Quality of Linked Bibliographic Data: The Models, Vocabularies, and Links of Datasets Published by Four National Libraries

Kim Tallerås

Department of Archivistics, Library, and Information Science,

Oslo and Akershus University College of Applied Sciences

Abstract

Little effort has been devoted to the systematic examination of published Linked data in the library community. This paper examines the quality of Linked bibliographic data published by the national libraries of Spain, France, the United Kingdom, and Germany. The examination is mainly based on a statistical study of the vocabulary usage and interlinking practices in the published datasets. The study finds that the national libraries successfully adapt established Linked data principles, but issues at the data level can limit the fitness of use. In addition, the study reveals that these four libraries have chosen widely different solutions to all the aspects examined.

Introduction

Since Berners-Lee (2006) introduced principles for Linked data, large quantities of bibliographic descriptions have been published on the Web, resulting in Linked bibliographic data (LBD). Linked data principles are intended to facilitate a semantic web of data, enabling a variety of novel applications. A satisfactory level of output quality is essential to realize this vision. The library community continuously discusses issues concerning involved operations, such as data modelling, transformation, and interlinking. Less effort, however, has been devoted to systematic examination of the actual output, particularly the organization of data and various aspects of data quality.

This paper examines bibliographic metadata published as Linked data by four European national libraries: the Bibliothèque Nationale de France (BNF), British Library (BNB), Biblioteca Nacional de España (BNE), and Deutsche Nationalbibliothek (DNB). The study is motivated by the lack of systematic analysis of LBD and by the pioneering nature of these particular datasets. The study is aimed at answering the following research questions:

- How do prominent agents (and experts) in the library community organize and represent bibliographic collections of metadata when they publish these collections as Linked data on the Web?

- How do these Linked datasets conform to established measurements of Linked data quality for vocabulary usage and interlinking?

To answer these questions, concrete dimensions of Linked data quality are analyzed statistically. A qualitative close reading of selected corpus samples supplements the statistical data. The first section of this paper presents background information on LBD data and quality dimensions, clarifying the scope of the study. The following sections summarize previous research and present the corpus data and methodological considerations. The remaining sections provide the findings and concluding remarks.

Background and Motivation

Linked Data

Berners-Lee (2006) first described Linked data with four principles to help support bottom-up adoption of the semantic web:

- Use Uniform Resource Identifiers (URIs) as names for things.
- Use HTTP URIs so that people can look up those names.
- When someone looks up a URI, provide useful information, using the standards (Resource Description Framework (RDF), SPARQL protocol and RDF query language (SPARQL)).
- Include links to other URIs, so that users can discover more things.

To further "encourage people along the road to good Linked data," Berners-Lee (2006) later added a rating system of five stars reflecting these principles. The principles have since evolved into comprehensive collections of best practice recommendations, both as general guidelines (see, e.g., Heath & Bizer, 2011; Hyland, Atemezing, & Villazón-Terrazas, 2014) and as guidelines targeting data providers in specific domains (e.g., van Hooland & Verborgh, 2014). Summarized, they advocate open publication of structured data in non-proprietary formats based on W3C standards on the Web. Widely mentioned Web standards in this context, as exemplified by the principles developed by Berners-Lee (2006), are URIs which identify and address specific resources, RDF which provide the structure for the organization of those resources, and SPARQL which is used to retrieve RDF data. The emphasis on standards and transparency indicates a lingua franca approach to solving heterogeneity conflicts across domains and datasets.

Despite these detailed guidelines, studies show that Linked datasets are compliant with best practice principles to varying degrees (see the "Previous Studies of Linked (Bibliographic) Data Quality" section for details). Such studies mostly investigate Linked data at the cloud level by analyzing huge amounts of data obtained from curating sources such as Data Hub (https://datahub.io/) and collected by specialized crawlers. The studies include but seldom highlight or directly address LBD. An examination (Villazón-Terrazas et al., 2012) of the Linked data publishing process (including the initial work on the publishing of Linked data conducted by the BNE) shows that there is no one-size-fits-all formula. Each domain represents a set of data types, data formats, data models, licensing contexts, and languages, forming individual problem areas. Thus, although it is crucial to analyze Linked data as a whole, it can also be useful to isolate and study parts of the cloud belonging to publishers that share contextual perspectives. The study reported herein examines and compares the quality of a particular type of Linked data, *bibliographic descriptions*, originating from the relatively uniform library field.

Linked Bibliographic Data

W3C's Library Linked Data Incubator Group (2011) published a final report that, in addition to listing pro-Linked data arguments, states that "relatively few bibliographic datasets have been made available as Linked data," and "the level of maturity or stability of available resources varies greatly." Since then, following the National Library of Sweden's publication of its catalogue as Linked data in 2009 (Malmsten, 2009), prominent institutions, such as OCLC (Fons, Penka, & Wallis, 2012), the Library of Congress (http://id.loc.gov/), and several national libraries, have made LBD openly available on the Web. Alongside these publishing endeavors, much work has been put into Linked data-oriented metadata models, such as BIBFRAME (Library of Congress, 2012) and FRBRoo (LeBoeuf, 2012).

In the Library of Congress's presentation of the goals for BIBFRAME in 2012, meeting the need to make "interconnectedness commonplace" is a clearly expressed ambition (Library of Congress, 2012). The emphasis on outreach and interoperability is also evident in European countries' national libraries' expressed motivation for publishing LBD:

- BNB: "One of our aims was to break away from library-specific formats and use more cross-domain XML-based standards in order to reach audiences beyond the library world" (Deliot, 2014, p. 1).
- BNF: "The BnF sees Semantic Web technologies as an opportunity to weave its data into the Web and to bring structure and reliability to existing information" (Simon, Wenz, Michel, & Di Mascio, 2013, p. 1).
- DNB: "The German National Library is building a Linked data service that in the long run will permit the semantic web community to use the entire stock of national bibliographic data, including all authority data. It is endeavouring to make a contribution to the global information infrastructure." (Hentschke, 2017)
- BNE: "The use of Linked Open Data to build a huge set of data, described according to best practices of LOD publication, transforming library data into models, structures and vocabularies appropriate for the Semantic Web environment, making it more interoperable, reusable and more visible to the Web, and effectively connecting and exchanging our data with other sources" (Santos, Manchado, & Vila-Suero, 2015, p. 2).

Some of these quotations also address the need to renew formats, data structures, and other organizational legacy features. The BNE documentation further highlights that it has used the opportunity to implement entity types from the FRBR model (IFLA Study Group on the Functional Requirements for Bibliographic Records, 1998; Santos et al., 2015), and the BNF reports that it has had to "transform data from non-interoperable databases into structured and exchangeable data" (Simon et al., 2013, p. 3).

Following from the reported work on organizational features, an interesting characteristic of the corpus sets selected for this study is that they all represent different, local, bottom-up approaches to modernizing bibliographic data and organization. The lingua franca aspects of Linked data principles may be interpreted as a (liberal) continuation of widely adopted principles of global standardization in the library community, often referred to as universal bibliographic control. However, when the national libraries transformed their data and published the corpus examined here as Linked data, they applied such principles more or less in parallel, and in line with the interoperability methodology of application profiles, mixing metadata elements from several standards (Heery &

Patel, 2000). Lately, there has been discussion on whether the plethora of new approaches and their resulting models really help lift bibliographic data out of their legacy silos or if these parallel publishing activities merely create new Linked data silos filled with heterogenic data (Suominen, 2017).

Quality Dimensions and the Study Scope

Data quality is commonly defined as fitness for use (van Hooland, 2009; Wang & Strong, 1996), and this notion of quality has been related to different dimensions in various fields. In the library domain, (meta)data quality has been related to completeness, accuracy, provenance, logical consistency and coherence, timeliness, accessibility, and conformance to expectations (Bruce & Hillmann, 2004).

The Linked data community has similar quality dimensions. In an analysis of the adoption of best practice principles, Schmachtenberg, Bizer, and Paulheim (2014) group quality issues into three categories: linking, vocabulary usage, and the provision of (administrative) metadata. Hogan et al. (2012) analyze the implementation of 14 best practice principles found in an expansion of Heath and Bizer (2011), categorized as issues related to naming (e.g., avoiding blank nodes[1] and using HTTP URIs), linking (e.g., using external URIs and providing `owl:sameAs` links), describing (e.g., re-using existing terms), and dereferencing (e.g., dereferencing back and forward links). Radulovic, Mihindukulasooriya, García-Castro, and Gómez-Pérez (2017) categorize aspects of Linked data quality into two groups: those related to inherent data and those related to the technical infrastructure. Inherent quality is further divided into the aspects of domain data, metadata, RDF model, interlinks, and vocabulary. Infrastructure aspects involve Linked data server, SPARQL, Linked data Fragments, and file servers. Zaveri et al. (2015) conduct a comprehensive literature review of studies published between 2002 and 2012 focusing on Linked data quality. They find 23 quality dimensions and group them as accessibility, intrinsic, trust, dataset dynamicity, contextual, and representational dimensions (Zaveri et al., 2015). Each dimension is connected to one or more procedures for measuring it (metrics). Interlinking is listed as a dimension in the accessibility group and is connected to metrics such as out- and indegree. Vocabulary usage is part of several dimensions in the representational group, with metrics such as re-use of existing vocabulary terms and dereferenced representation.

The scope and the research questions of this study are determined by the motivations expressed by the institutions publishing LBD, as outlined in the preceding section, to improve interoperability and to facilitate (re-)organization. Accordingly, the study primarily considers *interlinking* and *vocabulary usage*, which can be directly related to those motivations. The study does not take into consideration aspects of, for example, administrative metadata provision or the technical infrastructure.

