

# Metadata Structures of the Bibliographic Universe: Transformation, Interoperability, Conceptualizations, and Quality

Kim Tallerås

Dissertation for the degree of philosophiae doctor (PhD)  
Department of Archivistics, Library and Information Science  
Faculty of Social Sciences  
OsloMet – Oslo Metropolitan University

Spring 2018

CC-BY-SA OsloMet – storbyuniversitetet

OsloMet Avhandling 2018 nr 18

ISSN 2535-471X

ISBN 978-82-8364-100-4

OsloMet – storbyuniversitetet

Læringscenter og bibliotek,

Skriftserien

St. Olavs plass 4,

0130 Oslo,

Telefon (47) 64 84 90 00

Postadresse:

Postboks 4, St. Olavs plass

0130 Oslo

Trykket hos Byråservice

Trykket på Scandia 2000 white, 80 gram på materiesider/200 gram på coveret

## Forord

Dette prosjektet ble påstartet for seks år siden, i 2012. I løpet av disse årene har prosjektet tatt form og blitt til noe som ligner på det som ble forespeilet i den opprinnelige prosjektskissa. Det handler fortsatt om metadata og om interoperabilitet. Mitt personlige engasjement for tematikken er også i behold. For selv om seks år er lenge innenfor et såpass teknologisk tema, viser den inneværende avhandlingen at utfordringer og behov for gode løsninger er like vesentlige i dag som da avhandlingen ble planlagt. Om enn på litt nye måter.

I 2012 hadde kulturarvssektoren så vidt begynt å eksperimentere med såkalt lenkede åpne data. I løpet av seks år har eksperimentene blitt til store prosjekter og prinsipper om lenkede åpne data preger mye av den teknologiske utviklingen, fra konvertering av eksisterende data til utviklingen av nye modeller og standarder. Innimellom har det vært krevende å stå midt oppi denne utvikling og orientere seg med et litt mer langsomt forskningsperspektiv. Prosjektet har da også tatt nye retninger, og rygget ut av flere blindgater. Først og fremst har det likevel vært spennende å forsøke å plassere fingeren på pulsen. I den grad prosjektet har lyktes med akkurat det, er det takket være velvillige samtaler, veiledninger og hjelpsomme svar på famlende korrespondanse underveis. Og selvsagt generøse rammevilkår fra arbeidsplass, venner og familie.

Noen personer har i så måte vært uvurderlige. Veileder Nils Pharo har lest utallige skisser og geleidet prosjektet i havn med størst mulig tålmodighet, og med en godt utviklet gjennomføringsevne. Takk!

Kollegaer David Massey og Jørn-Helge Dahl har holdt ut med mange og lange forsøk på å finne røde tråder og bidratt med kunnskap og perspektiver som har løst opp vesentlige floker. De har også sammen med Nils bidratt til to av artiklene. Med sin faglig bredde og sosiale teft har mine øvrige OsloMet- og PhD-kollegaer gitt prosjektet friske innspill fra omkringliggende fagfelt, men også fått tankene over på noe annet når det har vært nødvendig. Instituttledere Liv Gjestrum og Tor Arne Dahl har gitt meg rause vilkår underveis.

Prosjektet også fått verdifulle innspill fra andre miljøer. I sine første år, før det ble flyttet til det ferske PhD-programmet ved Insitutt for arkiv-, bibliotek- og informasjonsvitenskap, var prosjektet plassert ved det som den gang het Det Informationsvidenskabelige Akademi, ved Københavns universitet. Der var Birger Larsen bi-veileder. Han organiserte blant annet et

besøk til Galway og DERI, som lenge har vært det ledende forskningsmiljøet innenfor Semantisk web. Der fikk problemstillinger virkelig brynt seg på teknisk kompetanse og solid forskningserfaring.

Osma Suominen fra det finske nasjonalbiblioteket var en god opponent og samtalepartner i forbindelse med prosjektets sluttseminar. Ricardo Santos Muñoz fra det spanske nasjonalbiblioteket og Asgeir Rekkevik fra Deichmanske bibliotek i Oslo har svart på e-poster og diskutert datamessige finurligheter. Oddrun Ohren og Elise Conradi fra det norske nasjonalbiblioteket har invitert meg og andre OsloMet-kollegaer inn i et faglig nettverk som har vært lærerikt og inspirerende. Det har vært en målsetning for prosjektet at valg av forskningsobjekt og problemstillinger skal resultere i kunnskap som er relevant og som kanskje til og med kan brukes til noe. Kontakten med dyktige folk "ute i felten" har vært avgjørende for forsøket på å strekke seg etter denne ambisjonen.

PhD-perioden muliggjorde også et opphold ved Information Science Institute på University of Southern California. Her ga Craig Knobloch, Pedro Szekely og deres PhD-studenter meg en rekke eksempler på nettopp "real problems" og ikke minst gode eksempler på hvordan man kan møte noen av disse.

Til oppholdet i USA fulgte det med en liten familie, som har fått et nytt medlem siden den gang. Siri, Lars og Anniken. Den største takken skylder jeg dere. Alltid til stede, alltid aller viktigst. I hverdag som på fest.

Jeg må også nevne gode venner og øvrig familie som kanskje har lurt på hva jeg har holdt på med i disse årene, men som like fullt har fungert som svært nyttige sparringspartnere når jeg har forsøkt å forklare akkurat det.

Oslo, 22. mai

Kim Tallerås

## Abstract (English)

This PhD dissertation examines different aspects of the quality of bibliographic metadata structures. In the library field, there is a long tradition of using bibliographic metadata to organize document collections. It essentially involves describing documents and structuring these descriptions in a way that optimizes fitness for use. Fitness for use applies to both the end users of metadata-based information systems and the computers that interpret metadata algorithmically (e.g., a search or a recommendation system) or in contexts where metadata are exchanged across systems.

Metadata descriptions and structures are developed according to standards based on the opportunities and limitations in their technological environment. These standards cover a variety of use cases and purposes. Consequently, significant resources are being put into modernizing standards and metadata practices to exploit technological innovations. In the library sector (and other sectors where metadata are business critical), much of this work in recent years has been inspired by the principles of Linked Data, which encourage metadata producers to publish data on the Web according to Web standards.

After years of Linked Data oriented experimentation and development, evaluations from several perspectives are required. The main purpose of this thesis, therefore, is to provide updated knowledge in this field of work, based on three main research questions: What are the main challenges in transforming bibliographic metadata according to Linked Data principles? What qualities characterize bibliographic metadata published as Linked Data? How do current users conceptualize entities and relationships in the bibliographic universe?

The questions are examined through four studies. The main challenges of Linked Data transformations are investigated through a literature review and an experimental case study. Sets of Linked Data published by four European national libraries are examined through a statistical study of their structural and semantic characteristics. User conceptualizations are explored in a study where informants used concept mapping to conceptualize relationships between selected documents.

The findings show that both published Linked Data and user conceptualizations vary. The national libraries have chosen different strategies when creating their Linked Data sets. The

data conform to Linked Data principles on a general level, but the divergent implementations can hinder interoperability across data sets and with the outside world. Some datasets are also characterized by significant quality problems in completeness and consistency.

A cluster analysis of the user conceptualizations, group participants into two main clusters and five subclusters. The two main clusters represent conceptualizations applying an abstracted multi-entity model to relate documents and conceptualizations that relate documents directly, respectively.

The review of main challenges in Linked Data transformations shows that a significant challenge concerns the choice of target vocabularies, which must be adapted to the purpose of the metadata. The experimental case study also shows that the existing data to be transformed can be characterized by inconsistencies, further affecting the results. This finding is confirmed by the study of the published Linked Data sets.

The findings indicate that the quality of the large amounts of existing data facilitating access to cultural heritage collections must be improved and that new practices and standards must be developed and implemented to prevent new inconsistencies. The variations in user conceptualizations and models for publishing Linked Data indicate that the further development of standards and practices should be closely monitored for relevant purposes and use-case scenarios.

## Abstract (Norwegian)

PhD-avhandlingen undersøker ulike kvalitetsaspekter ved bibliografiske metadata. I bibliotekfeltet har man lange tradisjoner for å bruke bibliografiske metadata til å organisere dokumentsamlinger. Det innebærer i hovedsak å beskrive dokumenter og om å strukturere disse beskrivelsene på en måte som optimaliserer brukskvaliteten. Brukskvalitet gjelder både for sluttbrukere av metadatabaserte informasjonssystemer, for datamaskiner som fortolker metadata algoritmisk (for eksempel et søke- eller anbefalingssystem) eller i sammenhenger hvor metadata utveksles på tvers av systemer.

Metadatabeskrivelser og -strukturer utvikles i henhold til standarder. Slike metadatastandarder har på sin side blitt utviklet på bakgrunn av muligheter og begrensninger i de teknologiske omgivelsene. De skal gjerne dekke flere bruksområder og nye kommer stadig til. Det legges derfor vesentlige ressurser inn i å modernisere standarder og metadatapraksis for å kunne utnytte teknologiske nyvinninger. I biblioteksektoren (og i andre sektorer hvor metadata er virksomhetskritisk) har mye av dette arbeidet de senere årene latt seg inspirere av prinsipper for såkalt *lenkede data* (Linked Data). Dette er prinsipper som oppfordrer metadataprodusenter til å publisere data på Weben i henhold til gitte Web-standarder. Moderniseringsarbeidet består både i å utvikle nye lenkede data-vennlige metadatastandarder og i å transformere eksisterende data i henhold til disse.

Etter noen år med mye eksperimentering, etterlyses evalueringer fra flere hold. Denne avhandlingens hovedformål er derfor å fremskaffe oppdatert kunnskap på dette feltet. Utviklingen undersøkes hovedsakelig ut fra tre hovedspørsmål: Hva er hovedutfordringene ved overgangen til nye bibliografiske metadatapraksiser basert på prinsipper for lenkede data? Hvilken kvalitet har bibliografiske metadata som er publisert på Weben som lenkede data? Hvordan konseptualiserer brukere bibliografiske strukturer?

Spørsmålene blir undersøkt gjennom fire studier. Hovedutfordringer ved en overgang til nye praksiser er undersøkt gjennom en litteraturstudie og et case studie av en eksperimentell metadatatransformasjon. Eksisterende samlinger med lenkede data, publisert av fire europeiske nasjonalbibliotek, undersøkes gjennom en statistisk studie.

Brukerkonseptualiseringer er undersøkt gjennom en studie hvor informanter gjennom en concept mapping-oppgave ble bedt om å angi sammenhenger mellom utvalgte dokumenter.

Funnene viser at både publiserte data og brukerkonseptualiseringer varierer.

Nasjonalbibliotekene har valgt nokså ulike modeller for sine «nye» data. Dataene er gode *lenkede* data på et overordnet nivå, men valgene av ulike modeller kan begrense interoperabiliteten mellom samlingene, og samlet sett mot omverdenen. Enkelte av datasettene er også preget av betydelige kvalitetsproblemer når det gjelder fullstendighet og konsistens.

Brukerkonseptualiseringene kan deles inn i to hovedmodeller, men videre i fem nokså ulike undermodeller av disse. De to hovedmodellene skiller konseptualiseringer som bruker en multi-entitetsmodell til å relatere dokumenter på et abstrahert nivå, og konseptualiseringer som relaterer dokumenter direkte. Undersøkelsen av hovedutfordringer ved transformasjonene viser at en vesentlig utfordring nettopp angår valg av modell, og at dette valget må tilpasses metadataenes formål. Den viser også at eksisterende data som skal transformeres preges av inkonsistenser som videre påvirker resultatet. Dette funnet underbygges av studien av de publiserte lenkede dataene.

Avhandlingen indikerer dermed at kvaliteten ved de store mengdene av eksisterende metadata må forbedres og at nye praksiser og standarder må utvikles og innføres på en slik måte av de best mulig forhindrer nye inkonsistenser. Variasjonen i brukerkonseptualiseringer og modeller for å publisere lenkede data, indikerer at videreutviklingen av standarder og praksiser bør vurderes nøye opp mot (nye) formål og bruksscenarioer.



## Table of Contents

Original Publications.....	3
1 Introduction.....	5
1.1 Disposition.....	9
2 Summary of Motivations, Objectives, Main Research Questions and Contributions of the Thesis.....	10
2.1 Motivations.....	10
2.2 Objectives.....	10
2.3 Research questions.....	10
2.4 Contributions.....	11
3 Background: The Practical and Theoretical Landscape of Bibliographic Data.....	12
3.1 Metadata: The Bigger Picture and Some Definitions.....	12
3.2 Bibliographic Structures.....	14
3.2.1 Bibliographic Metadata Standards.....	19
3.2.2 Bibliographic Control.....	20
3.2.3 Interoperability.....	22
3.2.4 Semantic Web and Linked Data.....	24
3.2.5 Conceptualizations.....	28
3.3 Transitions of Bibliographic Practices.....	30
3.4 Metadata Quality.....	33
4 Academic Affiliation and Theoretical Perspectives.....	36
5 Previous Research.....	40
5.1 Research on Metadata Quality.....	40
5.2 Research on Metadata Transformations and Linked Bibliographic Data Quality.....	41
5.3 Research on User Conceptualizations of Bibliographic Entities.....	43
6 Study design.....	44
6.1 Overall Considerations.....	44
6.2 Examining Challenges in Metadata Transformations.....	44
6.3 Evaluating the Quality of Bibliographic Linked Data.....	46
6.4 Elicitation of User Conceptualizations.....	48

6.5 Ethical considerations.....	50
7 Summary of the Main Findings.....	51
7.1 Challenges of Metadata Transformations.....	51
7.2 Quality of Bibliographic Linked Data.....	52
7.3 User Conceptualizations of Derivative Relationships.....	54
8 Discussion.....	57
9 Conclusions and Further Research.....	65
10 Paper Summaries.....	67
A. Quality of Linked Bibliographic Data: The Models, Vocabularies, and Links of Data Sets Published by Four National Libraries.....	67
B. User Conceptualizations of Derivative Relationships in the Bibliographic Universe.....	68
C. From Many Records to One Graph: Heterogeneity Conflicts in the Linked Data Restructuring Cycle.....	69
D. Ordo ad Chaos—Linking Norwegian Black Metal.....	70
E. Mediation Machines: How Principles from Traditional Knowledge Organization Have Evolved into Digital Mediation Systems.....	71
References.....	72
Appendix A-E	

## Original Publications

This thesis is submitted for the PhD degree at the Department of Archivistics, Library, and Information Science, OsloMet – Oslo Metropolitan University. Professor Nils Pharo has been the main supervisor.

It consists of an introduction and four papers reporting research findings (A-D). One additional paper is included in the appendixes (E).

- A. Tallerås, K. (2017). Quality of linked bibliographic data: The models, vocabularies, and links of data sets published by four national libraries. *Journal of Library Metadata*, 17(2), 126–155.
- B. Tallerås, K., Dahl, J. H. B., & Pharo, N. (2018). User conceptualizations of derivative relationships in the bibliographic universe. Revised and accepted version in the process of being published in *Journal of Documentation*.
- C. Tallerås, K. (2013). From many records to one graph: Heterogeneity conflicts in the Linked Data restructuring cycle. *Information Research*, 18(3).
- D. Tallerås, K., Massey, D., Dahl, J. H. B., & Pharo, N. (2013). Ordo ad chaos: Linking Norwegian black metal. In *Libraries, black metal and corporate finance: Current research in Nordic Library and Information Science* (pp. 136–150). Borås: University of Borås.
- E. Tallerås, K., & Pharo, N. (2017). Mediation machines: How principles from traditional knowledge organization have evolved into digital mediation systems. *Information Research*, 22(1).



