thesis entitled “DBpedia linkage analysis leveraging on entity semantics“. I duly marked out all quotations. The used literature and sources are stated in the attached list of references.

In Prague on 04.05.2020

Signature

Student’s name

## Acknowledgement

I hereby wish to express my appreciation and gratitude to the supervisor of my thesis, *prof. Ing. Vojtěch Svátek, Dr.*

## Abstract

This thesis focuses on the analysis of interlinking of Linked Open Data resources in various data silos and DBpedia, the hub of the Semantic Web. It also attempts to analyse the consistency of bibliographic records related to artwork in the two major encyclopaedic datasets, DBpedia and Wikidata, in terms of internal consistency of artwork in Wikidata, which models its entries in compliance with the Functional Requirements for Bibliographic Records (FRBR), as well as the consistency of interlinking from DBpedia to Wikidata.

The first part of the thesis describes the background of the topic, focusing on the concepts important for this thesis: Semantic Web, Linked Data, Data quality, knowledge representations in use on the Semantic Web, interlinking and two important ontologies (OWL and SKOS).

The second part is dedicated to the analysis of various data quality features of interlinking with DBpedia. The results of this analysis of interlinking between various sources of LOD and DBpedia has led to some concerns over duplicate and inconsistent entities, but the real problem appears to be the currency of data with only half of the datasets linking DBpedia being updated at most five years before the data collection for this thesis took place (October through November 2019). It is also concerning that almost 14 % of the interlinked datasets are not available through standard Semantic Web technologies (SPARQL, dereferenceable URIs, RDF dump). The third part starts with the description of the approach to modelling artwork entities in Wikidata in compliance with FRBR and then continues with the analysis of internal consistency of this part of Wikidata and the consistency of interlinking of annotated entities from DBpedia and their counterparts from Wikidata. The percentage of FRBR entities in Wikidata found to be affected by inconsistencies is 1.5 %, but this figure may be higher due to technological constraints that prevented several queries from finishing. To compensate for the failed queries, the number of inconsistent entities was estimated by a calculation to be 22 %. The inconsistency rate of interlinking between DBpedia and Wikidata was found to be about 16 % according to the annotators.

The last part aims to provide a holistic view of the problem domain, describing how the inconsistencies in different parts of the interlinking chain could lead to severe consequences unless pre-emptive measures are taken. A by-product of the research is a web application designed to facilitate the annotation of DBpedia resources with FRBR typing information, which was used to enable the analysis of interlinking between DBpedia and Wikidata. The key choices made during its development process are documented in the annex.



## Keywords

linked data quality, interlinking consistency, Wikidata consistency, Wikidata artwork, Wikidata FRBR, DBpedia linking Wikidata, linguistic datasets linking DBpedia, linked open datasets linking DBpedia

# Content

# List of Figures

# List of tables

# List of abbreviations

AMIE
: Association Rule Mining under Incomplete Evidence

API
: Application Programming Interface

ASCII
: American Standard Code for Information Interchange

CDA
: Confirmation data analysis

CL
: Code lists

CSV
: Comma-separated values

EDA
: Exploratory data analysis

FOAF
: Friend of a Friend

FRBR
: Functional Requirements for Bibliographic Records

GPLv3
: Version 3 of the GNU General Public License

HTML
: Hypertext Markup Language

HTTP
: Hypertext Transfer Protocol

IFLA
: International Federation of Library Associations and Institutions

IRI
: Internationalized Resource Identifier

JSON
: JavaScript Object Notation

KB
: Knowledge bases

KG
: Knowledge graphs

KML
: Keyhole Markup Language

KR
: Knowledge representation

LD
: Linked Data

LLOD
: Linguistic LOD

LOD
: Linked Open Data

OCLC
: Online Computer Library Center

OD
: Open Data

ON
: Ontologies

OWL
: Web Ontology Language

PDF
: Portable Document Format

POM
: Project object model

RDF
: Resource Description Framework

RDFS
: RDF Schema

ReSIST
: Resilience for Survivability in IST

RFC
: Request For Comments

SKOS
: Simple Knowledge Organization System

SMS
: Short message service

SPARQL
: SPARQL query language for RDF

SPIN
: SPARQL Inferencing Notation

UI
: User interface

URI
: Uniform Resource Identifier

URL
: Uniform Resource Locator

VIAF
: Virtual International Authority File

W3C
: World Wide Web Consortium

WWW
: World Wide Web

XHTML
: Extensible Hypertext Markup Language

XLSX
: Excel Microsoft Office Open XML Format Spreadsheet file

XML
: eXtensible Markup Language

# 1 Introduction

The encyclopaedic datasets DBpedia and Wikidata serve as hubs and points of reference for many datasets from a variety of domains. Because of the way these datasets evolve, in case of DBpedia through the information extraction from Wikipedia while Wikidata is being directly edited by the community, it is necessary to evaluate the quality of the datasets and especially the consistency of the data to help both maintainers of other sources of data and the developers of applications that consume this data.

To better understand the impact that data quality issues in these encyclopaedic datasets could have, we also need to know how exactly the other datasets are linked to them by exploring the data they publish to discover cross-dataset links. Another area which needs to be explored is the relationship between Wikidata and DBpedia, because having two major hubs on the Semantic Web may lead to compatibility issues of applications built for the exploitation of only one of them or it could lead to inconsistencies accumulating in the links between entities in both hubs. Therefore, the data quality in DBpedia and in Wikidata needs to be evaluated both as a whole and independently of each other, which corresponds to the approach chosen in this thesis.

Given the scale of both DBpedia and Wikidata though, it is necessary to restrict the scope of the research so that it can finish in a short enough timespan that the findings would still be useful for acting upon them. In this thesis, the analysis of datasets linking to DBpedia is done over linguistic linked data and general cross-domain data, while the analysis of the consistency of DBpedia and Wikidata focuses on bibliographic data representation of artwork.

## 1.1 Goals

The goals of this thesis are twofold. Firstly, the research focuses on the interlinking of various LOD datasets that are interlinked with DBpedia, evaluating several data quality features. Then the research shifts its focus to the analysis of artwork entities in Wikidata and the way DBpedia entities are interlinked with them. The goals themselves are to:

1. Quantitatively analyse the connectivity of linked open datasets with DBpedia using the public endpoint.
2. Study in depth the semantics of a specific kind of entities (artwork), analyse the internal consistency of Wikidata and the consistency of interlinking of DBpedia with Wikidata regarding the semantics of artwork entities and develop an empirical model allowing to predict the variants of this semantics based on the associated links.

## 1.2 Structure of the thesis

The first part of the thesis introduces the concepts in section 2 that are needed for the understanding of the rest of the text: Semantic Web, Linked Data, Data quality, knowledge representations in use on the Semantic Web, interlinking and two important ontologies (OWL and SKOS). The second part, which consists of section 3, describes how the goal to analyse the quality of interlinking between various sources of linked open data and DBpedia was tackled.

The third part focuses on the analysis of consistency of bibliographic data in encyclopaedic datasets. This part is divided into two smaller tasks, the first one being the analysis of typing of Wikidata entities modelled accordingly to the Functional Requirements for Bibliographic Records (FRBR) in subsection 4.1 and the second task being the analysis of consistency of interlinking between DBpedia entities and Wikidata entries from the FRBR domain in subsections 4.2 and 4.3.

The last part, which consists of section 5, aims to demonstrate the importance of knowing about data quality issues in different segments of the chain of interlinked datasets (in this case it can be depicted as: $various\ LOD\ datasets \rightarrow DBpedia \rightarrow Wikidata$) by formulating a couple of examples, where an otherwise useful application or its feature may misbehave due to low quality of data with consequences of varying levels of severity.

A by-product of the research conducted as part of this thesis is the Annotator for FRBR on DBpedia, an application developed for the purpose of enabling the analysis of consistency of interlinking between DBpedia and Wikidata by providing FRBR information about DBpedia resources, which is described in Annex B.

# 2 Research topic background

This section explains the concepts relevant to the research conducted as part of this thesis.

## 2.1 Semantic Web

The World Wide Web Consortium (W3C) is the organization standardizing technologies used to build the World Wide Web (WWW). In addition to helping with the development of the classic Web of documents, W3C is also helping build the Web of linked data, known as the Semantic Web, to enable computers to do useful work that leverages the structure given to the data by vocabularies and ontologies, as implied by the vision of W3C. The most important parts of the W3C's vision of the Semantic Web is the interlinking of data, which leads to the concept of Linked Data (LD), and machine-readability, which is achieved through the definition of vocabularies that define the semantics of the properties used to assert facts about entities described by the data.[1]

## 2.2 Linked Data

According to the explanation of linked data by W3C, the standardizing organisation behind the web, the essence of LD lies in making relationships between entities in different datasets explicit so that the Semantic Web becomes more than just a collection of isolated datasets that use a common format.[2]

LD tackles several issues with publishing data on the web at once according to the publication of Heath & Bizer (2011):

- The structure of HTML makes the extraction of data complicated and dependent on text mining techniques which are error prone due to the ambiguity of natural language.
- Microformats have been invented to embed data in HTML pages in a standardized and unambiguous manner. Their weakness lies in their specificity to a small set of types of entities and in that they often do not allow modelling relationships between entities.
- Another way of serving structured data on the web are Web APIs, which are more generic than microformats in that there is practically no restriction on how the provided data is modelled. There are, however, two issues, both of which increase the effort needed to integrate data from multiple providers:
    - the specialized nature of web APIs and

[1] Introduction of Semantic Web by W3C: https://www.w3.org/standards/semanticweb/
[2] Introduction of Linked Data by W3C: https://www.w3.org/standards/semanticweb/data

- local only scope of identifiers for entities, preventing the integration of multiple sources of data.

In LD, however, these issues are resolved by the Resource Description Framework (RDF) language as demonstrated by the work of Heath & Bizer (2011). The RDF Primer, authored by Manola & Miller (2004), specifies the foundations of the Semantic Web, the building blocks of RDF datasets, called triples, because they are composed of three parts that always occur as part of at least one triple. The triples are composed of a *subject*, a *predicate* and an *object* which gives RDF the flexibility to represent anything, unlike microformats, while at the same time ensuring that the data is modelled unambiguously. The problem of identifiers with local scope is alleviated by RDF as well because it is encouraged to use any Uniform Resource Identifier (URI), which also includes the possibility to use an Internationalized Resource Identifier (IRI), for each entity.

### 2.2.1 Uniform Resource Identifier

The specification of what constitutes a URI is written in RFC 3986 (see Berners-Lee et al. 2005) and it is described in the rest of part 2.2.1.

A URI is a string which adheres to the specification of URI syntax. It is designed to be a simple yet extensible identifier of resources. The specification of a generic URI does not provide any guidance as to how the resource may be accessed, because that part is governed by more specific schemas such as HTTP URIs. This is the strength of uniformity. The specification of a URI also does not specify what a resource may be – a URI can identify an electronic document available on the web as well as a physical object or a service (e.g. HTTP-to-SMS gateway). A URIs purpose is to distinguish a resource from all other resources and it is irrelevant how exactly it is done, whether the resources are distinguishable by names, addresses, identification numbers or from context.

In the most general form, a URI has the form specified like this:

```
URI = scheme ":" hier-part [ "?" query ] [ "#" fragment ]
```

Various URI schemes can add more information similarly to how HTTP scheme splits the hier-part into parts authority and path, where authority specifies the server holding the resource and path specifies the location of the resource on that server.

### 2.2.2 Internationalized Resource Identifier

The IRI is specified in RFC 3987 (see Duerst et al., 2005). The specification is described in the rest of the part 2.2.2 in a similar manner to how the concept of a URI was described earlier.

A URI is limited to a subset of US-ASCII characters. URIs are widely incorporating words of natural languages to help people with tasks such as memorization, transcription, interpretation and guessing of URIs. This is the reason why URIs were extended into IRIs by creating a specification that allows the use of non-ASCII characters. The IRI specification was also designed to be backwards compatible with the older specification of a URI through

a mapping of characters not present in the Latin alphabet by what is called percent encoding, a standard feature of the URI specification used for encoding reserved characters.

An IRI is defined similarly to a URI:

```
IRI = scheme ":" ihier-part [ "?" iquery ] [ "#" ifragment ]
```

The reason why IRIs are not defined solely through their transformation to a corresponding URI is to allow for direct processing of IRIs.

### 2.2.3 List of prefixes

Some RDF serializations (e.g. *Turtle*) offer a standard mechanism for shortening URIs, by defining a prefix. This feature makes the serializations that support it more understandable to humans and helps with manual creation and modification of RDF data. Several common prefixes are used in this thesis to illustrate the results of the underlying research and the prefix are thus listed below.

```
PREFIX dbo:  <http://dbpedia.org/ontology/>
PREFIX dc:   <http://purl.org/dc/terms/>
PREFIX owl:  <http://www.w3.org/2002/07/owl#>
PREFIX rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX wd:   <http://www.wikidata.org/entity/>
PREFIX wdt:  <http://www.wikidata.org/prop/direct/>
PREFIX wdrs: <http://www.w3.org/2007/05/powder-s#>
PREFIX xhv:  <http://www.w3.org/1999/xhtml/vocab#>
```

## 2.3 Linked Open Data

Linked Open Data (LOD) are LD that are published using an open license. Hausenblas described the system for ranking Open Data (OD) based on the format they are published in, which is called 5-star data (Hausenblas, 2012). One star is given to any data published using an open license regardless of the format (even a PDF is sufficient for that). To gain more stars, it is required to publish data in formats that are (in this order from two stars to five stars): machine-readable, non-proprietary, standardized by W3C, linked with other datasets.

## 2.4 Functional Requirements for Bibliographic Records

The FRBR is a framework developed by the International Federation of Library Associations and Institutions (IFLA). The relevant materials have been published by the IFLA Study Group (1998), the development of FRBR was motivated by the need for increased effectiveness in the handling of bibliographic data due to the emergence of automation,

electronic publishing, networked access to information resources and economic pressure on libraries. It was agreed upon that the viability of shared cataloguing programs as a means to improve effectiveness requires a shared conceptualization of bibliographic records based on the re-examination of the individual data elements in the records in the context of the needs of the users of bibliographic records. The study proposed the FRBR framework consisting of three groups of entities:

1. Entities that represent records about the intellectual or artistic creations themselves belong to either of these classes:
    - work,
    - expression,
    - manifestation or
    - item.
2. Entities responsible for the creation of artistic or intellectual content are either:
    - a person or
    - a corporate body.
3. Entities that represent subjects of works can be either members of the two previous groups or one of these additional classes:
    - concept,
    - object,
    - event,
    - place.

To disambiguate the meaning of the term subject, all occurrences of this term outside this subsection dedicated to the definitions of FRBR terms will have the meaning from the linked data domain as described in section 2.2, which covers the LD terminology.

### 2.4.1 Work

IFLA Study Group (1998) defines a work is an abstract entity which represents the idea behind all its realizations. It is realized through one or more expressions. Modifications to the form of the work are not classified as works, but rather as expressions of the original work they are derived from. This includes revisions, translations, dubbed or subtitled films and musical compositions modified for new accompaniments.

### 2.4.2 Expression

IFLA Study Group (1998) defines an expression is a realization of a work which excludes all aspects of its physical form that are not a part of what defines the work itself as such. An expression would thus encompass the specific words of a text or notes that constitute a musical work, but not characteristics such as the typeface or page layout. This means that every revision or modification to the text itself results in a new expression.

### 2.4.3 Manifestation

IFLA Study Group (1998) defines a manifestation is the physical embodiment of an expression of a work which defines the characteristics that all exemplars of the series should possess, although there is no guarantee that every exemplar of a manifestation has all these characteristics. An entity may also be a manifestation even if it has only been produced once with no intention for another entity belonging to the same series (e.g. author's manuscript). Changes to the physical form that do not affect the intellectual or artistic content (e.g. change of the physical medium) results in a new manifestation of an existing expression. If the content itself is modified in the production process, the result is considered as a new manifestation of a new expression.

### 2.4.4 Item

IFLA Study Group (1998) defines an item as an exemplar of a manifestation. The typical example is a single copy of an edition of a book. A FRBR item can, however, consist of more physical objects (e.g. a multi-volume monograph). It is also notable that multiple items that exemplify the same manifestation may, however, be different in some regards due to additional changes after they were produced. Such changes may be deliberate (e.g. bindings by a library) or not (e.g. damage).

## 2.5 Data quality

According to article *The Evolution of Data Quality: Understanding the Transdisciplinary Origins of Data Quality Concepts and Approaches* (see Keller et al., 2017), data quality has become an area of interest in 1940s and 1950s with Edward Deming's Total Quality Management, which heavily relied on statistical analysis of measurements of inputs. The article differentiates three different kinds of data based on their origin. They are designed data, administrative data, and opportunistic data. The differences are mostly in how well the data can be reused outside of its intended use case, which is based on the level of understanding of the structure of data. As it is defined, the designed data contains the highest level of structure, while opportunistic data (e.g. data collected from web crawlers or a variety of sensors) may provide very little structure, but compensate for it by abundance of datapoints. Administrative data would be somewhere between the two extremes, but its structure may not be suitable for analytic tasks.

The main points of view from which data quality can be examined are those of the two involved parties – the data owner (or publisher) and the data consumer according to the work of Wang & Strong (1996). It appears that the perspective of the consumer on data quality has started gaining attention during the 1990s. The main differences in the views lies in the criteria that are important to different stakeholders. While the data owner is mostly concerned about the accuracy of the data, the consumer has a whole hierarchy of criteria that determine the fitness for use of the data. Wang & Strong have also formulated how the criteria of data quality can be categorized:

- accuracy of data, which includes the data owner's perception of quality, but also other parameters like objectivity, completeness, and reputation,
- relevancy of data, which covers mainly the appropriateness of the data and its amount for a given purpose, but also its time dimension,
- representation of data, which revolves around the understandability of data and its underlying schema and
- accessibility of data, which includes for example cost and security considerations.

### 2.5.1 Data quality of Linked Open Data

It appears that data quality of LOD has started being noticed rather recently since most progress on this front has been done within the second half of the last decade. One of the earlier papers dealing with data quality issues of the Semantic Web authored by Fürber & Hepp was trying to build a vocabulary for data quality management on the Semantic Web (2011). At first, it produced a set of rules in the SPARQL Inferencing Notation (SPIN) language, a predecessor to Shapes Constraint Language (SHACL) specified in 2017. Both SPIN and SHACL were designed for describing dynamic computational behaviour, which contrasts with languages created for describing static structure of data like the Simple Knowledge Organization System (SKOS), RDF Schema (RDFS) and OWL as described by Knublauch et al. (2011) and Knublauch & Kontokostas (2017) for SPIN and SHACL respectively.

Fürber & Hepp (2011) released the data quality vocabulary at http://semwebquality.org/, as they indicated in their publication later on as well as the SPIN rules that were completed earlier. Additionally, at http://semwebquality.org/, Fürber (2011) explains the foundations of both the rules and the vocabulary. They have been laid by the empirical study conducted by Wang & Strong in 1996. According to that explanation, of the original twenty criteria, five have been dropped for the purposes of the vocabulary, but the groups into which they were organized were kept under new category names: intrinsic, contextual, representational and accessibility.

The vocabulary developed by Albertoni & Isaac and standardized by W3C (2016) that models data quality of datasets is also worth mentioning. It relies on the structure given to the dataset by *The RDF Data Cube Vocabulary* and the *Data Catalog Vocabulary* with the *Dublin Core Metadata Initiative* used for linking to standards that the datasets adhere to.

Tomčová also mentions in her master thesis (2014) dedicated to the data quality of open and linked data the lack of publications regarding LOD data quality and also the quality of OD in general with the exception of the Data Quality Act and an (at that time) ongoing project of the Open Knowledge Foundation. She proposed a set of data quality dimensions specific for LOD and synthesized another set of dimensions that are not specific to LOD, but that can nevertheless be applied to LOD. The main reason for using the dimensions proposed by her thus was that those remaining dimensions were either designed for this kind of data that is dealt with in this thesis or were found to be applicable for it. The translation of her results is presented as Table 1.

### 2.5.2 Data quality dimensions

With regards to Table 1 and the scope of this work the following data quality features, which represent several points of view from which datasets can be evaluated, have been chosen for further analysis:

- *accessibility* of datasets, which has been extended to partially include the *versatility* of those datasets through the analysis of access mechanisms,
- *uniqueness* of entities that are linked to DBpedia measured both in absolute numbers of affected entities or concepts and relatively to the number of entities and concepts interlinked with DBpedia,
- *consistency* of typing of FRBR entities in DBpedia and Wikidata,
- *consistency* of interlinking of entities and concepts in datasets interlinked with DBpedia measured in both absolute numbers and relatively to the number of interlinked entities and concepts,
- *currency* of the data in datasets that link to DBpedia.

The analysis of the accessibility of datasets was required to enable the evaluation of all the other data quality features and therefore had to be carried out. The need to assess the currency of datasets became apparent during the analysis of accessibility, because of a rather large portion of datasets that are only available through archives which called for a closer investigation of the recency of the data. Finally, the uniqueness and consistency of interlinked entities were found to be an issue during the exploratory data analysis further described in section 3.

Additionally, the consistency of typing of FRBR entities in Wikidata and DBpedia has been evaluated to provide some insight into the influence of hybrid knowledge representation consisting of an ontology and a knowledge graph on the data quality of Wikidata and the quality of interlinking between DBpedia and Wikidata.

Features of data quality based on the other data quality dimensions were not evaluated mostly because of the need for either extensive domain knowledge of each dataset (e.g. accuracy, completeness), administrative access to the server (e.g. access security), or a large scale survey among users of the datasets (e.g. relevancy, credibility, value-added).