Previous Studies of Linked (Bibliographic) Data Quality

Previous studies highlight several quality issues. The following review presents the findings from a selection of studies which include LBD.

[1] Blank nodes are nodes in an RDF graph which indicates the existence of a thing without using an URI or literal to identify that thing. Blank nodes are typically used to describe reifications or lists. Linked data principles recommend avoiding use of blank nodes due to their limited alignment to Linked data tools such as SPARQL (Hogan et al., 2012).

Hogan et al. (2012) analyze and statistically rank 188 pay level-domains (PLD)[2] harvested through a web crawl for conformance to 14 best practice principles. The study includes the Library of Congress loc.gov domain, which is the only domain to directly represent elements of LBD in the study sample (Hogan et al., 2012). The loc.gov domain has excellent scores for its RDF structure (avoids blank nodes) and acceptable scores for its use of stable HTTP URIs but poor scores for its re-use and mixing of well-known vocabularies (Hogan et al., 2012). It is overall ranked quite low, at number 182 (of 188).

Schmachtenberg et al. (2014) analyze a corpus of Linked datasets harvested through a web crawl and find that 56% of the analyzed datasets provide links to at least one external set, while the remaining 44% are mere target sets. Only 15.8% of the corpus sets link to more than 6 external sets (Schmachtenberg et al., 2014). Almost all of the sets (99.9%) use elements from non-proprietary vocabulary, while 23.2% of the sets also use vocabulary elements not used by others (from a proprietary vocabulary), and 72.8% of the proprietary vocabularies are not dereferencable (enabling "applications to retrieve the definition of vocabulary terms"). Schmachtenberg et al. (2014) further divide the corpus sets into 8 topical domains. Most interesting in the context of the present study is what is called the publication domain, which includes LBD sets. Some sets in this domain are among the overall top 10 with the highest in- and outdegree of interlinks, but none is a LBD sets.

Kontokostas et al. (2014) propose a test-driven approach to the evaluation of Linked data quality, using SPARQL queries in a variety of test patterns. The queries are used to test accuracy issues at the literal level (e.g., whether the birth date of a person comes before the death date) and that datasets do not violate restrictions on properties (e.g., regarding their domain and range) (Kontokostas et al., 2014). As proof of concept, Kontokostas et al. (2014) test five datasets, including LBD from the BNE and Library of Congress. The test shows that most errors in the datasets, including the LBD sets, come from violations on domain and range restrictions.

Papadakis, Kyprianos, and Stefanidakis (2015) investigate URIs used in LBD, including in the sets from the four national libraries studied here, and focus on the preconditions for designing URIs based on (UNI)MARC fields in legacy records. In addition, they provide an overview of the existing links between URIs across datasets from several LBD providers (Papadakis et al., 2014). Hallo, Luján-Mora, Maté, and Trujillo (2015) also investigate the quality of datasets that are part of the corpus studied in this paper. They identify vocabularies used and review the reported benefits and challenges of LBD (Hallo et al., 2015). Neither of these two studies includes detailed statistical analysis of the interlinking practice or vocabulary usage.

Data and Methods

Data Selection

The datasets assessed in the study must contain directly available, comparable, and non-experimental bibliographic data published by a library institution. Based on these criteria, the following datasets were selected.

[2] A PLD is a sub-domain of the public, top-level domain which users usually pay to access.

BNB

The British National Bibliography was first published as Linked data in 2011. It includes both books and serial publications made available in separate datasets. In this evaluation, only the book set is considered.

BNE

The Biblioteca Nacional de España has published LBD since 2011. This dataset covers "practically all the library's materials, including ancient and modern books, manuscripts, musical scores and recordings, video recordings, photographs, drawings and maps." (Biblioteca Nacional de España, 2014)

BNF

The Bibliothèque nationale de France has also published Linked data since 2011, including bibliographic data from the main catalogue (BnF Catalogue Général). The data are available through a searchable interface and RDF dumps for download. Different dumps separate the data into a variety of types. This study is based on the full RDF dump.

DNB

The Deutsche National Bibliothek has published Linked data since 2010 and included bibliographic data since 2012. For this evaluation, two datasets are downloaded and combined: the Deutsche Nationalbibliografie (DNBTitel) and the Integrated Authority File (GND).

Other datasets may also fit the selection criteria described here, but an analysis of the chosen datasets provided by significant agents in the library field is considered to give an adequate picture of the LBD sets available on the Web in 2016 for a variety of potential data consumers.

The national libraries offer their data through different sub sets. Most of these are complimentary and interlinked through common URIs. For example, the DNBTitel dataset mainly contains detailed information about documents, including references to URIs from the GND set where authors and other persons related to the documents are described in detail. To avoid loss of significant bibliographic information, most subsets are included in the corpus sets. The exception is the relatively small set of BNB Serials, which was considered to be out of the scope in this research.

The selected datasets were downloaded as dumps of RDF triples and ingested into a local Virtuoso triple store (https://virtuoso.openlinksw.com/). Table 1 shows the subset names, download and last modified dates, and license information of the four corpus sets analyzed. The sets were downloaded from late February to early April 2016 and were the most recently updated sets commonly available for download at that time.

	Downloaded	***Modified***	***License***	***Set names***
BNB	March 1, 2016	January 6, 2016	CC0 1.0	BNB LOD Books
BNE	March, 3, 2016	March 3, 2016	CC0 1.0	Registros de autoridad + Registros bibliográficos + Encabezamientos de Materias de la Biblioteca Nacional en SKOS
BNF	April 6, 2016	November 24–December 5, 2015	Open License 1.0	All documents (complete description)
DNB	February 29, 2016	October 23, 2015	CC0 1.0	DNBTitel + GND

Table 1. Download date, last modified date, license information, and set names of the four corpus sets.

RDF Data

The W3C recommendation (Cyganiak, Wood, & Lanthaler, 2014) defines the core structure of RDF as a graph-based data model in which sets of triples, each consisting of a subject, a predicate, and an object, form a RDF graph. The subject of a triple can be either a URI or a blank node. The predicate must be a URI, while the object can be an URI, a blank node, or a literal.

The URIs in the RDF graph represent *entities* (or resources) that can belong to various classes (i.e., a person, book, or publication event) and have various relationships (a person *is the author* of a book). RDF itself does not provide the terms to describe specific classes or relationships, so each graph must apply terms from locally or externally minted vocabularies. The following triple from the BNB set uses the property `dct:creator` from the DCMI Metadata Terms[3] vocabulary (expressed with the namespace `dct`[4]) to apply a relationship stating that a URI representing a certain book is created by a URI representing Bob Dylan:

```
http://bnb.data.bl.uk/id/resource/013220704 dct:creator
http://bnb.data.bl.uk/id/person/DylanBob1941-
```

The following triple states that Dylan (his URI representation) is a person using the class `foaf:Person` from the FOAF vocabulary:[5]

```
http://bnb.data.bl.uk/id/person/DylanBob1941- rdf:type foaf:Person
```

RDF graphs contain two types of triples, *literal triples* and *RDF links* (Heath & Bizer, 2011). A literal triple describes the properties of a given entity, with a literal string, number, or date as the object. An RDF link connects two URIs. An internal RDF links connects the URIs within a RDF graph (as illustrated in the triple with the URIs representing Dylan and his book from the BNB set). An external RDF link connects a local URI with a URI from an external dataset. An example is a triple from the BNE stating that the URI representing Dylan is the same as the URI representing Dylan in the VIAF dataset (https://viaf.org/):

```
http://datos.bne.es/resource/XX821701 owl:sameAs http://viaf.org/viaf/111894442
```

Further RDF definitions are used in line with Hogan et al. (2012). The RDF constants C are defined by the union of all the distinct URIs (U), blank nodes (B), and literals (L) of an RDF graph, formally denoted as $C \coloneqq U \cup B \cup L$. Data-level positions in triples are defined as subjects and objects, with the exception of the objects of `rdf:type` triples, which are schema-level class terms. Table 2 shows the numbers of triples, unique entities, and RDF constants on the data level in the four corpus sets. Regardless of internal differences, these sets are neither the smallest nor the largest in a Linked data context where prominent sets like DBpedia (https://datahub.io/dataset/dbpedia) and GeoNames (https://datahub.io/dataset/geonames-semantic-web) contain 1.2 billion and 94 million triples, respectively.

Set	*Triples*	*Entities*	*Data-level constants*
BNB	104,139,477	10,126,344	52,671,707
BNE	71,199,698	5,763,188	56,681,387
BNF	304,587,809	30,671,400	192,224,487

[3] http://dublincore.org/documents/dcmi-terms/
[4] All name spaces used throughout the paper are listed in Appendix I.
[5] http://xmlns.com/foaf/spec/

DNB	329,261,459	32,673,901	250,613,437
Average	202,297,111	19,808,708	138,047,754.5

Table 2. Number of triples, entities, and data-level constants.

All the corpus sets are described as bibliographic data by their publishers and, therefore, should be comparable due to their contents. However, it should be assumed that the datasets are tailored for particular user tasks, transformed into RDF from different types of legacy data, or differ in other aspects that make it inappropriate to compare them. To demonstrate the validity of the corpus sets (that they are comparable representatives of bibliographic data), samples of triples describing the authorship of Nobel laureates in literature from 2006 to 2016 are extracted from each set based on strict generic extraction procedures and selection criteria for the data. These samples are compared to the characteristics of the overall sets. Details on the extraction method and the results are presented in the analysis section. Data from this analysis are also used in the following case study.