## 1 Introduction

In the process of reviewing existing research, the reviewer is dependent on metadata. Research papers, scientific books, and conference proceedings (*data*) must be described (by *metadata*) in a way that allows them to be retrieved and identified as relevant or out of scope. The same prerequisites apply, for example, to parents seeking an age-matched computer game for a 7 year old or, when the day is over, searching for a television series that meets the desire to “Netflix and chill”. In these situations, confronting information systems with clear or vague needs, we, are dependent on detailed, accurate descriptions of content. We explore, retrieve, consume, purchase, and in other ways interact with information systems that rely heavily upon adequate, sufficient delineations of content. The kind of metadata used in such systems—structured descriptions of resources intended to aid finding and understanding (Riley, 2017) —thus has become a key element in our digital surroundings.

Libraries and other cultural heritage institutions continuously strive to increase the findability and access of their collections, particularly in light of the growing amount of digitized and born-digital resources. These collections are based on *bibliographic* metadata. The overarching subject of this research project is the different qualities characterizing such bibliographic metadata in light of the ongoing transitions in the library domain.

Used for various purposes and tasks, metadata have in common that, to some extent, they conform to metadata *standards*. Such standards guide data structures and semantics by defining the entities and relationships in a relevant universe of discourse. For example, in the *bibliographic universe*. Metadata standards affect *fitness for use* when they enable or constrain the information architecture of system interfaces (ideally resembling *user conceptualizations* of entities and relationships) and represent the basic logical schema for databases storing content for retrieval (providing effective *data models*). These standards are also essential tools for facilitating interoperability across collections of metadata.

In any role and use case, metadata standards provide and mandate different levels of complexity, granularity, coverage, and semantic expressiveness. One standard can organize bibliographic

entities hierarchically in an abstract manner so that a translation of a book represents an *expression* entity of an abstract *work* entity (e.g., a Norwegian translation of Shakespeare's original work *Romeo and Juliet*). Another standard may formalize a flat list of editions without any collocating device to organize them. Some standards are old, more or less stable, and widely adopted, whereas other standards are being developed continuously and applied locally.

Currently, in many metadata domains, we are seeing a transition from the use of standards developed under previous technological regimes to new and emerging standards adapting current needs and purposes. This is especially true for cultural heritage institutions. For example, libraries were early adopters of technologies for global data exchange in the 1960s but have struggled to adapt to later technologies, such as relational databases and the Web (Thomale, 2010).

Metadata standards formalize both conceptual and logical data models facilitating retrieval in specific systems. However, one of their main rationales is to support interoperability between metadata systems and providers. This is clearly expressed in the recently revised cataloguing principles issued by the International Federation of Library Associations and Institutions (IFLA) (IFLA Cataloguing Section & IFLA Meetings of Experts on an International Cataloguing Code, 2016), which describe only user convenience as more important than interoperability. These two aspects of bibliographic data—user convenience and interoperability—are also the main aspects investigated in this thesis.

The interoperability challenge for bibliographic metadata traditionally has been solved by the development and adaptation of universal standards following the principle of *universal bibliographic control*, which calls for a seamless global exchange of descriptions (Willer & Dunsire, 2013) to facilitate efficient global registration (“a document should only be described once”). The ability to distribute and integrate data across domains and disciplines on the Web requires more flexible, context-aware standards, capable of sharing bibliographic metadata not only between library systems and institutions but also between the library domain and potential external stakeholders. In addition, the Web has introduced new user expectations regarding

findability and usability, which have proven hard to serve based on dated standards geared towards retrieval tools such as the card catalogue (Clarke, 2015).

The solution for such needs in recent years has pointed towards what Tim Berners-Lee envisioned as the *Semantic Web* (Berners-Lee, Hendler, & Lassila, 2001) and later concretized through a stack of technologies as a set of *Linked Data* principles (Berners-Lee, 2006). The basic idea is that, in the same manner as webpages, raw data should be published and linked on the Web for novel applications based on common Web standards. Structure and semantic representation are assigned via shared standards, in the form of *ontologies*, that offer a vocabulary of defined entity types and relationships. Interoperability thus is established by both common ontologies and direct linking of resources. In the bibliographic area, prominent library institutions have largely embraced these ideas and published bibliographic data on the Web based on their own and others' ontologies. Significant efforts have also been put into (further) developing ontologies adapted to best-practice technologies and methodologies of Web publishing.

Throughout this large-scale, resource-intensive transition to new ways of managing metadata, it is important to assess the quality and results. Data quality can generally be defined as fitness for use (van Hooland, 2009), and as outlined above, the use cases may vary. This PhD project assesses two perspectives on use inspired by theories of ontology development and evaluation (Gómez-Pérez, 2004; Vrandečić, 2010):

- Verification perspective, which concerns whether a metadata standard is built or implemented correctly (according to a set of benchmarks and criteria)
- Validation perspective, which concerns whether the correct metadata standard has been built or implemented (according to the conceptualizations the standard is meant to specify)

Both perspectives are essential to legitimize the use of resources and ensure continued development in the right direction. However, little empirical research on (Linked) metadata quality has been conducted within the bibliographic domain. In the research field of digital libraries, where metadata stands as an essential research object, Saracevic (2005) claims that the

conceptual discussions on quality assessment outnumber the works actually reporting evaluations.

Many papers (see Section 3.3) arguing for the potential upsides of implementing Linked Data principles in libraries have been published. Other papers document various transformation processes. Although researchers have observed major weaknesses in Linked Data quality at a global level (e.g., Hogan et al., 2012), few have reported empirical assessments of bibliographic Linked Data. Regarding verification, this thesis provides an extensive examination of Linked Data sets published by four major European national libraries, reported in paper A.

Regarding aspects of validation, it has repeatedly been stated that bibliographic standards are insufficiently related to users (see e.g., Coyle, 2016; Pisanski & Žumer, 2010a; Zhang & Salaba, 2009). Paper B reports the second main research contribution of this thesis from a study on how 98 participants conceptualize certain entities and relationships in the bibliographic universe.

The transition of bibliographic practices in the library domain from widespread use of common legacy standards to the development of new standards and methods to support interoperability involves several phases. The phases are related but also represent specific challenges. Figure 1 from paper C illustrates a simplified transition process typical of institutions that have published their bibliographic Linked Data. First, they need to develop an ontology fit for the desired purpose. Second, the legacy data need to be mapped and transformed according to the new ontology. Before publication, the data should be linked to external data sets. After publication, an iterative process of evaluation, remodeling, and republishing starts.

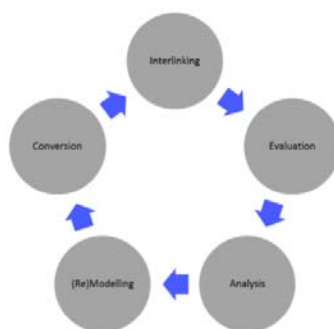


Figure 1. Transition process from legacy standards to Linked-Data-conformant metadata frameworks, used in papers C and D.



The research contributions reported in papers A and B are relevant to all phases but most clearly to the evaluation phase (paper A) and the (re)modelling phase (paper B). Papers C and D bring together a variety of challenges and provide a holistic perspective on the transition of bibliographic metadata standards and can be related to the challenges of each phase, providing the bigger picture. Paper C presents a conceptual literature review of research and bibliographic transition projects, examining the potential obstacles in each phase. Paper D reports on an experimental effort to transform a set of metadata describing musical recordings into best-practices Linked Data and to match the data with a prominent external data source (MusicBrainz).

In addition, the thesis includes paper E among the appendixes, which sheds light on the context of the PhD project. This is a conceptual contribution arguing that the traditional tools and methods for knowledge organization (KO) in libraries, including metadata production, can be viewed as mediating tools that have become essential elements in the algorithmic dissemination taking place in current search and recommender systems.

## 1.1 Disposition

The remainder of the thesis introductory is structured in the following way:

Section 2 summarizes motivations, objectives, research questions and the contribution of the PhD project.

Section 3 elaborate theoretical and practical aspects of the problem area examined.

Section 4 positions the project in the academic landscape and discusses the overall theoretical perspective.

Section 5 summarizes relevant previous research.

Section 6 outlines and discuss aspects of the chosen study designs.

Section 7 provide an overview of main findings from the different research efforts.

Section 8 discuss the implications of the main findings.

Section 9 conclude and outlines future research.

Section 10 provide summaries of the included papers.

## 2 Summary of Motivations, Objectives, Main Research Questions and Contributions of the Thesis

### 2.1 Motivations

In recent years, there has been increased interest in renewing the bibliographic landscape, in particular to implement new methods for interoperability, access, and discovery and following from that to develop new standards. The current thesis is motivated by the need for knowledge of this transition and addresses two evaluation needs insufficiently covered in the empirical literature:

- The lack and need for evaluations of the substantial efforts invested in transforming bibliographic data according to new standards and principles
- The lack and need for user testing of bibliographic standards

### 2.2 Objectives

The main objective of the PhD project is to provide knowledge about the outcomes of significant efforts to modernize bibliographic practices and to map out some needs for the road ahead. Through the use of different methods to study independent (but complementary) research objects, combined with overarching conceptual studies, the project also aims at contributing to a better understanding of the bigger picture and the contextual factors of the current transition process and challenges.

### 2.3 Research questions

The conducted research is based on the following questions:

1. What are the main challenges in transforming bibliographic metadata according to Linked Data principles?
2. What qualities characterize bibliographic metadata published as Linked Data?
3. How do users conceptualize entities and relationships in the bibliographic universe?

## 2.4 Contributions

The field of bibliographic metadata is characterized by transition and experimentation. The writings and documentation of efforts in this field are often based on professional reflections and discussions, which are important and necessary for the development of new metadata practices and provide insights from highly experienced, competent professionals. The main contribution of this PhD project is to add empirical knowledge to these ongoing efforts and bibliographic discourses.

## 3 Background: The Practical and Theoretical Landscape of Bibliographic Data

### 3.1 Metadata: The Bigger Picture and Some Definitions

In 2018, we use so-called artificial intelligence (AI) to perform everyday tasks; for example, we can ask the virtual assistant on our mobile phone to send a message to home, sharing our geographic location and informing our family that we likely will be late for dinner due to traffic jams. This form of AI may have a personal name, such as Apple’s Siri and Amazon’s Alexa, but does not (for the time being) come in the form of human-like robots as portrayed in science fiction movies. AI is primarily based on algorithms that utilize huge amounts of data about, for example, geographical conditions, user behavior, and sound waves from voices. Textbook definitions of *metadata* are broad: devices “encapsulating the information that describes any information-bearing entity” (Zeng & Qin, 2016, p. 11) or simply “information resource description” (Hider, 2012, p. 4). In practice, these definitions include metadata facilitating AI, understood as descriptions of particular phenomena—*data about data*—such as geocodes in a map of Oslo and information about the characteristics of a particular voice. Although, metadata are essential to the emerging technologies commonly described with labels such as Big Data, machine learning, and AI, they also belong to a long-standing tradition evolving from libraries (such as the great one in Alexandria) and serve an extensive variety of other use cases. Descriptions of information resources are useful in all contexts where we need to handle quantities of resources that exceed our cognitive capacity for memory storage and organization.

Since metadata can be used for different purposes, they are often categorized into different types (see e.g., Hider, 2012; Riley, 2017; Zeng & Qin, 2016). *Administrative* metadata typically provide provenance information, *technical* metadata relate to system functionality, and *descriptive* metadata facilitate retrieval and discovery. This PhD project is based on the long-standing tradition of producing and utilizing descriptive metadata in the library domain. In this tradition, the information-bearing entities typically are some form of documents entered into the physical, and eventually digital, collections of library institutions. These documents are described, and their descriptions—sometimes referred to as document *surrogates*—are organized

so that they, with different levels of success, facilitate access to increasingly larger document collections. These descriptions and their organization have been adapted to technological developments, from book catalogs and card catalogs to today's digital catalogs.

Machine learning algorithms, the accumulation of Big Data, and other phenomena enabling AI, may seem to belong to a slightly different domain. Zeng and Qin (2016, p. 393), however, claim that one of the research frontiers within the metadata field is related to what they refer to as *data-driven X*, where X stands for “any discipline or field of research, practice, and learning” (e.g., data-driven healthcare). The main challenge in this field of research is the great need for data processing and caching. The library community's endeavors to restructure and publish metadata openly available on the Web, ideally to support various forms of novel data-driven applications and AI, may very well be interpreted into this context. Paper E, entitled “Mediation Machines: How Principles from Traditional Knowledge Organization Have Evolved into Digital Mediation Systems,” also shows that new, sophisticated applications of tools and methods originating in the library realm (including classification principles, bibliometrics, information-retrieval techniques and evaluation metrics) have formed the technological foundation for game-changing search and recommender systems, such as those developed by Google and Netflix. The library domain, including the field of bibliographic metadata, may not be in the driver seat of this development but is highly relevant in both technology and motivations.

As described, metadata can be used by intelligent machines “under the hood” in different types of information systems but often appear “over the hood” in the interfaces presented to intelligent humans as part of the information systems architecture (the “structural design of shared information environments” (Rosenfeld, Morville, & Arango 2015, p. 24)). An example of a metadata structure could include the following statements: this is an “author,” this is a “scientific paper,” and this is a “peer,” and the author “has written” the paper “cited” by the peer. This kind of metadata structure, supported with the immense amounts of existing bibliometric data, can be utilized both in an AI context to analyze research networks for various purposes and in a user interface to support navigation, such as for the researcher performing a systematic literature review. In this dissertation, the structural characteristics of metadata—which are often emphasized in the definition of the term (see, e.g., Greenberg, 2005, 2009)—are examined from

different perspectives as they are developed and applied in a bibliographic domain continuously adapted to new technologies and users.

### 3.2 Bibliographic Structures

Following the outlined scope on metadata, the problem area—the *universe of discourse*—of this thesis can be defined as *the bibliographic universe*. In *Two Kinds of Power*, a highly influential book discussing this complex concept space, Wilson (1968) describe its constituents and composition. Smiraglia (2014, p. 10) summarizes the discussion in a rather poetic way:

Wilson sees points in the universe orbiting and clustering and crossing the bibliographical macrocosmos, in concert with each other according to specifiable (if so far unspecified) relationships patterns. Just as the physical universe reels with gravity and physical forces that propel, impel, and compel planets, stars, asteroids and other bodies to exist in relation to each other, so Wilson sees the bibliographical universe as a mulita-dimensional, relational system.

Wilson (1968) construes the orbiting points as texts, defined broadly to include everything from printed books to street signs, audiovisual material, and texts stored in the “memories of machines” (p. 12). The totality of texts can thus be outlined as a concept space containing all recorded knowledge (Smiraglia, 2014).<sup>1</sup> Since Wilson (1968), many have discussed the complexity of bibliographic objects and suggested other concepts to describe them, such as *documents* in the tradition of Paul Otlet and Suzanne Briet (Buckland, 1997) and *boundary objects* in the tradition of Susan Leigh Star and James R. Griesemer (Huvila, Anderson, Jansen, McKenzie, & Worrall, 2017). Most suggestions imply that the objects have latent informational value and are often followed by a discussion on how to organize (or control) such objects in a way that optimizes this value. Describing them with metadata is one such way (“descriptive control” in the terminology of Wilson, 1968).

---

<sup>1</sup> Glushko (2013, p. 142) points out that the *biblio-* root does not limit what is part of the *bibliographic* universe; instead, it is populated by all the resources typically contained in libraries.