Table 1: Data quality dimensions (source: (Tomčová, 2014) – compiled from multiple original tables and translated)

| Kind of data | Dimension | Consolidated definition | Example of measurement | Frequency |
|---|---|---|---|---|
| General data | Accuracy, Free-of-error, Semantic accuracy, Correctness | Data must precisely capture real-world objects. | Ratio of values that fit the rules for a correct value | 11 |
| General data | Completeness | A measure of how much of the requested data is present. | The ratio of the number of existing and requested records. | 10 |
| General data | Validity, Conformity, Syntactic accuracy | A measure of how much the data adheres to the syntactical rules. | The ratio of syntactically valid values to all the values | 7 |
| General data | Timeliness | A measure of how well the data represent the reality at a certain point in time. | The time difference between the time the fact is applicable from and the time when it was added to the dataset. | 6 |
| General data | Accessibility, Availability | A measure of how easy it is for the user to access the data. | Time to response. | 5 |
| General data | Consistency, Integrity | Data capturing the same parts of reality must be consistent across datasets. | The ratio of records consistent with a referential dataset. | 4 |
| General data | Relevancy, Appropriateness | A measure of how well the data align with the needs of the users. | A survey among users. | 4 |
| General data | Uniqueness, Duplication | No object or fact should be duplicated. | The ratio of unique entities. | 3 |
| General data | Interpretability | A measure of how clearly the data is defined and to which it is possible to understand their meaning. | The usage of relevant language, symbols, units, and clear definitions for the data. | 3 |
| General data | Reliability | The data is reliable if the process of data collection and processing is defined. | Process walkthrough. | 3 |

| Kind of data | Dimension | Consolidated definition | Example of measurement | Frequency |
|---|---|---|---|---|
| General data | Believability | A measure of how generally acceptable the data is among its users. | A survey among users. | 3 |
| General data | Access security, Security | A measure of access security. | The ratio of unauthorized access to the values of an attribute. | 3 |
| General data | Ease of understanding, Understandability, Intelligibility | A measure of how comprehensible the data is to its users. | A survey among users. | 3 |
| General data | Reputation, Credibility, Trust, Authoritative | A measure of reputation of the data source or provider. | A survey among users. | 2 |
| General data | Objectivity | The degree to which the data is considered impartial. | A survey among users. | 2 |
| General data | Representational consistency, Consistent representation | The degree to which the data is published in the same format. | Comparison with a referential data source. | 2 |
| General data | Value-added | The degree to which the data provides value for specific actions. | A survey among users. | 2 |
| General data | Appropriate amount of data | A measure of whether the volume of data is appropriate for the defined goal. | A survey among users. | 2 |
| General data | Concise representation, Representational conciseness | The degree to which the data is appropriately represented with regards to its format, aesthetics, and layout. | A survey among users. | 2 |
| General data | Currency | The degree to which the data is out-dated. | The ratio of out-dated values at a certain point in time. | 1 |
| General data | Synchronization between different time series | A measure of synchronization between different timestamped data sources. | The difference between the time of last modification and last access. | 1 |

| Kind of data | Dimension | Consolidated definition | Example of measurement | Frequency |
|---|---|---|---|---|
| General data | Precision, Modelling granularity | The data is detailed enough. | A survey among users. | 1 |
| General data | Confidentiality | Customers can be assured that the data is processed with confidentiality in mind that is defined by legislation. | Process walkthrough. | 1 |
| General data | Volatility | The weight based on the frequency of changes in the real-world. | Average duration of an attribute's validity. | 1 |
| General data | Compliance, Conformance | The degree to which the data is compliant with legislation or standards. | The number of incidents caused by non-compliance with legislation or other standards. | 1 |
| General data | Ease of manipulation | It is possible to easily process and use the data for various purposes. | A survey among users. | 1 |
| OD | Licensing, Licensed | The data is published under a suitable license. | Is the license suitable for the data? | - |
| OD | Primary | The degree to which the data is published as it was created. | Checksums of aggregated statistical data. | - |
| OD | Processability | The degree to which the data is comprehensible and automatically processable. | The ratio of data that is available in a machine-readable format. | - |
| LOD | History | The degree to which the history of changes is represented in the data. | Are there recorded changes to the data alongside the person who made them? | - |
| LOD | Isomorphism | A measure of consistency of models of different datasets during the merge of those datasets. | Evaluation of compatibility of individual models and the merged models. | - |

| Kind of data | Dimension | Consolidated definition | Example of measurement | Frequency |
|---|---|---|---|---|
| LOD | Typing | Are nodes correctly semantically described or are they only labelled by a datatype?<br>This improves the search and query capabilities. | The ratio of incorrectly typed nodes (e.g. typos). | - |
| LOD | Boundedness | The degree to which the dataset contains irrelevant data. | The ratio of out-dated, undue, or incorrect data in the dataset. | - |
| LOD | Attribution | The degree to which the user can assess the correctness and origin of the data. | The presence of information about the author, contributors, and the publisher in the dataset. | - |
| LOD | Interlinking, Connectedness | The degree to which the data is interlinked with external data and to which such interlinking is correct. | The existence of links to external data (through the usage of external URIs within the dataset). | - |
| LOD | Directionality | The degree of consistency when navigating the dataset based on relationships between entities. | Evaluation of the model and the relationships it defines. | - |
| LOD | Modelling correctness | Determines to what degree the data model is logically structured to represent the reality. | Evaluation of the structure of the model. | - |
| LOD | Sustainable | A measure of future provable maintenance of the data. | Is there a premise that the data will be maintained in the future? | - |

| Kind of data | Dimension | Consolidated definition | Example of measurement | Frequency |
|---|---|---|---|---|
| LOD | Versatility | The degree to which the data is potentially universally usable. (e.g. The data is multi-lingual, it is represented in a format not specific to any locale, there are multiple access mechanisms.) | Evaluation of access mechanisms to retrieve the data. (e.g. RDF dump, SPARQL endpoint) | - |
| LOD | Performance | The degree to which the data provider's system is efficient and how efficiently can large datasets be processed. | Time to response from the data provider's server. | - |

## 2.6 Hybrid knowledge representation on the Semantic Web

This thesis, being focused on the data quality aspects of interlinking datasets with DBpedia, must consider different ways in which knowledge is represented on the Semantic Web. The definitions of various knowledge representation (KR) techniques have been agreed upon by participants of the Internal Grant Competition (IGC) project: *Hybrid modelling of concepts on the semantic web: ontological schemas, code lists and knowledge graphs* (HYBRID). The three kinds of KR in use on the semantic web are:

- ontologies (ON),
- knowledge graphs (KG) and
- code lists (CL).

The shared understanding of what constitutes which kinds of knowledge representation has been written down by Nguyen (2019) in an internal document for the IGC project. Each of the knowledge representations can be used independently or in a combination with another one (e.g. KG-ON) as portrayed in Figure 1. The various combinations of knowledge, often including an engine, API or UI to provide support, are called knowledge bases (KB).

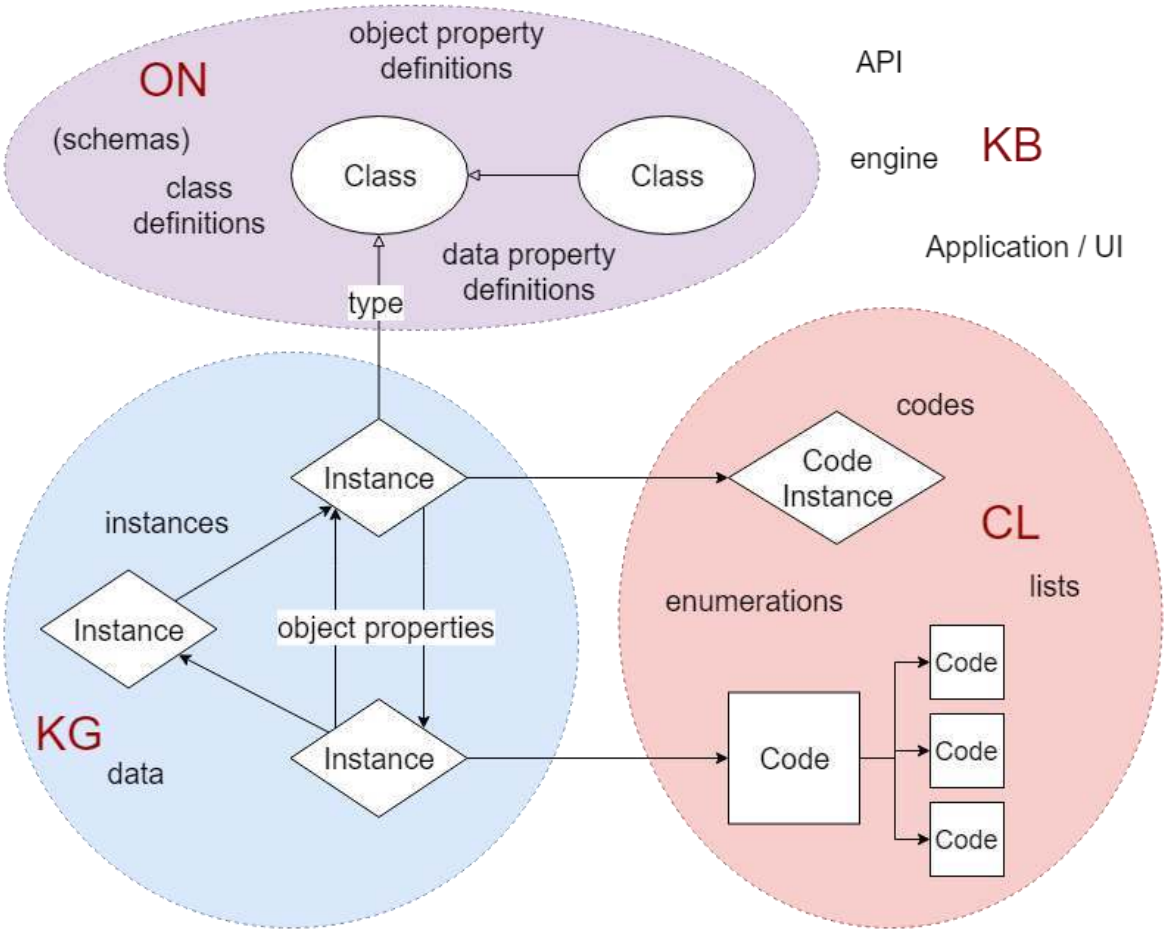


Figure 1: Hybrid modelling of concepts on the semantic web (source: (Nguyen, 2019))

Given that one of the goals of this thesis is to analyse the consistency of Wikidata and DBpedia with regards to artwork entities, it was necessary to accommodate the fact that both Wikidata and DBpedia are hybrid knowledge bases of the type KG-ON.

Because Wikidata is composed of a knowledge graph and an ontology, the analysis of the internal consistency of its representation of FRBR entities is necessarily an analysis of the interlinking of two separate datasets that utilize two different knowledge representations. The analysis relies on the typing of Wikidata entities (the assignment of instances to classes) and the attachment of properties to entities regardless of whether they are object or datatype properties.

The analysis of interlinking consistency in the domain of artwork with regards to FRBR typing between DBpedia and Wikidata is essentially the analysis of two hybrid knowledge bases, where the properties and typing of entities in both datasets provide vital information about how well the interlinked instances correspond to each other.

The subsection that explains the relationship between FRBR and Wikidata classes is 4.1. The representation (or more precisely the lack of representation) of FRBR in DBpedia ontology is described in subsection 4.2 which contains subsection 4.3 that offers a way to overcome the lack of representation of FRBR in DBpedia.

The analysis of the usage of code lists in DBpedia and Wikidata has not been conducted during this research, because code lists are not expected in DBpedia or Wikidata due to the difficulties associated with enumerating certain entities in such vast and gradually evolving datasets.

### 2.6.1 Ontology

The internal document (2019) for the IGC HYBRID project defines an ontology as a formal representation of knowledge and a shared conceptualization used in some domain of interest. It also specifies the requirements a knowledge base must fulfil to be considered an ontology:

- it is defined in a formal language such as the Web Ontology Language (OWL),
- it is limited in scope to a certain domain and some community that agrees with its conceptualization of that domain,
- it consists of a set of classes, relations, instances, attributes, rules, restrictions, and meta-information,
- its rigorous dynamic and hierarchical structure of concepts enables inference and
- it serves as a data model that provides context and semantics to the data.

### 2.6.2 Code list

The internal document (2019) recognizes the code lists as such lists of values from a domain that aim to enhance consistency and help to avoid errors by offering an enumeration of a predefined set of values so that they can then be linked to from knowledge graphs or

ontologies. As noted in *Guidelines for the Use of Code Lists* (see Dekkers et al., 2018), code lists used on the Semantic Web are also often called controlled vocabularies.

### 2.6.3 Knowledge graph

According to the shared understanding of the concepts described by the internal document supporting IGC HYBRID project (2019), the concept of knowledge graph was first used by Google but has since then spread around the world and that multiple definitions of what constitutes a knowledge graph exist alongside each other. The definitions of the concept of knowledge graph are these (Ehrlinger & Wös, 2016):

1. *"A knowledge graph (i) mainly describes real world entities and their interrelations, organized in a graph, (ii) defines possible classes and relations of entities in a schema, (iii) allows for potentially interrelating arbitrary entities with each other and (iv) covers various topical domains."*

2. *"Knowledge graphs are large networks of entities, their semantic types, properties, and relationships between entities."*

3. *"Knowledge graphs could be envisaged as a network of all kind things which are relevant to a specific domain or to an organization. They are not limited to abstract concepts and relations but can also contain instances of things like documents and datasets."*

4. *"We define a Knowledge Graph as an RDF graph. An RDF graph consists of a set of RDF triples where each RDF triple (s, p, o) is an ordered set of the following RDF terms: a subject $s \in U \cup B$, a predicate $p \in U$, and an object $U \cup B \cup L$. An RDF term is either a URI $u \in U$, a blank node $b \in B$, or a literal $l \in L$."*

5. *"[…] systems exist, […], which use a variety of techniques to extract new knowledge, in the form of facts, from the web. These facts are interrelated, and hence, recently this extracted knowledge has been referred to as a knowledge graph."*

The most suitable definition of a knowledge graph for this thesis is the 4th definition, which is focused on LD and is compatible with the view described graphically by Figure 1.

## 2.7 Interlinking on the Semantic Web

The fundamental foundation of LD is the ability of data publishers to create links between data sources and the ability of clients to follow the links across datasets to obtain more data. It is important for this thesis to discern two different aspects of interlinking, which may affect data quality either on their own or in a combination of those aspects.

Firstly, there is the semantics of various predicates which may be used for interlinking which is dealt with in part 2.7.1 of this subsection. The second aspect is the process of creation of links between datasets as described in part 2.7.2.

Given the information gathered from studying the semantics of predicates used for interlinking and the process of interlinking itself, it is clear that there is a possibility to trade-off well defined semantics to make the interlinking task easier by choosing a less reliable process or vice versa. In either case the richness of the LOD cloud would increase, but each of those situations would pose a different challenge to application developers that would want to exploit that richness.

### 2.7.1 Semantics of predicates used for interlinking

Although there are no constraints on which predicates may be used to interlink resource, there are several common patterns. The predicates commonly used for interlinking are revealed in *Linking patterns* (Faronov, 2011) and *How to Publish Linked Data on the Web* (Bizer et al., 2008). Two groups of predicates used for interlinking have been identified in the sources. Those that may be used across domains, which are more important for this work, because they were encountered in the analysis in a lot more cases then the other group of predicates, are:

- `owl:sameAs`, which asserts identity of the resources identified by two different URIs. Because of the importance of OWL for interlinking, there is a more thorough explanation of it in subsection 2.8,
- `rdfs:seeAlso`, which does not have the semantic implications of the `owl:sameAs` predicate and therefore does not suffer from data quality concerns over consistency to the same degree,
- `rdfs:isDefinedBy` states that the subject (e.g. a concept) is defined by object (e.g. an organization),
- `wdrs:describedBy` from the Protocol for Web Description Resources (POWDER) ontology is intended for linking instance-level resources to their descriptions,
- `xhv:prev`, `xhv:next`, `xhv:section`, `xhv:first` and `xhv:last` are examples of predicates specified by the XHTML+RDFa vocabulary that can be used for any kind of resource,
- `dc:format` is a property defined by Dublin Core Metadata Initiative to specify the format of a resource in advance to help applications achieve higher efficiency by not having to retrieve resources that they cannot process,
- `rdf:type` to reuse commonly accepted vocabularies or ontologies and
- a variety of Simple Knowledge Organization System (SKOS) properties, which is described in more detail in subsection 2.9 because of its importance for datasets interlinked with DBpedia.

The other group of predicates is tightly bound to the domain, which they were created for. While both Friend of a Friend (FOAF) and DBpedia properties occasionally appeared in the interlinking between datasets, they were not used on a significant enough number of entities to warrant further analysis. The FOAF properties commonly used for interlinking are: `foaf:page`, `foaf:homepage`, `foaf:knows`, `foaf:based_near` and `foaf:topic_interest` are used for describing resources that represent people or organizations.

Heath & Bizer (2011) highlight the importance of using commonly accepted terms to link to other datasets, and for cases when it is necessary to link to another dataset by a specific or

proprietary term they recommend that it is at least defined as a `rdfs:subPropertyOf` of a more common term.

The following questions can help when publishing LD (Heath & Bizer, 2011):

1. *"How widely is the predicate already used for linking by other data sources?"*
2. *"Is the vocabulary well maintained and properly published with dereferenceable URIs?"*

### 2.7.2 Process of interlinking

The choices available for interlinking of datasets are well described in the paper: *Automatic Interlinking of Music Datasets on the Semantic Web* (Raimond, et al., 2008). According to that the first choice when deciding to interlink a dataset with other data sources is the choice between a manual and an automatic process. The manual method of creating links between datasets is said to be practical only at a small scale such as for a FOAF file.

For the automatic interlinking, there are essentially two approaches.

- The naïve approach which assumes that datasets that contain data about the same entity describe that entity using the same literal and it therefore creates links between resources based on the equivalence (or more generally the similarity) of their respective text descriptions.
- The graph matching algorithm at first finds all triples in both graphs $D_1$ and $D_2$ with predicates used by both graphs such that $(s_1, p, o_1) \in D_1$ and $(s_2, p, o_2) \in D_2$.
  After that, all possible mappings $(s_1, s_2)$ and $(o_1, o_2)$ are generated and a simple similarity measure is computed similarly to the naïve approach.
  In the end, the final graph similarity measure is the sum of simple similarity measures across the set of possible pair mappings where the first resource in the mapping is the same which is then normalized by the number of such pairs. This is more formally described by formula (2.7.2.1).

$$graphSimilarity = \sum_{s_1} \frac{similarity(s_1, s_2)}{count(s_1, s_2)} \qquad (2.7.2.1)$$

## 2.8 Web Ontology Language

The language is specified by the document *OWL 2 Web Ontology Language* (see Hitzler et al., 2012). It is a language that was designed to take advantage of the description logics to model some part of the world. Because it is based on formal logic, it can be used to infer knowledge implicitly present in the data (e.g. in a knowledge graph) and make it explicit. It is, however, necessary to understand that an ontology is not a schema and cannot be used for defining integrity constraints unlike an XML Schema or database structure.

In the specification, Hitzler et al., state that in OWL, the basic building blocks are axioms, entities, and expressions. Axioms represent the statements that can be either true or false

and the whole ontology can be regarded as a set of axioms. The entities represent the real-world objects that are described by axioms. There are three kinds of entities: objects (individuals), categories (classes) and relations (properties). In addition, entities can also be defined by expressions (e.g. a complex entity may be defined by a conjunction of at least two different simpler entities).

The specification written by Hitzler et al. also says that when some data is collected and the entities described by that data are typed appropriately to conform to the ontology, the axioms can be used to infer valuable knowledge about the domain of interest.

Especially important for this thesis is the way the `owl:sameAs` predicate is treated by reasoners, because of its widespread use in interlinking. The DBpedia knowledge graph, which is central to the analysis this thesis is about, is mostly interlinked using `owl:sameAs` links and thus needs to be understood in depth, which can be achieved by studying the article *Web of Data and Web of Entities: Identity and Reference in Interlinked Data in the Semantic Web* (Bouquet et al., 2012). It is intended to specify individuals that share the same identity. The implications of this in practice are that the URIs that denote the underlying resource can be used interchangeably, which makes the `owl:sameAs` predicate comparatively more likely to cause problems due to issues with the process of link creation.

## 2.9 Simple Knowledge Organization System

The authoritative source for SKOS is the specification *SKOS Simple Knowledge Organization System Reference* (Miles & Bechhofer, 2009), according to which SKOS aims to stimulate the exchange of data representing the organization of collections of objects such as books or museum artifacts. These collections have been created and organized by librarians and information scientists using a variety of knowledge organization systems including thesauri, classification schemes and taxonomies.

With regards to RDFS and OWL which provide a way to express meaning of concepts through a formally defined language, Miles & Bechhofer imply that SKOS is meant to construct a detailed map of concepts over large bodies of especially unstructured information, which is not possible to carry out automatically.

The specification of SKOS by Miles & Bechhofer continues by specifying that the various knowledge organization systems are called concept schemes. They are essentially sets of concepts. Because SKOS is a LD technology, both concepts and concept schemes are identified by URIs. SKOS allows:

- the labelling of concepts using preferred and alternative labels to provide human-readable descriptions,
- the linking of SKOS concepts via semantic relation properties,
- the mapping of SKOS concepts across multiple concept schemes,
- the creation of collections of concepts which can be labelled or ordered for situations where the order of concepts can provide meaningful information,

- the use of various notations for compatibility with already in use computer systems and library catalogues and
- the documentation with various kinds of notes (e.g. supporting scope notes, definitions, and editorial notes).

The main difference between SKOS and OWL with regards to knowledge representation as implied by Miles & Bechhofer in the specification is that SKOS defines relations at the instance level while OWL models relations between classes which are only subsequently used to infer properties of instances.

From the perspective of hybrid knowledge representations as depicted in Figure 1, SKOS is an OWL ontology which describes structure of data in a knowledge graph, possibly using a code list defined through means provided by SKOS itself. Therefore, any SKOS vocabulary is necessarily a hybrid knowledge representation of either type KG-ON or KG-ON-CL.

# 3 Analysis of interlinking towards DBpedia

This section demonstrates the approach to tackling the second goal (to quantitatively analyse the connectivity of DBpedia with other RDF datasets).

Linking across datasets using RDF is done by including a triple in the source dataset such that its subject is an IRI from the source dataset and the object is an IRI from the target dataset. This makes the outgoing links readily available while the incoming links are only revealed through crawling the semantic web, much like how this works on the WWW.

The options for discovering incoming links to a dataset include:

- the LOD cloud's information pages about datasets (for example information page for DBpedia: https://lod-cloud.net/dataset/dbpedia),
- DataHub (https://datahub.io/) and
- specifically for DBpedia, its wiki page about interlinking, which features a list of datasets that are known to link to DBpedia (https://wiki.dbpedia.org/services-resources/interlinking).

The LOD cloud and DataHub are likely to contain more recent data in comparison with a wiki page that does not even provide information about the date when it was last modified, but both sources would need to be scraped from the web. This would be an unnecessary overhead for the purpose of this project. In addition, the links from the wiki page can be verified, the datasets themselves can be found by other means including the Google Dataset Search (https://datasetsearch.research.google.com/), assessed based on their recency if it is possible to obtain such information as date of last modification, and possibly corrected at the source.

## 3.1 Method

The research of the quality of interlinking between LOD sources and DBpedia relies on quantitative analysis, which can take the form of either confirmation data analysis (CDA) or exploratory data analysis (EDA).

The paper *Data visualization in exploratory data analysis: An overview of methods and technologies*, Mao (2015) formulates the limitations of the CDA, known as statistical hypothesis testing. Namely the fact that the analyst must:

1. understand the data and
2. be able to form a hypothesis beforehand based on his knowledge of the data.

This approach is not applicable when the data to be analysed is scattered across many datasets which do not have a common underlying schema, which would allow the researcher to define what should be tested for.

This variety of data modelling techniques in the analysed datasets justifies the use of EDA as suggested by Mao in an interactive setting with the goal to better understand the data and to extract knowledge about linking data between the analysed datasets and DBpedia.

The tools chosen to perform the EDA is Microsoft Excel because of its familiarity and the existence of an opensource plugin named RDFExcelIO with source code available on Github at https://github.com/Fuchs-David/RDFExcelIO, developed by the author of this thesis (Fuchs, 2018) as part of his Bachelor's thesis for the conversion of RDF data to Excel for the purpose of performing interactive exploratory analysis of LOD.

## 3.2 Data collection

As mentioned in the introduction to section 3, the chosen source for discovering datasets containing links to DBpedia resources is DBpedia's wiki page dedicated to interlinking information.

Table 10 presented in Annex A is the original table of interlinked datasets. Because not all links in the table led to functional websites, it was augmented with further information collected by searching the web for traces leading to those datasets as captured in Table 11 in Annex A as well. Table 2 displays the eleven datasets to present concisely the structure of Table 11. The example datasets are those that contain over 100,000 links to DBpedia. The meaning of the columns added to the original table is described on the following lines:

- data source URL, which may differ from the original one if the dataset was found by alternative means,
- availability flag indicating if the data is available for download,
- data source type to provide information about how the data can be retrieved,
- date when the examination was carried out,
- alternative access method for datasets that are no longer available on the same server[3],
- the DBpedia inlinks flag to indicate if any links from the dataset to DBpedia were found and
- last modified field for the evaluation of recency of data in datasets that link to DBpedia.