Statistics and Limitations

The statistics on vocabulary usage and interlinking are retrieved by SPARQLing the local triple store containing the downloaded corpus data. The SPARQL queries used are based on the COUNT expression with the necessary filter conditions[6]. To design efficient queries, previous research and projects concerning Linked data statistics and providing concrete examples are used as a starting point (see, e.g., Auer, Demter, Martin, & Lehmann, 2012; Cyganiak, 2105)

Regarding vocabulary usage, all the terms applied in the corpus sets are examined without any limitations. Some limitations are applied in the examination of interlinking. Previous studies use the term *outdegree* to denote the number of external datasets to which a source dataset links, independent of the predicate used in those links (Schmachtenberg et al., 2014). Two datasets are considered to be linked if at least one RDF link exists between resources belonging to those sets. This study follows this general notion of interlinking but with three limitations. First, internal linking is not examined. This limitation applies to links in a corpus set where the subject and object of the triple share the same PLD and to triples in which the object URI is interpreted to be part of the institutional context of the particular set (e.g., links from the DNB to the ZDB database of serial titles hosted and maintained by the DNB).

Second, the analysis considers only external datasets providing RDF data. In practice, this means that links to DBpedia but not Wikipedia are counted. This is in line with previous studies (Hogan et al., 2012) and Linked data principles. Third, for each particular predicate used in external RDF links (e.g., `owl:sameAs` or `rdfs:seeAlso`), the analysis is limited to RDF triples counting more than 300 distinct subject URIs pointing to a particular external dataset. In other words, for an external corpus set to be considered in the study, it needs to have links from more than 300 entities to it. The corpus sets contain millions of external RDF links to a great variety of domains, and there is a long tail of domains targeted only once or a few times (e.g., companies' homepages). A corpus with fewer than 300 links to a particular dataset, therefore, is considered to be outside the scope of the analysis for two main reasons. A minimum of 300 triples containing URIs from an external set ensures that the external set has a minimum level of substantiality (to be part of the widely referred to Linked data

[6] https://www.w3.org/TR/sparql11-query/

cloud requires at least 1,000 triples[7]). A reduced amount of external datasets ensures that the analysis is becoming more manageable.

To exemplify and provide a better understanding of the organizational principles of each set in the corpus, the statistics are supplemented with a brief qualitative case study of comparable samples describing the authorship of the most recent Nobel prize winner in literature, Bob Dylan.

Analysis

General RDF Model and the Content of the Corpus Sets

Figure 1 shows that the distribution of literals, URIs, and blank nodes among the RDF data-level constants differs across the corpus sets. The BNE and the DNB have larger shares of literals, indicating a structure with more entities labeled directly. An example is the representation of publishing events: the BNB set provides URIs for each unique event, each event year, and each publisher involved in those events, whereas the other sets relate publishing information directly to the manifestations as literal values. Figure 1 also shows that the BNE and the DNB violate a Linked data best practice by using significant amounts of blank nodes (Mallea, Arenas, Hogan, & Polleres, 2011).

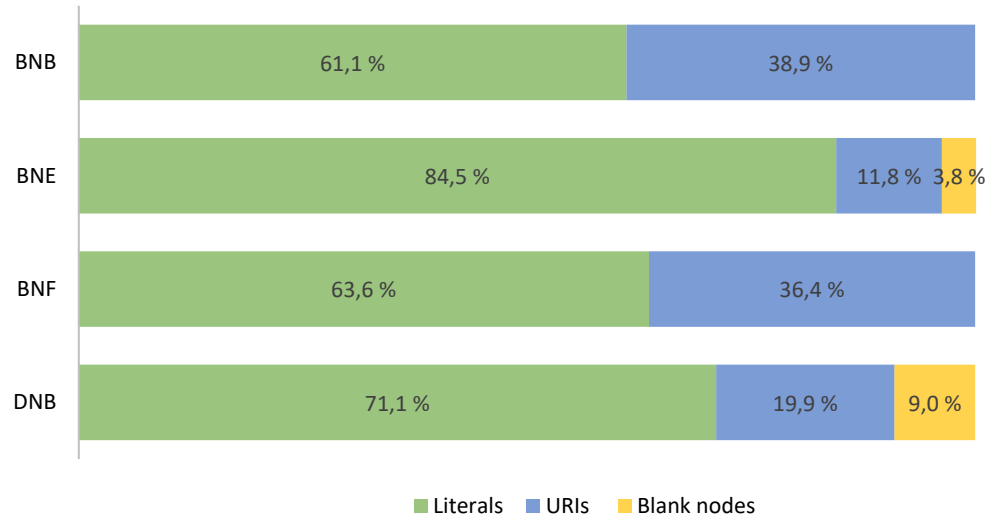

Figure 1. Distribution of literals, URIs, and blank nodes among the RDF data-level constants in the four sets.

An analysis of the content types in the four sets based on the class memberships of the entities related to persons, manifestations, and subjects (Table 3) shows that the sets have similar compositions. Approximately 30% of the entities in each set belong to classes used to represent manifestations. With the exception of the DNB, all the sets contain a large share of subject data, and with the exception of the BNF, a large quantity of entities represent persons. The most notable difference among the sets, not shown in Table 3, is the distribution of the FRBR entities *work* and *expression*, which are only part of the BNE and the BNF. Along with persons and subjects, works, expressions, and manifestations (W/E/M entities in FRBR lingo) account for more than 50% of class memberships in all the sets. In fact, entities related to all kinds of responsibility for the documents

[7] http://lod-cloud.net/

described by one or more W/E/M entities, such as the publisher and the year and place of publication, constitute almost 100% of all the entities in all the sets.

	Persons		*Manifestations*		*Concepts*	
	Class	***% of distinct entities with membership***	***Class***	***% of distinct entities with membership***	***Class***	***% of distinct entities with membership***
BNB	foaf:Person	12.5%	bibo:Book	29.2%	skos:Concept	18.8%
BNE	bneo:C1005	21.4%	bneo:C1003	33.8%	skos:Concept	8.6%
BNF	foaf:Person	5.2%	rdafrbr:Manifestation	27.4%	skos:Concept	9.0%
DNB	gndo:Person*	26.2%	bibo:Documents	30.7%	gndo:SubjectHeading*	0.6%

*Table 3. Distribution of entities as member of classes representing person, manifestation, and concept data. *In the DNB set, these numbers include members of a variety of classes which are subclasses of gndo:Person and gndo:SubjectHeading.*

As described in the methods section, samples from each set describing the authorship of Nobel laureates are extracted to demonstrate the validity of the corpora. The samples include URIs and their literal descriptions retrieved with SPARQL CONSTRUCT queries. The extraction procedure took the authors' URIs as the starting point and retrieved information about the documents (and their different W/E/M representations) for which the authors were responsible, contributed to, or were the subject of, along with information about other agents with responsibility for those documents. A common starting point across the sets is ensured by retrieving the VIAF identifiers for the relevant Nobel laureates from Wikidata. All the sets turn out to contain these identifiers. The models in the datasets differ, so specific queries mirroring these models could be developed to retrieve the information mentioned. Instead, to treat the datasets as neutrally as possible, all the queries are based on a generic RDF graph taking its starting point in the neighborhood of the nodes surrounding the author URIs.

The generic structures need some minor adaptations to the models in the BNB and BNF datasets. To extract the desired information, three nodes, in addition to the generic RDF graph, are included for the BNF set, while the BNB needs one additional node. To avoid overloading the information represented by a particular node (e.g., if common topical term were included in the sample, the procedure would need to avoid including every other document related to that term in the overall set), restrictions on properties are needed in one case each for the BNB and BNF sets.

The ratios between the triples, entities, and data-level constants in the sample data turn out to match the ratios in their respective overall datasets, as does the composition of RDF components. This indicates that the full datasets do not contain significant amounts of data not directly related to the bibliographic entities which could skew the comparative perspectives of the following analysis in this study. Moreover, while the modelling practices differ, all the sets clearly share a bibliographic nature centered on published documents, their topical contents, and the agents responsible for them.

Vocabulary Usage

Previous research on Linked data quality in vocabulary usage primarily investigates whether datasets *re-use* existing vocabularies and vocabulary terms. A consistent representation based on well-known vocabulary terms is considered to be a Linked data best practice that supports interoperability and increases usability for third-party consumers (Hogan et al., 2012). Studies also look at other aspects of quality, such as the *dereferencability* of applied terms. Dereferencability implies that in a Linked

data best practice to enable applications to retrieve and understand terms, the URIs identifying them should provide meaningful descriptions in response to HTTP requests (Schmachtenberg et al., 2014). This study is aimed at identifying the general character of the chosen bibliographical models and how they are realized by the use of vocabularies, as well as examining re-use, dereferencability, and other aspects of quality.

Vocabulary Models

By listing four different class terms used for manifestations, Table 3 indicates that the four publishers chose quite different vocabulary strategies in both their general modeling approach and re-use. Table 4, which provides an overview of the top 10 most used terms in each set, shows that all the W/E/M entities are described with different, exclusive terms.