This thesis primarily uses the terms bibliographic *entity* to describe objects in the bibliographic universe and bibliographic *relationships* to describe the gravitational forces between them. More concretely, bibliographic entities refer to instances of recorded knowledge (e.g., a novel, movie, or piece of music). Entities, or “information-bearing messages” as Svenonius (2000, p. 34) calls their most singular and particular form, can further be viewed from different levels of abstractions. In everyday language, we use phrases like “a good book” or “a new movie,” regardless of whether the book is translated from another language or the movie has been accessed via a cinema or a streaming service. This way of speaking from a perspective where the bibliographic universe consists of single entities (*works* abstract from, for example, a specific carrier type) makes sense in conversations but grossly simplifies existing relationships in the bibliographic universe. Although a single-entity perspective may be fruitful in some information systems, like those operating in a lexical context such as the IMDb and Wikipedia, which typically answer questions involving at least one known entity (e.g., Who played in a particular movie? Which books were written by a particular author?), this most certainly decreases the *fitness for use* of, for example, information systems helping us find a specific translation of a text in a library. Specific information needs may require detailed information about publishing history, and the most popular areas of the bibliographic universe are constantly populated with new editions, updates, translations, and derivations (Smiraglia & Leazer, 1999), which complicates things from an organizational point of view. Wilson (1968, p. 13), therefore, poses a fundamental question, which has since been central to the bibliographic debate: “Suppose that we could make an exhaustive inventory of the contents of the bibliographic universe. What should we want to count as one item in that universe?”

In pre-digital metadata practices in libraries, a physical document in the collection was represented by a physical card in a card catalogue. This main card was further filed at a particular place in the catalogue based on its determined main point of entry (typically the surname of the first listed author). The documents were placed on the shelves according to their main entries or classification numbers. This physical assumption leads to the situation where each physical document forms the basis of a publishing-oriented response to Wilson’s question; one release, one item, one card (and eventually one digital record). These cards can be linked through added entries (e.g., via co-authors and topicality) but are rarely linked to entities that collect documents

at a higher level of abstraction, such as what we refer to as a “work” in our daily conversations. Two editions result in two cards, collected (or spread) by alphabetical or numerical order. Current digital technologies have made it easier for bibliographic entities to be filed or discovered according to different structural models; in other words, the physical straightjacket is off.

In this context, Svenonius (2000) suggests a set theoretic perspective grouping bibliographic entities of different abstraction levels into sets of entity types (e.g., *editions* and *works*). In practice, this means there may exist different sets of editions (groups of documents “sharing the same information”), which can further be part of a work set (collocating editions “sharing essentially the same information” (Svenonius, 2000, p. 35). This results in a *multi-entity model* (Baker, Coyle, & Petiya, 2014), providing a variety of entity types that function as collocating devices at different levels of abstraction. Each entity type supports specific organizational purposes. The division between editions and works, as suggested by Svenonius (2000), could, for example, enable both effective work-level retrieval in a library system (e.g., by downsizing the hit list to contain single works rather than an insurmountable collection of various editions of the same work (Westrum, Rekkavik, & Tallerås, 2012)) and edition-level lending (e.g., identification of the right edition). In this way, bibliographic structures potentially contribute both to the systems handling data (e.g., indexing a database for managing loans) and user discovery by forming the information architecture of the interface.<sup>2</sup>

Traditional and widely used bibliographic standards, such as the Anglo-American Cataloguing Rules (AACR) and the MARC family of exchange formats,<sup>3</sup> are oriented toward single entities, materializing records describing particular releases of books (Clarke, 2015). Such records may group multiple physical exemplars of an edition, but for different reasons, it has proved challenging to use these standards to derive entities on a higher level of abstraction (Aalberg & Žumer, 2013). This is also the case in some of the national libraries’ efforts evaluated in this thesis (paper A) to transform legacy records according to bibliographic standards containing

---

<sup>2</sup> For an interesting discussion of the relationship (or the lack of reflection on the relationship) between the data model and the mental model in newer bibliographic models, see Coyle (2017).

<sup>3</sup> <https://www.loc.gov/marc/>



entity types similar to Svenonius's (2000) definition of works. In addition to the set theoretic approach suggested by Svenonius, current discussions and attempts at implementing multi-entity models are based on a long history of research on bibliographical entities (e.g., Carlyle, 1997; Lee, 1994; Tillet, 1991; Smiraglia, 1999) and, not the least, efforts to develop so-called reference models by prominent agents, such as the International Federation of Library Associations and Institutions (IFLA) and the Library of Congress. Most notable of such models is the *FRBR model*, which was first presented in 1998 (IFLA Study Group on the Functional Requirements for Bibliographic Records, 1998) and has recently culminated in the *Library Reference Model* (LRM) (Riva, Bœuf, & Žumer, 2017). Emerging standards, such as the *Resource, Description and Access* (RDA),<sup>4</sup> *BIBFRAME*,<sup>5</sup> and *FRBRoo* (LeBoeuf, 2012), all build on an interpretation of *W/E/M/I*-entities (Work/Expression/Manifestation/Items), introduced in the original FRBR model. In LRM (Riva et al., 2017, pp. 21–17) they are defined as:

- *Work*: “The intellectual or artistic content of a distinct creation.”
- *Expression*: “A distinct combination of signs conveying intellectual or artistic content.”
- *Manifestation*: “A set of all carriers that are assumed to share the same characteristics as to intellectual or artistic content and aspects of physical form.”
- *Item*: “An object or objects carrying signs intended to convey intellectual or artistic content.”

These entities have proven to be essential in all the research efforts discussed in this thesis. In addition, one project investigating user conceptualizations (paper B) includes yet another abstraction level that Svenonius (2000, p. 35) refers to as a *superwork* (“the set of all documents descended from a common origin”). Svenonius (2000, p. 38) explains that superworks collocate works similar “by virtue of emanating from the same ur-work.” As a concrete example of a *Hamlet* superwork, she lists collocated works such as the “original text, motion pictures, sound recordings of readings, analyses of the play, commentaries, playbills, derivative works like

---

<sup>4</sup> <https://www.loc.gov/aba/rda/>

<sup>5</sup> <https://www.loc.gov/bibframe/>

*Rosencrantz and Guildenstern Are Dead*” (Svenonius, 2000, p. 38). Smiraglia (2007) discusses *the bibliographic family*, a concept similar to superwork introduced by Wilson (1968). A bibliographic family collocates kindred works. The family structures are all “unique in the relationship the members bear to the originating work [...] yet distinct patterns occur among the members” (Smiraglia, 2007, p. 74). Others, including Carlyle (1999) and Yee (1994), have also touched upon the idea of a high-level collocating device. Vukadin (2014), in line with the empirical findings reported in paper B, points out the need for even more complex entities and bibliographic relationships to describe the structures emerging through so-called transmedial storytelling, in which fictional characters and narratives are shared and expanded across documents and media types.

The introduction of collocating devices at different abstraction levels exemplifies how bibliographic structures can vary in depth and other dimensions. In a study on the differences between the BIBFRAME standard developed by Library of Congress (“to determine a transition path for the MARC 21 formats”) (Library of Congress, n.d.) and the Schema.org vocabulary (“a collaborative, community activity with a mission to create, maintain, and promote schemas for structured data on the Internet, on web pages, in email messages, and beyond”) (W3C Schema.org Community Group, n.d.) used by the OCLC to publish its bibliographic data as Linked Data, Godby and Denenberg (2015) and Godby (2013) emphasize differences in both depth and breadth. While BIBFRAME is a bibliographic standard built to cover the bibliographic universe in depth, Schema.org is designed to support metadata embedded in Web pages in a more general manner and, therefore, is characterized by significant coverage across domains. Other structural varieties between bibliographic standards can be related to divergent naming and modeling practices dependent on specific technologies (e.g., structures based on lists, relational databases, hierarchical markup languages like XML, or network theory such as the Resource Description Framework (RDF)). These are referred to as potential *meta-level discrepancies* by Haslhofer and Klas (2010, p. 16).

### 3.2.1 Bibliographic Metadata Standards

Bibliographic structures can vary and be modelled in different ways, they are nevertheless formalized and guided by bibliographic standards.<sup>6</sup> The International Organization for Standardization (ISO, n.d.) defines standards as “documents that provide requirements, specifications, guidelines or characteristics that can be used consistently to ensure that materials, products, processes and services are fit for their purpose.” Bibliographic standards are designed to serve multiple purposes, often at the same time. Based on these purposes, Zeng and Qin (2016) divide metadata standards into four categories: those made to serve data contents, data values, data structures, and data exchange. Hider (2012, p. 6) lists four aspects of metadata standards similar to these categories: elements, values, format, and transmission.

The current metadata practice of libraries, at least regarding the descriptions of their physical collections, comprises a widely adopted set of bibliographic metadata standards involving all of these aspects. The *Machine-Readable Cataloging* (MARC-)format, developed as an exchange format for bibliographic data between libraries in the early 1960s (Avram, 1975), also specifies a particular data structure by providing a (numerically coded) *vocabulary* of mandatory or non-mandatory elements. The Anglo-American Cataloguing Rules (AACR) and the standard currently superseding it—Resource Description and Access (RDA)—are data content standards guiding which (MARC-)elements to use and how to form the content. The Dewey decimal classification system provides controlled values that can be used as common access points. This is a simplified summary of a complex domain. To a certain extent, AACR, for example, also guides the data structure, but these examples of different types of bibliographic standards show how various aspects are approached in practice.

In this context, we can add the already mentioned reference models. These models provide an abstract framework defining and clarifying core entities intended for inspiration and alignment,

---

<sup>6</sup> In the literature, we find alternative but more or less overlapping terms used to denote such formalizations, such as bibliographic *frameworks* (Glushko, 2013) and bibliographic *languages* (Svenonius, 2000).

for example, across domains to harmonize the development of domain-specific standards (OASIS SOA Technical Committee, n.d.). A prominent example of a reference model within the bibliographic landscape is LRM, which defines a set of conceptual entities that have, for example, been incorporated in RDA. FRBRoo is another reference model that harmonizes ways of organizing entities from the library and museum sectors through FRBR entities.

The terms model, structure and standard are used to describe somewhat overlapping phenomena throughout this introductory and in the included papers. A model is a generic term that name and describe a variety of metadata constructions (such as in *bibliographic models*, *the FRBR model*, *data models*, *reference models* etc). The term structure is used to denote the concrete arrangement and semantic organization of the granular building blocks of models. Standards are formalized models.

### 3.2.2 Bibliographic Control

The theoretical and, to some extent, empirical investigations of bibliographic structures and, not the least, the efforts to formalize standards like LRM and to publish bibliographic data conforming with new Web standards relate in different ways to the overall objective of achieving *bibliographic control*. Bibliographic control is defined by the ODLIS dictionary (Reitz, n.d.):

A broad term encompassing all the activities involved in creating, organizing, managing, and maintaining the file of bibliographic records representing the items held in a library or archival collection, or the sources listed in an index or database, to facilitate access to the information contained in them. Bibliographic control includes the standardization of bibliographic description and subject access by means of uniform catalog code, classification systems, name authorities, and preferred headings; the creation and maintenance of catalogs, union lists, and finding aids; and the provision of physical access to the items in the collection.

Moreover, for contextual purposes, it is worth mentioning that the traditional approach to bibliographic control in the library sector—descriptive cataloguing of the formal features of books and other documents based on a given metadata standard—is often justified by the need to support specific user tasks. Already in the late 1800s, Charles Cutter formulated a set of such

tasks that, with some reforms and extensions, is explicitly recognized as the starting point for both LRM (Riva et al., 2017, p. 14) and IFLA's recently reformulated version of Statement of International Cataloging Principles. Here, in the wording of the latter principles (IFLA Cataloging Section & IFLA Meetings of Experts on an International Cataloging Code, 2016, p. 10) [emphasis added]:

The catalogue should be an effective and efficient instrument that enables a user:

6.1 *to find* bibliographic resources in a collection as the result of a search using attributes or relationships of the entities:

to find a single resource or sets of resources representing:

- all resources realizing the same work
- all resources embodying the same expression
- all resources exemplifying the same manifestation
- all resources associated with a given person, family, or corporate body
- all resources on a given theme
- all resources defined by other criteria (language, place of publication, publication date, content form, media type, carrier type, etc.), usually as a secondary limiting of a search result;

6.2 *to identify* a bibliographic resource or agent (that is, to confirm that the described entity corresponds to the entity sought or to distinguish between two or more entities with similar characteristics);

6.3 *to select* a bibliographic resource that is appropriate to the user's needs (that is, to choose a resource that meets the user's requirements with respect to medium, content, carrier, etc., or to reject a resource as being inappropriate to the user's needs);

6.4 *to acquire* or obtain access to an item described (that is, to provide information that will enable the user to acquire an item through purchase, loan, etc., or to access an item electronically through an online connection to a remote source); or to access, acquire, or obtain authority data or bibliographic data;

6.5 *to navigate and explore*

within a catalogue, through the logical arrangement of bibliographic and authority data and the clear presentation of relationships among entities beyond the catalogue, to other catalogues and in non-library contexts.

The classical user tasks emphasized in this definition permit interpreting bibliographic control from slightly different perspectives. Wilson (1968) describes two distinct kinds of control. The first, *descriptive control*, relates directly to the first four user tasks by providing the means, traditionally by cataloging, to create (arbitrary) lists that enable retrieval of all the entities characterized by certain attributes ("All (available) plays by Ibsen"). *Exploitative control*, in contrast, is the ability to procure the best entities available serving a specific purpose (e.g., a text outlining contemporary interpretations of *A Doll's House* for a school paper). The first kind of control is evaluative neutral from a user's perspective, whereas the second involves user appraisal

(Wilson, 1968). According to Wilson (1968), exploitative control is more important, but descriptive control is a precondition for achieving exploitative control. To identify the best entities, these entities must be known, and to be known, they must be described accordingly. The last user task mentioned, “to navigate or explore,” is a relatively recent addition directly related to exploitative control. One could argue that the problem of descriptive control has more or less been solved today, especially given the amount of digitized content enabling automated generation of descriptions. However, the ability to exploit these descriptions—to exercise exploitative control—can be defined as the core business (and perhaps the core challenge) of the world’s leading tech companies. Some reasons for the changes we see in the library sector, described in more detail in the following sections, may also be attributed to a desire to facilitate exploratory retrieval, for example, by introducing FRBR entities or by converting data into more Web-oriented formats.

### 3.2.3 Interoperability

Another motivation underlying the transition processes is a desire to increase the level of interoperability. As mentioned, this is a clear priority in the recently reformed cataloguing principles. A separate paragraph concerning interoperability states that “all efforts should be made to ensure the sharing and reuse of bibliographic and authority data within and outside the library community” (IFLA Cataloguing Section & IFLA Meetings of Experts on an International Cataloguing Code, 2016, p. 5). Interoperability can be related directly to bibliographic control via the extended concept of *universal* bibliographic control (UBC) based on the objective to promote “a world-wide system for the control and exchange of bibliographic information” (Willer & Dunsire, 2013, p. 3). An essential part of this system is the development and use of standards: “the comprehensive bibliographic record of a publication is made once in a country of its origin, in accordance with the international standards which are applicable in both manual and in mechanized systems; and is then available speedily, in a physical form which is also internationally acceptable” (Willer & Dunsire, 2013, p. 4).

UBC represents a domain-specific approach to interoperability. In a general metadata context, Riley (2017, p. 4) defines interoperability as “the effective exchange of content between systems”

and further argues that interoperability “relies on metadata describing that content so that the systems involved can effectively profile incoming material and match it to their internal structures.” There are different suggestions for the categorization of types (or levels) of interoperability; see, for example, Chan and Zeng (2006); Nilsson, Baker, and Johnston (2009); Tolk and Muguira (2003); and the overall ICT principles of the Norwegian Agency for Public Management and eGovernment (Direktoratet for forvaltning og ikt, 2016). Such categorizations often include or at least touch upon *technical* interoperability, or the infrastructures for data exchange (e.g., enabled by HTTP protocol); *structural* interoperability, or shared data formats (e.g., XML, JSON, and RDF); and *semantic* interoperability, or common interpretations of meaning. The main perspective on interoperability in this thesis concerns this last category: how systems understand, communicate, and utilize data.