The relatively high number of datasets that are no longer available, but whose data is, thanks to the existence of the Internet Archive (https://archive.org/), led to the addition of *last modified* field in an attempt to map the recency[4] of data as it is one of the factors of data quality. According to Table 6, the most up to date datasets have been modified during the year 2019, which is also the year when the dataset availability and the date of last

[3] Alternative access method is usually filled with links to an archived version of the data that is no longer accessible from its original source, but occasionally there is a URL for convenience to save time later during the retrieval of the data for analysis.
[4] Also used interchangeably with the term currency in the context of data quality.

modification were determined. In fact, six of those datasets were last modified during the two-month period from October to November 2019 when the dataset modification dates were being collected. The topic of data currency is more thoroughly covered in subsection part 3.3.4.

Table 2: List of interlinked datasets with added information and more than 100,000 links to DBpedia (source: Author)

| Data Set | Number of Links | Data source | Availability | Data source type | Date of assessment | Alternative access | DBpedia inlinks | Last modified |
|---|---|---|---|---|---|---|---|---|
| Linked Open Colors | 16,000,000 | http://linkedopencolors.appspot.com/ | false | | 04/10/2019 | | | |
| dbpedia lite | 10,000,000 | http://dbpedialite.org/ | false | | 27/09/2019 | | | |
| flickr wrappr | 3,400,000 | http://wifo5-03.informatik.uni-mannheim.de/flickrwrappr/ | false | | 04/10/2019 | | | 27/04/2009 |
| Freebase | 3,348,530 | https://developers.google.com/freebase/ | true | dump | 04/10/2019 | | | 09/06/2013 |
| YAGO | 2,625,671 | https://datahub.io/collections/yago | true | SPARQL, dump | 16/10/2019 | | | 08/01/2019 |
| Twarql | 981,415 | https://old.datahub.io/dataset/twarql | false | SPARQL, dump | 16/10/2019 | | | 30/07/2016 |
| DBpedia in Portuguese | 365,839 | http://pt.dbpedia.org/ | true | SPARQL, dump | 27/09/2019 | | | 03/04/2017 |
| CORDIS | 285,256 | https://data.europa.eu/euodp/data/dataset/cordisref-data | true | SPARQL | 27/09/2019 | | true | 10/12/2018 |
| EU: fintrans.publicdata.eu | 199,168 | https://old.datahub.io/dataset/beneficiaries-of-the-european-commission | false | SPARQL | 02/10/2019 | | | 30/07/2016 |
| TaxonConcept | 147,877 | https://old.datahub.io/dataset/taxonconcept | partial | SPARQL, dump | 16/10/2019 | | | 30/07/2016 |

## 3.3 Data quality analysis

The EDA of datasets that link to DBpedia is focused on discovering the properties of the data quality dimensions in compliance with those chosen in subsection Data quality dimensions.

Accessibility and currency data quality dimensions have been analysed for all datasets, while the other dimensions uniqueness and consistency of interlinking were analysed on a small sample of datasets, because the analysis of these dimensions of data quality is more time consuming than the analysis of accessibility and currency. The sample consists of the following datasets:

- Alpine Ski Racers of Austria (https://old.datahub.io/dataset/austrian_ski_racers),
- BBC Music (https://archive.org/download/kasabi/bbc-music.gz),
- BBC Wildlife Finder (https://archive.org/download/kasabi/bbc-wildlife.gz),
- Classical (DBtune) (http://dbtune.org/classical/),
- EARTh (https://old.datahub.io/dataset/environmental-applications-reference-thesaurus),
- lexvo (http://www.lexvo.org/linkeddata/resources.html),
- lingvoj (http://www.linkedvocabs.org/lingvoj/data.ttl),
- Linked Clean Energy Data (reegle.info) (http://poolparty.reegle.info/PoolParty/sparql/glossary),
- OpenData Thesaurus (http://vocabulary.semantic-web.at/PoolParty/sparql/OpenData),
- SSW Thesaurus (http://vocabulary.semantic-web.at/PoolParty/sparql/semweb) and
- STW (https://zbw.eu/stw/version/latest/download/about).

The sample is topically centred on linguistic LOD (LLOD) with the exception of the first five datasets, that are focused on describing the real-world objects rather than abstract concepts. The reason for focusing so heavily on LLOD datasets is to contribute to the start of the NexusLinguarum project. The description of the project's goals from the project's website (COST Association, ©2020) is in the following two paragraphs:

*"The main aim of this Action is to promote synergies across Europe between linguists, computer scientists, terminologists, and other stakeholders in industry and society, in order to investigate and extend the area of linguistic data science. We understand linguistic data science as a subfield of the emerging "data science", which focuses on the systematic analysis and study of the structure and properties of data at a large scale, along with methods and techniques to extract new knowledge and insights from it. Linguistic data science is a specific case, which is concerned with providing a formal basis to the analysis, representation, integration and exploitation of language data (syntax, morphology, lexicon, etc.). In fact, the specificities of linguistic data are an aspect largely unexplored so far in a big data context.*

*In order to support the study of linguistic data science in the most efficient and productive way, the construction of a mature holistic ecosystem of multilingual and semantically interoperable linguistic data is required at Web scale. Such an ecosystem, unavailable today, is needed to foster the systematic cross-lingual discovery, exploration, exploitation, extension, curation and quality control of linguistic data. We argue that linked data (LD) technologies, in combination with natural language processing (NLP) techniques and multilingual language resources (LRs) (bilingual dictionaries, multilingual corpora, terminologies, etc.), have the potential to enable such an ecosystem that will allow for transparent information flow across linguistic data sources in multiple languages, by addressing the semantic interoperability problem."*

The role of this work in the context of the NexusLinguarum project is to provide an insight into which linguistic datasets are interlinked with DBpedia as a data hub of the Web of Data, and how high the quality of interlinking with DBpedia is.

One of the first steps of the Workgroup 1 (WG1) of the NexusLinguarum project is the assessment of the current state of the LLOD cloud and especially of the quality of data, metadata, and documentation of the datasets it consists of. This was agreed upon by the NexusLinguarum WG1 members (2020) participating on the teleconference on March 13th, 2020.

The datasets can be informally split into two groups:

- The first kind of datasets focuses on various subdomains of encyclopaedic data. This kind of data is specific because of its emphasis on describing physical objects and their relationships and because of their heterogeneity in the exact subdomain that they describe. In fact, most of the datasets provide information about noteworthy individuals. These datasets are:
  - Alpine Ski Racers of Austria,
  - BBC Music,
  - BBC Wildlife Finder and
  - Classical (DBtune).
- The other kind of analysed datasets belong to the lexico-linguistic domain. Datasets belonging to this category focus mostly on the description of concepts rather than objects that they represent as is the case of the concept of carbohydrates in the EARTh dataset (`http://linkeddata.ge.imati.cnr.it/resource/EARTh/17620`). The lexico-linguistic datasets analysed in this thesis are:
  - EARTh,
  - lexvo,
  - lingvoj,
  - Linked Clean Energy Data (reegle.info),
  - OpenData Thesaurus,
  - SSW Thesaurus and
  - STW.

Of the four features evaluated for the datasets, two (the uniqueness of entities and the consistency of interlinking) are computable measures. In both cases, the most basic measure is the absolute number of affected distinct entities. To account for different sizes

of the datasets, this measure needs to be normalized in some way. Because this thesis focuses only on the subset of entities, those that are interlinked with DBpedia, a decision was made to compute the ratio of unique affected entities relative to the number of unique interlinked entities. The alternative would have been to count the total number of entities in the dataset, but that would have been potentially less meaningful due to the different scale of interlinking in datasets that target DBpedia.

A concise overview of data quality features uniqueness and consistency is presented by Table 3. The details of identified problems as well as some additional information are described in parts 3.3.2 and 3.3.3 that are dedicated to uniqueness and consistency of interlinking respectively. There is also Table 4, which reveals the totals and averages for the two analysed domains and even across domains. It is apparent from both tables that more datasets are having problems related to consistency of interlinking than with uniqueness of entities. The scale of the two problems as measured by the number of affected entities, however, clearly demonstrates that there are more duplicate entities spread out across fewer datasets then there are inconsistently interlinked entities.

Table 3: Overview of uniqueness and consistency (source: Author)

| Domain | Dataset | Number of unique interlinked entities or concepts | Affected entities | | | |
|---|---|---|---|---|---|---|
| | | | Uniqueness | | Consistency | |
| | | | Absolute | Relative | Absolute | Relative |
| encyclopaedic data | Alpine Ski Racers of Austria | 70 | 0 | 0.0% | 1 | 1.4% |
| | BBC Music | 25359 | 351 | 1.4% | 1 | 0.0% |
| | BBC Wildlife Finder | 1402 | 0 | 0.0% | 0 | 0.0% |
| | Classical (DBtune) | 3169 | 32 | 1.0% | 0 | 0.0% |
| lexico-linguistic data | EARTh | 1861 | 0 | 0.0% | 0 | 0.0% |
| | lexvo | 4483 | 0 | 0.0% | 1 | 0.0% |
| | lexvo (including minor problems) | 4483 | - | - | 18 | 0.4% |
| | lingvoj | 7874 | 0 | 0.0% | 0 | 0.0% |
| | Linked Clean Energy Data (reegle.info) | 611 | 12 | 2.0% | 0 | 0.0% |
| | Linked Clean Energy Data (reegle.info) (including minor problems) | 611 | - | - | 14 | 2.3% |
| | OpenData Thesaurus | 54 | 0 | 0.0% | 0 | 0.0% |
| | SSW Thesaurus | 333 | 0 | 0.0% | 3 | 0.9% |
| | STW | 2614 | 0 | 0.0% | 2 | 0.1% |

Table 4: Aggregates for analysed domains and across domains (source: Author)

| Domain | Aggregation function | Number of unique interlinked entities or concepts | Affected entities | | | |
|---|---|---|---|---|---|---|
| | | | Uniqueness | | Consistency | |
| | | | Absolute | Relative | Absolute | Relative |
| encyclopaedic data | Total | 30000 | 383 | 1.3% | 2 | 0.0% |
| | Average | | 96 | 0.3% | 1 | 0.0% |
| lexico-linguistic data | Total | 17830 | 12 | 0.1% | 6 | 0.0% |
| | Average | | 2 | 0.0% | 1 | 0.0% |
| | Average (including minor problems) | | - | - | 5 | 0.0% |
| both domains | Total | 47830 | 395 | 0.8% | 8 | 0.0% |
| | Average | | 36 | 0.1% | 1 | 0.0% |
| | Average (including minor problems) | | - | - | 4 | 0.0% |

### 3.3.1 Accessibility

The analysis of dataset accessibility revealed that only about half of the datasets are still available. Another revelation of the analysis, apparent from Table 5, is the distribution of various access mechanisms. It is also clear from the table that SPARQL endpoints and RDF dumps are the most widely used methods for publishing LOD with 54 accessible datasets providing a SPARQL endpoint and 51 providing a dump for download. The third commonly used method for publishing data on the web is the provisioning of resolvable URIs, employed by a total of 26 datasets.

In addition 14 of the datasets that provide resolvable URIs are accessed through the RKBExplorer (http://www.rkbexplorer.com/data/) application developed by the European Network of Excellence Resilience for Survivability in IST (ReSIST). ReSIST is a research project from 2006, which ran up to the year 2009, aiming to ensure resilience and survivability of computer systems against physical faults, interaction mistakes, malicious attacks, and disruptions. (Network of Excellence ReSIST, n.d.)

Table 5: Usage of various methods for accessing LOD resources (source: Author)

| Count of Data Set | Available | | | | |
|---|---|---|---|---|---|
| Access method | fully | partially* | paid** | undetermined*** | not at all |
| SPARQL | 53 | 1 | | | 48 |
| dump | 52 | 1 | | | 33 |
| dereferenceable URIs | 27 | | | | 1 |
| web search | 18 | | | | |
| API | 8 | | | | 5 |
| XML | 4 | | | | |
| CSV | 3 | | | | |
| XLSX | 2 | | | | |
| JSON | 2 | | | | |
| SPARQL (authentication required) | 1 | | | | 1 |
| web frontend | 1 | | | | |
| KML | 1 | | | | |
| (no access method discovered) | | | 2 | 3 | 29 |
| RDFa | | | | | 1 |
| RDF browser | | | | | 1 |

* Partially available datasets are specific in that they publish data as a set of multiple dumps for download, but not all the dumps are available, effectively reducing the scope of the dataset. It was only considered when no alternative method (e.g. a SPARQL endpoint) was functional.

** Two datasets were identified as paid and therefore not available for analysis.

*** Three datasets were found where no evidence could be discovered as to how the data may be accessible.

### 3.3.2 Uniqueness

The measure of the data quality feature of uniqueness is the ratio of the number of entities that have a duplicate in the dataset (each entity is counted only once) and the total number of unique entities that are interlinked with an entity from DBpedia.

As far as encyclopaedic datasets are concerned, high numbers of duplicate entities were discovered in these datasets:

- DBtune, a non-commercial site providing structured data about music according to LD principles. At 32 duplicate entities interlinked DBpedia, it is just above 1 % of the interlinked entities. In addition, there are twelve entities that appear to be duplicates, but there is only indirect evidence through the form that the URI takes. This is, however, only a lower bound estimate because it is based only on entities that are interlinked with DBpedia.
- BBC Music, which has slightly above 1.4 % of duplicates out of the 24,996 unique entities interlinked with DBpedia.

An example of an entity that is duplicated in DBtune is the composer and musician André Previn whose record on DBpedia is `<http://dbpedia.org/resource/André_Previn>`. He is present in DBtune twice with these identifiers that when dereferenced lead to two different RDF subgraphs of the DBtune knowledge graph:

- `<http://dbtune.org/classical/resource/composer/previn_andre>` and
- `<http://dbtune.org/classical/resource/conductor/andre_previn>`.

Similarly, the BBC Music dataset contains among others two records about the rock band Acid Mothers Temple interlinked with `<http://dbpedia.org/resource/Acid_Mothers_Temple>`. The duplicate record URIs, each resolving to a slightly different page, are:

- `<https://www.bbc.co.uk/music/artists/49f03c14-8aa9-426c-a7f4-8e36409451a0#artist>` and
- `<https://www.bbc.co.uk/music/artists/41984dda-1f0e-436d-88d1-decb8d787122#artist>`.

On the opposite side, there are datasets BBC Wildlife, and Alpine Ski Racers of Austria, that do not contain any duplicate entities.

With regards to datasets containing LLOD, there were six datasets with no duplicates:

- EARTh,
- lingvoj,
- lexvo,
- the Open Data Thesaurus,
- the SSW Thesaurus and
- the STW Thesaurus for Economics.

Then, there is the reegle dataset, which focuses on the terminology of clean energy. It contains 12 duplicate values, which is about 2 % of the interlinked concepts. Those concepts are mostly interlinked with DBpedia using `skos:exactMatch` (in 11 cases) as opposed to the remaining one entity which is interlinked using `owl:sameAs`.

### 3.3.3 Consistency of interlinking

The measure of the data quality feature of consistency of interlinking is calculated as the ratio of different entities in a dataset that are linked to the same DBpedia entity using a predicate whose semantics is identity (`owl:sameAs`, `skos:exactMatch`) and the number of unique entities interlinked with DBpedia.

Problems with the consistency of interlinking have been found in five datasets. In the cross-domain encyclopaedic datasets no inconsistencies were found in:

- DBtune,
- BBC Wildlife.

While the dataset of Alpine Ski Racers of Austria does not contain any duplicate values, it has a different, but related problem. It is caused by using percent encoding of URIs even

when it is not necessary. An example when this becomes an issue is resource `http://vocabulary.semantic-web.at/AustrianSkiTeam/76` which is indicated to be the same as the following entities from DBpedia:

- http://dbpedia.org/resource/Fischer_%28company%29,
- http://dbpedia.org/resource/Fischer_(company).

The problem is that while accessing DBpedia resources through resolvable URIs just works, it prevents the use of SPARQL, possibly because of RFC 3986, which standardizes the general syntax of URIs. The RFC states that implementations must not percent-encode or decode the same string twice (Berners-Lee, et al., 2005). This behaviour can thus make it difficult to retrieve data about resources, whose URI has been unnecessarily encoded.

In the BBC Music dataset, the entities representing composer Bryce Dessner and songwriter Aaron Dessner are both linked using `owl:sameAs` property to the DBpedia entry about `http://dbpedia.org/page/Aaron_and_Bryce_Dessner` that describes both. A different property, possibly `rdfs:seeAlso`, should have been used when the entities do not match perfectly.

Of the lexico-linguistic sample of datasets, only EARTh was not found to be affected by consistency of interlinking issues at all.

The lexvo dataset contains 18 ISO 639-5 codes (or 0.4 % of interlinked concepts) linked to two DBpedia resources, which represent languages or language families, at the same time using `owl:sameAs`. This is, however, mostly not an issue. In 17 out of the 18 cases, the DBpedia resource is linked by the dataset using multiple alternative identifiers. This means that only one concept, `http://lexvo.org/id/iso639-3/nds`, has a consistency issue, because it is interlinked with two different German dialects:

- http://dbpedia.org/resource/West_Low_German and
- http://dbpedia.org/resource/Low_German.

This also means that only 0.02 % of interlinked concepts are inconsistent with DBpedia, because the other concepts that at first sight appeared to be inconsistent were in fact merely superfluous.

The reegle dataset contains 14 resources linking a DBpedia resource multiple times (in 12 cases using the `owl:sameAs` predicate while the `skos:exactMatch` predicate is used twice). Although it affects almost 2.3 % of interlinked concepts in the dataset, it is not a concern for application developers. It is just an issue of multiple alternative identifiers and not a problem with the data itself (exactly like most of the findings in the lexvo dataset).

The SSW Thesaurus was found to contain three inconsistencies in the interlinking between itself and DBpedia and one case of incorrect handling of alternative identifiers. This makes the relative measure of inconsistency between the two datasets come up to 0.9 %. One of the inconsistencies is that both the concepts representing “Big data management systems” and “Big data” were both linked to the DBpedia concept of “Big data” using `skos:exactMatch`. Another example is the term “Amsterdam” (`http://vocabulary.semantic-web.at/semweb/112`), which is linked to both the city and the 18th century ship of the Dutch East India Company

using `owl:sameAs`. A solution of this issue would be to create two separate records which would each link to the appropriate entity.

The last analysed dataset was STW, which was found to contain 2 inconsistencies. The relative measure of inconsistency is 0.1 %. There were these inconsistencies:

- the concept of "Macedonians" links to the DBpedia entry for "Macedonian" using `skos:exactMatch`, which is not accurate, and
- the concept of "Waste disposal", a narrower term of "Waste management", is linked to the DBpedia entry of "Waste management" using `skos:exactMatch`.

### 3.3.4 Currency

Figure 2 and Table 6 provide insight into the recency of data in datasets that contain links to DBpedia. The total number of datasets for which the date of last modification was determined is ninety-six. This figure consists of thirty-nine datasets whose data is not available[5], one dataset which is only partially[6] available and fifty-six datasets that are fully[7] available.

The fully available datasets are worth a more thorough analysis with regards to their recency. The freshness of data within half (that is twenty-eight) of these datasets did not exceed six years. The three years during which the most datasets were updated for the last time are 2016, 2012 and 2009. This mostly corresponds with the years when most of the datasets that are not available were last modified which might indicate that some events during these years caused multiple dataset maintainers to lose interest in LOD.

[5] Those are datasets whose access method does not work at all. (e.g. a broken download link or SPARQL endpoint)
[6] Partially accessible datasets are those that still have some working access method, but that access method does not provide access to the whole dataset. (e.g. A dataset with a dump split to multiple files, some of which cannot be retrieved.)
[7] The datasets that provide an access method to retrieve any data present in them.

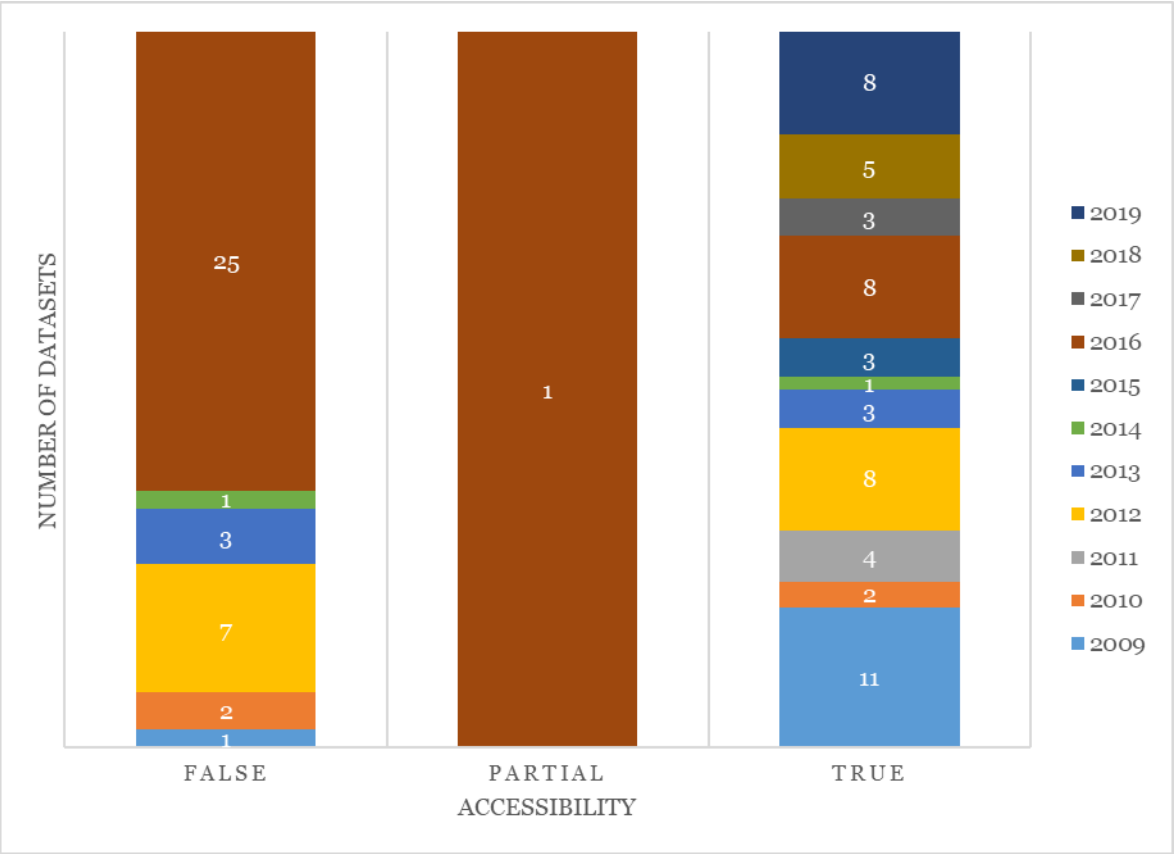


Figure 2: Number of datasets by year of last modification (source: Author)

Table 6: Dataset recency (source: Author)

| Count | Year of last modification | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Available** | **2009** | **2010** | **2011** | **2012** | **2013** | **2014** | **2015** | **2016** | **2017** | **2018** | **2019** | **Total** |
| not at all* | 1 | 2 | | 7 | 3 | 1 | | 25 | | | | 39 |
| partially** | | | | | | | | 1 | | | | 1 |
| fully | 11 | 2 | 4 | 8 | 3 | 1 | 3 | 8 | 3 | 5 | 8 | 56 |
| **Total** | **12** | **4** | **4** | **15** | **6** | **2** | **3** | **34** | **3** | **5** | **8** | **96** |

* Those are datasets which are not accessible through their own means. (e.g. Their SPARQL endpoints are not functioning, RDF dumps are not available etc.)