	Property	*% of triples*	*Class*	*% of rdf:type triples*	*Property*	*% of triples*	*Class*	*% of rdf:type triples*
No.	**BNB**				**BNE**			
1	rdfs:label	14.6%	dcterms:BibliographicResource	17.4%	rdf:type	9.0%	bneo:C1003	33.5%
2	rdf:type	12.5%	blt:PublicationEvent	16.2%	rdfs:label	8.5%	bneo:C1001	24.9%
3	owl:sameAs	8.8%	bibo:Book	16.1%	bneo:id	8.1%	bneo:C1005	21.2%
4	event:place	3.7%	skos:Concept	10.4%	bneo:P1011	4.6%	skos:Concept	8.6%
5	blt:bnb	3.7%	foaf:Agent	8.7%	rdf:first	3.3%	bneo:C1002	5.5%
6	dcterms:title	3.7%	dcterms:Agent	8.7%	rdf:rest	3.3%	bneo:C1006	5.3%
7	dcterms:language	3.7%	foaf:Person	6.9%	bneo:P3002	3.0%	madsrdf:Topic	1.0%
8	event:agent	3.7%	blt:TopicLCSH	6.7%	bneo:P3064	3.0%	skos:ConceptScheme	0.0%
9	blt:publication	3.7%	blt:TopicDDC	2.8%	bneo:P3004	3.0%		
10	isbd:P1053	3.7%	bio:Birth	1.7%	bneo:P3003	3.0%		
No.	**BNF**				**DNB**			
1	rdf:type	11.1%	foaf:Document	27.4%	rdf:type	10.6%	bibo:Document	27.6%
2	owl:sameAs	7.6%	rdafrbr:Manifestation	27.4%	owl:sameAs	6.7%	rdf:Seq	25.2%
3	dcterms:created	4.1%	rdafrbr:Expression	27.4%	gndo:surname	4.2%	gndo:UndifferentiatedPerson	13.2%
4	rdfs:seeAlso	4.1%	skos:Concept	9.1%	gndo:forname	4.2%	bibo:Collection	10.4%
5	bnfo:FRBNF	4.0%	foaf:Person	5.2%	dcterms:medium	3.9%	gndo:DifferentiatedPerson	10.4%
6	dcterms:modified	4.0%	rdafrbr:Work	1.7%	dcterms:issued	3.9%	gndo:CorporateBody	3.1%
7	bnfo:firstYear	3.3%	foaf:Organization	1.2%	rdf:_1	3.7%	bibo:Issue	2.6%
8	dcterms:title	3.2%	geo:SpatialThing	0.4%	dce:title	3.7%	gndo:ConferenceOrEvent	1.6%
9	rdafrbr:expressionManifested	3.0%	dcmit:Event	0.2%	gndo:gndIdentifier	3.6%	bibo:Periodical	1.3%
10	foaf:primaryTopic	3.0%	bnfo:expositionVirtuelle	0.0%	dce:identifier	3.4%	bibo:Article	0.8%

Table 4. Top 10 property and class terms by the number of triples and class memberships for each corpus set.

The BNB and DNB use the same vocabulary to represent manifestations of books, albeit with different levels of abstraction (the `bibo:Book` used by the BNB is a sub-class of the `bibo:Document` used by DNB). The BNE and the BNF have both works and expressions in their sets, but the BNE uses a local vocabulary (built on terms from existing RDA vocabularies but hosted and presented as a local ontology with, for example, `bneo:C1001` as the class for works), and the BNF uses the now deprecated *FRBR Entities for RDA* vocabulary (prefix `rdafrbr:`). The sets are a bit more consistent in their representation of persons and concepts. The BNB and the BNF both use the FOAF vocabulary for persons, and the BNB, the BNE, and the BNF use `skos:Concept` for topical entities. Nevertheless, the remaining vocabulary terms in the sets reflect idiosyncratic vocabulary practices. The leftmost column in Table 5 shows the total numbers of terms used. Among the 1,141 unique property and class terms used by the four publishers, only three are shared by all the sets (`owl:sameAs`, `rdf:type`, `dct:language`). Thirteen terms are shared by three sets,

and 34 by two sets. The BNB and the BNF share 27 terms, while the DNB and the BNE only share three.

	No. of		*No. of unique vocabularies used for*		
	class terms used in set	*property terms used in set*	*class terms*	*property terms*	*all terms*
BNB	25	47	9	13	16
BNE	8	138	3	7	7
BNF	13	671	8	22	24
DNB	58	248	5	13	15
All	98	1043	19	32	38

Table 5. Number of vocabularies and vocabulary terms used in the sets.

The corpus sets can also be distinguished by other characteristics. The BNF uses 24 different vocabularies to describe its bibliographic data, but the BNE only uses seven. Each entity in the BNB set, on average, belongs to 1.8 classes (e.g., `bibo:Book` AND `dct:BibliographicResource`), whereas the entities in the three other sets very seldom belong to more than one (BNE average: 1.01, BNF average: 1.001). This implies, for example, that `bibo:Book` is used for 29.2% of the entities in the BNB set that have a class membership (Table 3) but represents only 16.1% of all the BNB `rdf:type` membership links (Table 4). In the other three sets, the figures from the two tables (3 and 4) are close to equal. Table 5 shows that the BNF uses 671 different property terms, whereas the BNB uses 47, primarily because the BNF set applies a much more detailed structure for representing the roles between responsible agents and documents. The BNF set also supports the interoperability of this detailed system by using existing properties with overlapping semantics in parallel. For example, the set has one triple with a property term from its local vocabulary and a parallel triple with a matching relator code from the MARC21 relator code vocabulary[8] (for examples, see Appendix IV).

Vocabulary Re-use

Schmachtenberg et al. (2014) consider a vocabulary to be *proprietary* "if it is used only by a single dataset". Although this might be true of some of the vocabularies used by one of the four corpus sets examined in this study, these terms very well could be applied by other datasets outside this context. This study, therefore, uses the more moderate term *local vocabulary*. A local vocabulary is further defined by an institutional connection to the publisher of the particular dataset in which it is used. For instance, in the following triple from the BNE set, the entity and the class term share the PLD:

`http://datos.bne.es/resource/XX821701 rdf:type http://datos.bne.es/def/C1005`

Thus, these vocabularies are not necessarily proprietary but neither are they examples of re-use. All the sets use one local vocabulary, except for the BNF which uses two. Table 6 shows the percentage of local vocabulary terms used and the percentage of the triples using them.

[8] https://www.loc.gov/marc/relators/relacode.html

	% of local				
	class terms	*property terms*	*vocabulary terms in total*	*class terms in rdf:type triples*	*property terms of data level triples*
BNB	40.0%	12.8%	22.2%	26.6%	10.6%
BNE	62.5%	84.7%	83.6%	90.4%	76.0%
BNF	8.0%	71.7%	70.5%	0.0%	15.3%
DNB	79.3%	74.6%	75.5%	30.8%	36.0%
All	59.6%	74.5%	70.4%	23.4%	29.0%

Table 6. Percentage of local vocabularies and vocabulary terms and the percentage of the triples in the sets using these terms.

The BNE in particular but also the DNB use local terms to a much greater extent than the BNB and the BNF. The BNF uses many local terms but applies them in a relatively small percentage of the `rdf:type` triples and data-level triples. The BNB uses more local class terms than the BNF but fewer local property terms. On the class level, the BNE uses almost exclusively local terms, with the distinct exception of `skos:Concept`, which represents more than 8.6% of the BNE's classes (Table 4). The DNB uses fewer local terms than the BNE, but still more than 30% of both its class and property terms are locally developed.

Data providers apply local terms for several reasons, for example, to facilitate logical consistency in a given dataset or to express semantic relationships not covered by existing vocabularies. In the case of the BNE, its predominant use of local terms is probably due to intrinsic consistency issues. The three other sets, however, all primarily use local terms to express rather specific, granular relationships. For example, the BNB uses local terms to represent a complex modeling of publishing data (e.g., `blt:PublicationEvent` and `blt:publication`), while the BNF uses local terms to express a large number of detailed role statements (e.g., `bnfrel:r550` represents a person or organization responsible for an introduction or preface). The DNB uses local terms for several purposes but primarily to express quite specific semantics. The corpus sets do not use local terms in a clear or systematic way to express complex semantics within overlapping bibliographic areas. It, therefore, is hard to identify a common semantic area in the corpus where the use of local terms indicates a lack of existing generic bibliographic vocabulary terms.

Since Linked data principles recommend using existing vocabulary terms when publishing data on the Web, it would be interesting to examine whether there exist matching vocabulary terms which could be used instead of the local terms in the corpus sets. That, however, is a substantial task which future studies should investigate.

Other Quality Aspects of Vocabulary Usage

Table 6 shows that, on average, less than 30% of property and class terms applied across the corpus sets is local, while more than 70% of the usage consists of re-use of external vocabulary terms. Many best practice guidelines for Linked data contain explicit criteria for selecting such external vocabularies (see, e.g. Hyland et al., 2014) and Janowicz, Hitzler, Adams, Kolas, and Vardeman (2014) propose a dedicated five-star rating system for Linked data vocabularies. In such guidelines, it is often stressed that the vocabularies should be well known or at least used by others. Other quality

criteria include meaningful documentation, long-term accessibility, dereferencability and language support. Figure 2 shows the scores for the 38 vocabularies used by the four sets on five heuristic measurements derived from a selection of best practice recommendations: dereferencability, adoption in the Linked data community, provision of human readable documentation, provision of vocabulary restriction, and links to other vocabularies. The vocabularies were tested in March 2017, a year after the datasets were downloaded. A sixth measurement thus could be long-term accessibility.

The first bar in Figure 2 shows that six, or 15.8%, of the vocabularies returned a 404 not found response to a HTTP GET request. Manual examination of the vocabulary URLs reveals that four of these six vocabularies are actually dereferencable but are applied in the sets with slightly different URI names. This could be due to name changes over time or misspellings of URIs. The number of positive responses nevertheless is satisfying, especially considering the long-term accessibility. The remaining measurements answer the question of whether the publishers choose vocabularies that possess certain qualities but not the question of whether the publishers address vocabulary terms correctly. The four vocabularies initially returning a 404 response but later manually identified are therefore included in the examinations.