Even more concretely, semantic interoperability can be described (negatively) as the absence of specific heterogeneity conflicts. In the literature, such conflicts (or just *heterogeneities*) are presented from different application perspectives seeking to integrate or align data between for example geographical information systems (Bishr, 1998), XML data (Pluempitiwiriyawej & Hammer, 2000), databases (Doan & Halevy, 2005), e-government services (Peristeras, Loutas, Goudos, & Tarabanis, 2008), and software architecture (Davis, Flagg, Gamble, & Karatas, 2003). Heterogeneity conflicts are also problematized from a more theoretical perspective as challenges occurring due to domain evolution (Ventrone & Heiler, 1991) and worth considering in methodological frameworks, for example, as a set of challenges to overcome in the evaluation of instance matching tools (Ferrara, Lorusso, Montanelli, & Varese, 2008). The theoretical perspective on semantic interoperability in this thesis is based on the general definitions of metadata heterogeneities provided by Haslhofer and Klas (2010). Most authors mentioned distinguish between heterogeneities that occur due to incompatibilities between structural characteristics and to language use. An example of a typical structural heterogeneity is *abstraction-level incompatibility* (Haslhofer & Klas, 2010, p. 17), as in cases when we want to integrate data conforming to metadata structures with and without a bibliographic *work* entity. A typical language-use heterogeneity is a terminological mismatch due to the use of synonyms or homonyms in the naming of metadata elements. Challenges reported in all thesis papers relate directly to many of Haslhofer and Klas’s (2010) categories, such as the examination and

comparison of abstraction levels in user conceptualizations reported in paper B and the comparison of terminology conflicts between metadata standards used to publish bibliographic data on the Web (paper A).

Alongside the UBC tradition, other suggestions on how to solve heterogeneity conflicts have been presented in recent decades. These include the aforementioned reference models, development of crosswalks and mappings between standards (Chan & Zeng, 2006), and merging of metadata elements from existing standards in *application profiles* (Heery & Patel, 2000).

### 3.2.4 Semantic Web and Linked Data

After inventing the essential components of the Web architecture, Tim Berners-Lee introduced the idea of an extension of the Web enabling relationships between not only documents but also the things documents are about: in practice, a graph of interlinked data objects published and exposed on the Web. The idea was first presented as a *Semantic Web* (Berners-Lee et al., 2001), then connected to a concrete technological infrastructure and a set of best practices for publishing, and revitalized as *Linked Data*<sup>7</sup> (Berners-Lee, 2006) principles to support bottom-up adoption of the Semantic Web.

The original Linked Data principles advocate open publication of structured data on the Web in non-proprietary formats based on World Wide Web Consortium (W3C) standards. The standards explicitly listed are uniform resource identifiers (URIs), which identify and address specific resources; the HTTP protocol for exchanging URIs; RDF, which provides the structure for the organization of the resources; and SPARQL, a query language used to retrieve RDF data. The original principles (Berners-Lee, 2006) are:

- Use URIs as names for things
- Use HTTP URIs so that people can look up those names.

---

<sup>7</sup> *Linked Open Data* (LOD) is a common term used to emphasize that the relevant data are both linked and licensed in way that make them freely available for reuse and distribution. Linked Data and LOD are used interchangeably in the literature. The original term *Linked Data* (Berners-Lee, 2006) is used consistently throughout this introductory to avoid confusion.



- When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL)
- Include links to other URIs so that users can discover more things.

The principles have since evolved into comprehensive collections of best practice recommendations, including both general guidelines (see, e.g., Heath & Bizer, 2011; Hyland, Atemez, & Villazón-Terrazas, 2014) and guidelines targeting data providers in specific domains (e.g., van Hooland & Verborgh, 2014). The emphasis on standards and transparency indicates a lingua franca approach to solving heterogeneity conflicts across domains and datasets.

Semantic Web and Linked Data thus have a slightly different approach to interoperability and data integration than that typical in the library sector. In the library sector, UBC has led to top-down development of common and widely adopted standards. Within the Semantic Web, it is essential to use standards to support both semantics and technical data exchange, and a central Linked Data principle is to reuse existing standards; therefore, this approach can be interpreted as a continuation of substantive UBC principles. At the same time, the Semantic Web represents a more flexible, heterogeneous approach to KO: data providers are encouraged to extend the Web by adding and linking more data but are not controlled in how to do so. The Semantic Web, therefore, is *incomplete* by default and based on an open world assumption (OWA). This implies that “absence of information is interpreted as unknown information, not as negative information” (Keet, 2013). There may be relevant information that for some reason is not represented in a given (Linked) dataset. Zero instances, therefore, do not mean that one can logically exclude a phenomenon. For example, one cannot exclude that Henrik Ibsen has written more or other books: these works simply are not part of the Semantic Web, yet. This challenges a UBC regime and any bibliographic systems based on an inventory function with the purpose of gathering or *controlling* a complete collection of metadata. In contrast, the purpose of UBC is based on the *closed* world assumption (CWA) that missing data do not exist; in the preceding example, depending on the context, it is presumed that Ibsen has not written other books than those listed, or that no other books is part of a given collection. In discussions, some OWA advocates has claimed that CWA has led to “over-engineered schema, too-complicated architectures and massive specification efforts” (Bergman, 2009) and that the OWA can lead to more viable development along the premises of an extensible Web.

Although OWA certainly offers flexibility, this flexibility can be challenging for data quality. As Pattuelli and Rainbow (2013, pp. 1–2) describe it, “the literature on Linked Data development has just begun to address the implications of dealing with loosely formalized knowledge structures that produce significant amounts of ‘noisy’ data. [...] Nonetheless, there is not yet a substantial enough body of research with which to frame an articulate and cohesive discussion on LOD quality data.” Paper A of this thesis contributes research examining the data quality of bibliographic Linked Data, including some discussions on handling completeness as a quality criterion.

Another challenge that arises in the merger of metadata from the library sector and the Semantic Web is the basic organizational model of the Web. Libraries traditionally have organized metadata as records, as described in Section 3.2. On the Web, documents and metadata are organized in networks and graphs, similar to the *giant global graph* (Berners-Lee, 2007) or *Web of data* (Bizer, Heath, & Berners-Lee, 2009) in Berners-Lee’s vision of the Semantic Web. The recommended RDF format specially designed to support this vision thus is defined in the specification as a “graph based data model” (Cyganiak, Wood, & Lanthaler, 2014). The sets of Linked Data examined in paper A are accordingly published as RDF graphs. The experimental transformation described in paper D also targets an RDF graph.

A metadata record organizes information in relatively isolated units collocating the descriptive details of a document (e.g., author(s), titles, and publishers). Applying RDF as the data model for these descriptions makes it part of a connected network of nodes and edges. The smallest unit in an RDF graph is a resource (or entity) that “denotes something in the world” (Cyganiak et al., 2014). According to the RDF specification, anything can be a resource, “including physical things, documents, abstract concepts, numbers and strings” (Cyganiak et al., 2014). Thus, Ibsen can be a resource, his play *Peer Gynt* another one, and the publisher yet another one. Such resources can be further connected via so-called *RDF triples*, in which each resource is a node (the *subject* and *object* of the triple) connected by a *predicate* providing the semantics of the relationship, as illustrated in Figure 2. In the figure, `:peer_gynt` and `:henrik_ibsen` are examples of URIs representing the author and the play. The predicate is formalized according to

Linked data principles<sup>8</sup> by applying the `author` property from the `schema.org`-ontology. Paper C outlines and discusses challenges to the transition from “many records to one graph.”

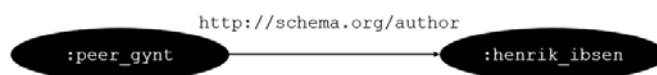


Figure 2. A visualized example of an RDF triple where `:peer_gynt` is the subject, `:henrik_ibsen` the object, and `http://schema.org/author` the property.

Although huge amounts of data have been published on the Web according to Linked Data principles,<sup>9</sup> this methodology has been criticized for not supporting a “killer app.” Others, though, have argued that Google’s increasing use of structured data in its search interface (see, e.g., Isidoro, 2013) and the world’s prominent search engines’ call to embed structured data in Web sites applying the `schema.org`-ontology recall the vision of a Semantic Web. In the library domain, institutions have started to build discovery tools directly on top of RDF data, as in the case of Oslo Public Library (Westrum, 2014). Some of the most recent developments in the Linked Data landscape focus on increasing the usefulness of Linked Data. Two examples are the *JSON for Linking Data* format<sup>10</sup> (JSON-LD), which harmonizes the widely adopted JSON format for data exchange with the graph-based data RDF model, and Linked Data fragments (Verborgh et al., 2016), a project intended to create a more scalable search experience over Linked Data collections.

---

<sup>8</sup> “Standardized vocabularies should be reused as much as possible to facilitate inclusion and expansion of the Web of data” (Hyland, Atemezing, & Villazón-Terrazas, 2014b).

<sup>9</sup> For concrete statistics, see the LODStat website (<http://stats.lod2.eu/>), which crawls the Web of interlinked data and provides an overview of the total numbers of sets, triples (at the time of writing, 149,423,660,620 triples from 2,973 datasets were retrieved) and vocabularies used. *The Linking Open Data cloud diagram* (Abele, McCrae, & Buitelaar, 2017) is also built on crawling Linked Data sets and visualizing the retrieved datasets and the links between them.

<sup>10</sup> <https://www.w3.org/TR/json-ld/>

### 3.2.5 Conceptualizations

García-Castro and Gómez-Pérez (2011, p. 14) define semantic interoperability as “the ability that semantic systems have to interchange ontologies and use them.” This relates directly to the expressed goals of library institutions to transform data to Semantic Web and Linked Data conformant standards (as described in section 3.3). Gruber (1993, p. 199) defines an ontology as an “explicit specification of a conceptualization.” Studer, Benjamins, and Fensel (1998) refine Gruber’s (1993) ontology definition by stating that these conceptualizations should be *shared*. Smith (2004, p. 161) elaborates the implications of such conceptualizations:

As we engage with the world from day to day, we participate in rituals, and we tell stories. We use information systems, databases, specialized languages, and scientific instruments. [...] Each of these ways of behaving involves, we can say, a certain conceptualization. What this means is that it involves a system of concepts in terms of which the corresponding universe of discourse is divided up into objects, processes, and relations in different sorts of ways. [...] Tools can be developed to specify and to clarify the concepts involved and to establish their logical structure.

In this sense, an ontology can be regarded as a concrete tool utilized for KO. As concrete tools, formal ontologies are often referred to as mere *vocabularies*, *metadata schemas*, or *data models*. W3C<sup>11</sup> argues that there is no clear division between what is, for example, a vocabulary and an ontology, but ontologies tend to be used for more complex, formal structures. Nevertheless, following Gruber’s (1993) general definition, vocabularies, schemas, and data models can all be classified as ontologies. McCuinness (2002) provides a useful definition of a *simple* ontology as a finite, controlled vocabulary providing unambiguous interpretations of classes and term relationships and strict hierarchical subclass relationships between classes. BIBFRAME and other standards developed for use in the bibliographic domain meet these criteria. Some can even be included among the more complex, *structured* ontologies by providing a minimum of constraints on classes and relationships. As outlined in Section 3.2.1, formal ontologies are included in a broad definition of metadata standards in this thesis.

---

<sup>11</sup> <http://www.w3.org/standards/semanticweb/ontology>

The notion of *shared* conceptualizations corresponds to what is often referred to as *mental models* in human–computer interaction and related research fields. Norman (2013, p. 25), a leading proponent of this tradition, defines mental models as “the conceptual models in people’s minds that represent their understanding of how things work,” and further points out that “people create mental models of themselves, others, the environment, and the things with which they interact.” Theories on mental models derive from psychology but since the 1940s have been used to support different interpretations in a number of fields (Westbrook, 2006). Like previous studies on users’ internal models of bibliographic structures (e.g., Pisanski & Žumer, 2010a), this thesis is based on Norman’s (2013) perspective on mental models. More specifically, such models are interpreted as conceptualizations, which can be formalized by ontologies.

Many have claimed that the new standards are insufficiently tested on users (see, e.g., Coyle, 2016; Pisanski & Žumer, 2010; Zhang & Salaba, 2009). These standards, which model conceptualizations of entities and relationships in the bibliographic universe, are first and foremost based on experts’ assumptions about users’ conceptualization of these entities and relationships. An example is the development of the original FRBR model. Olivia Madison (2005, p. 29), chair of the FRBR study group, explains:

One obvious option was to query, using a systematic methodology, a broad range of users and draw conclusions from this analysis. Another option, was to use our collective knowledge of the various types of users from the working group membership and commentators, as well as to draw upon experts in the fields to provide necessary user perspectives and conclusions. The Study Group decided in favor of the latter approach.

Coyle (2016, p. 109) comments on this by stating that “the study of user needs was done without studying users.” She further notes that “librarians are free to develop an expert meaning for the term [*Work*], but cannot expect that meaning to be shared perfectly with the others” (Coyle, 2016, p. 19). This thesis addresses the lack of user testing, as reported in paper B, not by user testing an existing standard, but by investigating user conceptualizations independent of existing expert based conceptualizations.

### 3.3 Transitions of Bibliographic Practices

Svenonius (2000, p. 64) argues that “the role of the bibliographic record in a digital environment is not yet clear. Especially unclear is what exactly a bibliographic record should describe.” She further claims that the bibliographic record has both served an *inventory* function and a *conceptual* or *information* function, before concluding: “It is hardly surprising that using one device to serve several functions should lead to trouble in times of technological change” (Svenonius, 2000, p. 64).

Throughout history, from ancient tablets to the (digital) infrastructure of the Web, the practices and purposes of bibliographic description have been adapted to innovations in technology. Such adaptations obviously involve trouble. The literature on the library community thoroughly discusses trouble arising from the meeting between traditions of bibliographic metadata and technologies of the current digital paradigm. The report *On the Record* (Library of Congress Working Group on the Future of Bibliographic Control, 2008, p. 24) states that the “library community's data carrier, MARC, is based on forty-year-old techniques for data management and is out of step with programming styles of today.” Others detail some challenges:

- The cataloguing code, AACR2, and the accompanying data format, MARC, which are widespread and still heavily used today, mirror the structure and logic of the human-readable card catalogue. Hence, the data are often represented as text strings difficult for computers to interpret and utilize (Coyle & Hillmann, 2007).
- Legacy standards represent early adoptions of non-analogue techniques of data handling but lack the functionality and innovations of digital technologies, such as relational databases and the Web (Thomale, 2010). They are further criticized for representing a “jungle of intertwined formats” (Andresen, 2004; Tennant, 2002), being too document centric (Alemu, Stevens, Ross, & Chandler, 2012), being inflexible (Gonzales, 2014), imposing replicate cataloguing (Yee, 2013), and building on a conceptual record structure that “affects the affordances and limitations of the data, especially in digital environments” (R. I. Clarke, 2015, p. 287).