** In this case the RDF dump is split into multiple files, but only not all of them are still available.

# 4 Analysis of the consistency of bibliographic data in encyclopaedic datasets

Both the internal consistency of DBpedia and Wikidata datasets and the consistency of interlinking between them is important for the development of the semantic web. This is the case because both DBpedia and Wikidata are widely used as referential datasets for other sources of LOD, functioning as the nucleus of the semantic web.

This section thus aims at contributing to the improvement of the quality of DBpedia and Wikidata by focusing on one of the issues raised during the initial discussions preceding the start of the *GlobalFactSyncRE* project in June 2019, specifically the *Interfacing with Wikidata's data quality issues in certain areas*. GlobalFactSyncRE, as described by Hellmann (2018), is a project of the *DBpedia Association* which aims at improving the consistency of information among various language versions of Wikipedia and Wikidata.

The justification of this project according to Hellmann (2018) is that DBpedia has a near complete information about facts in Wikipedia infoboxes and the usage of Wikidata in Wikipedia infoboxes, which allows DBpedia to detect and display differences between Wikipedia and Wikidata and different language versions of Wikipedia to facilitate reconciliation of information. The *GlobalFactSyncRE* project treats the reconciliation of information as two separate problems:

- Lack of information management on a global scale affects the richness and the quality of information in Wikipedia infoboxes and in Wikidata.
  The *GlobalFactSyncRE* project aims to solve this problem by providing a tool that helps editors decide whether better information exists in another language version of Wikipedia or in Wikidata and offer to resolve the differences.
- Wikidata lacks about two thirds of facts from all language versions of Wikipedia. The *GlobalFactSyncRE* project tackles this by developing a tool to find infoboxes that reference facts according to Wikidata properties, find the corresponding line in such infoboxes and eventually find the primary source reference from the infobox about the facts that correspond to a Wikidata property.

The issue *Interfacing with Wikidata's data quality issues in certain areas* created by user Jc86035 (2019) brings attention to Wikidata items, especially those of bibliographic records of books and music, that are not conforming to their currently preferred item models based on FRBR. The specifications for these statements are available at:

- https://www.wikidata.org/wiki/Wikidata:WikiProject_Books and
- https://www.wikidata.org/wiki/Wikidata:WikiProject_Music.

The quality issues mentioned by Jc86035 in the initial discussion regarding the start of the project are that:

- data which only applies to an edition is used to describe the written work itself,
- items in Wikidata might not have been edited to match the models,
- other items might have similar issues despite not representing creative works while some creative works (e.g. video games for multiple operating systems) do not have this issue due to the difference in how they are modelled in Wikidata,
- the structure of some Wikipedia articles could result in incorrect references for items in infoboxes,
- some data may be difficult to verify, or its verification may to legal issues,
- fixing items requires creating new items and transferring non-conforming properties to the newly created items as well as fixing links that led to original amalgamation instead of a correctly modelled item.

## 4.1 FRBR representation in Wikidata

The style how Wikidata items of bibliographic records of books and music are modelled are described by:

- https://www.wikidata.org/wiki/Wikidata:WikiProject_Books,
- https://www.wikidata.org/wiki/Wikidata:WikiProject_Music.

In Wikidata, the FRBR classes are modelled differently for books and for music, although in both cases Wikidata collapses two of the classes into one class. While the bibliographic records of books collapse FRBR classes Expression and Manifestation, those of music collapse FRBR classes Work and Expression instead.

In addition, these subdomains of bibliographic records treat the FRBR class Item differently. In the subdomain dedicated to music, items are not modelled at all, because they are deemed to not be worth a Wikidata entry. In the subdomain of books, items have a special category called Exemplars and even a separate category for Manuscripts, which according to FRBR are Manifestation.

More formally, Wikidata items describing musical works are mapped to FRBR classes as described by a set comprised of formulas (4.2.1) and (4.2.2), that present the identity relations between corresponding classes or unions of classes in Wikidata and the FRBR framework.

$$Composition \equiv Work \sqcup Expression \tag{4.2.1}$$

$$Release \equiv Manifestation \tag{4.2.2}$$

Similarly, Wikidata items describing books are mapped to FRBR, but the way FRBR classes are collapsed for the purposes of Wikimedia projects is different. The mapping between FRBR and Wikidata is described by a set comprised of formulas (4.2.3), (4.2.4), (4.2.5) and (4.2.6). Similarly how the modelling of music related classes is specified, the book related

classes are defined using the identity relation between the Wikidata class and the corresponding FRBR class or union of FRBR classes with the exception of Edition and Manuscript classes both correspond partially to the FRBR Manifestation and as a result are defined as concepts included within the FRBR concept they are paired with.

$$Work \equiv Work \tag{4.2.3}$$

$$Edition \sqsubseteq Expression \sqcup Manifestation \tag{4.2.4}$$

$$Manuscript \sqsubseteq Manifestation \tag{4.2.5}$$

$$Exemplar \equiv Item \tag{4.2.6}$$

### 4.1.1 Determining the consistency of FRBR data in Wikidata

Regarding data quality of items modelled according to the FRBR specification, two kinds of issues were identified that need to be examined separately. Those two kinds of issues are:

- a combination of properties that should be used with different kinds of entities, because such combination implies that the data is inconsistent, and
- a k-permutation without repetition of Wikidata class and property which implies that the entity should belong to a different class.

Both kinds of issues fit into the data quality dimension concerned with consistency as described in subsection dedicated to Data quality and especially in Table 1 which represents a compilation of general data quality dimensions, data quality dimensions for OD and data quality dimensions for LOD.

Wikidata provides a public SPARQL endpoint which is everything that is needed to examine the consistency issues mentioned in Determining the consistency of FRBR data in Wikidata. The inconsistencies should be revealed by a set of SPARQL queries that each focus on a specific combination of properties or a k-permutation of a class and property.

The number of Wikidata classes in the FRBR domain is four for books and two for music. Therefore, according to formula (4.2.2.1) the number of combinations of the sets of properties intended for different classes is $C(2,4) = 6$ and $C(2,2) = 1$ for books and music, respectively. This means there are seven possible kinds of inconsistencies (6 for books, 1 for music) tied to the use of properties not intended to describe the very same entity.

Similarly, by using formula (4.2.2.2) we get the number of k-permutations without repetition $P(2,4) = 12$ and $P(2,2) = 2$ for books and music, respectively. This is the number of kinds of inconsistencies that manifest themselves as a mismatch between assignment of an entity to a class and the property attached to the entity, where the property belongs to a set of properties intended for use with a different class.

The relatively small number of queries required for an exhaustive examination makes it feasible to cover all possible cases.

$$C(k,n) = \frac{n!}{k!(n-k)!} \quad (4.2.2.1)$$

$$P(k,n) = \prod_{i=0}^{k-1}(n-i) \quad (4.2.2.2)$$

In the subdomain of bibliographic records dedicated to music, three queries are enough to cover all possible kinds of inconsistencies. This is the result of the combinatorial analysis covered earlier, which demonstrates that in the musical subdomain, the mismatches between the assigned class and the attached properties are of two kinds, while the mismatch of properties intended for instances of different classes can only happen in one form because the ordering of properties is not relevant for this analysis.

The first query demonstrated as snippet Code 4.1.1.1 focuses on finding entities that are described by properties that are not intended to be used for the description of instances of the same class.

Code 4.1.1.1: Query to check the existence of an inconsistency in the combination of properties (source: Author)

```
ask{
  ?entity wdt:P31 ?class .
  values ?class {
     wd:Q2031291 wd:Q207628
  }.
  ?entity ?composition_property [].
  ?entity ?release_property [].
  values ?composition_property {
    wdt:P1236 wdt:P1243 wdt:P1994 wdt:P2624 wdt:P2908 wdt:P3736 wdt:P3839 wdt:P3996
wdt:P4035 wdt:P435 wdt:P4860 wdt:P4932 wdt:P5241 wdt:P5262 wdt:P6080 wdt:P6218 wdt:P6348
wdt:P6431 wdt:P839 wdt:P1191 wdt:P144 wdt:P1625 wdt:P179 wdt:P3030 wdt:P3931 wdt:P5059
wdt:P51 wdt:P5202 wdt:P6116 wdt:P6439 wdt:P6670 wdt:P6686 wdt:P6883 wdt:P826 wdt:P87
wdt:P870
  }.
  values ?release_property {
    wdt:P1729 wdt:P1954 wdt:P2205 wdt:P2281 wdt:P2513 wdt:P2723 wdt:P2819 wdt:P3483
wdt:P4027 wdt:P4041 wdt:P4199 wdt:P436 wdt:P4518 wdt:P4748 wdt:P5144 wdt:P5153 wdt:P5680
wdt:P5749 wdt:P5813 wdt:P7175 wdt:P1303 wdt:P155 wdt:P156 wdt:P162 wdt:P1638 wdt:P175
wdt:P264 wdt:P483 wdt:P527 wdt:P5707 wdt:P658 wdt:P736
  }.
}
```

The second snippet, Code 4.1.1.2, presents a query intended to check whether the items assigned to the Wikidata class Composition, which is a union of FRBR types Work and Expression in the musical subdomain of bibliographic records, are described by properties intended for use with Wikidata class Release representing a FRBR Manifestation. If the query finds an entity for which it is true, it means that an inconsistency is present in the data.

Code 4.1.1.2: Query to check the presence of inconsistencies between an assignment to class representing the amalgamation of FRBR types work and expression and properties attached to such item (source: Author)

```
ask{
  ?entity wdt:P31 ?class .
  values ?class {
     wd:Q2031291 wd:Q207628
  }.
  ?entity wdt:P31 wd:Q207628.
  ?entity ?release_property [].
  values ?release_property {
    wdt:P1729 wdt:P1954 wdt:P2205 wdt:P2281 wdt:P2513 wdt:P2723 wdt:P2819 wdt:P3483
wdt:P4027 wdt:P4041 wdt:P4199 wdt:P436 wdt:P4518 wdt:P4748 wdt:P5144 wdt:P5153 wdt:P5680
wdt:P5749 wdt:P5813 wdt:P7175 wdt:P1303 wdt:P155 wdt:P156 wdt:P162 wdt:P1638 wdt:P175
wdt:P264 wdt:P483 wdt:P527 wdt:P5707 wdt:P658 wdt:P736
  }.
}
```

The last snippet, Code 4.1.1.3, introduces the third possibility of how an inconsistency may manifest itself. It is rather similar to query from Code 4.1.1.2, but differs in one important aspect, which is that it checks for inconsistencies from the opposite direction. It looks for instances of the class representing a FRBR Manifestation described by properties that are appropriate only for a Work or Expression.

Code 4.1.1.3: Query to check the presence of inconsistencies between an assignment to class representing FRBR type manifestation and properties attached to such item (source: Author)

```
ask{
  ?entity wdt:P31 ?class .
  values ?class {
     wd:Q2031291 wd:Q207628
  }.
  ?entity wdt:P31 wd:Q2031291.
  ?entity ?composition_property [].
  values ?composition_property {
    wdt:P1236 wdt:P1243 wdt:P1994 wdt:P2624 wdt:P2908 wdt:P3736 wdt:P3839 wdt:P3996
wdt:P4035 wdt:P435 wdt:P4860 wdt:P4932 wdt:P5241 wdt:P5262 wdt:P6080 wdt:P6218 wdt:P6348
wdt:P6431 wdt:P839 wdt:P1191 wdt:P144 wdt:P1625 wdt:P179 wdt:P3030 wdt:P3931 wdt:P5059
wdt:P51 wdt:P5202 wdt:P6116 wdt:P6439 wdt:P6670 wdt:P6686 wdt:P6883 wdt:P826 wdt:P87
wdt:P870
  }.
}
```

When an ask query proves that there exists at least one inconsistency between the data and the model, it is possible to check how common the problem is by replacing the reserved word *ask* with snippet from Code 4.1.1.4. It is also possible to make a list of affected entities by replacing the reserved word *ask* with snippet from Code 4.1.1.5.

Code 4.1.1.4: Counting entities affected by an inconsistency (source: Author)

```
select (count(distinct ?entity) as ?number_of_affected_entities)
```

Code 4.1.1.5: Making a list of affected entities (source: Author)

```
select distinct ?entity
```

### 4.1.2 Results of Wikidata examination

Table 7 provides a concise summary of the results of SPARQL queries based on the examples presented as code snippets Code 4.1.1.1 through Code 4.1.1.3.

It is apparent from Table 7 that with two thirds of the queries not running till the end, because they are stopped by the SPARQL endpoint, the number of inconsistencies is likely to be much higher than what was discovered. One way to potentially overcome the issue of most of the queries not finishing could be to carefully reorder the triple patterns in the queries, because the order of triple patterns matters as pointed out at the GitHub wiki page of Blazegraph (Bebee, 2020), the database system used by Wikidata (Wikidata, 2019).

The 3,062 inconsistent entities are therefore just the minimal possible number of inconsistencies. The highest number of inconsistent entities returned by a query was 2,933. Therefore, by assuming that the query complexity does not differ much and that the highest number of inconsistencies successfully obtained is the limit of a query of such complexity to return in time, it is possible to calculate an estimate by adding a one to the highest number which the endpoint returned and multiplying it by the number of queries that timed out: $(2{,}933 + 1) \cdot 14 = 41{,}076$. Because the calculations behind the estimate are based on the lowest possible number of entries that would cause a query of a certain complexity to time out, it would be best described as a conservative lower bound estimate.

Given the total number of FRBR entities in Wikidata of 201,495 as discovered by the query from Code 4.1.2.1, it would mean that about 22 % of all Wikidata entries regarding FRBR entities are inconsistent. This is, however, just an estimate and the ratio of entities that are undoubtedly inconsistent is 1.5 %.

Code 4.1.2.1: Count of all FRBR entities (source: Author)

```
select (count(distinct ?entity) as ?number_of_entities)
{
  ?entity wdt:P31 ?class .
  values ?class {
     wd:Q87167 wd:Q213924 wd:Q1440453 wd:Q834459 wd:Q2217259 wd:Q274076 wd:Q1754581
wd:Q690851 wd:Q284465 wd:Q53731850 wd:Q3331189 wd:Q47461344 wd:Q2031291 wd:Q207628
  }
}
```

Table 7: Inconsistently typed Wikidata entities by the kind of inconsistency (source: Author)

| Category of inconsistency | Subdomain | Classes | Properties | Is inconsistent | Number of affected entities |
|---|---|---|---|---|---|
| properties | music | | Composition, Release | TRUE | timeout |
| class with properties | music | Composition | Release | TRUE | 2,933 |
| class with properties | music | Release | Composition | TRUE | 18 |
| properties | books | | Work, Edition | TRUE | timeout |
| class with properties | books | Work | Edition | TRUE | timeout |
| class with properties | books | Edition | Work | TRUE | timeout |
| properties | books | | Edition, Exemplar | TRUE | timeout |
| class with properties | books | Exemplar | Edition | TRUE | 22 |
| class with properties | books | Edition | Exemplar | TRUE | 23 |
| properties | books | | Edition, Manuscript | TRUE | timeout |
| class with properties | books | Manuscript | Edition | TRUE | timeout |
| class with properties | books | Edition | Manuscript | TRUE | timeout |
| properties | books | | Exemplar, Work | TRUE | timeout |
| class with properties | books | Exemplar | Work | TRUE | 13 |
| class with properties | books | Work | Exemplar | TRUE | 31 |
| properties | books | | Manuscript, Work | TRUE | timeout |
| class with properties | books | Manuscript | Work | TRUE | timeout |
| class with properties | books | Work | Manuscript | TRUE | timeout |
| properties | books | | Manuscript, Exemplar | TRUE | timeout |
| class with properties | books | Manuscript | Exemplar | TRUE | timeout |
| class with properties | books | Exemplar | Manuscript | TRUE | 22 |

## 4.2 FRBR representation in DBpedia

FRBR is not specifically modelled in DBpedia, which complicates both the development of applications that need to distinguish entities based on FRBR types and the evaluation of data quality with regards to consistency and typing.

One of the tools that tried to provide information from DBpedia to its users based on the FRBR model was FRBRpedia. It is described in the article *FRBRPedia: a tool for FRBRizing web products and linking FRBR entities to DBpedia* (Duchateau, et al., 2011) as a tool for FRBRizing web products tailored for Amazon bookstore. Even though it is no longer available, it still illustrates the effort needed to provide information from DBpedia based on FRBR by utilizing several other data sources:

- the Online Computer Library Center (OCLC) classification service to find works related to the product,
- xISBN[8], which is another OCLC service, to find related Manifestations and infer the existence of Expressions based on similarities between Manifestations,
- the Virtual International Authority File (VIAF) for identification of actors contributing to the Work and
- DBpedia which is queried for related entities that are then ranked based on various similarity measures and eventually presented to the user to validate the entity. Finally, the FRBRized data enriched by information from DBpedia is presented to the user.

The approach in this thesis is different in that it does not try to overcome the issue of missing information regarding FRBR types by employing other data sources, but relies on annotations made manually by annotators using a tool specifically designed, implemented, tested and eventually deployed and operated for exactly this purpose. The details of the development process are described in section An, which is also the name of the tool, whose source code is available on GitHub under the GPLv3 license at the following address: https://github.com/Fuchs-David/Annotator.

## 4.3 Annotating DBpedia with FRBR information

The goal to investigate the consistency of DBpedia and Wikidata entities related to artwork, requires both datasets to be comparable. Because DBpedia does not contain any FRBR information, it is therefore necessary to annotate the dataset manually.

The annotations were created by two volunteers together with the author, which means there were three annotators in total. The annotators provided feedback about their user

[8] According to issue https://github.com/xlcnd/isbnlib/issues/28, the xISBN service has been retired in 2016, which may be the reason why FRBRpedia is no longer available.

experience with using the applications. The first complaint was that the application did not provide guidance about what should be done with the displayed data, which was resolved by adding a paragraph of text to the annotation web form page. The second complaint, however, was only partially resolved by providing a mechanism to notify the user that he reached the pre-set number of annotations expected from each annotator. The other part of the second complaint was not resolved, because it requires a complex analysis of the influence of different styles of user interface on the user experience in the specific context of an application gathering feedback based on large amounts of data.

The number of created annotations is 70, about 2.6 % of the 2,676 of DBpedia entities interlinked with Wikidata entries from the bibliographic domain. Because the annotations needed to be evaluated in the context of interlinking of DBpedia entities and Wikidata entries, they had to be merged with at least some contextual information from both datasets.

More information about the development process of the FRBR Annotator for DBpedia is provided in Annex B.

### 4.3.1 Consistency of interlinking between DBpedia and Wikidata

It is apparent from Table 8 that majority of links between DBpedia to Wikidata target entries of FRBR Works. Given the Results of Wikidata examination, it is entirely possible that the interlinking is based on the similarity of properties used to describe the entities rather than on the typing of entities. This would therefore lead to the creation of inaccurate links between the datasets, which can be seen in Table 9.

Table 8: DBpedia links to Wikidata by classes of entities (source: Author)

| Wikidata class | Label | Entity count | Expected FRBR class |
|---|---|---|---|
| http://www.wikidata.org/entity/Q213924 | codex | 2 | Item |
| http://www.wikidata.org/entity/Q3331189 | version, edition, or translation | 3 | Expression or Manifestation |
| http://www.wikidata.org/entity/Q47461344 | written work | 25 | Work |
| http://www.wikidata.org/entity/Q207628 | musical composition | 2642 | Work or Expression |
| http://www.wikidata.org/entity/Q87167 | manuscript | 3 | Item |
| http://www.wikidata.org/entity/Q2217259 | manuscript codex | 1 | Item |

Table 9 reveals the number of annotations of each FRBR class grouped by the type of the Wikidata entry to which the entity is linked. Given the knowledge of mapping of FRBR classes to Wikidata which is described in subsection 4.1 and displayed together with the distribution of the classes Wikidata in Table 8, the FRBR classes Work and Expression are the correct classes for entities of type `wd:Q207628`. The 11 entities annotated as either Manifestation or Item though, point to a potential inconsistency that affects almost 16 % of annotated entities randomly chosen from the pool of 2,676 entities representing bibliographic records.

Table 9: Number of annotations by Wikidata entry (source: Author)

| Wikidata class | FRBR class | Count |
|---|---|---|
| wd:Q207628 | frbr:term-Item | 1 |
| wd:Q207628 | frbr:term-Work | 47 |
| wd:Q207628 | frbr:term-Expression | 12 |
| wd:Q207628 | frbr:term-Manifestation | 10 |

### 4.3.2 RDFRules experiments

An attempt was made to create a predictive model using the RDFRules tool available on GitHub: https://github.com/propi/rdfrules.

The tool has been developed by Václav Zeman from the University of Economics, Prague. It uses an enhanced version of Association Rule Mining under Incomplete Evidence (AMIE) system named AMIE+ (Zeman, 2018), designed specifically to address issues associated with rule mining in the open environment of the semantic web.

Snippet Code 4.2.1.1 demonstrates the structure of the rule mining workflow. This workflow can be directed by the snippet Code 4.2.1.2 which defines the thresholds and the pattern that provides is searched for in each rule in the ruleset. The default thresholds of minimal head size 100, minimal head coverage 0.01 could not have been satisfied at all, because the minimal head size exceeded the number of annotations. Thus, it was necessary to allow weaker rules to be considered and so the thresholds were set to be as permissive as possible, leading to the minimal head size of 1, minimal head coverage of 0.001 and the minimal support of 1.

The pattern restricting the ruleset to only include rules whose head consists of a triple with `rdf:type` as predicate and one of `frbr:term-Work`, `frbr:term-Expression`, `frbr:term-Manifestation` and `frbr:term-Item` as object therefore needed to be relaxed. Because the FRBR resources are only used in the dataset in instantiation, the only meaningful relaxation of the mining parameters was to remove the FRBR resources from the pattern.

Code 4.2.1.1: Configuration to search for all rules (source: Author)

```
[{
    "name": "LoadDataset",
    "parameters": {
      "url": "file:/// ... \\DBpediaAnnotations.nt",
      "format": "nt"
}},{
    "name": "Index",
    "parameters": {}
},{
    "name": "Mine",
    "parameters": {
      "thresholds": [],
      "patterns": [],
```

```
      "constraints": []
}},{
    "name": "GetRules",
    "parameters": {}
}]
```

Code 4.2.1.2: Patterns and thresholds for rule mining (source: Author)

```
"thresholds": [{
    "name": "MinHeadSize",
    "value": 1
  },{
    "name": "MinHeadCoverage",
    "value": 0.001
  },{
    "name": "MinSupport",
    "value": 1
}],
"patterns": [{
    "head": {
      "subject": { "name": "Any" },
      "predicate": {
        "name": "Constant",
        "value": "<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>"
      },
      "object": {
        "name": "OneOf",
        "value": [{
            "name": "Constant",
            "value": "<http://vocab.org/frbr/core.html#term-Work>"
          },{
            "name": "Constant",
            "value": "<http://vocab.org/frbr/core.html#term-Expression>"
          },{
            "name": "Constant",
            "value": "<http://vocab.org/frbr/core.html#term-Manifestation>"
          },{
            "name": "Constant",
            "value": "<http://vocab.org/frbr/core.html#term-Item>"
          }]},
      "graph": { "name": "Any" }},
    "body": [],
    "exact": false
}]
```

After dropping the requirement for the rules to contain a FRBR class in the object position of a triple in the head of the rule, two rules were discovered. They both highlight the relationship between a connection between two resources by a `dbo:wikiPageWikiLink` and the assignment of both resources to the same class. The following qualitative metrics of the rules have been obtained: $HeadCoverage = 0.02$, $HeadSize = 769$ and $support = 16$. Neither of them could, however, possibly be used to predict the assignment of a DBpedia resource to a FRBR class, because the information the `dbo:wikiPageWikiLink` predicate carries does not have any specific meaning in the domain modelled by the FRBR framework. It only means that a specific wiki page links to another wiki page, but the relationship between the two pages is not specified in any way.