Whether (other) dataset publishers adopt a vocabulary is an indication that it is well known. The numbers in this study are based on statistical data from LODStats (http://stats.lod2.eu/) and LOV (http://lov.okfn.org/dataset/lov), two services providing information about published Linked datasets. Both services provide a search interface for vocabularies and return the number of datasets identified as using a particular vocabulary. Each of the 38 vocabularies is tested using these two services. Both find that 13 vocabularies, nine of them overlapping, are not used by datasets other than those in the corpus. On average, 65.8% of the vocabularies are used at least by one other dataset. Furthermore, a manual investigation of the vocabularies shows that almost all include human readable descriptions in the form of comments and labels. More than 90% of the vocabularies have restrictions on domain and range (which is one of the axiomizations mentioned, for example, by Jonawicz et al., 2014), and according to the LOV service, almost 90% of the vocabularies contain alignments to external vocabularies. There are no significant differences between the datasets for any of these measurements.

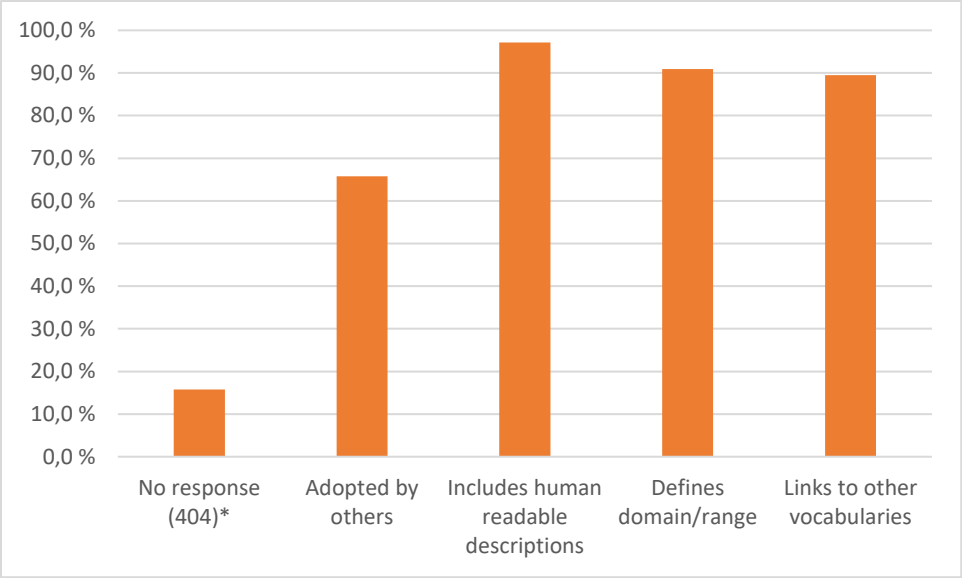

Figure 2. Five quality measurements showing the overall score for all 38 external vocabularies used in the corpus.

Interlinking

Re-using quality vocabularies ensures interoperability by increasing the use of common semantics. Another core interoperability practice of Linked data is interlinking, or the provision of direct relationships across published datasets. Interlinking is formally defined as an external RDF link in which the subject URI represents a local entity, and the object URI an entity from an external dataset. The external RDF links in the corpus sets are counted in line with the limitations listed in the methods section[9]. The analysis of linking practices is based on the main components of external RDF links: the the properties used and the external target datasets. These components correspond to metrics from earlier Linked data quality research. Counting external datasets allows comparing the outdegree of a particular dataset and looking at the properties permits evaluating representational aspects.

General Numbers

Table 7 shows that the BNB has most external RDF links relative to its number of triples, as well as the highest ratio of interlinked entities. Linked data guidelines tend to favor `owl:sameAs` links (external RDF links using the property `sameAs` from the OWL ontology) for their ability to facilitate browsing and consolidation of additional information related to URI aliases (Hogan et al., 2012). The DNB provides slightly more `owl:sameAs`-links than the other sets relative to both triples and entities.

Set	*External RDF links of all triples*	*owl:sameAs links of all triples*	*External RDF links per entity*	*owl:sameAs links per entity*
BNB	14.5%	1.1%	1.5	0.1
BNE	3.6%	0.8%	0.4	0.1
BNF	5.2%	1.4%	0.5	0.1
DNB	7.8%	2.5%	0.8	0.3
Avg.	7.7%	1.5%	0.8	0.2

[9] The limitations do not lead to the exclusion of significant amounts of RDF triples, with some notable exceptions. Nearly all the sets have links to Wikipedia, and the DNB provide nearly 150,000 links to filmportal.de. These two sites do not offer RDF data and, therefore, are not included in the analysis.

Table 7. External RDF links for all triples and per entity.

Outdegree

The metric outdegree is defined as the number of unique external datasets to which a given corpus set links. To count the outdegree precisely, previous studies count the links between unique PLDs. In this study, which has a manageable amount of data, PLDs and unique datasets sharing the same PLD are counted separately. Thus, <http://id.loc.gov/authorities/subjects/> and <http://id.loc.gov/vocabulary/countries/> are counted as one PLD, but as two datasets even though they belong to the same PLD. This approach allows comparing the numbers from this study with those of previous studies while also getting a more detailed picture of linking practices. In addition, the institutional context of the external datasets is analyzed, particularly their origin in the library domain, defined as being hosted by a library institution. Figure 2 shows the full network of links between the corpus sets and the external datasets. The thickness of lines indicates the number of RDF links between the datasets. Table 8 lists the outdegree of each set. Table 9 provides an overview of the ten datasets that are the targets of most RDF links, along with the distribution in each corpus set.

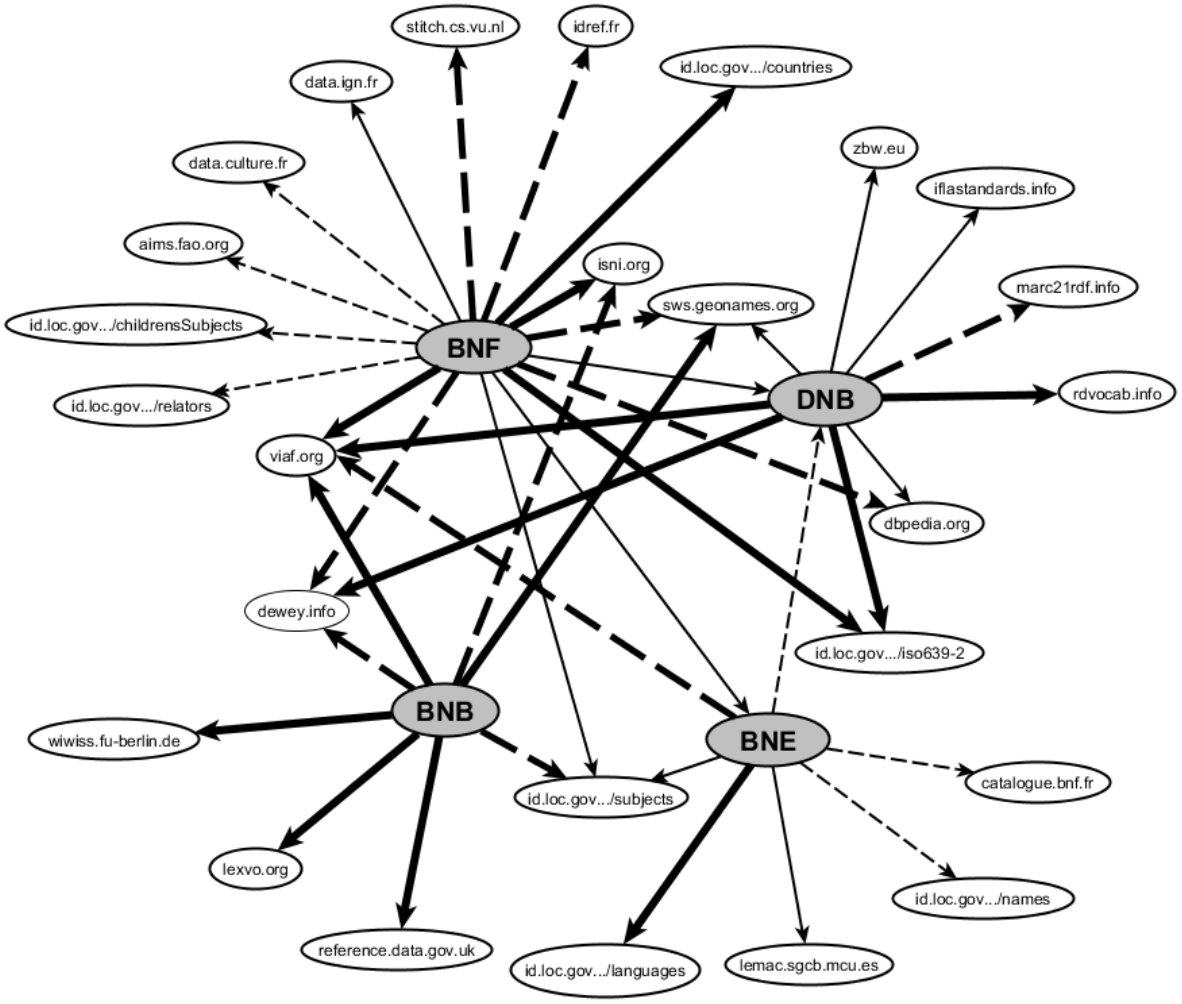

Figure 3. The four corpus sets and the external datasets targeted by their external RDF links. Thick lines: more than 1 million links; thick dotted lines: 100,000–1 million links, thin lines: 10,000–100,000 links; thin dotted lines: fewer than 10,000 links.