Recent attempts and efforts to adapt library data to state-of-the-art technology have increasingly pointed towards Semantic Web technologies and Linked Data. Arguments for this approach are well documented in the literature. Linked Data, for example, is claimed to

- increase contextualization and discoverability (and serendipity) by providing connections to external data (Alemu et al., 2012; Lindquist, Dulock, Törnroos, Hyvönen, & Mäkelä, 2013; Schreur, 2012)
- help overcome the problems of “degrees of equivalency” and “round-tripability” by exploiting inferences on ontologies (by mapping and “dumbing down”-techniques) (Dunsire, Hillmann, & Phipps, 2012; Dunsire, Hillmann, Phipps, & Coyle, 2011)
- reduce cataloguing costs through shared and decentralized cataloguing (Gardašević, 2013; Tillett, 2013; Yee, 2013)
- break the “tyranny” of the flat record structure (Schreur, 2012)
- facilitate interoperability and federated search (Byrne & Goddard, 2010)

In 2011, the W3C’s Library Linked Data Incubator Group published a final report summarizing the potential benefits of Linked Data, most mentioned. The report, though, also 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” (Library Linked Data Incubator Group, 2011). Since then, a vast number of prominent actors, such as the OCLC, Library of Congress, and a variety of public and national libraries, have published significant amounts of bibliographic data in RDF format. A survey collecting data on existing Linked Data projects carried out by the OCLC in 2015 drew responses from 90 institutions in 20 countries (Smith-Yoshimura, 2016). Parallel to transformations and publishing of data, a number of standards are being developed top-down as ontologies in the formalization of RDF. In a primer, the Library of Congress presents its ambitions with BIBFRAME related to the wider public. A recurrent objective mentioned in the text relates to the need to make “interconnectedness commonplace” (Library of Congress, 2012, p. 4):

In short, the BIBFRAME model is the library community’s formal entry point for becoming part of a much larger web of data. [...] Focus will shift from capturing and recording descriptive details about library resources to identifying and establishing more relationships between and

among resources. This includes related resources found on the Web, and especially those beyond the traditional bounds of the library universe. These relationships—these links—drive the web, transforming the information space from many independent silos to a network graph that branches out in every direction.

The emphasis on outreach is shared by the bottom-up publishers of Linked Data, such as the OCLC, which declares that “it’s not enough to have pages ‘on the web’; library data must be ‘woven into the web’ and integrated into the sites and services that library users frequent daily—Google, Wikipedia, social networks.”<sup>12</sup> Paper A quotes similar claims found in documentation published by the national libraries that have published their data as Linked Data. Interpreting these statements in the light of the criticism of legacy practices and the arguments for Semantic Web and Linked Data solutions found in the literature reveals a common objective: to increase *metadata quality* and *interoperability* by liberating metadata from local data silos and merging them with data from other domains on the Web.

With the utilization of Semantic Web technologies in the library community, there are some tension between local (bottom-up) and global (top-down) approaches to transition. At the same time the national libraries have applied different ontologies for their Linked data sets, holistic ontologies, with unique naming of classes and properties (e.g., BIBFRAME), are being built from scratch. The latter approach is viewed by many as an incentive for an “exclusive ontology instantiation” (Vrandečić, 2010, p. 43), using only property names and class names from a single ontology. This practice does not conform with Linked Data principles recommending the reuse of existing ontology elements and the idea of RDF as a scheme-less OWA model (see, e.g., the BIBFRAME listserv-thread “Reuse (or Not) of Existing Ontologies”<sup>13</sup> and the blogpost “Of Records and RDF” (Brinxmat, 2015) for discussions on the topic). Nevertheless, the diversity in approaches represents interesting aspect of data modelling, which could have long-term effects on the ultimate goal of interoperability.

---

<sup>12</sup> <http://www.oclc.org/data.en.html>

<sup>13</sup> <http://listserv.loc.gov/cgi-bin/wa?A1=ind1303&L=bibframe>



### 3.4 Metadata Quality

Metadata quality, in its widest sense, is often defined as *fit for purpose* or *fitness for use* in accordance with relevant ISO standards and practical use cases (Guy, Powell, & Day, 2004; van Hooland, 2009; Wang & Strong, 1996). In a heavily cited paper, Wang and Strong (1996, p. 6) outline the essence of the fitness-for-use-perspective: “It emphasizes the importance of taking a consumer viewpoint of quality because ultimately it is the consumer who will judge whether or not a product is fit for use.” As mentioned, the formalization of purposes and user tasks has a long tradition in the library community, from Cutter’s (1904) “Rules for a Dictionary Catalog” to the IFLA’s *Statement of International Cataloguing Principles* (2016). In addition, the literature of the library community provides more detailed definitions, and a handful of the sources of these definitions are frequently cited.

Bruce and Hillman (2004) list seven characteristics of quality metadata<sup>14</sup>.

- Completeness
- Accuracy
- Provenance
- Conformance with expectations
- Logical consistency and coherence
- Timeliness
- Accessibility

Shreeves et al. (2005) merge Bruce and Hillman’s (2004) characteristics with the taxonomy of information quality (IQ) dimensions developed by Gasser and Stvilia (2001), Stvilia et al. (2004), and Stvilia et al. (2007). The IQ taxonomy consists of three dimensions (Shreeves et al., 2005, pp. 224–225):

---

<sup>14</sup> Hillmann and Phipps (2007) further extend this list with some concrete measurements based on the use of application profiles as described by Heery and Patel (2000).

- Intrinsic IQ: “includes dimensions that can be assessed by measuring internal attributes or characteristics of information in relation to some reference standard in a given culture.”  
Examples include spelling mistakes and standard conformance.
- Relational or contextual IQ: “measures relationships between information and some aspects of its usage in context”
- Reputational IQ: “measures the position of an information quantity in a cultural or activity structure”

Duval, Hodgins, Sutton, and Weibel (2002) list four principles “judged to be common to all domains of metadata,” which address interoperability issues, in particular<sup>15</sup>.

- Modularity: “In a modular metadata world, data elements from different schemas as well as vocabularies and other building blocks can be combined in a syntactically and semantically interoperable way.”
- Extensibility: “Metadata architectures must easily accommodate the notion of a base schema with additional elements that tailor a given application to local needs or domain-specific needs without unduly compromising the interoperability provided by the base schema.”
- Refinement: “There are two notions of refinement to consider. The first is the addition of qualifiers that refine or make more specific the meaning of an element. [...] For general interoperability purposes, the values of such elements can be thought of as subtypes of a broader element. A second variety of refinement involves the specification of particular schemes or value sets that define the range of values for a given element.”
- Multilingualism: “[...] unless such resources can be made available to users in their native languages, in appropriate character sets, and with metadata appropriate to management of the resources, the Web will fail to achieve its potential as a global information system”

---

<sup>15</sup> Nilsson, Johnston, Naeve, and Powell (2007) extend this list with the principle of *machine processability*.

The Linked Data community has applied similar quality criteria, which are detailed in paper A. For this PhD project, a comprehensive literature review of studies on Linked Data quality published between 2002 and 2012 conducted by Zaveri et al. (2015) proves useful. Zaveri et al. (2015) find 23 quality dimensions and group them as accessibility, intrinsic, trust, dataset dynamicity, contextual, and representational dimensions. 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, and its metrics include the re-use of existing vocabulary terms and dereferenced representation. Vocabulary usage and interlinking, as well as more generic quality aspects (e.g., interoperability, modularity, and extensibility), are explicitly included as quality aspects in the Linked Data study reported in paper A. Other aspects or notions, such as both intrinsic and relational quality perspectives and conformance with community standards, are implicitly considered in both the Linked Data study (paper A) and the other studies.

## 4 Academic Affiliation and Theoretical Perspectives

The PhD project examines the qualities of bibliographic metadata in light of emerging standards and practices highly influenced by the principles of Semantic Web and Linked Data. At its core, Semantic Web and Linked Data are about connecting data across heterogeneous domains to enable computers to understand the *meaning* of data. Hence, the PhD project is related to semantics in a data context and, more specifically, to an interoperability perspective of semantics, or *shared* meaning (Shreeves et al., 2005; Studer, Benjamins, & Fensel, 1998). In addition, the study on users in paper B is concerned with semantic aspects of the conceptualizations of metadata structures.

Hjørland (2007, p. 367) argues that semantic issues in general “underlie all research questions” in *library and information science* (LIS), especially the subfield of *knowledge organization* (KO). Broughton, Hansson, Hjørland, and López-Huertas (2005, p. 133) define KO as the handling of “knowledge organizing systems (KOS) such as bibliographical records, classification systems (e.g., DDC, LCC, and UDC), thesauri, semantic networks,” while Clarke (2009) emphasizes that semantic interoperability has become a key challenge and requirement of such systems. Other definitions of KO—for, example, as an overarching label collecting different traditions (Hjørland, 2008)—are only vaguely including bibliographic descriptions and (descriptive) metadata as part of the field. Smiraglia (2014, p. 66), however, states that “metadata for resource description are considered to play a role in knowledge organization when they are used to provide order to a set of such descriptions.” He relates the notion of order to the theoretical discussions and organizational devices, such as Wilson’s (1968) coining of the terms “descriptive and exploitative control,” and Svenonius’s (2000) suggestion of a set theoretical approach to the categorization of bibliographic entities. These theoretical “yardsticks,” are, as outlined in the background section, essential for the research conducted in this thesis. Furthermore, based on the premise concerning the conceptual ordering role, Smiraglia (2014, p. 69) states that “metadata schemas” (explicitly including RDF, which is at the core of Linked Data principles) clearly represent a form of KO. The current PhD project leans toward this understanding of metadata and KO.

Crotty (1998, p. 3) defines a theoretical perspective as the “philosophical stance informing the methodology and thus providing a context for the purpose for the [research] process and grounding its logic and criteria.” From a historical perspective, KO can be interpreted as a *metatheoretical* perspective in itself: a “science of science.” Dahlberg (2006) points to early 1900s books by Henry Evelyn Bliss on the general classification of the sciences, such as *The Organization of Knowledge and the System of the Science* (1929), as the inspiration for the initial naming of the field.<sup>16</sup> However, as a subfield of LIS, KO research can be interpreted from a variety of theoretical perspectives.

Frohmann (1992, p. 365) describes the theoretical debates in LIS as “waged as a confrontation between rival epistemological positions, each claiming to provide the most fruitful theoretical foundation for knowledge production in a contested field.” A diversity of theoretical perspectives is also present in KO research. One such perspective is described by Bates (2005, p. 9) as the “constructionist or discourse-analytic approach [...] in which it is assumed that the discourse of a society predominately conditions the responses of individuals within that society, including the social understanding of information.” To this tradition belongs Campbell’s (2007) Foucauldian reading of the Semantic Web, comparing the transition from classification to clinical diagnosis in medicine (as argued upon by Foucault, 1973) with the transition from rigorous classification and cataloguing rules to the flexible Semantic Web in libraries. Radford’s (2003) use of classification systems to exemplify *discursive formations* is another constructionist example. This thesis’s view on the Semantic Web can be interpreted from a similar perspective as a study on (historical) change and disruptions, for example, from the ideas of UBC to OWA.

Nevertheless, while Hjørland (2007) claims that every problem in KO can be related to semantic and thus to linguistics issues, the semantic problem area does not necessarily bind a constructionist or any language-aware theoretical perspective. While the constructionist

---

<sup>16</sup> Smiraglia (2006, p. 8) distinguishes between the professional and the scientific origins of KO: “knowledge organization used to be called classification, and classification has primarily played a distinct role in science and librarianship. In science, classification is the primary product of research, providing terms and their definition [...]. In librarianship, of course, classification is used to render the subject content of documents and to enhance information retrieval.”

approaches are interesting, the research questions of the current project point in another direction, towards another kind of knowledge: they do not ask “does this artifact or phenomenon represent a discursive formation?”, but rather “do they work?”

Floridi’s (2005, 2010) semantic definition of information provides a theoretical model of the examined problem area:

$\sigma$  is an instance of information, understood as semantic content, if and only if:

$\sigma$  consists of  $n$  data, for  $n \geq 1$ ;

the data are *well* formed;

the well-formed data are meaningful.

Floridi further emphasizes that being well formed is facilitated by a rule or syntax and that well-formed data are meaningful in the contexts of systems. In the current PhD project, bibliographic data and structures are examined in the contexts of particular rules of well formedness supported by standards and Linked Data principles. The ontologies are developed with the intention to increase semantic interoperability, or meaningful data in the context of data consumers.

Accordingly, do the standards work? Are the data they describe meaningful?<sup>17</sup>

Bates (2005, p. 12) relates such questions to the engineering tradition of LIS “in which it is assumed that humans needs and usage of information can best be accommodated by successive development and testing of indigenous systems and devices to improve information retrieval and services.” The term *design science* refers to a similar, complementary tradition in *information systems*, a neighboring research field of LIS. According to Hevner, March, Park, and Ram (2004, p. 77), design science “creates and evaluates IT artifacts intended to solve identified organizational problems.” This PhD project is related to a rather practical problem area encountered by cultural heritage institutions and information professionals organizing data with the overarching objective to “facilitate knowledge creation” (Lankes, 2011, p. 15). Information

---

<sup>17</sup> For an interesting, further discussion of Floridi’s overarching philosophy of information and its potential as a metatheory in library and information science, see Bawden and Robinson (2018). Floridi’s semantic model is also further discussed in paper E.

technology (IT) artifacts are, in the context of a design science, broadly defined as “constructs (vocabulary and symbols), models (abstractions and representations), methods (algorithms and practices, and instantiations (implemented prototypes systems)” (Hevner et al., 2004, p. 77). The research efforts presented in this thesis relate to IT artifacts in most of these senses.

To summarize, the PhD project is situated within LIS and KO, approaching organizational challenges in the bibliographic universe with theoretical perspectives regarding motivations and purposes similar to those of the design science tradition. As in most professional fields, LIS and KO are based upon a heterogeneous knowledge base (Grimen, 2008). This implies that methods and theoretical perspectives from different areas of knowledge and scientific research are often applied. In addition, the project situated within KO builds on and refers to state-of-the-art research in a variety of subfields of computer science.

## 5 Previous Research

The two most recent articles of this thesis (papers A and B) cover previous research in detail. This introductory section, therefore, summarizes and briefly outlines the main research designs and results applied in the relevant fields. Notable research related to the research efforts reported in the other papers and conducted after their publication is also mentioned. The purpose of this outline (and the previous research sections in the papers) is to provide an overview of relevant empirical research. This implies that the numerous papers and documented efforts related to development projects conducted in libraries mostly are left out.

### 5.1 Research on Metadata Quality

Park (2009) examines the current state of research and practice on metadata quality in digital environments and argues that the most common metrics utilized are *completeness*, *accuracy* (or *correctness*), and *consistency*. The main methodological approach is to count the use of metadata elements (e.g., Dushay & Hillmann, 2003; Greenberg, Pattuelli, Parsia, & Robertson, 2001; Park, 2006; Shreeves et al., 2005; Ward, 2003; Yasser, 2012). Hillmann (2008) gives examples of the typical research question of such counting studies:

- Which metadata fields are present?
- What percentage of the total number of records is in each field?
- How consistent are the metadata within those fields?
- What patterns can be detected?

Yasser (2011) compares and analyzes reported problems in metadata records in a selection of studies in this tradition and reports the five most common errors:

- Incorrect values
- Incorrect elements
- Missing information
- Information loss
- Inconsistent value representation

Wisser (2014) studies errors found in headings from MARC records and confirms some of these categories, especially incorrect and inconsistent value representation. Such counting studies are



important to detect common errors and ineffective metadata practices. The problems summarized by Yasser (2011) can also be directly related to some heterogeneity conflicts mentioned in section 3.2.2, such as naming conflicts. Wisser (2014) also relates metadata quality to challenges in the interoperability and aggregation of data.