Code 4.2.1.4:

```
( ?c <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> ?b )
^ ( ?c <http://dbpedia.org/ontology/wikiPageWikiLink> ?a )
⇒ ( ?a <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> ?b )
```

Code 4.2.1.3:

```
( ?a <http://dbpedia.org/ontology/wikiPageWikiLink> ?c )
^ ( ?c <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> ?b )
⇒ ( ?a <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> ?b )
```

### 4.3.3 Results of interlinking of DBpedia and Wikidata

Although the rule mining did not provide the expected results, interactive analysis of annotations did reveal at least some potential inconsistencies. Overall, 2.6 % of DBpedia entities interlinked with Wikidata entries about items from the FRBR domain of interest were annotated. The percentage of potentially incorrectly interlinked entities has come up close to 16 %. If this figure is representative of the whole dataset, it could mean over 420 inconsistently modelled entities.

# 5 Impact of the discovered issues

The outcomes of this work can be categorized into three groups:

- data quality issues associated with linking to DBpedia,
- consistency issues of FRBR categories between DBpedia and Wikidata and
- consistency issues of Wikidata itself.

DBpedia and Wikidata represent two major sources of encyclopaedic information on the Semantic Web and serve as a hub, supposedly because of their vast knowledge bases[9] and sustainability[10] of their maintenance.

The Wikidata project is focused on the creation of structured data for the enrichment of Wikipedia infoboxes while improving their consistency across different Wikipedia language versions. DBpedia on the other hand extracts structured information both from the Wikipedia infoboxes and the unstructured text. The two projects are according to Wikidata page about the relationship of DBpedia and Wikidata (2018) expected to interact indirectly through the Wikipedia's infoboxes with Wikidata providing the structured data to fill them and DBpedia extracting that data through its own extraction templates. The primary benefit is supposedly less work needed for the development of extraction, which would allow the DBpedia teams to focus on higher value-added work to improve other services and processes. This interaction can also be used for feedback to Wikidata about the degree to which structured data originating from it is already being used in Wikipedia though, as suggested by the *GlobalFactSyncRE* project, to which this thesis aims to contribute.

## 5.1 Spreading of consistency issues from Wikidata to DBpedia

Because the extraction process of DBpedia relies to some degree on information that may be modified by Wikidata, it is possible that the inconsistencies found in Wikidata and described by section 4.1.2 have been transferred to DBpedia and discovered through the analysis of annotations in section 4.3.3. Given that the scale of the problem with internal consistency of Wikidata with regards to artwork is different than the scale of a similar problem with consistency of interlinking of artwork entities between DBpedia and Wikidata, there are several explanations:

1. In Wikidata, only 1.5 % of entities are known to be affected, but according to annotators, about 16 % of DBpedia entities could be inconsistent with their Wikidata counterparts. This disparity may be caused by the unreliability of text extraction.

[9] This may be considered as fulfilling the data quality dimension called: Appropriate amount of data.
[10] Sustainability is itself a data quality dimension, which considers the likelihood of a data source being abandoned.

2. If the estimated number of affected entities in Wikidata is accurate, the consistency rate of DBpedia interlinking with Wikidata would be higher than the internal consistency measure of Wikidata. This could mean that either the text extraction avoids inconsistent infoboxes or that the process of interlinking avoids creating links to inconsistently modelled entities. It could, however, also mean that the inconsistently modelled entities have not yet been widely applied to Wikipedia infoboxes.
3. The third possibility is a combination of both phenomena, in which case it would be hard to decide what the issue is.

Whichever case it is though, cleaning-up Wikidata of the inconsistencies and then repeating the analysis of its internal consistency as well as the annotation experiment would likely provide a much clearer picture of the problem domain together with valuable insight into the interaction between Wikidata and DBpedia.

Repeating this process without the delay to let Wikidata get cleaned-up may be a way to mitigate potential issues with the process of annotation, which could be biased in some way towards some classes of entities for unforeseen reasons.

## 5.2 Effects of inconsistency in the hub of the Semantic Web

High consistency of data in DBpedia and Wikidata is especially important to mitigate the adverse effects that inconsistencies may have on applications that consume the data or on the usability of other datasets that may rely on DBpedia and Wikidata to provide context for their data.

### 5.2.1 Effect on a text editor

To illustrate the kind of problems an application may run into, let us assume that in the future, checking the spelling and grammar is a solved problem for text editors and that to stand out among the competing products, the better editors should also check the pragmatic layer of the language. That could be done by using valency frames together with information retrieved from a thesaurus (e.g. SSW Thesaurus) interlinked with a source of encyclopaedic data (e.g. DBpedia as is the case of the SSW Thesaurus).

In such case, issues like the one which manifests itself by not distinguishing between the entity representing the city of Amsterdam and the historical ship Amsterdam, could lead to incomprehensible texts being produced. Although this example of inconsistency is not likely to cause much harm, more severe inconsistencies could be introduced in the future unless appropriate action is taken to improve the reliability of the interlinking process or the consistency of the involved datasets. The impact of not correcting the writer may vary widely depending on the kind of text being produced from mild impact such as some passages of a not so important document being unintelligible, through more severe consequence such as the destruction of somebody's reputation, to the most severe consequences which could lead to legal disputes over the meaning of the text (e.g. due to mistakes in a contract).

### 5.2.2 Effect on a search engine

Now, let us assume that some search engine would try to improve the search results by comparing textual information in the documents on the regular web with structured information from curated datasets such as DBtune or BBC Music. In such case, searching for a specific release of a composition that was performed by a specific artist with a DBtune record could lead to inaccurate results due to either inconsistencies in the interlinking of DBtune and DBpedia, inconsistencies of interlinking between DBpedia and Wikidata or finally due to inconsistencies of typing in Wikidata.

The impact of this issue may not sound severe, but for somebody who collects musical artworks it could mean wasted time or even money if he decided to buy a supposedly rare release of an album to only later discover that it is in fact not as rare as he expected it to be.

# 6 Conclusions

The first goal of this thesis, which was to quantitatively analyse the connectivity of linked open datasets with DBpedia was fulfilled in section 2.6 and especially its last subsection 3.3 dedicated to describing the results of analysis focused on data quality issues discovered in the eleven assessed datasets. The most interesting discoveries with regards to data quality of LOD is that:

- recency of data is a widespread issue, because only half of the available datasets have been updated within the five years preceding the period during which the data for evaluation of this dimension was being collected (October and November 2019),
- uniqueness of resources is an issue which affects three of the evaluated datasets. The volume of affected entities is rather low, tens to hundreds of duplicate entities, as well as the percentages of duplicate entities which is between 1 and 2 % of the whole, depending on the dataset,
- consistency of interlinking affects six datasets, but the degree to which they are affected is low, merely up to tens of inconsistently interlinked entities, as well as the percentage of inconsistently interlinked entities in a dataset – at most 2.3 % – and
- applications can mostly get away with standard access mechanisms for semantic web (SPARQL, RDF dump, dereferenceable URI), although some datasets (almost 14 % of those interlinked with DBpedia) may force the application developers to use non-standard web APIs or handle custom XML, JSON, KML or CSV files.

The second goal was to analyse the consistency (an aspect of data quality) of Wikidata entities related to artwork. This task was dealt with in two different ways. One way was to evaluate the consistency within Wikidata itself as described in part 4.1.2 of the subsection dedicated to FRBR in Wikidata. The second approach to evaluating the consistency was aimed at the consistency of interlinking, where Wikidata was the target dataset and DBpedia the linking dataset. To tackle the issue of the lack of information regarding FRBR typing at DBpedia, a web application has been developed to help annotate DBpedia resources. The annotation process and its outcomes are described in section 4.3. The most interesting results of consistency analysis of FRBR categories in Wikidata are that:

- the Wikidata knowledge graph is estimated to have an inconsistency rate of around 22 % in the FRBR domain while only 1.5 % of the entities are known to be inconsistent and
- the inconsistency of interlinking affects about 16 % of DBpedia entities that link to a Wikidata entry from the FRBR domain.
- The part of the second goal that focused on the creation of a model that would predict which FRBR class a DBpedia resource belongs to, did not produce the desired results, probably due to an inadequately small sample of training data.

## 6.1 Future work

Because the estimated inconsistency rate within Wikidata is rather close to the potential inconsistency rate of interlinking between DBpedia and Wikidata, it is hard to resist the thought that inconsistencies within Wikidata propagate through Wikipedia's infoboxes to DBpedia. This is, however, out of scope of this project and would therefore need to be addressed in subsequent investigation that should be conducted with a delay long enough to allow Wikidata to be cleaned-up of the discovered inconsistencies.

Further research also needs to be carried out to provide a more detailed insight into the interlinking between DBpedia and Wikidata, either by gathering annotations about artwork entities at a much larger scale than what was managed by this research or by assessing the consistency of entities from other knowledge domains.

More research is also needed to evaluate the quality of interlinking on a larger sample of datasets than those analysed in section 3. To support the research efforts, a considerable amount of automation is needed. To evaluate the accessibility of datasets as understood in this thesis, a tool supporting the process should be built, that would incorporate a crawler to follow links from certain starting points (e.g. the DBpedia's wiki page on interlinking found at https://wiki.dbpedia.org/services-resources/interlinking) and detect presence of various access mechanisms, most importantly links to RDF dumps and URLs of SPARQL endpoints. This part of the tool should also be responsible for the extraction of the currency of the data, which would likely need to be implemented using text mining techniques. To analyse the uniqueness and consistency of the data, the tool would need to use a set of SPARQL queries, some of which may require features not available in public endpoints (as was occasionally the case during this research). This means that the tool would also need access to a private SPARQL endpoint to upload data extracted from such sources to and this endpoint should be able to store and efficiently handle queries over large volumes of data (at least in the order of gigabytes (GB) – e.g. for VIAF's 5 GB RDF dump).

As far as tools supporting the analysis of data quality are concerned, the tool for annotating DBpedia resources could also use some improvements. Some of the improvements have been identified as well as some potential solutions at a rather high level of abstraction:

- The annotators who participated in annotating DBpedia were sometimes confused by the application layout. It may be possible to address this issue by changing the application such that each of its web pages is dedicated to only one purpose (e.g. introduction and explanation page, annotation form page, help pages).
- The performance could be improved. Although the application is relatively consistent in its response times, it may improve the user experience if the performance was not so reliant on the performance of the federated SPARQL queries, which may also be a concern for reliability of the application due to the nature of distributed systems. This could be alleviated by implementing a preload mechanism such that a user does not wait for a query to run, but only for the data to be processed, thus avoiding a lengthy and complex network operation.
- The application currently retrieves the resource to be annotated at random, which becomes an issue when the distribution of types of resources for annotation is not

uniform. This issue could be alleviated by introducing a configuration option to specify the probability of limiting the query to resources of a certain type.

- The application can be modified so that it could be used for annotating other types of resources. At this point, it appears that the best choice would be to create an XML document holding the configuration as well as the domain specific texts. It may also be advantageous to separate the texts from the configuration to make multi-lingual support easier to implement.
- The annotations could be adjusted to comply with the Web Annotation Ontology (https://www.w3.org/ns/oa). This would increase the reusability of data, especially if combined with the addition of more metadata to the annotations. This would, however, require the development of a formal data model based on web annotations.

# List of references


1 Albertoni, R. & Isaac, A., 2016. *Data on the Web Best Practices: Data Quality Vocabulary* [Online]. Available at: https://www.w3.org/TR/vocab-dqv/ [Accessed 17 MAR 2020].

2 Balter, B., 2015. *6 motivations for consuming or publishing open source software* [Online]. Available at: https://opensource.com/life/15/12/why-open-source [Accessed 24 MAR 2020].

3 Bebee, B., 2020. *In SPARQL, order matters*. [Online]. Available at: https://github.com/blazegraph/database/wiki/SPARQL_Order_Matters [Accessed 20 APR 2020].

4 Berners-Lee, T. et al., 2005. *Uniform Resource Identifier (URI): Generic Syntax* [Online]. Available at: https://tools.ietf.org/html/rfc3986 [Accessed 23 MAR 2020].

5 Bizer, C., Cyganiak, R. & Heath, T., 2008. *How to Publish Linked Data on the Web* [Online]. Available at: http://wifo5-03.informatik.uni-mannheim.de/bizer/pub/LinkedDataTutorial/ [Accessed 2 APR 2020].

6 Bouquet, P., Stoermer, H. and Vignolo, M., 2012. *Web of Data and Web of Entities: Identity and Reference in Interlinked Data in the Semantic Web*. Philosophy & Technology, 25(1), pp. 9-16. DOI: https://doi.org/10.1007/s13347-010-0011-6.

7 COST Association, ©2020. *CA18209 - European network for Web-centred linguistic data science* [Online]. Available at: https://www.cost.eu/actions/CA18209/#tabs|Name:overview [Accessed 22 MAR 2020].

8 Crispin, L., 2011. *Using the Agile Testing Quadrants* [Online]. Available at: https://lisacrispin.com/2011/11/08/using-the-agile-testing-quadrants/ [Accessed 25 MAR 2020].

9 DBpedia Association, ©2019. *Interlinking* [Online]. Available at: https://wiki.dbpedia.org/services-resources/interlinking [Accessed 11 MAR 2020].

10 Dekkers, M., Brule, D., Droscariu, A. & Novacean, I., 2018. *Guidelines for the Use of Code Lists* [Online]. Available at: https://joinup.ec.europa.eu/sites/default/files/document/2018-05/Guidelines%20for%20the%20Use%20of%20Code%20Lists%20v1.00.pdf [Accessed 1 APR 2020].

11 Duchateau, F., Takhirov, N. & Aalberg, T., 2011. *FRBRPedia: a tool for FRBRizing web products and linking FRBR entities to DBpedia*. In Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries (JCDL '11). Association for Computing Machinery, New York, NY, USA, 455–456. DOI: https://doi.org/10.1145/1998076.1998183.

12 Duerst, M., W3C, Suignard, M. & Microsoft Corporation, 2005. *Internationalized Resource Identifiers (IRIs)* [Online]. Available at: https://tools.ietf.org/html/rfc3987 [Accessed 6 APR 2020].

13 Ehrlinger, L. & Wös, W., 2016. *Towards a Definition of Knowledge Graphs*. In Joint Proceedings of the Posters and Demos Track of the 12th International Conference on Semantic Systems - SEMANTiCS2016, Leipzig, Germany, September 12-15, 2016, CEUR-WS.org, online CEUR-WS.org/Vol-1695/paper4.pdf.

14 Faronov, V., 2011. *Linking patterns* [Online]. Available at: https://www.w3.org/2001/sw/wiki/Linking_patterns [Accessed 2 APR 2020].
15 Fuchs, D., 2018. *Data v RDF a jejich zpracování v MS Excel* (Data in RDF and their processing in MS Excel). Bachelor thesis. University of Economics, Prague. Ing. Jan Kučera, Ph.D. Also available at: https://vskp.vse.cz/id/1344781.
16 Fürber, C., 2011. *Data Quality* [Online]. Available at: http://semwebquality.org/mediawiki/index.php?title=Data_Quality [Accessed 17 MAR 2020].
17 Fürber, C. & Hepp, M., 2011. *Towards a Vocabulary for Data Quality Management in Semantic Web Architectures*. In Proceedings of the 1st International Workshop on Linked Web Data Management (LWDM '11). Association for Computing Machinery, New York, NY, USA, 1–8. DOI: https://doi.org/10.1145/1966901.1966903.
18 Hausenblas, M., 2012. *5 ★ OPEN DATA* [Online]. Available at: https://5stardata.info/ [Accessed 16 MAR 2020].
19 Heath, T. & Bizer, C., 2011. *Linked Data: Evolving the Web into a Global Data Space* [Online]. Available at: http://linkeddatabook.com/editions/1.0/ [Accessed 12 MAR 2020].
20 Hellmann, S., 2018. *Grants:Project/DBpedia/GlobalFactSyncRE* [Online]. Available at: https://meta.wikimedia.org/wiki/Grants:Project/DBpedia/GlobalFactSyncRE [Accessed 4 MAR 2020].
21 Hitzler, P. et al., 2012. *OWL 2 Web Ontology Language* [Online]. Available at: https://www.w3.org/TR/owl2-primer/ [Accessed 6 APR 2020].
22 IFLA Study Group, 1998. *Functional Requirements for Bibliographic Records* [Online]. Available at: https://www.ifla.org/files/assets/cataloguing/frbr/frbr.pdf [Accessed 7 MAR 2020].
23 Jc86035, 2019. *Interfacing with Wikidata's data quality issues in certain areas* [discussion post] In: Grants talk:Project/DBpedia/GlobalFactSyncRE [Online]. Available at: https://meta.wikimedia.org/w/index.php?title=Grants_talk:Project/DBpedia/GlobalFactSyncRE&oldid=19522842#Interfacing_with_Wikidata's_data_quality_issues_in_certain_areas [Accessed 4 MAR 2020].
24 Keller, S. et al., 2017. *The Evolution of Data Quality: Understanding the Transdisciplinary Origins of Data Quality Concepts and Approaches*. Annual Review of Statistics and Its Application, Vol. 4:86-90 (Volume publication date March 2017). DOI: https://doi.org/10.1146/annurev-statistics-060116-054114.
25 Knublauch, H., Hendler, J. A. & Idehen, K., 2011. *SPIN - Overview and Motivation* [Online]. Available at: https://www.w3.org/Submission/spin-overview/ [Accessed 17 MAR 2020].
26 Knublauch, H. & Kontokostas, D., 2017. *Shapes Constraint Language (SHACL)* [Online]. Available at: https://www.w3.org/TR/shacl/ [Accessed 17 MAR 2020].
27 Manola, F. & Miller, E., 2004. *RDF Primer* [Online]. Available at: https://www.w3.org/TR/rdf-primer/ [Accessed 12 MAR 2020].
28 Mao, Y., 2015. *Data visualization in exploratory data analysis: An overview of methods and technologies* [Online]. Available at: https://rc.library.uta.edu/uta-ir/bitstream/handle/10106/25475/MAO-THESIS-2015.pdf?sequence=1&isAllowed=y [Accessed 10 MAR 2020].
29 Miles, A. & Bechhofer, S., 2009. *SKOS Simple Knowledge Organization System Reference* [Online]. Available at: https://www.w3.org/TR/skos-reference/ [Accessed 6 APR 2020].

30 Network of Excellence ReSIST, n.d. *ReSIST - Project Summary* [Online]. Available at: http://www.resist-noe.org/overview/summary.html [Accessed 20 MAR 2020].
31 NexusLinguarum WG1 members, 2020. *Teleconference* [Interview] (13 MAR 2020).
32 Nguyen, V. B., 2019. *'Definice pojmů a východisek pro výzkum*'. Unpublished internal document.
33 Raimond, Y., Sutton, C. & Sandler, M., 2008. *Automatic Interlinking of Music Datasets on the Semantic Web* [Online]. Available at: http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-369/paper18.pdf [Accessed 9 APR 2020].
34 Tomčová, L., 2014. *Datová kvalita v prostředí otevřených a propojitelných dat* (Data quality on the context of open and linked data). Master thesis. University of Economics, Prague, Ing. Dušan Chlapek, Ph.D. Also available at: https://vskp.vse.cz/id/1264666.
35 Wang, R. Y. & Strong, D. M., 1996. *Beyond accuracy: What data quality means to data consumers* [Online]. Available at: http://mitiq.mit.edu/Documents/Publications/TDQMpub/14_Beyond_Accuracy.pdf [Accessed 16 MAR 2020].
36 Wikidata, 2018. *Wikidata/Notes/DBpedia and Wikidata* [Online]. Available at: https://meta.wikimedia.org/wiki/Wikidata/Notes/DBpedia_and_Wikidata [Accessed 25 APR 2020].
37 Wikidata, 2019. *Wikidata:SPARQL query service* [Online]. Available at: https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service [Accessed 20 APR 2020].
38 Zeman, V., 2018. *RDFRules* [Online]. Available at: https://github.com/propi/rdfrules [Accessed 2 APR 2020].