	BNB	BNE	BNF	DNB	Total	*Avg.*
No. of datasets	8	7	17	9	28	10.5
No. of PLDs	8	5	13	9	22	8.75
No. of PLDs not hosted by a library institution	4	1	7	4	11	4
No. of PLDs linked with predicate owl:sameAs	3	1	5	3	7	3
No. of non-library PLDs linked with predicate owl:sameAs	0	0	3	1	4	1

Table 8. Various aspects of outdegree in each set, average and total for the corpus.

In total, the corpus sets link to 28 unique datasets and 22 unique PLDs. Eleven PLDs originate from outside the library domain (e.g., dbpedia.org, sws.geonames.org, and isni.org). Seven PLDs are linked via the `owl:sameAs` property, and four of these are non-library datasets (the aforementioned three and idref.org). At the dataset level, 20 of the 28 datasets are linked to only one corpus set. Three datasets have links to two corpus sets, and four datasets to three corpus sets and only one external dataset (viaf.org) have links to all four sets. Of the 22 PLDs, 15 datasets are linked to one corpus set, three are linked to two corpus sets, and another two datasets from three corpus sets and two datasets (id.loc.gov and viaf.org) to all four sets. All the corpus sets provide `owl:sameAs` links, with an average outdegree of three.

The property used in most external RDF triples throughout the corpus sets is `dct:language`, applied by all sets to represent relationships with external language authorities. Each of the sets uses this property to link to one external data set (BNB: http://lexvo.org/id/iso639-3/), BNE: http://id.loc.gov/vocabulary/languages/, BNF and DNB: http://id.loc.gov/vocabulary/iso639-2/). Other popular property terms used for interlinking include `rdfs:seeAlso` and different terms from the SKOS vocabulary. The latter is mostly used to relate local topics to Dewey numbers (http://dewey.info/class/).

	% of RDF links, total and individual sets	Distinct objects of RDF links, total and individual sets
http://id.loc.gov/vocabulary/iso639-2/	**23.1%**	**486**
BNF	59.5%	446
DNB	15.7%	486
http://rdvocab.info/termList/RDACarrierType/	**20.3%**	**4**
DNB	46.3%	4
http://viaf.org/viaf/	**19.7%**	**10,341,459**
DNB	31.5%	8,141,903
BNE	22.0%	559,783
BNF	11.3%	1,807,538
BNB	7.2%	1,040,851
http://www4.wiwiss.fu-berlin.de/bookmashup/books/	**5.6%**	**3,262,475**
BNB	23.0%	3,262,475
http://sws.geonames.org/	**5.4%**	**148,845**
BNB	20.7%	156
BNF	0.7%	101,629
DNB	0.2%	47,104
http://lexvo.org/id/iso639-3/	**5.2%**	**272**
BNB	21.0%	272

http://reference.data.gov.uk/id/year/		**5.0%**	**224**
	BNB	20.6%	224
http://id.loc.gov/vocabulary/languages/		**3.3%**	**256**
	BNE	75.8%	256
http://isni.org/		**3.3%**	**1,651,998**
	BNF	7.5%	1,196,185
	BNB	5.0%	725,148
http://dewey.info/class/		**3.1%**	**215,059**
	BNB	1.4%	1733
	BNF	2.1%	63
	DNB	5.1%	214,005

Table 9. Ten external datasets that are the targets for the most RDF links for all four sets and for each individual corpus set. The rightmost column shows the number of distinct URIs targeted in total and in each set.

Table 9 shows the ten most popular external datasets as measured by RDF links. Among these, viaf.org has more than 11.5 million RDF links across the corpus set and accounts for a significant amount of the RDF links in each set. The links from viaf.org point to 10.3 million distinct objects, which suggest that the overlap in entities represented by viaf.org between the sets is not that high. This does not reflect any quality issue but, rather, indicates the national characteristics of the sets. Most sets only link to persons in VIAF, except for the BNE, which also provides VIAF links to works and expressions. Table 10 shows the overlap of VIAF entities between the sets, limited to person entities of owl:sameAs-links. Overall, 0.2% of the distinct VIAF entities from such links (22,621 persons) are represented in RDF links from all sets.

Set combinations	**Overlap**
BNF–BNB	12.7%
BNF–DNB	6.5%
BNF–BNE	5.6%
DNB–BNB	4.3%
BNB–BNE	2.6%
DNB–BNE	1.1%
BNF–BNE–BNB–DNB	0.2%

Table 10. Overlapping viaf.org entities limited to person entities and owl:sameAs links in different set combinations and between all sets.

Case Study

To get an even clearer idea of the quality of the corpus sets, especially their organizational features, limited samples are retrieved using the previously described methodology based on generic SPARQL queries. The samples describe Dylan, the most recent Nobel laureate in literature, and his single fiction novel *Tarantula*. Dylan has a limited authorship, making a case study feasible, and it is likely that his book is represented in the four datasets. In addition, Dylan does not come from any of the four countries that published the datasets studied. It, therefore, is less likely that the data describing him and his book are given special treatment as might be the case for bibliographic data describing famous writers sharing the same nationality as the dataset publishers. The samples thus are not necessarily representative of the collections but can provide insight into how the publishers represent author sets. The samples contain triples describing Dylan and all kinds of W/E/M entities representing his book and other persons who might have shared responsibility for some of those W/E/M entities. The samples are visualized as graphs with nodes and edges in Appendices II–V.

All the corpus sets contain representations of Dylan and the novel *Tarantula*. The BNB has three different manifestations in English. The BNE has two different works, but only one work has an expression (in Spanish), which has two manifestations. In this case, the BNF has no works but four expressions (in French) with four manifestations. The DNB has two German manifestations.

The visualizations clarify some of the differences between the sample sets related to the amounts of information provided about people and documents and related to structure and granularity. The following list provides some concrete examples:

- The BNF contains detailed information about the "country associated with the person", which none of the others provides.
- The BNB and the BNE chose to include inverse triples for many relationships (e.g., `blt:hascreated` from author to book AND `dct:creator` from book to author in the BNB set).
- All the sets, except the BNE, provide both the full name "Bob Dylan" and the name split into his given and family names.

The particular BNF sample lacks the expected work entities, so it does not illustrate the relationships between the W/E/M entities that are actually part of the BNF corpus set. Taking such relationships into consideration, nevertheless, it can be concluded that the BNF and the BNE organize W/E/M entities quite differently. Figure 4 provides a simplified overview of the main W/E/M entities with the responsible persons and their relationships in each set.

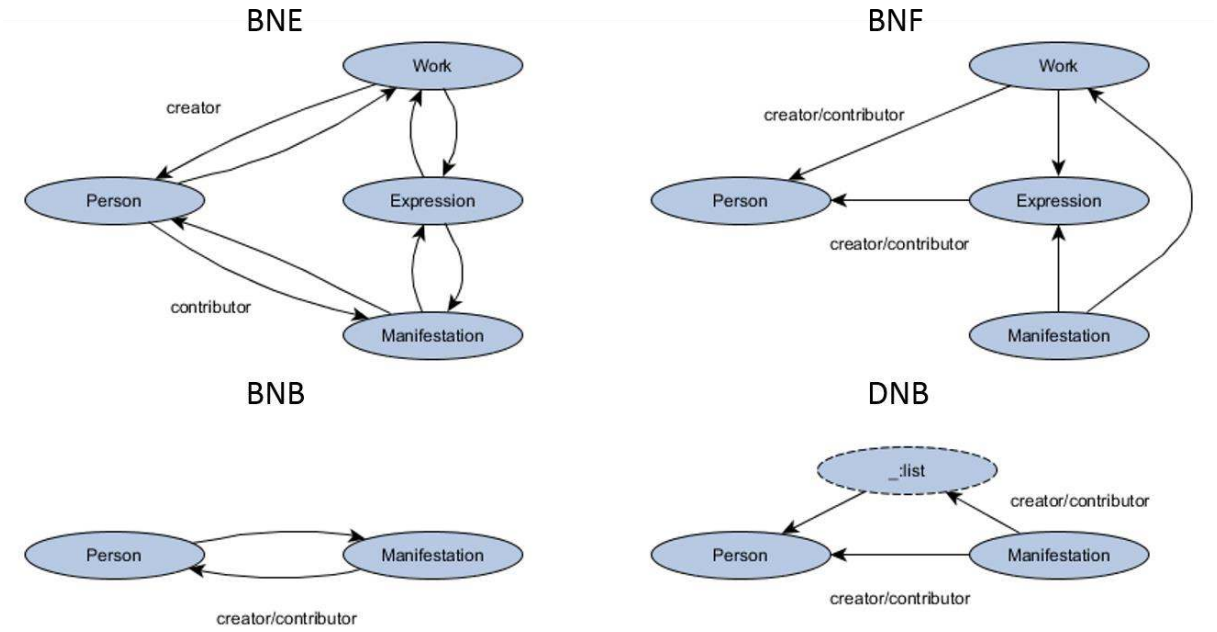

Figure 4. W/E/M models, including the relationships to the persons responsible in each set.

The BNE follows a standard structure from works via expressions to manifestations, as outlined in the original FRBR specification (IFLA Study Group on the Functional Requirements for Bibliographic Records, 1998). Creators (in almost all cases) are further related to works (`bneo:OP5001/OP1001` is the creator of/is created by). Other contributors such as translators (as in the sample) are related

to manifestations (bneo:OP3006/OP3005 has a contributor/contributes to)[10]. All the relationships in the BNE have inverse counterparts. The BNF set also contains the standard W/E/M entities, but they are related somewhat differently. Both works and manifestations are directed toward expressions. In addition, the models include possible relationships between manifestations and works. The BNF has very detailed representation of responsibility attributes, using 470 different properties to describe roles (e.g., bnfrel:r70 for authors and bnfrel:r680 for translators in the sample). These properties are defined in the local BNF vocabulary as sub-properties of dcterms:contributor and related to the corresponding properties in the MARC relator code vocabulary. Roles are mostly related to expressions, as in the sample data, but occasionally also to works when they exist. The BNB and the DNB are, as described, oriented toward manifestations but use slightly different models. The BNB includes inverse relationships between creators/contributors and manifestations. The DNB includes a system based on RDF Sequence containers[11] for listing multiple creators/contributors in an ordered way.