The effort to examine Linked Data quality in this thesis (paper A) counts metadata occurrences in environments where aggregation and interoperability are essential objectives. In addition, the experimental transformation of MARC data reported in paper D reveals severe accounts of inconsistent legacy data. These studies thus can be related to the counting tradition of assessing metadata quality. However, many studies in this tradition primarily address instance data (Ochoa & Duval, 2009), not bibliographic structures at the standard or semantic level, which is the main focus of the current project.

## 5.2 Research on Metadata Transformations and Linked Bibliographic Data Quality

Zaveri et al. (2015) review 30 articles relating to Linked Data quality published from 2002 to 2014. They find that many of these apply quantitative design by counting occurrences, in line with the tradition described above: “we notice that most of the metrics take the form of a ratio, which measures the occurrences of observed entities out of the occurrence of the desired entities” (Zaveri et al., 2015, p. 23). Some studies include but seldom highlight or directly address sets of bibliographic Linked Data. Two prominent examples of such studies are conducted by Hogan et al. (2012) and Schmachtenberg, Bizer, and Paulheim (2014). These studies assess and compare Linked Data sets at a global level by analyzing huge amounts of interlinked data obtained from curating sources, such as Data Hub,<sup>18</sup> with specialized crawlers. They use conformance with Linked Data principles as an assessment criterion, for example, determining whether the Linked Data sets reuse existing vocabularies and to what extent they interlink to external sources. Both of these studies expose quality issues. The study reported in paper A, which assesses only bibliographic Linked Data, partly builds on the metrics and quantitative design of the mentioned studies. However, instead of just counting used vocabularies and external interlinking targets,

---

<sup>18</sup> <https://datahub.io/>

paper A studies the composition and bibliographic structures these vocabularies and interlinks represent.

Kontokostas et al. (2014) suggest a test-driven evaluation framework focusing on the accuracy and consistency of Linked Data and include bibliographic data published by Biblioteca Nacional de España (BNB) and the Library of Congress in a proof-of-concept test. This evaluation exposes violations of the use of ontology restriction (*domain* and *range* in particular). Papadakis, Kyprianos, and Stefanidakis (2015) and Hallo, Luján-Mora, Maté, and Trujillo (2015) examine interlinks, URIs, and vocabularies used in bibliographic Linked Data sets. Neither study includes detailed statistical analysis of applied practices, as the current thesis does. However, by summarizing reported problems related to Linked Data from a selection of writings, the latter study (Hallo et al., 2015) partially updates and confirms knowledge on the overall landscape of metadata transitions in the library domain covered in paper C. Among the most prominent problems mentioned is interoperability issues due to the application of “too many vocabularies for the same metadata” and “mapping problems; for example, not all basic relations of FRBR could be extracted from MARC records” (Hallo et al., 2015, p. 125). Another contribution to the discussion of Linked Data challenges in the library domain is provided by McKenna (2017), who emphasizes that Linked Data transformation and publication are technically demanding operations requiring competences that must be increased among librarians in the sector.

Regarding experimental transformations and interlinking, much research happened simultaneous with and after the efforts reported in Paper D. The reported experiment in the paper were performed in a controlled environment to test certain technical frameworks and to identify challenges in the Linked Data publishing process. A number of general frameworks and methods have been developed by different research communities. For example, *Karma*, a system developed at the University of Southern California, is a graphical interface that takes different data sources as input to transform them into an RDF graph according to the user’s chosen ontology (Knoblock et al., 2012; Szekely et al., 2013). The European Union-funded LOD2

project,<sup>19</sup> which ran from 2010-2014, introduced a set of tools for both RDF publishing and interlinking. *Silk*,<sup>20</sup> developed by a team at the University of Mannheim, is a tool that generates RDF links by taking into account the structures surrounding data instances in a manner similar to the procedure proposed for interlinking bibliographic data in paper D. Rajabi, Sicilia, and Sanchez-Alonso (2014, p. 646) evaluate and compare several interlinking tools, including Silk, and conclude “that using an interlinking tool is an effective way of linking between two datasets or from a data collection.”

### 5.3 Research on User Conceptualizations of Bibliographic Entities

As described in Paper B, studies on user conceptualizations of bibliographic entities primarily verify entities from existing standards, especially FRBR entities. These studies are mainly of two types: those examining users’ interactions with such entities in interfaces of information systems and those simulating interactions with bibliographic entities. The first type includes the work by Zhang and Salaba (2009), examining how users succeed in performing different tasks in three FRBR-inspired catalogs, and Merčun, Žumer, and Aalberg (2016, 2017), comparing user interactions in a non-FRBRized system and FRBRized prototype system. An example of the second type is research conducted by Pisanski and Žumer (2010, 2010, 2012) combining different methods, such as card sorting and concept mapping, to verify FRBR entities. All of these studies, to some extent, report tendencies of user preferences for FRBR entities and structures.

Unlike previous research, the participants in the study reported in paper B are not presented with existing solutions or bibliographic records but, instead, are asked to conceptualize bibliographic families based on their own understandings of the documents’ important characteristics. This seems to represent a novel approach in this research field. Another novel feature of this study is the focus on derivative relationships.

---

<sup>19</sup> <http://lod2.eu/Welcome.html>

<sup>20</sup> <http://silkframework.org/>

## 6 Study design

### 6.1 Overall Considerations

The research questions are investigated using different methods, from experimental development via statistical evaluation and case studies to user studies based on concept mapping. The analytical approaches are mostly quantitative. The user study requires a close reading of detailed concept maps, which are ultimately analyzed through a cluster analysis. The analysis of published Linked Data sets includes a case study, but its results are interpreted in light of the statistical features of the complete data sets. Beyond their quantitative characteristics, these methods are quite varied, primarily because they are applied to investigate strongly related but different research objects, requiring a tailored approach. Although the research objects originate from the same problem area and can be related to common problems, such as interoperability issues, the different methods do not perform triangulation (Miller, 1997) in the sense that they investigate common objects from different standpoints. The varied research objects analyzed in the project rather contribute different perspectives on a coherent but diverse problem area: metadata structures in the bibliographic universe.

The selection and application of methods is largely based on relevant previous research. These methods are described in the papers, but the following section provides brief summaries and some clarifications of the initial choices related to the study designs and operationalizations of the research questions. The summaries, in certain cases, have some descriptions not included in the papers.

### 6.2 Examining Challenges in Metadata Transformations

To gain knowledge about the main challenges characterizing the transition from legacy standards and legacy metadata to Linked Data-based metadata frameworks, two approaches were chosen: an exploratory literature review (reported in paper C) and an experimental case study (reported in paper D). In the exploratory literature review, prominent writings reporting research findings, theoretical discussions, and documentation of experimental metadata transformations were collected and analyzed. Writings were included based on their relevance in a bibliographic

context and their relation to defined phases in a transformation cycle. The different phases of such a cycle are identified in existing literature (e.g., Hyland, 2010) and confirmed by the experimental study transforming MARC data into RDF, reported in paper D.

The case study (paper D) set up a realistic scenario with the goals to transform MARC data describing musical recordings into a Linked-Data-conformant data set and to link the transformed resources to corresponding resources in the MusicBrainz data set.<sup>21</sup> For the experiment, 99 MARC records from the Norwegian national discography<sup>22</sup> describing albums in the black metal genre were collected. The black metal genre was chosen due to its complex structures between bands and musicians and the musicians' extensive use of pseudonyms. These characteristics ensured sufficiently challenging source material. These metadata records were also assessed to represent typical metadata describing cultural products with a high level of international outreach. The experiment included the design of a Linked-Data-conformant target ontology for the MARC data and the development of an interlinking procedure, inspired by Raimond, Sutton, and Sandler (2008), utilizing information found in the graph surrounding the interlinking candidates. The challenges arising during this process are reported in paper D and mentioned in Section 7.1.

The technical aspects of the interlinking procedure are only briefly described in paper D, so more details are provided in the following. The procedure was aimed at interlinking corresponding artists in two data sets. One set represented a source set to be published as Linked Data, while the other set was an already-published target set. The procedure took artist names in the source set as a starting point. They were retrieved with a SPARQL query adjusted to the relevant structure and semantics. In addition, the SPARQL query retrieved a list of titles related to those artists in the source set. Then, a second SPARQL query retrieved matching artist names and track titles from the target set. As outlined in the paper, an exact match for both artist names and track titles was required, but fuzzy matching meeting a certain threshold of proximity could also be applied. The final calculation of the similarity between artists thus was based on two components: the potential

---

<sup>21</sup> An in-depth analysis of metadata structures of MusicBrainz (<https://musicbrainz.org/>) and how they relate to traditional bibliographic data is provided by Dahl, Knutsen, and Tallerås (2012).

<sup>22</sup> [http://nabo.nb.no/trip?\\_t=1&\\_b=baser&navn=nordisko](http://nabo.nb.no/trip?_t=1&_b=baser&navn=nordisko)

name match and the matching of tracks. The similarity in tracks was defined as the number of matched tracks across the sets over the total of tracks identified in the source set. A fixed weight ( $\lambda$ ) between 0 and 1 was included to enable tuning of the balance between the two components. In practice, if  $\lambda$  is set to 0.3, this will be the weight given to the artist-name component, and following from that 0.7 ( $1 - \lambda$ ) is set as the weight of the track-title component. This is somewhat imprecisely illustrated in the flowchart and formula of paper D.

A proof-of-concept system showing promising results were built, but not tested sufficiently. A significant limitation of the interlinking effort was thus the lack of a systematic test of the procedure, for example, regarding a fruitful weighting between the components. Consequently, this contribution is primarily conceptual. However, similar frameworks taking advantage of graph structures, as mentioned, have proven to be effective (Rajabi et al., 2014).

### 6.3 Evaluating the Quality of Bibliographic Linked Data

Paper A reports on a comprehensive study on the quality of bibliographic Linked Data published by four prominent European national libraries. Data dumps of RDF graphs were downloaded from February to April in 2016. The study was mainly based on statistical methods and measurements found in or inspired by previous research (Hogan et al., 2012; Schmachtenberg et al., 2014). It was not straightforward to treat RDF data statistically, so the applied statistical measurements are described quite thoroughly in paper A. In addition, case studies of limited subgraphs from each data set were examined and compared to get a more detailed impression of the present structures and qualities.

Paper A presents two research question that can be regarded as operationalizations of the first main research question defined for the overall PhD project (as stated in section 2.3):

- 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 Data sets conform to established measurements of Linked Data quality for vocabulary usage and interlinking?

The selection of data sets included in the study was based on the following set of criteria:

1. The sets must be directly available in their entirety as a data dump.
2. The sets must contain bibliographical data and, as a minimum, provide information about authors and their intellectual products (*Author sets* in the terminology of Svenonius, 2000, p. 43).
3. The bibliographical information must have a global character in the sense that it contains factual data likely to be a potential interlinking candidate for external Linked Data publishers.
4. The sets must be official publications, not results from mere experimental case studies.
5. The sets must have been published by a library institution.
6. The sets must have been updated in 2015 or 2016.

These criteria were chosen based on the following methodological considerations:

- The data must be comparable due to a concise, consistent methodology (criteria 1, 2, and 3).
- The data must have comparable content (criteria 2 and 3).
- The data should represent typical state-of-the-art Linked Data from the library community (criteria 3, 4, 5, and 6).

The four data sets examined met these criteria and were considered to provide an adequate picture of the landscape of available bibliographic Linked Data in 2016. These sets were published by the Bibliothèque Nationale de France (BNF), British Library (BNB), BNE, and Deutsche Nationalbibliothek (DNB).

As described in paper B, a validity test of the sets was performed using SPARQL to construct subgraphs of author sets related to Nobel literature laureates. The SPARQL queries took the *Virtual International Authority File*<sup>23</sup> (VIAF) URIs of those authors collected from Wikidata as a

---

<sup>23</sup> <https://viaf.org/>

neutral starting point. Then, a generic RDF model was applied to the SPARQL query to construct a subgraph consisting of the neighboring nodes of the VIAF URIs in each set. To avoid overloading the subgraph, the generic RDF models in two cases were (slightly) adjusted to set specific graph structures. The composition of the subgraphs was eventually compared to the composition of the overall graphs and found to more or less match. This indicates that any of the full data sets could have contained significant amounts of data not directly related to the bibliographic entities, possibly skewing the comparative perspectives of the analysis. This technique was also used to construct samples for the additional case study of one author set.

The study design provided insights into the compositions and nature of the data sets, thereby helping to reveal quality issues in a variety of aspects. To investigate the causes of the identified issues, other complementary methods, such as qualitative interviews with people involved in the transformations, should be applied in further studies. The case study was helpful to provide more insight into the concrete bibliographic structures and shortcomings, but more samples are needed to get a more complete picture.

#### 6.4 Elicitation of User Conceptualizations

Gaining knowledge about how users conceptualize entities and relationships in the bibliographic universe, the second main research interest of the PhD project, required a proper method for the elicitation of such conceptualization. Additionally, getting a clear idea of such a phenomenon from a single study targeting a complex problem area necessitated limiting the scope. The study on user conceptualization reported in paper B of this thesis, therefore, was limited to the question of “how do users conceptualize *derivative relationships* between entities in the bibliographic universe?” The reason for focusing on derivative relationships was twofold. First, these relationships are interesting in a media landscape increasingly characterized by adaptations and transmedial storytelling (Jenkins, 2006; Vukadin, 2014). Second, derivative relationships represent an under-researched area of user-oriented studies.

The study used concept mapping as a method to elicit conceptualizations, which has proven to be a fruitful method in previous research in similar problem areas and other domains (e.g., Chang, 2007; Pisanski & Žumer, 2010a). A novel approach of the current study was that it did not follow



the attempts by previous studies to verify conceptualizations represented in existing standards. While such studies have contributed valuable insights, this study was instead designed to elicit conceptualizations as independent of such standards as possible. Accordingly, the primary objective was not to answer the question of whether established standards resemble user conceptualization but rather the opposite to gain knowledge of user conceptualizations that can inform their further developments. In the study, 107 participants were asked to map their conceptualizations based on handouts representing real documents. The involved documents, participants, and the mapping process are described in detail in paper B.

To analyze the resulting maps, two researchers interpreted them in two iterations. The first iteration provided an overview of the maps' common characteristics, such as the main nodes and the relationships between them. In the second iteration, the relationships between the document nodes were encoded in a spreadsheet as present or absent based on specific criteria described in paper B. This yielded a matrix with binary data used as input in a cluster analysis.

According to Kaufman and Rousseeuw (2009), cluster analysis is the art of finding groups in data with the aim to identify structures already present. A cluster analysis takes two input structures: a matrix with objects and their attributes and a collection of proximities for all the pairs of objects. The analysis requires one operation to calculate the proximities and another to cluster the results. These operations are dependent on the variable types in the input matrix. In this study, the input matrix contained symmetric binary data. The well-known, *simple matching coefficient* (Sokal & Michener, 1958) was used to calculate proximities. The cluster analysis applied the *average linkage method* (from the *hclust* package in R).<sup>24</sup> The result was visualized as a dendrogram (Figure 3).