# Annexes

## Annex A Datasets interlinked with DBpedia

Table 10: List of interlinked datasets (source: (DBpedia Association, ©2019))

| Data Set | Number of Links |
|---|---|
| ACM (RKBExplorer) | 5 |
| AEMET metereological dataset | 82 |
| AGROVOC | 993 |
| Airports | 9,761 |
| Alpine Ski Racers of Austria | 921 |
| Amsterdam Museum | 43 |
| BBC Music | 23 |
| BBC Programmes | 23,237 |
| BBC Wildlife Finder | 415 |
| BFS LD | 261 |
| BibBase | 53 |
| Bible Ontology | 371 |
| Brazilian Politicians | 1,500 |
| Bricklink | 25,797 |
| Chronicling America | 10 |
| CiteSeer (RKBExplorer) | 1 |
| Classical (DBtune) | 3 |
| Climbing | 300 |
| cnr.it | 34,706 |
| CORDIS | 285,256 |
| CORDIS (RKBExplorer) | 16 |
| Courseware (RKBExplorer) | 41 |
| DailyMed | 2,552 |
| DataGovIE | 70 |
| Datos.bcn.cl | 568 |
| datos.bne.es | 36,431 |
| DBLP (FU Berlin) | 100 |
| DBLP (RKBExplorer) | 2 |
| DBpedia in Portuguese | 365,839 |
| dbpedia lite | 10,000,000 |
| DBTropes | 6 |
| Didactalia (GNOSS) | 8,824 |
| Discogs in RDF | 5,169 |
| Diseasome | 1,943 |
| DrugBank | 422 |
| EARTh | 1,862 |

| Data Set | Number of Links |
|---|---|
| ECCO-TCP Eighteenth Century Texts Linked Data | 50 |
| ECS Southampton (RKBExplorer) | 11 |
| education.data.gov.uk | 1,697 |
| El Viajero's tourism dataset | 3,093 |
| Enipedia – Energy Industry Data | 1,365 |
| ERA (RKBExplorer) | 543 |
| ESD standards | 25 |
| EU: fintrans.publicdata.eu | 199,168 |
| EUNIS | 5,683 |
| EURES | 2,146 |
| Europeana | 1,304 |
| Eurostat (FU Berlin) | 129 |
| Eurostat (OntologyCentral) | 45 |
| EUTC Productions | 166 |
| EventMedia | 15,420 |
| FAO geopolitical ontology | 195 |
| FAO LD | 673 |
| farmers-markets-geographic-data-united-states | 52 |
| Finnish Municipalities | 336 |
| Fishes of Texas | 15,241 |
| flickr wrappr | 3,400,000 |
| Freebase | 3,348,530 |
| GBA Thesaurus | 100 |
| GEMET | 3,005 |
| GeoLinkedData | 51 |
| GeoSpecies Knowledge Base | 11,805 |
| GESIS | 5,024 |
| gnoss.com | 506 |
| Goodwin Family | 500 |
| GoogleArt wrapper | 1,632 |
| GovTrack | 470 |
| GovWILD | 5,845 |
| Greek DBpedia | 45 |
| GTAA | 25,844 |
| Hellenic FBD | 104,117 |
| Hellenic PD | 21,916 |
| Institutions and Bodies of the European Union | 154 |
| ISTAT Immigration (LinkedOpenData.it) | 319 |
| Italian Museums | 2,894 |
| John Peel (DBtune) | 1,143 |
| Klappstuhlclub | 50 |
| Last.FM (rdfize) | 23 |
| Lexvo | 2,577 |
| LIBRIS | 4,669 |

| Data Set | Number of Links |
|---|---|
| Lichfield District Council – Spending | 17 |
| lingvoj | 215 |
| Linked Clean Energy Data (reegle.info) | 330 |
| Linked Crunchbase (OntologyCentral) | 80 |
| LinkedCT | 25,476 |
| Linked EDGAR (OntologyCentral) | 50 |
| LinkedGeoData | 53,024 |
| LinkedLCCN | 10,911 |
| LinkedMDB | 30,354 |
| Linked Open Colors | 16,000,000 |
| Linked Open Numbers | 320 |
| lobid-organisations | 352 |
| lobid-Resources | 5,794 |
| lod.sztaki.hu | 13,034 |
| LODE | 10 |
| Lotico | 65 |
| Magnatune (DBtune) | 233 |
| MARC Codes List | 599 |
| meducator | 932 |
| morelab | 38 |
| Mortality (EnAKTing) | 5 |
| Moseley Folk | 18 |
| MusicBrainz (Data Incubator) | 76,171 |
| MusicBrainz (DBTune) | 64 |
| myExperiment | 2,586 |
| My Family Lineage | 2,254 |
| NASA (Data Incubator) | 61 |
| New York Times | 10,359 |
| Nomenclator Asturias 2010 | 78,859 |
| Norwegian Medical Subject Headings (MeSH) | 316 |
| NSF (RKBExplorer) | 1 |
| NSZL Catalog | 6,285 |
| NVD | 502 |
| Ocean Drilling – Codices | 3,022 |
| Ontos News Portal | 6,935 |
| OpenCalais | 1 |
| Open Corporates | 500 |
| OpenData Thesaurus | 50 |
| OpenEI.org | 52,546 |
| Open Election Data Project | 87 |
| Open Library (Talis) | 1,633 |
| Openly Local | 400 |
| Organisation for Economic Co-operation and Development (OECD) Linked Data | 2,613 |
| OS (RKBExplorer) | 156 |

| Data Set | Number of Links |
|---|---|
| P20 | 25 |
| PBAC | 1,607 |
| Pleiades | 127 |
| Pokedex (Data Incubator) | 493 |
| Poképédia | 493 |
| Polythematic Structured Subject Heading System | 3 |
| ProductDB | 193 |
| Product Types Ontology | 300 |
| Public Library of Veroia | 4,197 |
| radatana | 30,346 |
| RAE2001 (RKBExplorer) | 1 |
| RDFohloh | 1 |
| Rechtspraak.nl | 575 |
| reference.data.gov.uk | 22 |
| research.data.gov.uk | 3 |
| RESEX (RKBExplorer.com) | 11 |
| Revyu | 29 |
| Scholarometer | 1 |
| sears.com | 100 |
| SEC (rdfabout) | 86 |
| Semantic CrunchBase | 250 |
| Semantic XBRL | 63 |
| SIDER | 2,126 |
| smcjournals | 11 |
| Source Code Ecosystem Linked Data | 2,100 |
| SSW Thesaurus | 300 |
| STITCH | 123 |
| STW | 3 |
| Surge Radio | 1 |
| TaxonConcept | 147,877 |
| TCMGeneDIT Dataset | 1,400 |
| Telegraphis | 651 |
| Thesaurus W | 627 |
| The View From | 31 |
| totl.net | 500 |
| Transparency International LD | 183 |
| transport.data.gov.uk | 3,768 |
| Turismo de Zaragoza | 5,469 |
| Twarql | 981,415 |
| TWC LOGD | 2,039 |
| Uberblic.org | 1,196 |
| UK Legislation | 33 |
| UMBEL | 257 |
| UN/LOCODE (RKBExplorer) | 240 |

| Data Set | Number of Links |
|---|---|
| URIBurner | 1 |
| VIAF | 10 |
| VIVO Cornell | 58 |
| VIVO Indiana | 58 |
| VIVO UF | 58 |
| Weather Stations | 1,123 |
| Wiki (RKBExplorer) | 19 |
| WordNet (RKBExplorer) | 38 |
| World Bank LD | 380 |
| YAGO | 2,625,671 |
| Yahoo Geoplanet RDF | 248 |
| yovisto | 300 |
| Zhishi.me | 193 |

Table 11: List of interlinked datasets with added information (source: Author)

| Data Set | Number of Links | Data source | Availability | Data source type | Date | Alternative access | DBpedia inlinks | Last modified |
|---|---|---|---|---|---|---|---|---|
| ACM (RKBExplorer) | 5 | http://acm.rkbexplorer.com/ | true | web search, SPARQL, dump, dereferenceable URIs | 15/07/2019 | | false | 31/12/2009 |
| AEMET meteorological dataset | 82 | http://aemet.linkeddata.es/ | true | SPARQL | 15/07/2019 | | | 19/09/2011 |
| AGROVOC | 993 | http://aims.fao.org/standards/agrovoc/concept-scheme | true | SPARQL | 15/07/2019 | | false | |
| Airports | 9,761 | http://www.linklion.org/dataset/airports.dataincubator.org | false | | 15/07/2019 | https://archive.org/details/kasabi | | 20/07/2012 |
| Alpine Ski Racers of Austria | 921 | http://vocabulary.semantic-web.at/AustrianSkiTeam.html | true | SPARQL, dump, dereferenceable URIs | 15/07/2019 | | true | 25/11/2013 |
| Amsterdam Museum | 43 | http://semanticweb.cs.vu.nl/lod/am/ | false | SPARQL | 15/07/2019 | | | |
| BBC Music | 23 | http://www.bbc.co.uk/music | true | dereferenceable URIs | 15/07/2019 | https://archive.org/details/kasabi | true | 20/07/2012 |
| BBC Programmes | 23,237 | https://www.bbc.co.uk/programmes | false | | 15/07/2019 | https://archive.org/details/kasabi | | 20/07/2012 |
| BBC Wildlife Finder | 415 | http://www.bbc.co.uk/nature/wildlife | false | | 15/07/2019 | https://archive.org/details/kasabi | true | 20/07/2012 |

| Data Set | Number of Links | Data source | Availability | Data source type | Date | Alternative access | DBpedia inlinks | Last modified |
|---|---|---|---|---|---|---|---|---|
| BFS LD | 261 | http://linkeddatacatalog.dws.informatik.uni-mannheim.de/de/dataset/bfs-linked-data | false | SPARQL, dump | 15/07/2019 | | | |
| BibBase | 53 | http://linkeddatacatalog.dws.informatik.uni-mannheim.de/dataset/bibbase | false | SPARQL | 15/07/2019 | | | |
| Bible Ontology | 371 | http://bibleontology.com/ | false | | 15/07/2019 | | | |
| Brazilian Politicians | 1,500 | https://old.datahub.io/dataset/brazilian-politicians | false | | 25/11/2019 | | | |
| Bricklink | 25,797 | http://linkeddatacatalog.dws.informatik.uni-mannheim.de/de/dataset/bricklink | false | dump | 15/07/2019 | https://archive.org/details/kasabi | | 20/07/2012 |
| Chronicling America | 10 | http://chroniclingamerica.loc.gov/about/api/ | false | | 15/07/2019 | | | |
| CiteSeer (RKBExplorer) | 1 | http://citeseer.rkbexplorer.com/ | true | web search, SPARQL, dump, dereferenceable URIs | 15/07/2019 | | false | 29/12/2009 |
| Classical (DBtune) | 3 | http://dbtune.org/classical/ | true | SPARQL, dump, dereferenceable URIs | 15/07/2019 | | true | |

| Data Set | Number of Links | Data source | Availability | Data source type | Date | Alternative access | DBpedia inlinks | Last modified |
|---|---|---|---|---|---|---|---|---|
| Climbing | 300 | https://old.datahub.io/dataset/data-incubator-climb | false | SPARQL | 27/09/2019 | | | |
| cnr.it | 34,706 | http://data.cnr.it/site/ | false | SPARQL | 27/09/2019 | | | |
| CORDIS | 285,256 | https://data.europa.eu/euodp/data/dataset/cordisref-data | true | SPARQL | 27/09/2019 | | true | 10/12/2018 |
| CORDIS (RKBExplorer) | 16 | http://cordis.rkbexplorer.com/ | true | web search, SPARQL, dump, dereferenceable URIs | 27/09/2019 | | true | 28/12/2009 |
| Courseware (RKBExplorer) | 41 | http://courseware.rkbexplorer.com/ | true | web search, SPARQL, dump, dereferenceable URIs | 27/09/2019 | | | 23/12/2009 |
| DailyMed | 2,552 | https://dailymed.nlm.nih.gov/dailymed/spl-resources.cfm | true | web search | 27/09/2019 | | | |
| DataGovIE | 70 | https://data.gov.ie/ | true | API, CSV, JSON, XML, XLSX, KML | 27/09/2019 | | | |
| Datos.bcn.cl | 568 | http://datos.bcn.cl/es/ | true | SPARQL | 27/09/2019 | | false | |
| datos.bne.es | 36,431 | http://datos.bne.es/inicio.html | true | web search | 27/09/2019 | | | |
| DBLP (FU Berlin) | 100 | https://old.datahub.io/dataset/fu-berlin-dblp | false | SPARQL | 27/09/2019 | | | |

| Data Set | Number of Links | Data source | Availability | Data source type | Date | Alternative access | DBpedia inlinks | Last modified |
|---|---|---|---|---|---|---|---|---|
| DBLP (RKBExplorer) | 2 | http://dblp.rkbexplorer.com/ | true | web search, SPARQL, dump, dereferenceable URIs | 27/09/2019 | | | 18/02/2012 |
| DBpedia in Portuguese | 365,839 | http://pt.dbpedia.org/ | true | SPARQL, dump | 27/09/2019 | | | 03/04/2017 |
| dbpedia lite | 10,000,000 | http://dbpedialite.org/ | false | | 27/09/2019 | | | |
| DBTropes | 6 | skipforward.opendfki.de/wiki/DBTropes | true | dump | 27/09/2019 | | | 30/04/2015 |
| Didactalia (GNOSS) | 8,824 | https://didactalia.net/de/gemeinde/materialeducativo | true | web frontend | 27/09/2019 | | | |
| Discogs in RDF | 5,169 | https://old.datahub.io/dataset/data-incubator-discogs@2011-07-15T12:37:39.778095 | false | SPARQL | 27/09/2019 | https://archive.org/details/kasabi | | 20/07/2012 |
| Diseasome | 1,943 | https://old.datahub.io/dataset/fu-berlin-diseasome | false | SPARQL, dump | 27/09/2019 | | | |
| DrugBank | 422 | http://wifo5-03.informatik.uni-mannheim.de/drugbank/ | true | SPARQL, dump | 27/09/2019 | | | 31/12/2011 |

| Data Set | Number of Links | Data source | Availability | Data source type | Date | Alternative access | DBpedia inlinks | Last modified |
|---|---|---|---|---|---|---|---|---|
| EARTh | 1,862 | https://old.datahub.io/dataset/environmental-applications-reference-thesaurus | true | SPARQL, dump, dereferenceable URIs | 27/09/2019 | | true | 30/07/2016 |
| ECCO-TCP Eighteenth Century Texts Linked Data | 50 | https://old.datahub.io/dataset/ecco-tcp-linked-data | false | SPARQL | 27/09/2019 | https://archive.org/details/kasabi | | 20/07/2012 |
| ECS Southampton (RKBExplorer) | 11 | http://southampton.rkbexplorer.com/ | true | web search, SPARQL, dump, dereferenceable URIs | 27/09/2019 | | | 14/06/2012 |
| education.data.gov.uk | 1,697 | https://ckan.publishing.service.gov.uk/dataset | true | API, CSV, JSON, XML, XLSX | 27/09/2019 | | | |
| El Viajero's tourism dataset | 3,093 | http://webenemasuno.linkeddata.es/ | true | SPARQL | 02/10/2019 | | | |
| Enipedia – Energy Industry Data | 1,365 | https://old.datahub.io/dataset/enipedia | false | SPARQL, dump | 02/10/2019 | | | 30/07/2016 |
| ERA (RKBExplorer) | 543 | http://era.rkbexplorer.com/ | true | web search, SPARQL, dump, dereferenceable URIs | 02/10/2019 | | | 02/04/2010 |
| ESD standards | 25 | https://standards.esd.org.uk/? | true | API | 02/10/2019 | | | |

| Data Set | Number of Links | Data source | Availability | Data source type | Date | Alternative access | DBpedia inlinks | Last modified |
|---|---|---|---|---|---|---|---|---|
| EU: fintrans.publicdata.eu | 199,168 | https://old.datahub.io/dataset/beneficiaries-of-the-european-commission | false | SPARQL | 02/10/2019 | | | 30/07/2016 |
| EUNIS | 5,683 | http://eunis.eea.europa.eu/ | true | SPARQL, dump | 02/10/2019 | | | 31/12/2012 |
| EURES | 2,146 | https://old.datahub.io/dataset/fu-berlin-eures | false | SPARQL | 02/10/2019 | | | 30/07/2016 |
| Europeana | 1,304 | https://pro.europeana.eu/resources/datasets | true | SPARQL | 02/10/2019 | http://sparql.europeana.eu/ | | 03/12/2018 |
| Eurostat (FU Berlin) | 129 | https://old.datahub.io/dataset/fu-berlin-eurostat | true | SPARQL, dump | 02/10/2019 | | | 30/07/2016 |
| Eurostat (OntologyCentral) | 45 | http://ontologycentral.com/2009/01/eurostat/ | false | SPARQL, dump | 02/10/2019 | http://data.europa.eu/euodp/en/linked-data | | 18/12/2010 |
| EUTC Productions | 166 | https://bedlamtheatre.co.uk/ | false | | 02/10/2019 | | | |
| EventMedia | 15,420 | http://eventmedia.eurecom.fr/ | false | | 04/10/2019 | | | |
| FAO geopolitical ontology | 195 | http://www.fao.org/countryprofiles/geoinfo.asp?lang=en | false | | 04/10/2019 | http://www.fao.org/figis/flod/sparqlform.jsp | | |
| FAO LD | 673 | https://old.datahub.io/dataset/fao-linked-data | false | SPARQL, dump | 04/10/2019 | | | |

| Data Set | Number of Links | Data source | Availability | Data source type | Date | Alternative access | DBpedia inlinks | Last modified |
|---|---|---|---|---|---|---|---|---|
| farmers-markets-geographic-data-united-states | 52 | https://catalog.data.gov/dataset/farmers-markets-geographic-data | true | dump | 04/10/2019 | https://catalog.data.gov/dataset?q=farmers+markets&sort=views_recent+desc&res_format=RDF&as_sfid=AAAAAAWi2TcCYngAcxofu5aX7oMDmq4btjL67jg9Pc8-5RPT-Xlrwcp-bfdB-xjTruLk8-7fSh7_MOIzLDIjmbKgjciTTWhgM0UdZ39MDn6OmEm782EmRDliF2b-_ZTj94B4zvg%3D&as_fid=6b780e8862ee7064763cbddd6c8662d5c205566f&ext_location=&ext_bbox=&ext_prev_extent=-142.03125%2C8.754794702435617%2C-59.0625%2C61.773122864531446 | | 02/05/2019 |
| Finnish Municipalities | 336 | http://onki.fi/en/browser/overview/kunnat | true | API | 04/10/2019 | | | 31/12/2011 |
| Fishes of Texas | 15,241 | http://data.fishesoftexas.org/ | false | | 04/10/2019 | | | |
| flickr wrappr | 3,400,000 | http://wifo5-03.informatik.uni-mannheim.de/flickrwrappr/ | false | | 04/10/2019 | | | 27/04/2009 |
| Freebase | 3,348,530 | https://developers.google.com/freebase/ | true | dump | 04/10/2019 | | | 09/06/2013 |
| GBA Thesaurus | 100 | https://thesaurus.geolba.ac.at/ | true | SPARQL, dump | 04/10/2019 | | true | |

| Data Set | Number of Links | Data source | Availability | Data source type | Date | Alternative access | DBpedia inlinks | Last modified |
|---|---|---|---|---|---|---|---|---|
| GEMET | 3,005 | https://www.eionet.europa.eu/gemet/en/themes/ | true | dump, API | 04/10/2019 | | | |
| GeoLinkedData | 51 | http://www.oeg-upm.net/index.php/es/linkeddata/73-geolinkeddata/index.html | false | SPARQL | 04/10/2019 | | | |
| GeoSpecies Knowledge Base | 11,805 | http://lod.geospecies.org/ | true | dereferenceable URIs | 04/10/2019 | | | |
| GESIS | 5,024 | http://lod.gesis.org/thesoz/de.html | true | dump, dereferenceable URIs | 04/10/2019 | http://lod.gesis.org/thesoz-komplett.xml.gz | | |
| gnoss.com | 506 | https://www.gnoss.com/en/home | paid | | 04/10/2019 | | | |
| Goodwin Family | 500 | http://www.johngoodwin.me.uk/family/ | false | | 04/10/2019 | | | |
| GoogleArt wrapper | 1,632 | http://linkeddata.few.vu.nl/googleart/ | false | | 04/10/2019 | | | |
| GovTrack | 470 | https://www.govtrack.us/ | false | | 04/10/2019 | | | |
| GovWILD | 5,845 | http://govwild.hpi-web.de/project/govwild-sources.html | true | SPARQL | 04/10/2019 | http://govwild.hpi-web.de/sparql | | |
| Greek DBpedia | 45 | http://wiki.el.dbpedia.org/ | false | SPARQL, dump | 04/10/2019 | | | 14/10/2010 |
| GTAA | 25,844 | http://data.beeldengeluid.nl/gtaa/GTAA | true | API | 04/10/2019 | | | |

| Data Set | Number of Links | Data source | Availability | Data source type | Date | Alternative access | DBpedia inlinks | Last modified |
|---|---|---|---|---|---|---|---|---|
| Hellenic FBD | 104,117 | http://greek-lod.math.auth.gr/fire-brigade/ | false | | 04/10/2019 | | | |
| Hellenic PD | 21,916 | http://greek-lod.math.auth.gr/police/ | false | | 04/10/2019 | | | |
| Institutions and Bodies of the European Union | 154 | http://institutions.publicdata.eu/ | false | | 04/10/2019 | | | |
| ISTAT Immigration (LinkedOpenData.it) | 319 | https://linkedopendata.it/ | false | | 04/10/2019 | | | |
| Italian Museums | 2,894 | https://data.wu.ac.at/schema/datahub_io/NjNjZjg2YWMtMTI5Ni00ODdlLTg3ODItNzlkNzUzMTlkZTNl | true | SPARQL, API, web search | 04/10/2019 | | | 30/07/2016 |
| John Peel (DBtune) | 1,143 | http://dbtune.org/bbc/peel/ | true | SPARQL | 04/10/2019 | | | |
| Klappstuhlclub | 50 | http://www.klappstuhlclub.de/wp/ | true | dump | 04/10/2019 | | | 05/09/2018 |
| Last.FM (rdfize) | 23 | http://lastfm.rdfize.com/ | true | API | 04/10/2019 | | | |
| Lexvo | 2,577 | http://www.lexvo.org/ | true | SPARQL, dump, dereferenceable URIs | 04/10/2019 | | true | 31/12/2018 |
| LIBRIS | 4,669 | http://libris.kb.se/ | true | SPARQL | 04/10/2019 | http://libris.kb.se/sparql | | |

| Data Set | Number of Links | Data source | Availability | Data source type | Date | Alternative access | DBpedia inlinks | Last modified |
|---|---|---|---|---|---|---|---|---|
| Lichfield District Council – Spending | 17 | https://data.lichfieldc.gov.uk/View/spending/ | true | XML | 04/10/2019 | | | 31/10/2019 |
| lingvoj | 215 | http://linkedvocabs.org/lingvoj/ | true | dump, dereferenceable URIs | 04/10/2019 | | true | 12/06/2015 |
| Linked Clean Energy Data (reegle.info) | 330 | https://www.reeep.org/reegle-clean-energy-information-portal | true | SPARQL | 04/10/2019 | http://poolparty.reegle.info/PoolParty/sparql/glossary | true | |
| Linked Crunchbase (OntologyCentral) | 80 | http://km.aifb.kit.edu/services/crunchbase/ | true | SPARQL (authentication required), dump | 04/10/2019 | | | 06/02/2016 |
| Linked EDGAR (OntologyCentral) | 50 | http://edgarwrap.ontologycentral.com/ | true | dump | 04/10/2019 | | | 26/07/2014 |
| Linked Open Colors | 16,000,000 | http://linkedopencolors.appspot.com/ | false | | 04/10/2019 | | | |
| Linked Open Numbers | 320 | http://km.aifb.kit.edu/projects/numbers/ | true | dereferenceable URIs | 04/10/2019 | | | 01/04/2010 |
| LinkedCT | 25,476 | http://linkedct.org/ | true | dump | 04/10/2019 | http://www.cs.toronto.edu/~oktie/linkedct/linkedct-live-dump-latest.nt.bz2 | | |
| LinkedGeoData | 53,024 | http://linkedgeodata.org/About | true | SPARQL, dump | 04/10/2019 | | | 02/11/2015 |
| LinkedLCCN | 10,911 | http://purl.org/NET/lccn/ | false | SPARQL, RDFa | 04/10/2019 | https://old.datahub.io/cs_CZ/dataset/linkedlccn | | |

| Data Set | Number of Links | Data source | Availability | Data source type | Date | Alternative access | DBpedia inlinks | Last modified |
|---|---|---|---|---|---|---|---|---|
| LinkedMDB | 30,354 | https://old.datahub.io/dataset/linkedmdb | false | SPARQL, dump | 04/10/2019 | http://www.cs.toronto.edu/~oktie/linkedmdb/ | | 30/07/2016 |
| lobid-organisations | 352 | http://lobid.org/organisations/api/de | true | dump | 04/10/2019 | | false | 31/03/2017 |
| lobid-Resources | 5,794 | http://lobid.org/resources/api | true | dump | 04/10/2019 | | false | 31/03/2017 |
| lod.sztaki.hu | 13,034 | http://lod.sztaki.hu/ | true | SPARQL, dump | 04/10/2019 | | | |
| LODE | 10 | https://old.datahub.io/dataset/linked-open-data-of-ecology | false | SPARQL, dump | 07/10/2019 | | | 30/07/2016 |
| Lotico | 65 | http://www.lotico.com/index.php/Lotico | false | SPARQL, dereferenceable URIs, RDF browser | 07/10/2019 | | | |
| Magnatune (DBtune) | 233 | http://dbtune.org/magnatune/ | true | SPARQL, dump | 07/10/2019 | | | |
| MARC Codes List | 599 | https://github.com/rsinger/LinkedMARCCodes | true | dump | 07/10/2019 | | | 18/01/2012 |
| meducator | 932 | http://linkededucation.org/meducator | false | SPARQL, API | 07/10/2019 | | | |
| morelab | 38 | http://apps.morelab.deusto.es/teseo/sparql | false | SPARQL | 07/10/2019 | | | |
| Mortality (EnAKTing) | 5 | http://mortality.psi.enakting.org/ | false | | 07/10/2019 | | | |