As indicated, the samples reveal some inconsistencies concerning W/E/M data in the BNE and BNF samples. As mentioned, the BNE sample includes a work that is related to Dylan but not to an expression (and from that neither to a manifestation). The BNF sample contains no works. This study does not speculate about the reasons. Nevertheless, the overall datasets indicate that both cases of inconsistencies are quite typical.

The BNE set has 1,451,069 distinct works, but only 13% of these works are related to expressions. The set contains 1,950,465 distinct manifestations, of which 14% are connected to expressions. Thus, the majority of works and manifestations in the BNE set are not connected to each other. Consequently, a large number of manifestations are only connected to their main creators via literals and not to possible URI representations of these persons, who are connected only to the works.

The BNF set contains 520,671 works, of which only 103,342 (20%) are connected to expressions. The number of distinct expressions equals the number of distinct manifestations, and there is exactly one link between each of these two entities. Further, 409,792 (5%) distinct manifestations are connected to 103,342 distinct works, the exact same amount of distinct works as for expressions. This indicates the same inconsistent W/E/M realization as in the BNE, with a majority of works and manifestations only loosely connected to the author sets. In addition, the overlapping manifestation and expression numbers suggest that these two entities form one semantic cluster in reality.

Other Quality Issues

Some issues of data quality at the instance level are beyond the defined scope of this work are detected as a spinoff product from the analysis presented (e.g., issues of URI duplication). Since duplication issues and other forms of messy data can influence interoperability, which is within the defined scope of the study, these findings are briefly reported in the following paragraphs. However, it must be emphasized that the findings do not result from a systematic examination that could reveal even more issues or show that the findings are only representative for a limited number of

[10] The BNE ontology contain a property for expressing a relationship between manifestations and creators (bneo:OP5002/OP3003), and the publishers mention this in a paper documenting the publishing process (Santos et al., 2015), but in the analyzed corpus, this connection is applied only four times.
[11] https://www.w3.org/TR/rdf-schema/#ch_seq

triples. The findings, by all means, do exist in the sets, but more dedicated examination is needed to provide a clear picture of the amounts of errors and the reasons behind them.

Duplicate URIs are found in all the sets, for example, through the interlinking analysis. The analysis shows that there are several cases in which the number of distinct (local) subjects is higher than the number of distinct corresponding (external) objects. This implies that in these cases, more than one local entity is linked to the same external entity. This is natural if, for example, the entities represent topics but not necessarily if they represent people or places. Take an example from the BNB set:

```
http://bnb.data.bl.uk/id/person/LouisXIVKingofFrance1638-1715 owl:sameAs
http://viaf.org/viaf/268675767
```

```
http://bnb.data.bl.uk/id/person/LouisKingofFrance1638-1715 owl:sameAs
http://viaf.org/viaf/268675767
```

The human readable URIs make it is easy to detect the duplication of the two BNB entities. Another example can be drawn from the DNB where two URIs represent the actor Thomas Eckert and are related to the same external source:

```
http://d-nb.info/gnd/1072088207 owl:sameAs
http://www.filmportal.de/person/A64B48535A1641C5819E3A7F53DCE143
```

```
http://d-nb.info/gnd/1073848744 owl:sameAs
http://www.filmportal.de/person/A64B48535A1641C5819E3A7F53DCE143
```

The examination of overlaps of VIAF entities between the sets reveals some issues particular to the BNE set. This downloaded set contains 558,920 distinct VIAF URIs. In a check of the type of the subjects in the owl:sameAs triples linking to those VIAF entities, approximately 50,000 distinct subjects are proved to have no specified class membership. It can be unproblematic for URIs to have no class membership; they can serve structural purposes or have other specific functions in a Linked dataset. An analysis of a sample of these URIs, however, shows that they represent both work and person entities that should have class membership according to the logic of the BNE Linked dataset.

The test of subsets based on Nobel laurates also reveals other issues related to VIAF links common to all the sets. The subsets are generated with SPARQL CONSTRUCT queries taking VIAF URIs retrieved from Wikidata as the starting point. As part of the procedure, all the URIs across the corpus set matching the VIAF URIs from Wikidata are retrieved for all 113 persons ever to win the Nobel Prize in literature. The retrieved lists of URIs show that all the sets, except the BNF, lack `owl:sameAs` links to one or more of these persons. In many cases, the sets simply do not cover the relevant authorship. In other cases, it is proved to be due one of two issues:

- The set has an entity representing the author but lacks a VIAF link.
- The set has an entity representing the author but links it to another VIAF authority, which indicates a duplication issue in VIAF.

The analysis also uncovers duplication issues among the local URI representations. For example, the DNB set contains double sets of URIs for the authors Patrick Modiano and Svetlana Alexievich.

Summary

It is fair to conclude that all the sets studied generally conform to the five-star Linked data requirements because they are available on the Web, offer structured RDF data (despite the use of blank nodes by two sets), and provide substantial numbers of links to external sources. They also re-

use dereferencable and widely-adopted vocabularies. In addition, they perform well compared with the findings from previous studies of Linked data conformance. Without the limitations restricting this analysis (a minimum of 300 local entities linked to each external dataset), Hogan et al. (2012) find that the PLDs in their corpus link to an average of 20.4 external PLDs. The corpus sets of this study have an average outdegree of 8.75 external PLDs; however, Schmachtenberg et al. (2014) find that only 15.8% of the sets analyzed in their study have an outdegree higher than 6, and almost 44% have no external RDF links at all. Based on these findings, it can be concluded that the corpus sets studied here have fewer external links then the top linkers worldwide but are still among the sets with the most links. When isolating the `owl:sameAs` links, Hogan et al. (2012) report that only 29.8% of their datasets have such links, with an average outdegree of 1.79. In this study, all the corpus sets contain `owl:sameAs` links, with an average outdegree of 3. Overall, the list of external datasets represents a varied collection of potential linkage candidates for bibliographic data. The BNF, in particular, provides links to an impressive number of datasets. However, when combining the expressed goal of reaching outside the library field with the best practice of using the `owl:sameAs` property, the linking practices of the corpus set are less successful. Only the BNF and the DNB contain `owl:sameAs` links targeting a few external datasets not hosted by library institutions. The analysis also reveals that a high proportion of external datasets, nearly 70%, are unique to each corpus, regardless of counting method. The few overlapping linking targets show diverse interlinking practices that hinder the potential usage of RDF links to common datasets to facilitate interoperability between the sets. Regarding vocabulary usage, the vocabularies applied by the corpus sets more or less resemble those found to be most used at the cloud level by Schmachtenberg et al. (2014).

The BNB and the DNB sets retain the manifestation-oriented structure from the legacy data of their origin. The BNE and the BNF take greater risk with their FRBRisations. Based on the examined versions of the datasets, however, this study shows that these FRBRsations have limited value because they lack a significant number of the expected links between the various W/E/M entities. This is not necessarily erroneous in a Linked data context based on an open-world assumption, but it can decrease the fitness of use. To utilize this data, for example, through a SPARQL end point, data consumers depend on trustful information about the data models to formulate adequate queries. In the case of the BNE and the BNF, one expects a specified FRBR model, but the published data do not support that model by instantiating it properly. The BNB and the DNB, which have data only about manifestations, avoid this problem, but they also inherit problems related to manifestation-oriented legacy data.

Concluding Remarks and Future Research

This study approaches the examined datasets from the perspective of potential data consumers. Thus, the reasons behind the revealed issues are outside the scope of the research and should be pursued in later investigations. Nevertheless, it should be noted that many of these problems likely are due to difficulties transforming legacy data based on manifestation-oriented models into new models based on novel conceptualizations. More research, therefore, should also be devoted to transformation issues, which are shared globally among libraries using the same legacy standards.

An answer to the second research question of data quality raised initially in this paper can be summarized as follows: as mentioned, the Linked data quality is generally impeccable for all the corpus sets. They meet the basic Linked data best practices and follow more specific

recommendations, such as the re-use of widely adopted vocabularies. At the same time, the study reveals quality issues. The datasets are deficient and potentially quite messy. Regarding the latter, further studies are needed to gather more knowledge about the amounts and reasons. From the present study, one can only conclude that some quantities of messy data exist in the sets.

Regarding the first research question of how the four national libraries, all prominent agents in the library community, choose to organize their data, the study primarily shows that they all do it rather differently. They apply different vocabularies for data representation, largely link to different external sources, and chose different bibliographic models for their structures. These independent solutions might serve individual purposes perfectly well but can hamper interoperability across sets and institutions. Interoperability between datasets of bibliographic data is important for global data utilization not only internally within the library field but also externally among data consumers who want to compile data from complementary sources. The examined national libraries are not alone in publish Linked data or utilizing new bibliographic models (Suominen, 2017). More research on the preferences and the use cases of potential data consumers is crucial to provide insights that could inform the way forward.