---

<sup>24</sup> <https://stat.ethz.ch/R-manual/R-devel/library/stats/html/hclust.htm>

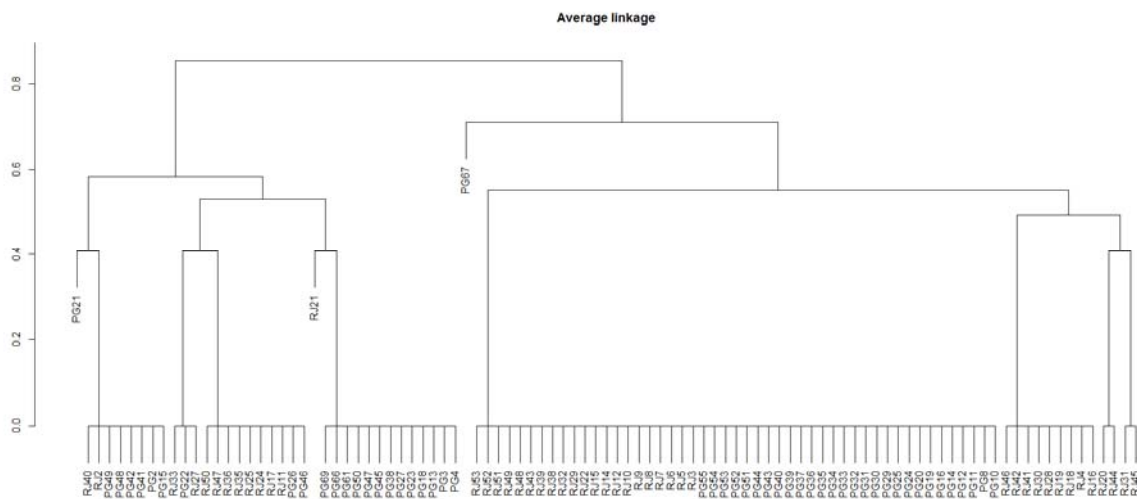


Figure 3. The resulting dendrogram from the cluster analysis.

The study design has some limitations, as discussed in the paper, primarily due to the number of document types tested and the composition of participants. Although further studies are needed to confirm and gain more knowledge, the study design proved to be beneficial for the elicitation and analysis of user conceptualizations.

## 6.5 Ethical considerations

Regarding the study of users reported in paper B, all participants remained anonymous throughout the study. Participants put finished drawings and questionnaire answers in closed envelopes without any personal information attached. The Linked Data sets reported in paper A were all licensed in a way that allows for reuse and experimentation.

## 7 Summary of the Main Findings

### 7.1 Challenges of Metadata Transformations

The main finding from the experimental transformation process (paper D) is the identification of some quality issues in the legacy data used as input. These data are MARC records describing black metal records harvested from the Norwegian national discography. The quality issues can be related to the typical criteria summarized in sections 3.1 and 5.1, such as missing information, accuracy, and inconsistencies. Due to the extensive use of pseudonyms in the black metal genre, the artists are registered under widely different names throughout the data sample. The tracks are also registered differently; some records leave them out, others mention them only in notes, and those records containing tracks properly registered at an instance level apply different MARC fields to do so. Figure 4 shows three records illustrating these inconsistencies.

Record A	Record B	Record C
=110\$aMayhem	=110\$aMayhem	=110\$aMayhem
=24510\$aMediolanum capta est	=24510\$aGrand declaration of war	=24510\$aDeathcrush
=700 0\$aManiac	=700 1\$aManiac	=700 1\$aManiac
=700 0\$aBlasphemer	=700 1\$aBlasphemer	=700 1\$aMessiah
=700 0\$aHellhammer	=700 1\$aHellhammer	=700 1\$aAvnskog, Erik
=700 0\$aNecrobutcher	=700 1\$aNecrobutcher	=700 1\$aButcher, Necro
=710 0\$aMayhem	=700 1\$aFinstad, Børge	=700 1\$aAarseth, Øystein
=7400 \$aCarnage	=700 0\$aManiac\$tTo Daimonion	=700 0\$aSchnitzler, Conrad\$tSilvester Anfang
=7400 \$aNecrolust	=700 0\$aManiac\$tA time to die	=710 0\$aMayhem
=7400 \$aDeathcrush	=700 0\$aManiac\$tView from Nihil	=710 0\$aMayhem\$tNecrolust
=7400 \$aAncient skin	=700 0\$aBlasphemer\$tTo Daimonion	=710 0\$aMayhem\$tDeathcrush
=7400 \$aFreezing moon	=700 0\$aBlasphemer\$tA time to die	=710 0\$aVenom\$tWitching hour
=7400 \$aFall of seraphs	=700 0\$aBlasphemer\$tView from Nihil	=710 0\$aMayhem\$tWeird Manheim
=7400 \$aSilvester Anfang	=700 0\$aManiac\$tA grand declaration of war	=710 0\$aMayhem\$tChainsaw gutsfuck
=7400 \$aChainsaw gutsfuck	=700 0\$aManiac\$tA bloodsword and a colder sun	=710 0\$aMayhem\$tPure fucking armageddon
=7400 \$aFrom the dark past	=700 0\$aBlasphemer\$tA grand declaration of war	
=7400 \$aI am thy labyrinth	=700 0\$aManiac\$tIn the lies where upon you lay	
=7400 \$aSymbols of bloodswords	=700 0\$aManiac\$tCompletion in science of agony	
=7400 \$aPure fucking Armageddon	=700 0\$aBlasphemer\$tA bloodsword and a colder sun	
=900 0\$aEriksen, Rune\$tBlasphemer	=700 0\$aBlasphemer\$tIn the lies where upon you lay	
=900 0\$aStubberud, Jørn\$tNecrobutcher	=700 0\$aManiac\$tCrystalized pain in deconstruction	
=900 0\$aKristiansen, Sven-Erik\$tManiac	=700 0\$aBlasphemer\$tCompletion in science of agony	
=900 0\$aBlomberg, Jan Axel\$tHellhammer	=700 0\$aBlasphemer\$tCrystalized pain indeconstruction	
	=710 0\$aMayhem	
	=900 1\$aNecro\$tNecrobutcher	

Figure 4. Three MARC records illustrating the inconsistent registration of artist names and tracks.

The inconsistencies of the input data further affect the quality of the output RDF data. A more sophisticated matching operation could increase the output quality to a certain level. However, a transformation process targeting an RDF-based output model that both encompasses another data structure (graphs) and is based on an extensive use of identifiers would be problematic if the

input data are not of sufficient quality. The interlinking procedure reported in paper D illustrates that the graph structure of RDF data can be used to qualify the matching of bibliographic data entities.

## 7.2 Quality of Bibliographic Linked Data

The study on the quality of bibliographic Linked Data reported in Paper A shows that the four data sets examined conform to essential Linked Data principles, such as using W3C standards, applying widely adopted vocabularies, and providing links to external data sets. These sets also compared well with results from other studies (Hogan et al., 2012; Schmachtenberg et al., 2014). For example, they have fewer external links than the “top linkers” worldwide but are still among the sets with the most links. Figure 5 shows the network of external links from each set. Thus, the quality of the Linked Data, assessed as isolated sets, was quite high on a general level.

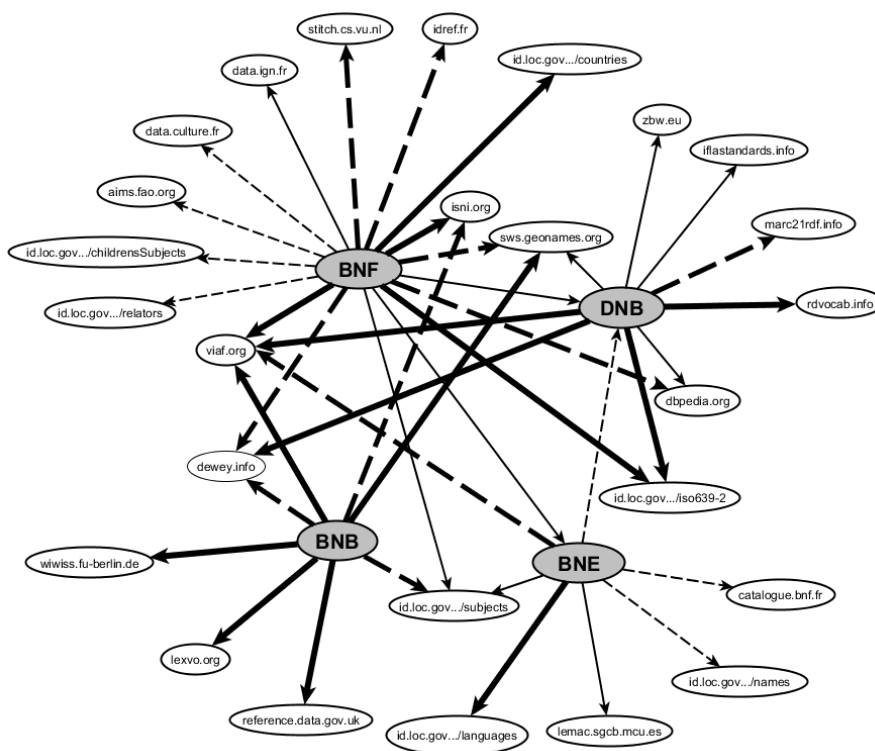


Figure 5. The four corpus sets and the external data sets targeted by their external RDF links. The thickness of the lines indicates the amount of links between the sets.

These sets represent bibliographic data published by institutions sharing similar mandates and objectives and furthermore belong to the same metadata traditions, so it is also interesting to compare the bibliographic structures on a more detailed level. This part of the study reveals highly different practices regarding interlinking targets (Figure 5), vocabulary usage, and bibliographic entities. Of the interlinking targets, only one (viaf.org) is shared by all sets. Of the 28 targets identified across the sets, only eight are shared by more than two sets. Regarding vocabulary usage, only three of 1,141 unique property and class terms are used by the four publishers (`owl:sameAs`, `rdf:type`, and `dct:language`). Thirteen terms are shared by three sets, and 34 by two sets. BNE and BNF include W/E/M entities from the FRBR model but implement them differently. Figure 6 shows how such entities are implemented across the sets.

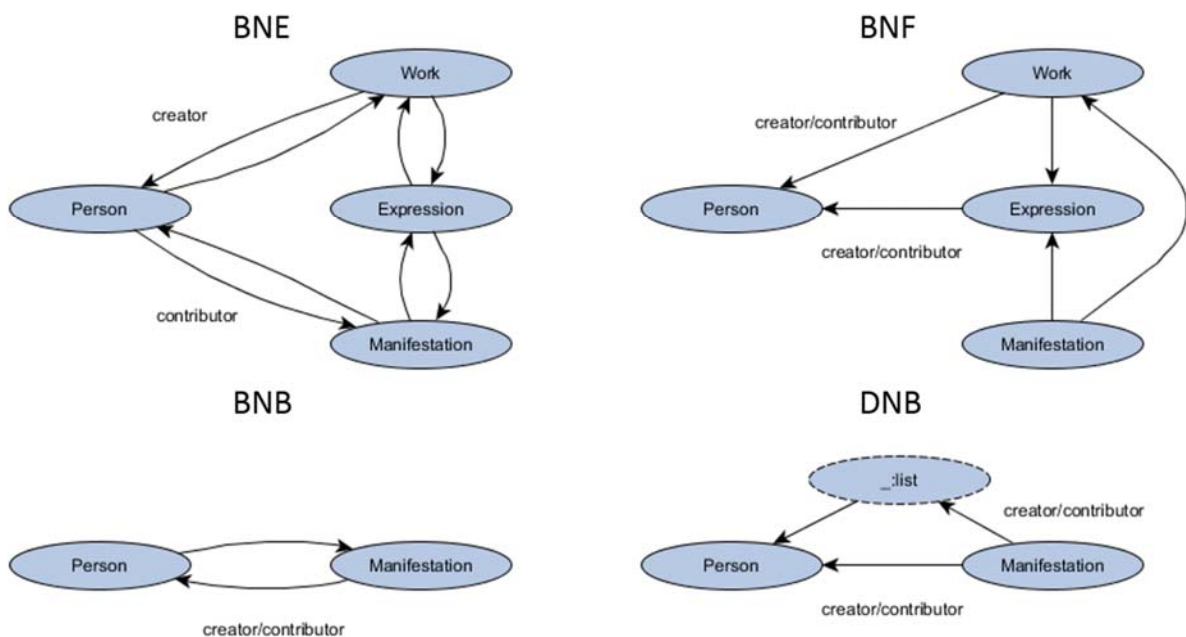


Figure 6. W/E/M entities implemented in the four corpus sets.

The diversity of vocabulary usage and the structural levels in the implementation of multi-entity models introduce some interoperability challenges to the potential integration of data between the sets.

Another main finding of this study is the presence of significant quality issues regarding the completeness of the instantiations of the FRBR entities. Of the work entities included in the BNE set, only 13% are related to an expression, and only 14% of the manifestation entities are connected to expressions. A similar problem of isolated entities on both sides of the “FRBR chain” is identified in the BNF set. As discussed in Section 8, the incompleteness of entities and relationships between them is most likely due to challenges in transforming legacy data but nonetheless is a characteristic of the data sets that risks not meeting the expectations of many potential data consumers.

### 7.3 User Conceptualizations of Derivative Relationships

In the study on user conceptualizations reported in paper B, the participants were asked to draw networks of nodes and relationships representing their conceptualizations of three related documents and their relationships. The participants were given documents belonging to different document families. Half of the participants were asked to conceptualize documents related to Ibsen’s play *Peer Gynt*, and the other half documents related to William Shakespeare’s play *Romeo and Juliet*. This process resulted in 98 concepts maps considered to be sufficiently interpretable. The following cluster analysis of these concepts maps resulted in two main groups with five subclusters. The cluster analysis was performed on a matrix comprising six variables, which were the relationships between the four main nodes identified in an initial analysis of the drawings; one node for each of the three documents; and a central node (Figure 7).

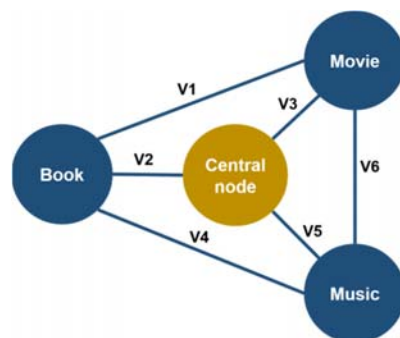


Figure 7. The main nodes identified in the concept maps and their relationships treated as binary variables (present/not present) in the cluster analysis.

The five subclusters proved to represent different ways of relating the documents. The participants in cluster A related the documents only through arbitrary links, for example, via shared characteristics, such as nodes representing common topics. The participants in clusters B and C directly related the documents, whereas the participants in clusters D and E applied a central node to collocate the documents. Clusters B, C, D, and E grouped the participants more or less according to the document families they conceptualized, and the differences between their concepts maps can be related to the specific characteristics of those documents. The cluster analysis was followed by an examination of the node names and attributes, which confirmed the identified groupings. The participants in clusters A, B, and C tended to prefer document-oriented naming of nodes and document-oriented attributes, whereas the participants in clusters D and E provided information on a higher abstraction level, such as data on characters and original languages. Based on these findings, the clusters were placed on a spectrum, as shown in Figure 8. The left side of the spectrum represents a document-oriented, single-entity approach, and the right side represents a multi-entity approach involving some form of abstraction.