| Data Set | Number of Links | Data source | Availability | Data source type | Date | Alternative access | DBpedia inlinks | Last modified |
|---|---|---|---|---|---|---|---|---|
| Moseley Folk | 18 | https://old.datahub.io/tr/dataset/data-incubator-moseley | false | SPARQL, dump | 07/10/2019 | https://archive.org/download/kasabi/moseley-folk-festival-data.gz | | 30/07/2016 |
| MusicBrainz (Data Incubator) | 76,171 | https://old.datahub.io/dataset/data-incubator-musicbrainz | false | SPARQL | 07/10/2019 | | | 30/07/2016 |
| MusicBrainz (DBTune) | 64 | http://dbtune.org/musicbrainz/ | true | SPARQL | 07/10/2019 | | | |
| My Family Lineage | 2,254 | https://old.datahub.io/dataset/my-family-lineage | false | SPARQL, dump | 07/10/2019 | | | |
| myExperiment | 2,586 | https://www.myexperiment.org/home | false | SPARQL | 07/10/2019 | | | |
| NASA (Data Incubator) | 61 | https://old.datahub.io/ne/dataset/data-incubator-nasa/resource/8e452e60-92eb-4081-b0a8-e68ca2c76525 | true | SPARQL, dump | 07/10/2019 | https://ia601601.us.archive.org/6/items/kasabi/nasa.gz | | 20/07/2012 |
| New York Times | 10,359 | https://old.datahub.io/dataset/nytimes-linked-open-data | false | dump | 07/10/2019 | | | 30/07/2016 |
| Nomenclator Asturias 2010 | 78,859 | https://old.datahub.io/dataset/nomenclator-asturias | false | SPARQL, dump | 07/10/2019 | | | 11/10/2013 |

| Data Set | Number of Links | Data source | Availability | Data source type | Date | Alternative access | DBpedia inlinks | Last modified |
|---|---|---|---|---|---|---|---|---|
| Norwegian Medical Subject Headings (MeSH) | 316 | http://folk.ntnu.no/greenall/nenmesh/ | false | | 07/10/2019 | | | |
| NSF (RKBExplorer) | 1 | http://nsf.rkbexplorer.com/ | true | web search, SPARQL, dump, dereferenceable URIs | 07/10/2019 | | | 29/12/2009 |
| NSZL Catalog | 6,285 | http://nektar.oszk.hu/wiki/Semantic_web | false | SPARQL | 07/10/2019 | | | |
| NVD | 502 | https://old.datahub.io/cs_CZ/dataset/nvd | true | SPARQL, dump | 07/10/2019 | | | 30/07/2016 |
| Ocean Drilling – Codices | 3,022 | https://old.datahub.io/ca/dataset/oceandrilling-codices | false | SPARQL, dump | 07/10/2019 | | | 30/07/2016 |
| Ontos News Portal | 6,935 | https://lod-cloud.net/dataset/ontos-news-portal | false | dump | 07/10/2019 | | | |
| Open Corporates | 500 | https://opencorporates.com/ | paid | | 07/10/2019 | | | |
| Open Election Data Project | 87 | https://old.datahub.io/dataset/open-election-data-project | false | dump | 07/10/2019 | | | |
| Open Library (Talis) | 1,633 | https://old.datahub.io/cs_CZ/dataset/talis-openlibrary | true | SPARQL, dump | 07/10/2019 | https://openlibrary.org/data/ol_dump_latest.txt.gz | | 30/07/2016 |

| Data Set | Number of Links | Data source | Availability | Data source type | Date | Alternative access | DBpedia inlinks | Last modified |
|---|---|---|---|---|---|---|---|---|
| OpenCalais | 1 | https://old.datahub.io/dataset/opencalais | false | dump | 07/10/2019 | | | |
| OpenData Thesaurus | 50 | http://vocabulary.semantic-web.at/PoolParty/wiki/OpenData | true | SPARQL, dereferenceable URIs | 07/10/2019 | | true | 15/05/2011 |
| OpenEI.org | 52,546 | https://openei.org/wiki/Main_Page | false | SPARQL | 07/10/2019 | | | |
| Openly Local | 400 | https://old.datahub.io/dataset/openlylocal | false | API | 07/10/2019 | | | 30/07/2016 |
| Organisation for Economic Co-operation and Development (OECD) Linked Data | 2,613 | http://oecd.270a.info/ | false | SPARQL | 07/10/2019 | https://github.com/csarven/oecd-linked-data | | |
| OS (RKBExplorer) | 156 | http://os.rkbexplorer.com/ | true | web search, SPARQL, dump, dereferenceable URIs | 07/10/2019 | | | 02/10/2009 |
| P20 | 25 | http://zbw.eu/beta/sparql-lab/about/ | true | SPARQL | 07/10/2019 | | | |
| PBAC | 1,607 | http://keithalexander.co.uk/pbac/ | true | web search | 07/10/2019 | | | |
| Pleiades | 127 | https://pleiades.stoa.org/downloads | true | dump | 07/10/2019 | | false | 25/11/2019 |

| Data Set | Number of Links | Data source | Availability | Data source type | Date | Alternative access | DBpedia inlinks | Last modified |
|---|---|---|---|---|---|---|---|---|
| Pokedex (Data Incubator) | 493 | https://old.datahub.io/cs_CZ/dataset/data-incubator-pokedex | true | SPARQL, dump | 07/10/2019 | https://archive.org/download/kasabi/pokedex-data-rdf.gz | | 20/07/2012 |
| Poképédia | 493 | https://www.pokepedia.fr/Portail:Accueil | false | | 07/10/2019 | | | |
| Polythematic Structured Subject Heading System | 3 | http://psh.ntkcz.cz/skos/home/html/en | false | | 07/10/2019 | | | |
| Product Types Ontology | 300 | http://www.productontology.org/ | true | dump | 07/10/2019 | | | 25/11/2019 |
| ProductDB | 193 | https://old.datahub.io/dataset/productdb | false | SPARQL | 07/10/2019 | | | |
| Public Library of Veroia | 4,197 | http://libver.math.auth.gr/sparql | true | SPARQL | 07/10/2019 | | | |
| radatana | 30,346 | https://old.datahub.io/dataset/radatana | false | SPARQL, dump | 07/10/2019 | | | |
| RAE2001 (RKBExplorer) | 1 | http://rae2001.rkbexplorer.com/ | true | web search, SPARQL, dump, dereferenceable URIs | 07/10/2019 | | | 30/12/2009 |
| RDFohloh | 1 | http://rdfohloh.wikier.org/ | not determined | | 07/10/2019 | | | |
| Rechtspraak.nl | 575 | https://old.datahub.io/cs_CZ/dataset/rechtspraak | false | SPARQL | 07/10/2019 | | | |

| Data Set | Number of Links | Data source | Availability | Data source type | Date | Alternative access | DBpedia inlinks | Last modified |
|---|---|---|---|---|---|---|---|---|
| reference.data.gov.uk | 22 | https://old.datahub.io/dataset/reference-data-gov-uk | false | SPARQL, dump, API | 07/10/2019 | | | |
| research.data.gov.uk | 3 | http://research.data.gov.uk/ | false | | 07/10/2019 | | | |
| RESEX (RKBExplorer.com) | 11 | http://resex.rkbexplorer.com/ | true | web search, SPARQL, dump, dereferenceable URIs | 07/10/2019 | | | 28/12/2009 |
| Revyu | 29 | http://revyu.com/ | true | SPARQL | 07/10/2019 | | | 30/04/2013 |
| Scholarometer | 1 | https://old.datahub.io/dataset/scholarometer | false | | 07/10/2019 | | | 30/07/2016 |
| sears.com | 100 | https://www.sears.com/en_intnl/dap/shopping-tourism.html | not determined | | 07/10/2019 | | | |
| SEC (rdfabout) | 86 | https://old.datahub.io/dataset/sec-rdfabout | false | SPARQL, dump | 07/10/2019 | | | 30/07/2016 |
| Semantic CrunchBase | 250 | http://cb.semsol.org/ | false | | 07/10/2019 | | | |
| Semantic XBRL | 63 | https://old.datahub.io/dataset/semantic-xbrl | true | dump | 07/10/2019 | | | |
| SIDER | 2,126 | https://old.datahub.io/dataset/fu-berlin-sider | false | SPARQL, dump | 07/10/2019 | | | 30/07/2016 |

| Data Set | Number of Links | Data source | Availability | Data source type | Date | Alternative access | DBpedia inlinks | Last modified |
|---|---|---|---|---|---|---|---|---|
| smcjournals | 11 | https://datahub.ckan.io/ar/dataset/data-incubator-smcjournals | false | SPARQL | 16/10/2019 | | | 30/07/2016 |
| Source Code Ecosystem Linked Data | 2,100 | https://sites.google.com/site/asegsecold/download | false | dump | 16/10/2019 | | | |
| SSW Thesaurus | 300 | http://vocabulary.semantic-web.at/PoolParty/wiki/semweb | true | SPARQL, dereferenceable URIs | 16/10/2019 | | true | 20/04/2012 |
| STITCH | 123 | https://old.datahub.io/dataset/fu-berlin-stitch | false | SPARQL | 16/10/2019 | | | 30/07/2016 |
| STW | 3 | http://zbw.eu/stw/version/latest/about | true | dump, dereferenceable URIs | 16/10/2019 | | true | 15/08/2018 |
| Surge Radio | 1 | https://datahub.ckan.io/sk/dataset/surge-radio | false | | 16/10/2019 | | | 30/07/2016 |
| TaxonConcept | 147,877 | https://old.datahub.io/dataset/taxonconcept | partial | SPARQL, dump | 16/10/2019 | | | 30/07/2016 |
| TCMGeneDIT Dataset | 1,400 | https://code.google.com/archive/p/junsbriefcase/wikis/TGDdataset.wiki | false | | 16/10/2019 | | | |
| Telegraphis | 651 | http://telegraphis.net/data/ | true | SPARQL, dump | 16/10/2019 | | | |

| Data Set | Number of Links | Data source | Availability | Data source type | Date | Alternative access | DBpedia inlinks | Last modified |
|---|---|---|---|---|---|---|---|---|
| The View From | 31 | https://datahub.ckan.io/de/dataset/the-view-from | false | SPARQL, API | 16/10/2019 | | | 11/10/2013 |
| Thesaurus W | 627 | https://old.datahub.io/dataset/thesaurus-w | false | SPARQL, dump | 16/10/2019 | | | 11/10/2013 |
| totl.net | 500 | http://data.totl.net/ | true | dump | 16/10/2019 | | | |
| Transparency International LD | 183 | https://old.datahub.io/dataset/transparency-linked-data | false | SPARQL, dump | 16/10/2019 | | | 30/07/2016 |
| transport.data.gov.uk | 3,768 | http://naptan.app.dft.gov.uk/datarequest/help | true | XML, CSV | 16/10/2019 | | | 15/11/2019 |
| Turismo de Zaragoza | 5,469 | https://datos.gob.es/es/catalogo/l01502973-planifica-tu-visita | true | SPARQL | 16/10/2019 | | | 24/11/2019 |
| Twarql | 981,415 | https://old.datahub.io/dataset/twarql | false | SPARQL, dump | 16/10/2019 | | | 30/07/2016 |
| TWC LOGD | 2,039 | https://old.datahub.io/dataset/twc-logd | false | SPARQL (authentication required), dump | 16/10/2019 | | | 30/07/2016 |
| Uberblic.org | 1,196 | https://old.datahub.io/dataset/uberblic | false | SPARQL, dump | 16/10/2019 | | | 30/07/2016 |
| UK Legislation | 33 | https://data.gov.uk/dataset/a2416481-271a-42b2-ace8-fc247dd251be/legislation-api | false | SPARQL, API | 16/10/2019 | | | 18/03/2014 |

| Data Set | Number of Links | Data source | Availability | Data source type | Date | Alternative access | DBpedia inlinks | Last modified |
|---|---|---|---|---|---|---|---|---|
| UMBEL | 257 | https://old.datahub.io/dataset/umbel | false | dump | 16/10/2019 | | | 30/07/2016 |
| UN/LOCODE (RKBExplorer) | 240 | http://unlocode.rkbexplorer.com/ | true | web search, SPARQL, dump, dereferenceable URIs | 16/10/2019 | | | 29/12/2009 |
| URIBurner | 1 | https://old.datahub.io/dataset/uriburner | true | SPARQL | 16/10/2019 | | | 30/07/2016 |
| VIAF | 10 | http://viaf.org/viaf/data/ | true | dump, dereferenceable URIs | 16/10/2019 | | | 04/11/2019 |
| VIVO Cornell | 58 | https://datahub.ckan.io/bg/dataset/vivo-cornell-university | false | dump | 16/10/2019 | | | 30/07/2016 |
| VIVO Indiana | 58 | https://old.datahub.io/dataset/vivo-indiana-university | false | dump | 16/10/2019 | | | 30/07/2016 |
| VIVO UF | 58 | https://old.datahub.io/dataset/vivo-university-of-florida | false | SPARQL, dump | 16/10/2019 | | | 30/07/2016 |
| Weather Stations | 1,123 | https://toolbox.google.com/datasetsearch/search?query=Weather%20Stations&docid=7SlQ11uksIRyxj7RAAAAAA%3D%3D | not determined | | 25/11/2019 | | | |

| Data Set | Number of Links | Data source | Availability | Data source type | Date | Alternative access | DBpedia inlinks | Last modified |
|---|---|---|---|---|---|---|---|---|
| Wiki (RKBExplorer) | 19 | http://wiki.rkbexplorer.com/ | true | web search, SPARQL, dump, dereferenceable URIs | 16/10/2019 | | | 28/12/2009 |
| WordNet (RKBExplorer) | 38 | http://wordnet.rkbexplorer.com/ | true | web search, SPARQL, dump, dereferenceable URIs | 16/10/2019 | | | 02/10/2009 |
| World Bank LD | 380 | https://old.datahub.io/dataset/world-bank-linked-data | false | SPARQL, dump | 16/10/2019 | | | 30/07/2016 |
| YAGO | 2,625,671 | https://datahub.io/collections/yago | true | SPARQL, dump | 16/10/2019 | | | 08/01/2019 |
| Yahoo Geoplanet RDF | 248 | https://datahub.ckan.io/he/dataset/yahoo_geoplanet | false | SPARQL, dump | 16/10/2019 | https://archive.org/download/kasabi/yahoo-geoplanet.gz | | 20/07/2012 |
| yovisto | 300 | https://datahub.ckan.io/ca/dataset/yovisto | true | SPARQL, dump | 16/10/2019 | | | 30/07/2016 |
| Zhishi.me | 193 | http://zhishi.me/ | true | SPARQL | 16/10/2019 | | | |

# Annex B Annotator for FRBR in DBpedia

This annex describes the end-to-end process of the development of an application that supports annotators in the task of choosing appropriate FRBR class to which a DBpedia entity belongs.

The development process has been started by requirement specification, continued through deciding the architecture of the solution, implementing the chosen solution and testing the implementation, eventually leading to the deployment to a production environment and the operation of the application.

## B.1 Requirements

The first step to successfully develop any program is to know what needs it is supposed to satisfy. In this case, the functional requirements are rather straightforward:

- It must integrate with a SPARQL endpoint which can retrieve data from DBpedia using a federated query. (Accomplished by using Apache Jena for retrieving data from the endpoint.)
- It must be a multi-user application. (Facilitated by using the web technologies as the underlying infrastructure.)
- It must require authentication for traceability. (Done by my own implementation of account creation and login component with secure storage of salted passwords.)
- It must provide its users the means necessary to select the FRBR type of the entity they are annotating. (Done by a web form displayed together with the data.)
- It must provide a way to persistently store annotations. (Achieved by using Apache Jena to insert triples into the database.)

The non-functional requirements that needed to be explicitly addressed were:

- User experience, which was accommodated by including a page explaining the differences between FRBR categories and by having online support via e-mail and other communication channels provided by the developer (author of this thesis).
- Performance, which was not specifically addressed, but which was tested using Apache JMeter™ as described in subsection dedicated to Testing.

## B.2 Architecture

Given the need for multi-user access to the application, it was clear that the application needs to be client-server and the need for its own triplestore combining data from DBpedia and Wikidata means that a classical 3-tier architecture suits this application the most.

Because of the limited scope of this project, it was decided that the application should be monolithic rather than based on microservices so that time can be spent on feature

development and testing rather than on integrating components. This decision also greatly simplified integration testing, which was reduced to the integration between the client programmed in JavaScript and server written in Java, and the integration of the server and Apache Jena Fuseki running as a standalone server. Had the application been developed as a collection of microservices, it would have increased the number of components at least twice, because it would have likely contained a dedicated service for authentication and a dedicated service for interfacing with Fuseki database server.

The entire process of annotating a DBpedia resource is described by Figure 3. The diagram can be summarized by a couple of steps:

1. Client requests data.
2. Server retrieves data and formats it such that the client can render it in a way that is meaningful to the annotator.
3. The annotator chooses to:
    - select a FRBR type and immediately submit it to the server,
    - select a FRBR type and request data about a new resource to annotate while the annotated resource is temporarily stored on the client,
    - select a FRBR type and request data about a previously annotated resources to change the annotation,
    - request data about a new resource without selecting a FRBR type,
    - request previously annotated resource without selecting FRBR type or
    - submit the data without selecting a FRBR type.
4. In either case the server responds accordingly either with data about a new resource, data about an already annotated resource or by confirming that data have been successfully annotated.

The motivation to specifically allow the annotator to choose to ignore a resource is that some resources may not be described well enough to decide about the correct FRBR type. Leaving the resource not annotated would thus be preferable to annotating it just to get past it to another resource, for which there would hopefully be enough information available

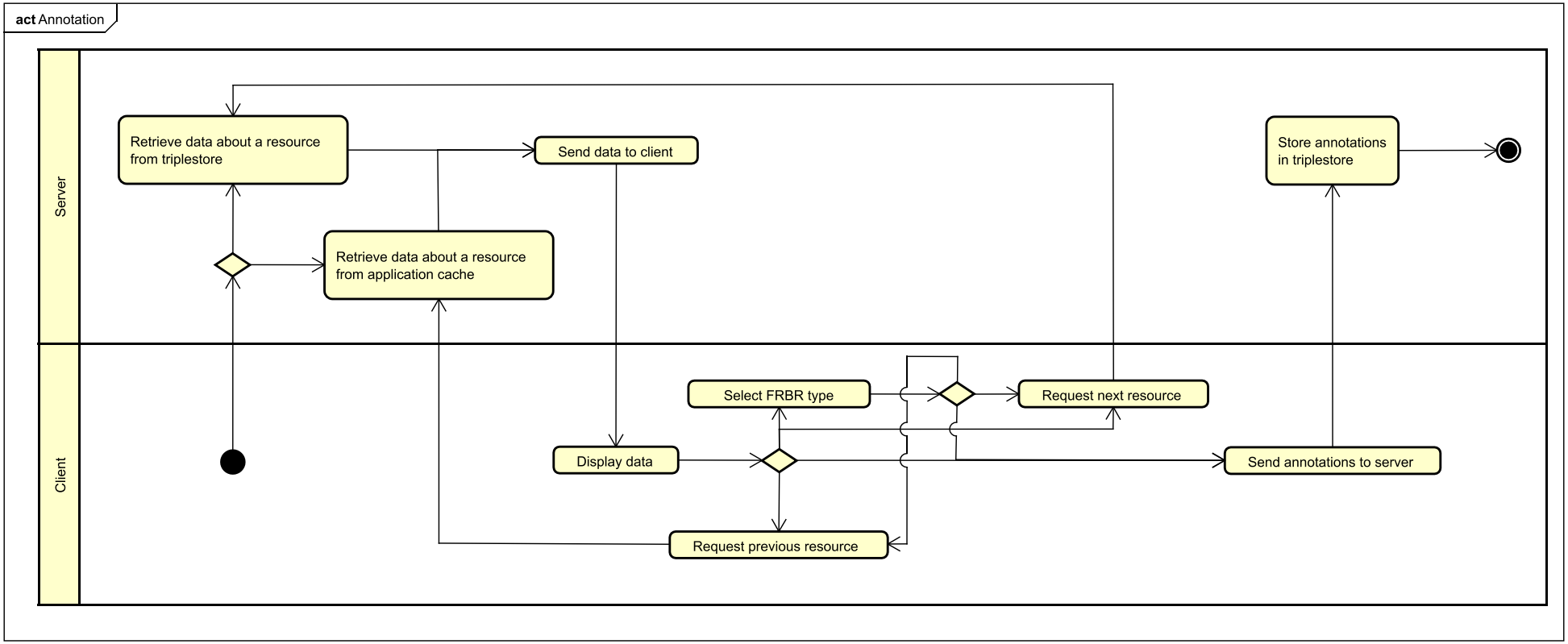


Figure 3: Diagram depicting the annotation process (source: Author)

## B.3 Implementation

The most crucial dependency of the application is the Apache Jena framework, which provides APIs needed by LD applications. It can be easily managed using Apache Maven, which is a tool supporting software project management. By specifying Apache Jena as a dependency in Maven project object model (POM) as appears in the XML snippet in Code B.3.1, it is possible to automate dependency management of the project.

Code B.3.1: Declaring Apache Jena as a dependency in Maven POM (source: Author)

```
<dependency>
  <groupId>org.apache.jena</groupId>
  <artifactId>apache-jena-libs</artifactId>
  <type>pom</type>
  <version>3.13.1</version>
</dependency>
```

What also helped with the development of this application is that some Java implementations, including OpenJDK 13, offer an HTTP server in package `com.sun.net.httpserver.HttpServer`.

With the help of this pre-existing software, the development could focus mainly on the application logic. The advantages gained from building upon the readily available Open-Source implementations are widely known in the industry. The most relevant ones for this project are that (Balter, 2015):

- developers can focus on high-value work, which in this context means more time spent on application logic and less time wasted on infrastructure related tasks, and
- higher quality software, because Open-Source software has been empirically proven to be of higher quality than even proprietary software, let alone software developed by only one rather inexperienced developer severely constrained by time.

Example Code B.3.2 demonstrates how the application is divided into subsystems for improved maintainability and readability. The two most important parts are handlers for requests, with the most generic one being used for filling the structure of the table, as depicted at snippet Code B.3.3, on the server and serving it is plain XHTML. The second handler is used to serve only data to the client which then updates the table without having to reload the whole page and the resources associated with it.

Code B.3.2: Starting the HTTP server (source: Author)

```
server = HttpServer.create(new InetSocketAddress(port),30);
server.createContext("/", Annotator::handleRequest);
server.createContext("/data", Annotator::handleDataRequest);
server.createContext("/auth", Annotator::handleAuthentication);
server.createContext("/s/", Annotator::doGetStaticFiles);
server.start();
```

Code B.3.3: Structure of the table presented to annotators (source: Author)

```
<table id="data">
    <caption></caption>
    <thead><tr><th>DBpedia predicate</th><th>DBpedia object</th></tr></thead>
    <tbody></tbody>
</table>
```

The reason for displaying only predicates and objects in the table body is that the subject is the same for all the triples as illustrated by query in Code B.3.4. The query also illustrates how the resources, which have already been annotated, are treated. They are simply filtered out by the SPARQL endpoint.