References

Auer, S., Demter, J., Martin, M., & Lehmann, J. (2012). LODStats — an extensible framework for high-performance dataset analytics. In *Proceedings of the 18th international conference on Knowledge Engineering and Knowledge Management* (pp. 353–362). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-33876-2_31

Berners-Lee, T. (2006). *Linked data: Design issues*. W3C. Retrieved from http://www.w3.org/DesignIssues/LinkedData.html

Biblioteca Nacional de España. (2014). datos.bne.es 2.0. Retrieved July 5, 2017, from http://www.bne.es/en/Inicio/Perfiles/Bibliotecarios/DatosEnlazados/datos2-0/

Bruce, T. R., & Hillmann, D. I. (2004). The continuum of metadata quality: Defining, expressing, exploiting. In D. I. Hillmann & E. L. Westbrooks (Eds.), *Metadata in practice* (pp. 203–222). Chicago, IL: American Library Association.

Cyganiak, R. (2105). SPARQL queries for statistics. Retrieved from https://github.com/cygri/void/blob/master/archive/google-code-wiki/SPARQLQueriesForStatistics.md

Cyganiak, R., Wood, D., & Lanthaler, M. (2014). *RDF 1.1 Concepts and abstract Syntax*. *W3C Recommendation*. Retrieved from https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/

Deliot, C. (2014). Publishing the British National Bibliography as Linked Open Data. *Catalogue & Index*, (174), 13–18.

Fons, T., Penka, J., & Wallis, R. (2012). OCLC's Linked Data initiative: using Schema.org to make library data relevant on the web. *Information Standards Quarterly*, *24*(2/3), 29–33.

Group, L. L. D. I. (2011). *Library Linked data incubator group: Final report*. W3C. Retrieved from https://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/

Hallo, M., Luján-Mora, S., Maté, A., & Trujillo, J. (2015). Current state of Linked Data in digital libraries. *Journal of Information Science*, *42*(2), 117–127. https://doi.org/10.1177/0165551515594729

Heath, T., & Bizer, C. (2011). *Linked Data: Evolving the Web into a global data space*. Morgan & Claypool.

Heery, R., & Patel, M. (2000). Application profiles: mixing and matching metadata schemas. *Ariadne*, (25). Retrieved from http://www.agi-imc.de/internet.nsf/0/f106435e0fd9ffc1c125699f002ddf31/$FILE/dubin_core.pdf

Hentschke, J. (2017). Linked data service of the German national library. Retrieved July 5, 2017, from http://www.dnb.de/EN/Service/DigitaleDienste/LinkedData/linkeddata_node.html

Hogan, A., Umbrich, J., Harth, A., Cyganiak, R., Polleres, A., & Decker, S. (2012). An empirical survey of Linked Data conformance. *Web Semantics: Science, Services and Agents on the World Wide Web*, *14*, 14–44. https://doi.org/10.1016/j.websem.2012.02.001

Hyland, B., Atemezing, G., & Villazón-Terrazas, B. (2014). *Best Practices for Publishing Linked Data: W3C Working Group Note 09 January 2014*. Retrieved from https://www.w3.org/TR/ld-bp/

IFLA Study Group on the Functional Requirements for Bibliographic Records. (1998). Functional requirements for bibliographic records: Final report. München: K.G. Saur.

Janowicz, K., Hitzler, P., Adams, B., Kolas, D., & Vardeman II, C. (2014). Five stars of Linked data vocabulary use. *Semantic Web Journal*, *5*(3), 173–176. https://doi.org/10.3233/SW-140135

Kontokostas, D., Westphal, P., Auer, S., Hellmann, S., Lehmann, J., Cornelissen, R., & Zaveri, A. (2014). Test-driven evaluation of Linked data quality. In *Proceedings of the 23rd international conference on World wide web - WWW '14* (pp. 747–758). New York: ACM Press. https://doi.org/10.1145/2566486.2568002

LeBoeuf, P. (2012). A strange model named FRBRoo. *Cataloging & Classification Quarterly*, *50*(5–7), 422–438. https://doi.org/10.1080/01639374.2012.679222

Library of Congress. (2012). *Bibliographic Framework as a Web of data: Linked data model and supporting services*. Washington DC. Retrieved from http://www.loc.gov/marc/transition/pdf/marcld-report-11-21-2012.pdf

Mallea, A., Arenas, M., Hogan, A., & Polleres, A. (2011). On Blank Nodes. In *International Semantic Web Conference* (pp. 421–437). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-25073-6_27

Malmsten, M. (2009). Exposing library data as linked data. *IFLA Satellite Preconference*.

Papadakis, I., Kyprianos, K., & Stefanidakis, M. (2015). Linked data URIs and libraries: The story so far. *D-Lib Magazine*, *21*(5/6). https://doi.org/10.1045/may2015-papadakis

Radulovic, F., Mihindukulasooriya, N., García-Castro, R., & Gómez-Pérez, A. (2017). A comprehensive quality model for Linked Data. *Semantic Web*, 1–22. https://doi.org/10.3233/SW-170267

Santos, R., Manchado, A., & Vila-Suero, D. (2015). Datos.bne.es: a LOD service and a FRBR-modelled access into the library collections. In *IFLA WLIC* (pp. 1–18). Cape Town. Retrieved from http://library.ifla.org/id/eprint/1085

Schmachtenberg, M., Bizer, C., & Paulheim, H. (2014). Adoption of the Linked data best practices in different topical domains. In P. Mika, T. Tudorache, A. Bernstein, C. Welty, C. Knoblock, D. Vrandečić, … C. Goble (Eds.), *ISWC 2014, LNCS 8796* (pp. 245–260). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-11964-9_16

Simon, A., Wenz, R., Michel, V., & Mascio, A. Di. (2013). Publishing Bibliographic Records on the Web of Data: Opportunities for the BnF (French National Library). In P. Cimiano, O. Corcho, V.

Presutti, L. Hollink, & S. Rudolph (Eds.), *The Semantic Web: Semantics and Big Data: 10th International Conference, ESWC 2013, Montpellier, France, May 26-30, 2013.* (Vol. 7882, pp. 563–577). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-38288-8

Suominen, O. (2017). From MARC silos to Linked Data silos? Data models for bibliographic Linked Data. In *DCMI/ASIS&T Joint Webinar*. Retrieved from http://dublincore.org/resources/training/#2017suominen

van Hooland, S. (2009). *Metadata quality in the cultural heritage sector : Stakes, problems and solutions*. Universite Libre De Bruxelles.

van Hooland, S., & Verborgh, R. (2014). *Linked Data for Libraries, Archives and Museums: How to clean, link and publish your metadata*. London: Facet publishing. Retrieved from http://difusion.ulb.ac.be/vufind/Record/ULB-DIPOT:oai:dipot.ulb.ac.be:2013/156413/TOC

Villazón-Terrazas, B., Vila-Suero, D., Garijo, D., Vilches-Blazquez, L., Poveda-Villalon, M., Mora, J., … Gómez-Pérez, A. (2012). Publishing Linked Data - There is no One-Size-Fits-All Formula. In *Proceedings of the European Data Forum 2012*. Copenhagen.

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: what data quality means to data consumers. *Journal of Management Information Systems*, *12*(4), 5–33. Retrieved from http://dl.acm.org/citation.cfm?id=1189570.1189572

Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S. (2015). Quality assessment for Linked Data: A Survey. *Semantic Web*, *7*(1), 63–93. https://doi.org/10.3233/SW-150175

Appendix I: Namespaces

```
bio:        <http://purl.org/vocab/bio/0.1/>
blt:        <http://www.bl.uk/schemas/bibliographic/blterms#>
bnb:        <http://bnb.data.bl.uk/id/>
bne:        <http://datos.bne.es/resource/>
bneo:       <http://datos.bne.es/def/>
bnf:        <http://data.bnf.fr/ark:/12148/>
bnfo:       <http://data.bnf.fr/ontology/bnf-onto/>
bnfrel:     <http://data.bnf.fr/vocabulary/roles/>
dce:        <http://purl.org/dc/elements/1.1/>
dcmit:      <http://purl.org/dc/dcmitype/>
dcterms:    <http://purl.org/dc/terms/>
dnb:        <http://d-nb.info/>
event:      <http://purl.org/NET/c4dm/event.owl#>
foaf:       <http://xmlns.com/foaf/0.1/>
frgeo:      <http://rdf.insee.fr/geo/>
geo:        <http://www.w3.org/2003/01/geo/wgs84_pos#SpatialThing>
geonames:   <http://www.geonames.org/ontology/ontology_v3.1.rdf/>
geosparql:  <http://www.opengis.net/ont/geosparql#>
gnd:        <http://d-nb.info/gnd/>
gndo:       <http://d-nb.info/standards/elementset/gnd#>
igno:       <http://data.ign.fr/ontology/topo.owl/>
interval:   <http://reference.data.gov.uk/def/intervals/>
isbd:       <http://iflastandards.info/ns/isbd/elements/>
```

```
library:     <http://purl.org/library/>
madsrdf:     <http://www.loc.gov/mads/rdf/v1#>
mo:          <http://musicontology.com/>
ore:         <http://www.openarchives.org/ore/terms/>
org:         <http://www.w3.org/ns/org#>
owl:         <http://www.w3.org/2002/07/owl#>
rdacarrier: <http://rdvocab.info/termList/>
rdafrbr:     <http://rdvocab.info/uri/schema/FRBRentitiesRDA/>
rdag1:       <http://rdvocab.info/Elements/>
rdag2:       <http://rdvocab.info/ElementsGr2/>
rdau:        <http://rdaregistry.info/Elements/u/>
rdfs:        <http://www.w3.org/2000/01/rdf-schema#>
schema:      <http://schema.org/>
sf:          <http://www.opengis.net/ont/sf#>
skos:        <http://www.w3.org/2004/02/skos/core#>
umbel:       <http://umbel.org/umbel#>
```

Appendix II: Case study of Dylan in BNB

Appendix III: Case study of Dylan in BNE

Appendix IV: Case study of Dylan in BNF

Appendix V: Case study of Dylan in DNB