Code B.3.4: SPARQL query to retrieve not yet annotated triples (source: Author)

```
BASE <http://github.com/Fuchs-David/Annotator/tree/master/src/ontology/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX wd: <http://www.wikidata.org/entity/>
CONSTRUCT {
  ?dbr ?dbp ?dbo.
}
WHERE {
  {
    SELECT ?dbr ?wdr
    WHERE {
      ?dbr owl:sameAs ?wdr.
      ?wdr wdt:P31 ?wdc.
      FILTER (!isBlank(?dbr)).
      FILTER(strstarts(str(?wdr),"http://www.wikidata.org/")
             && strstarts(str(?dbr),"http://dbpedia.org/")).
      OPTIONAL{
        ?dbr a ?frbr_category.
        FILTER(strstarts(str(?frbr_category),"http://vocab.org/frbr/core.html#")).
      }
      FILTER(!bound(?frbr_category)).
      VALUES ?wdc {
        wd:Q207628 wd:Q2031291 wd:Q47461344 wd:Q3331189 wd:Q53731850 wd:Q87167 wd:Q213924
        wd:Q1440453 wd:Q834459 wd:Q2217259 wd:Q274076 wd:Q1754581 wd:Q690851 wd:Q284465
      }.
    }
    ORDER BY ?dbr
    OFFSET ?offset
    LIMIT ?limit
  }

  SERVICE <http://dbpedia.org/sparql> {
```

```
    ?dbr ?dbp ?dbo.
    FILTER(!strstarts(str(?dbo),"http://www.wikidata.org/")
    && !strstarts(str(?dbo),"http://wikidata.dbpedia.org/")).
    FILTER (!isBlank(?dbo)).
  }
}
```

## B.4 Testing

The testing was mainly done manually, except for a unit test that checks whether the server manages to start successfully, which functions as a smoke test, and a performance test. This testing strategy complies with Figure 4, although if the development was to continue for a prolonged period of time, it would have been beneficial to automate at least the end-to-end test to check that annotations are being added into the Fuseki database and a couple of 0-switch and 1-switch tests to check the client's behaviour using the Selenium framework.

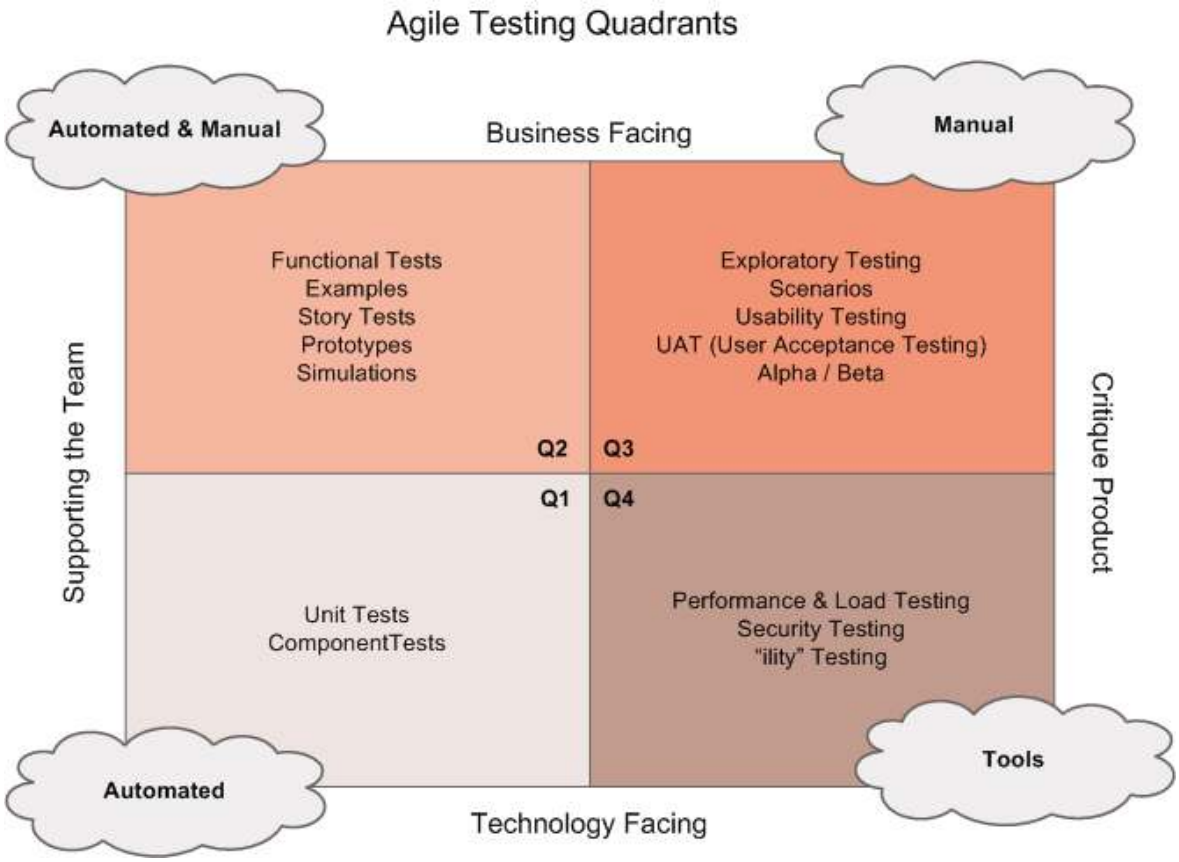


Figure 4: Automation quadrants in testing (source: (Crispin, 2011))

The automated tests were executed every time Maven reached the `test` phase, which meant that the application could be fixed quickly before any time was wasted trying to execute manual tests. Manual tests were carried out after every change in the code, but the exact extent of the tests varied widely depending on the part of the application that was modified. For example, changes to the authentication component required a new round of tests only for this component while changes to the code that serves DBpedia resources only meant that

the new round of tests would consist of the 0-switch and 1-switch tests as described later. On the other hand, modifications to code that stores annotations in Fuseki required an end-to-end test to be run.

Given that the application is to a large extent about keeping track of the state so that it can eventually associate annotations from the client with the underlying DBpedia resources, the mostly used method for test analysis and design was the state machine approach.

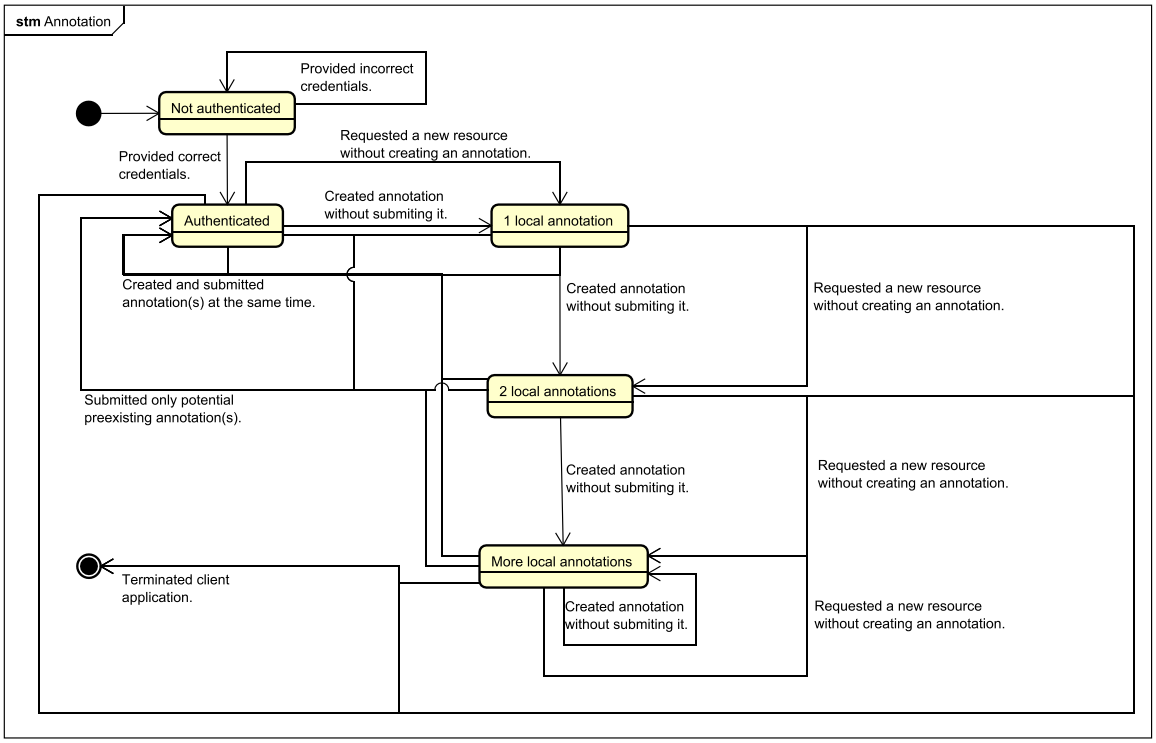


Figure 5: State machine diagram (source: Author)

### B.4.1 Functional testing

Given Figure 5, the only state that does not have a precondition is the initial Not authenticated state. The testing thus starts with test cases that validate the functionality of authentication, because other tests rely on this component to work to satisfy their prerequisites. The test cases for authentication and account creation, designed using the method of equivalence partitioning, are listed in the last parts of this annex B.6 and B.7.

When the positive test case for authentication passes, it is then possible to test application logic itself by exploratory testing based on the guidelines created using the state machine method of test analysis and design as listed on the next couple of lines.

- The 0-switch operations are:
    - creating an annotation which is immediately submitted to the server,
    - creating an annotation which is stored locally,
    - requesting an annotation without creating an annotation.
- The 1-switch operations include:

- the 0-switch operation of creating a locally stored annotation followed by the creation of another locally stored annotation to verify that the first annotation is not overwritten,
- the 0-switch operation of requesting a new resource without annotating the first one followed by the creation of an annotation for the second resource to verify that the annotation is associated to the correct resource.

### B.4.2 Performance testing

The test plan for performance tests is as follows:

1. Creation of accounts for virtual users is spread out across 128 seconds for 16 users.
2. Afterwards, each virtual user annotates four resources.
3. After all accounts are created, another thread group runs, which logs in the 16 users again spread out across a 128 second ramp-up time.
4. Each user proceeds to annotate additional four resources.

In Figure 6 the two easily distinguishable phases are apparent even though they partially overlap. It indicates that most of the activity during account creation phase occurred between 8:33 and 8:34, while activity in the login phase peaked during 8:37 and 8:38. When combined with information from Figure 7, we can see that the throughput peaked at 20 and 35 requests per second respectively for the two phases.


Figure 6: Thread count during performance test (source: Author)


Figure 7: Throughput in requests per second (source: Author)

An important measure of quality of software is the error rate of the system under load, which in this case remained at zero as displayed in Figure 8. Another measure determining quality of software under load is the time needed to respond to requests, which is presented by Figure 10. In addition, all these graphs need to be evaluated in the context of Figure 9, which displays the total numbers of requests over time.

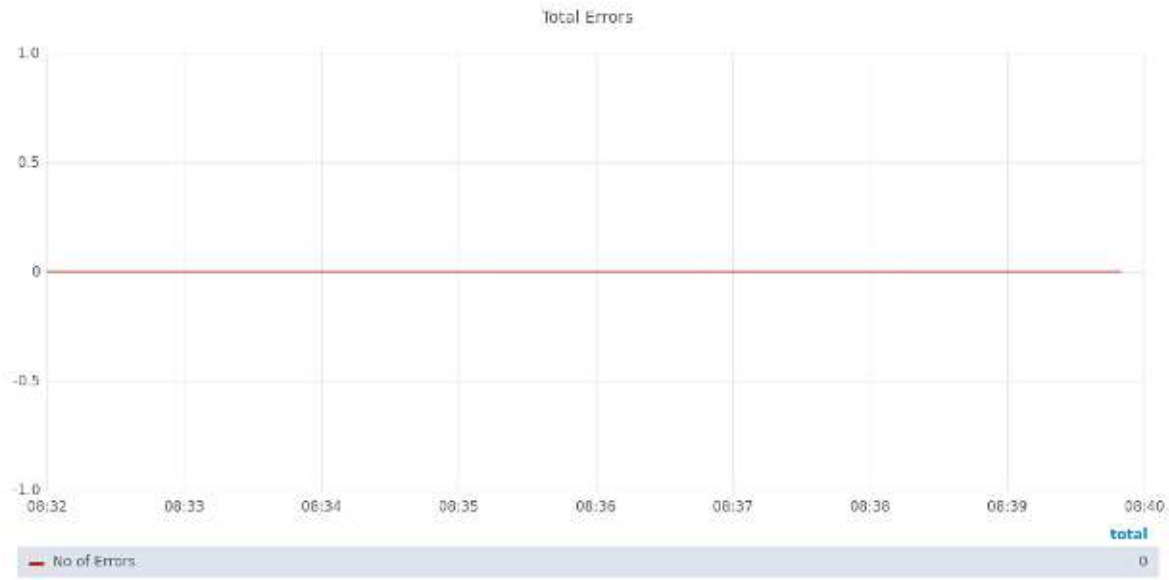


Figure 8: Error rate during test execution (source: Author)


Figure 9: Number of requests over time (source: Author)

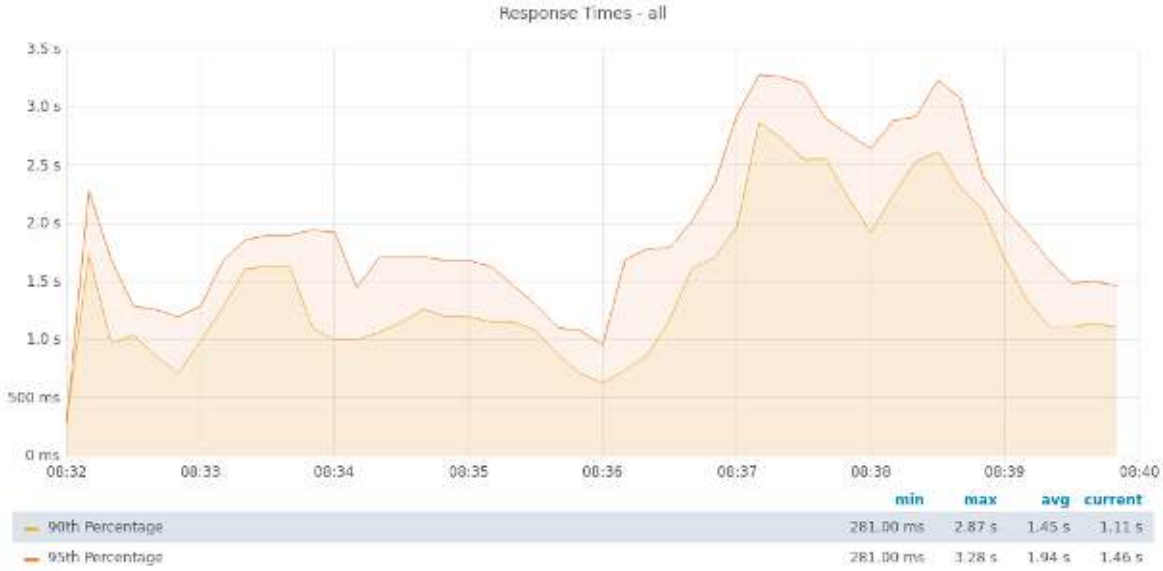


Figure 10: Response times over time (source: Author)

The performance tests revealed that the application can handle a load generated by 16 simultaneously working users without errors, although with increased response times. It was nevertheless required to release the application to a production environment where it would be freely accessible over the internet.

## B.5 Deployment and operation

Because the purpose of the development of this application is to collect data from volunteers who are willing to annotate DBpedia resources, it was necessary to plan the transition from the development phase of the project to operation and maintenance phase.

### B.5.1 Deployment

The deployment process has been automated mainly to reduce the time for test preparation, because that was the most time-consuming action during the process of test environment setup. It also had the benefit of rapid deployment to production, when the has come, and allowed for a quick reaction to user feedback. The automated deployment process now only consists of downloading the current version of source code and using `mvn install` command in the root of the extracted directory. The POM configuration is presented by snippets Code B.5.1.1 and Code B.5.1.3, while the script itself appears as Code B.5.1.2.

Code B.5.1.1: POM specification of setup script of the application (source: Author)

```
<configuration>
    <executable>./setup${script.extension}</executable>
    <workingDirectory>${basedir}</workingDirectory>
</configuration>
```

Code B.5.1.2: Setup script for UNIX systems (source: Author)

```
#!/bin/bash
mkdir -p ./target/security
cp ./security/auth.json ./target/security/
```

Code B.5.1.3: POM profiles for multi-platform support (source: Author)

```
<profiles>
    <profile>
        <id>Windows</id>
        <activation>
            <os><family>Windows</family></os>
        </activation>
        <properties>
            <script.extension>.bat</script.extension>
        </properties>
    </profile>
    <profile>
        <id>unix</id>
        <activation>
            <os><family>unix</family></os>
        </activation>
        <properties>
            <script.extension>.sh</script.extension>
        </properties>
    </profile>
</profiles>
```

### B.5.2 Operation

To ensure stability of the of the application, a cron job was set up to regularly check if the database is listening on its assigned port and to restart the database server as well as the application server if it is not running. Similarly, there is a cron job scheduled to run regularly on the application server which checks the health of the application by verifying that it responds to an HTTP request and restarts the application server when no response is obtained from the server.

Any issues with the application, like making sure that annotators understand the task correctly, were discussed using various online communication tools.

At the end, when it was time to retrieve the annotations, the query presented as Code B.5.2.1 provided all information DBpedia has about the annotated resources along with the linking triples between DBpedia and Wikidata. The data had to be retrieved in several iterations (by specifying different offsets), because the query would timeout otherwise.

Code B.5.2.1: Query to retrieve annotations and information about the annotated resources (source: Author)

```
BASE <http://github.com/Fuchs-David/Annotator/tree/master/src/ontology/>
CONSTRUCT {
  ?dbr owl:sameAs ?wdr.
  ?wdr wdt:P31 ?wdc.
  ?dbr <annotatedBy> ?annotator.
  ?dbr a ?frbr_category.
  ?dbr ?dbp ?dbo.
}
WHERE {
  {
    SELECT *
    WHERE {
      ?dbr owl:sameAs ?wdr.
      ?wdr wdt:P31 ?wdc.
      FILTER (!isBlank(?dbr)).

FILTER(strstarts(str(?wdr),"http://www.wikidata.org/")&&strstarts(str(?dbr),"http://dbpedi
a.org/")).
      ?dbr <annotatedBy> ?annotator.
      ?dbr a ?frbr_category.
      FILTER(strstarts(str(?frbr_category),"http://vocab.org/frbr/core.html#")).
      VALUES ?wdc {
        wd:Q207628 wd:Q2031291 wd:Q47461344 wd:Q3331189 wd:Q53731850 wd:Q87167 wd:Q213924
wd:Q1440453 wd:Q834459 wd:Q2217259 wd:Q274076 wd:Q1754581 wd:Q690851 wd:Q284465
      }.
    }
    ORDER BY ?dbr
```

```
    OFFSET 0
    LIMIT 25
  }
  SERVICE <https://dbpedia.org/sparql> {
    ?dbr ?dbp ?dbo.
  }
}
```

## B.6 Authentication test cases for application Annotator

Table 12: Positive authentication test case (source: Author)

| **Test case name** | **Authentication with valid credentials** | |
|---|---|---|
| **Test case type** | **positive** | |
| **Prerequisites** | Application contains a record with user *test@example.org* and password *testPassword*. | |
| **Step** | **Action** | **Result** |
| 1 | Navigate to the main page of the application. | You are redirected to the authentication page. |
| 2 | Fill in the e-mail address *test@example.org* and the password *testPassword* and submit the form. | The browser displays a message confirming a successfully completed authentication. |
| 3 | Press OK to continue. | You are redirected to a page with information about a DBpedia resource. |
| **Postconditions** | The user is authenticated and can use the application. | |

Table 13: Authentication with invalid e-mail address (source: Author)

| **Test case name** | **Authentication with invalid e-mail** | |
|---|---|---|
| **Test case type** | **negative** | |
| **Prerequisites** | Application contains a record with user *test@example.org* and password *testPassword*. | |
| **Step** | **Action** | **Result** |
| 1 | Navigate to the main page of the application. | You are redirected to the authentication page. |
| 2 | Fill in the e-mail address field with *test* and the password *testPassword* and submit the form. | The browser displays a message stating the e-mail is not valid. |
| **Postconditions** | The user is not authenticated and when accessing the main page is redirected to authenticate himself. | |

Table 14: Authentication with not registered e-mail address (source: Author)

| **Test case name** | **Authentication with not registered e-mail** | |
|---|---|---|
| **Test case type** | **negative** | |
| **Prerequisites** | Application does **not** contain a record with user *test@example.org* and password *testPassword*. | |
| **Step** | **Action** | **Result** |
| 1 | Navigate to the main page of the application. | You are redirected to the authentication page. |
| 2 | Fill in e-mail address *test@example.org* and password *testPassword* and submit the form. | The browser displays a message stating the e-mail is not registered or password is wrong. |
| **Postconditions** | The user is not authenticated and when accessing the main page is redirected to authenticate himself. | |

Table 15: Authentication with invalid password (source: Author)

| **Test case name** | **Authentication with invalid password** | |
|---|---|---|
| **Test case type** | **negative** | |
| **Prerequisites** | Application contains a record with user *test@example.org* and password *testPassword*. | |
| **Step** | **Action** | **Result** |
| 1 | Navigate to the main page of the application. | You are redirected to the authentication page. |
| 2 | Fill in the e-mail address *test@example.org* and password *wrongPassword* and submit the form. | The browser displays a message stating the e-mail is not registered or password is wrong. |
| **Postconditions** | The user is not authenticated and when accessing the main page is redirected to authenticate himself. | |

## B.7 Account creation test cases for application Annotator

Table 16: Positive test case of account creation (source: Author)

| **Test case name** | **Account creation with valid credentials** | |
|---|---|---|
| **Test case type** | **positive** | |
| **Prerequisites** | - | |
| **Step** | **Action** | **Result** |
| 1 | Navigate to the main page of the application. | You are redirected to the authentication page. |
| 2 | Select the option to create a new account, fill in e-mail address *test@example.org*, fill in password *testPassword* into both password fields, and submit the form. | The browser displays a message confirming a successful creation of an account. |
| 3 | Press OK to continue. | You are redirected to a page with information about a DBpedia resource. |
| **Postconditions** | Application contains a record with user *test@example.org* and password *testPassword*. The user is authenticated and can use the application. | |

Table 17: Account creation with invalid e-mail address (source: Author)

| **Test case name** | **Account creation with invalid e-mail address** | |
|---|---|---|
| **Test case type** | **negative** | |
| **Prerequisites** | - | |
| **Step** | **Action** | **Result** |
| 1 | Navigate to the main page of the application. | You are redirected to the authentication page. |
| 2 | Select the option to create a new account, fill in e-mail address field with *test*, fill in password *testPassword* into both password fields, and submit the form. | The browser displays a message that the credentials are invalid. |
| **Postconditions** | The user is not authenticated and when accessing the main page is redirected to authenticate himself. | |

Table 18: Account creation with non-matching password (source: Author)

| **Test case name** | **Account creation with not matching passwords** | |
|---|---|---|
| **Test case type** | **negative** | |
| **Prerequisites** | - | |
| **Step** | **Action** | **Result** |
| 1 | Navigate to the main page of the application. | You are redirected to the authentication page. |
| 2 | Select the option to create a new account, fill in e-mail address *test@example.org*, fill in password *testPassword* into password the password field and *differentPassword* into the repeated password field, and submit the form. | The browser displays a message that the credentials are invalid. |
| **Postconditions** | The user is not authenticated and when accessing the main page is redirected to authenticate himself. | |

Table 19: Account creation with already registered e-mail address (source: Author)

| **Test case name** | **Account creation with already registered e-mail** | |
|---|---|---|
| **Test case type** | **negative** | |
| **Prerequisites** | Application contains a record with user *test@example.org* and password *testPassword*. | |
| **Step** | **Action** | **Result** |
| 1 | Navigate to the main page of the application. | You are redirected to the authentication page. |
| 2 | Select the option to create a new account, fill in e-mail address *test@example.org*, fill in password *testPassword* into both password fields, and submit the form. | The browser displays a message stating that the e-mail is already used with an existing account. |
| **Postconditions** | The user is not authenticated and when accessing the main page is redirected to authenticate himself. | |