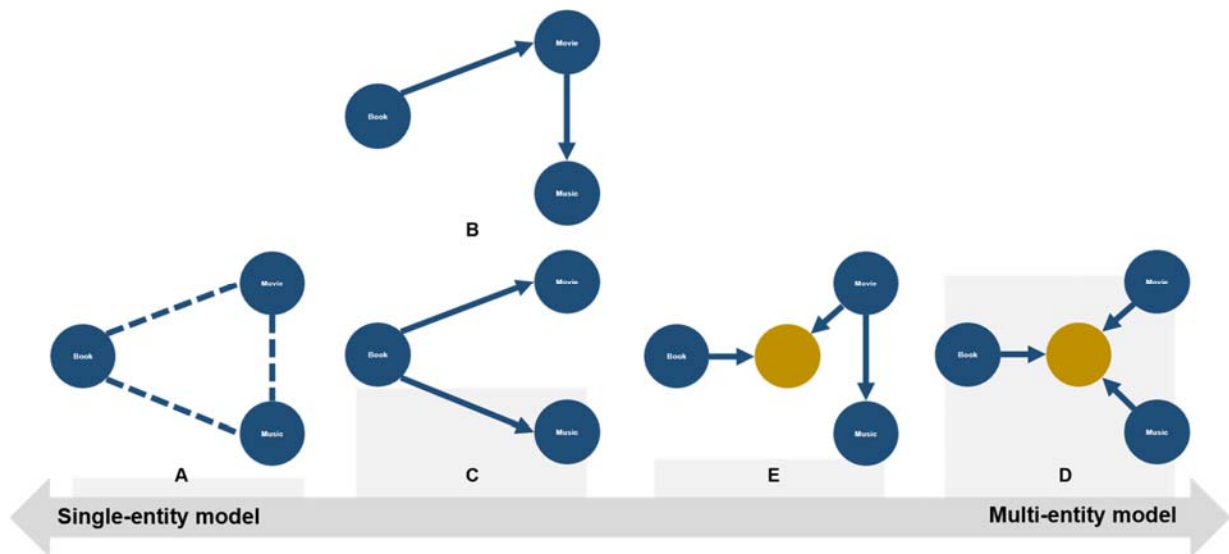


Figure 8. The five clusters along a spectrum from single-entity to multi-entity conceptualizations. The bars in the background indicate the number of conceptualizations distributed across the spectrum.

The main findings from the study on user conceptualization are thus twofold. At a detailed level, the participants provided great variations in their conceptualization, in terms of both structures and attributes. On a higher level, the conceptualization can be split into two main groups: one

whose participants applied a multi-entity model involving abstract relationships between documents and one with participants who directly related the documents. As discussed in paper B, the abstraction level of the first group can be interpreted as tendencies towards FRBR works as described in the LRM (Riva et al., 2017), a superwork as described by Svenonius (2000), and a fictional world as described in the literature on transmedial storytelling (Vukadin, 2014).



## 8 Discussion

The introductory section states that this PhD project adopts both a verification perspective and a validation perspective on bibliographic metadata structures. The verification perspective asks whether something has been built correctly, whereas the validation perspective asks whether the correct thing has been built (Gómez-Pérez, 2004). The first perspective relates to technical criteria, and the second perspective to fitness for use. The main objective of most standards and principles is to bring these perspectives together: if something is built correctly, according to specific standards or principles, it is believed to increase fitness for use. Correct cataloging according to a given cataloging code, for example, is believed to enable users to solve a set of given tasks in interactions with catalogs (assuming that the catalogs make the most of the cataloged metadata). These perspectives can be related to a pair of metadata quality dimensions (mentioned in section 3.4), coined by Shreeves et al. (2005, pp. 224–225): intrinsic quality “can be assessed by measuring attributes of information items themselves in relation to a reference standard,” whereas relational quality “depends on relationships between the information and some aspect of its usage context.” As described in the thesis papers and in this introductory, the reorganization of bibliographic data according to Linked Data principles is aimed at both increasing the fitness for use of well-established ideas of user needs and meeting new ones. The following discussion begins by addressing the intrinsic technical perspective of this process and then moves to the relational validation perspective.

In the library sector, significant efforts have been put into building things correctly, both building new standards and following the principles of existing ones, such as the Linked Data principles. The BIBFRAME standard, intended to replace the long-lasting MARC format, was released as a “Linked Data model” and a “Bibliographic framework for the Web of data” (Library of Congress, 2012). Since 2011, the *Bibliographic Framework Transition Initiative Forum* has received thousands of contributions from its email list.<sup>25</sup> Some discuss use-case scenarios and questions related to fitness for use, but many also engage in quite detailed discussions of how BIBFRAME

---

<sup>25</sup> <http://www.lsoft.com/scripts/wl.exe?SL1=BIBFRAME&H=LISTSERV.LOC.GOV>

should be built correctly, for example, according to RDF formalisms or the implementation of FRBR entities. While BIBFRAME and other emerging Linked-Data-oriented standards are built top down, others have built things locally in a bottom-up manner, as in the case of the four European national libraries that have published Linked Data on the Web according to their own standards.

These institutions are quite homogenous. They organize knowledge in the bibliographic universe, and as national libraries, they share mandates and objectives. In their common metadata domain, something new has long been awaited. The need for the MARC format “to die”, for example, was pronounced sixteen years ago (Tennant, 2002). The examination of the transformation of data finally taking place in these institutions consequently offers interesting insights into the desired bibliographic structures. That these efforts were conducted in parallel makes for a comprehensive and substantial research object. The PhD project shows that the institutions have succeeded in building correct Linked Data. As reported in paper A, one set could perhaps use fewer blank nodes (an RDF mechanism not recommended for Linked Data) (Mallea, Arenas, Hogan, & Polleres, 2011), and another could apply more external vocabularies, but in general, they conform to the Linked Data principles.

In the analogue world, bibliographic data have served the purpose of providing access to physical documents. As the metadata describing these documents has become a digital and online phenomenon, the possibility to connect them to other metadata, both within the bibliographic communities and outside targets, has gained interest. Such connections are dependent on interoperability. While the Internet and its protocols have provided an essential infrastructure for technical interoperability, the next step is to exchange *meaningful* data. The Semantic Web technologies and the Linked Data principles are concrete means proposed to meet this challenge. These proposed means have largely been embraced by the library community. The BnF (2018) website documenting its Linked Data effort describes a broad specter of interoperability gains:

We can optimize dissemination and reuse of data produced by the BnF, by pushing them out of our internal silos and giving them an enhanced audience and visibility on the Web. Potential usages are various and innovative. Other libraries can now not only retrieve data from the BnF

but also create links to it. Moreover, data is bound to get out of the library world in order to be broadly widespread.

These objectives are expressed similarly by other libraries publishing Linked Data, as documented in paper A, showing that interoperability has become a desired use case in its own right.

The effort to make data accessible has undoubtedly some positive effects, particularly increased visibility and presence on the Web and contributions to the transparency of “public action” (BNF, 2018). Although the PhD project finds that the libraries are successful in publishing useable data with respect to such high-level goals, it also indicates that it might take more than conformance with general Linked Data principles to achieve sufficient levels of semantic interoperability. The implementations conducted by the four libraries prove to be widely different, and the results suffer from significant data-level quality issues. Some issues, such as duplications of person entities, can be solved by a manageable data-cleansing process. Others, such as completeness and missing links between FRBR entities, are harder to overcome.

The missing links between works and manifestations in the BNE and BNF sets can most likely be explained by their chosen strategies for the transformation process. Legacy records can contain limited description of works, for example, titles listed as part of a complete work or songs on a record—a “component-to-whole relationship” (Riva et al., 2017, p. 72). Following a given transformation procedure, these titles can be transformed into independent work entities without any relationships to expression or manifestation entities. Expressions or manifestations are not evidenced in the records. There can also be records that, according to a given transformation rule, lack proper evidence (for example the presence of an original title), to generate a work entity, resulting in further independent manifestation entities.

Even with interoperability as the central fitness-for-use criteria, incompleteness is not necessarily problematic in Linked Data publication in the context of Semantic Web and the OWA. The idea is that the Semantic Web will gradually be extended in a distributed manner without any requirement for completeness in individual data sets. The whole is believed to become greater than the sum of its parts. This is also an intriguing idea for bibliographic data. Table 1, from

paper A, shows that person entities (based on an analysis of present VIAF identifiers) overlap across the four examined sets only to an inconsiderable degree. This implies that the complementary potential is very interesting.

<b>Set combinations</b>	<b>Overlap</b>
<i>BNF-BNB</i>	12.7 %
<i>BNF-DNB</i>	6.5 %
<i>BNF-BNE</i>	5.6 %
<i>DNB-BNB</i>	4.3 %
<i>BNB-BNE</i>	2.6 %
<i>DNB-BNE</i>	1.1 %
<i>BNF-BNE-BNB-DNB</i>	0.2 %

*Table 1. Overlapping VIAF entities limited to person entities and owl:sameAs links in different set combinations and between all sets.*

The OWA nonetheless would risk failure in many use-case scenarios based on the CWA. Individual data sets based on inventory data from libraries can provide proper answers to some questions regarding a specific collection (e.g., which manifestations are included in the collection?). Sets based on data from national bibliographies can provide lists of published manifestations with given authors. The published data sets, however, cannot meet probable user expectations of completeness at the work level. It, therefore, would be hard to use the work entity to build a novel app or service on top of these data, perhaps an IMDb for books, from which users would expect a high level of completeness. Bruce and Hillman (2004, pp. 5–6) list both completeness and user expectations as central quality criteria for the application of metadata standards: “it does little good to prescribe a particular element set if most of the elements are never used, or if their use cannot be relied upon across the entire collection. [...] They should not contain false promises, i.e., elements that are not likely to be used because they are superfluous, irrelevant, or impossible to implement.”

The complementary aspect may help meet this challenge but is furthermore dependent on a sufficient level of semantic interoperability between the sets. The four examined libraries have chosen to implement FRBR entities in widely different ways. The BNE and BNF provide W/E/M entities, while DNB and BNB only provide manifestations, and they all use different vocabularies or vocabulary elements to do so. A metadata expert, with in-depth knowledge about bibliographic

structures may be able to map some of these entities across the sets. Data consumers from other domains, however, would have a hard time doing so.

Another main finding of the PhD project is that the participants asked to conceptualize bibliographic structures did it quite differently. Paper B describes how the participants took different forms of document-oriented and abstract approaches to derivative relationships. Such relationships constitute an increasingly large share of the bibliographic universe. The seemingly ever-expanding superhero universes in the contemporary popular culture is a telling example as narratives and characters are adapted and further developed across a variety of digital and analog platforms. Classical works, such as Shakespeare's *Romeo and Juliet*, continuously accumulate adaptations, expanding their bibliographic families (Smiraglia & Leazer, 1999). Furthermore, derivative relationships can be linked to Wilson's notion of exploitative control, which prescribes the formalization and utilization of relationships between bibliographic entities beyond mere descriptive characteristics. To exercise exploitative control, in order to facilitate discovery and to improve user experience, represents a significant challenge in knowledge organization.

The PhD project does not directly evaluate how user conceptualizations correspond to conceptualizations built into existing standards. The main finding, based on a single study involving a limited number of documents (but a significant number of participants), is the great variation in the identified conceptualizations. The categorization of the participants on a higher level into two main groups by their concrete expression of derivative relationships, however, is interesting in light of existing structures. These groups comprise those applying multi-entity models involving abstraction and those directly relating documents. Multi-entity models are a long-awaited feature of bibliographic catalogs. The findings presented in paper B can be interpreted resembling a FRBR work but higher levels of abstraction like a superwork or even a contextualizing construction that organizes the narrative universe to which the documents relate. FRBR entities make up the essential components in several existing standards (LRM, FRBRoo, BIBFRAME and RDA), and they are implemented in the application profiles used by the BNE and BNF for Linked Data publishing. A superwork collocation can be expressed through the *F15 complex work* element in FRBRoo (Working Group on FRBR/CRM Dialogue, 2016, p. 26). It, however, is more challenging to formalize relationships between entities in narrative universes.

The findings indicate that the formalization of these relationships should be considered in the further development of existing standards. The underlying varieties of conceptualizations also indicate that a key challenge is implementing concrete formalizations of relationships in user interfaces. This is where the exploitative control ultimately needs to be applied.

All the research efforts in the PhD project can be related to the standardization of bibliographic metadata structures, whether concerning dependencies for standardization or evaluations of the use of standards. Svenonius (2000, pp. 80–81) mentions three dangers regarding standardization:

- One danger of excessive standardization is conflict with the principles of user convenience.
- A second danger of standardization carried too far is that the reasons and principles underlying a bibliographic code become obscured.
- A third danger of overstandardization is its tendency to inhibit change.

Regarding user convenience, the study illustrates at least three challenges in today's practice. First, publication of Linked Data without the necessary amount of metadata quality may struggle when facing a variety of use-case scenarios, especially those based on expectations of completeness. Second, the study shows that users might have highly varied conceptualizations of bibliographic structures. Third, derivative relationships in existing standards need more work.

Considering the risk that standards obscure underlying purposes, it is not unreasonable to assert that this has happened to metadata regimes in libraries. The current metadata regime in many countries (including Norway) is still a combination of the MARC format and the AACR. These standards developed based on the card catalogue are applied in a completely different digital context today. Something “new” has become a necessity. However, it should also be noted that the danger of obscurity may apply to newer metadata regimes. This thesis shows that the Linked Data principles alone do not necessarily provide the desired effects, such as increasing interoperability. To some extent, one might argue that at least some reasons for this are to be found in the legacy data, and that this regime works better for new descriptions. In that case, the library community must invest in the improvements of the vast amounts of existing data describing our common cultural heritage. Otherwise, we risk what Suominen and Hyvönen (2017) refer to as moving “from MARC silos to Linked Data silos,” transforming a multitude of

data that communicate neither internally within a homogenous domain nor externally with the outside world.

Regarding change, a key question is whether to embrace new, more flexible solutions for metadata organization, such as application profiles, or to stick with the traditional universal solution based on a set of shared standards (UBC). The latter seems to have a strong foothold. Both BIBFRAME, aimed at replacing the MARC format, and RDA, which has replaced AACR in many bibliographic communities, maintain continuity with the UBC perspective in their holistic approach to solving many needs at the same time. Svensson (2013, p. 12), with DNB, argues for a common model, emphasizing the importance of interoperability:

In order to replace the current records-based model with one that allows library information to be reused in other settings and also allows libraries to make better use of data originating outside of the library domain, it is necessary to agree on a common model that reduces the complexity of that data integration.

The variety of ways of publishing bibliographic Linked Data revealed in this thesis may ultimately indicate the need for different approaches. An examination of Linked Data projects conducted across different domains (including the initial work on BNE's publication of Linked Data) shows that there is no one-size-fits-all formula (Villazón-Terrazas et al., 2012). Each domain represents a set of data types, data formats, data models, licensing contexts, and languages, forming individual problem areas. This may also be the case within the bibliographic landscape. Although most of the Linked Data publishers studied in this thesis (paper A) express an overall interoperability objective, they may have utilized Linked Data to fit local purposes and use cases. Dunsire, Hillmann and Phipps (2012, p. 164) have proposed that Linked Data represent a flexible continuation of UBC, in that it allows both to preserve "local granularity, semantic focus, context, and the data itself" by applying an RDF version of the preferred local standard and to assimilate the data into a Web-scale environment by the use of links. This requires an extensive mapping of semantics between source and target standards. The PhD project shows that this infrastructure needs more work (and attention) in order to live up to its potentiality. Considering the examined sets of bibliographic Linked Data as a whole, it will be challenging to link up valuable data such as the W/E/M entities, to create a connected Web of data. The

examination also show that these entities only to a limited extent is linked to corresponding resources in external sets.

The study findings thus point out unresolved issues within knowledge organization. Linked Data principles, which have been the driving force behind recent efforts to reorganize a long-standing metadata regime, have not yet provided sufficiently useful results. Linked Data principles may point in the right direction, but more research and experimental work are needed to solve challenges related to data quality and semantic interoperability. The long-standing metadata regime also builds on some fixed ideas of user needs, which likely have become moving targets in the digital world with an increasingly contentious and complex bibliographic universe. The PhD project does not examine information needs but shows that achieving exploratory control over complex relationships is not a straightforward undertaking.

As described in the paper E, metadata and tools for knowledge organization have become key components of new methods for mediating information, from Big-Data-based AI to recommendation algorithms and modern information-retrieval systems. The findings of the PhD project thus should not only be interesting in a strict bibliographic sense but also relevant to any operation in which metadata and knowledge organization are critical for business.



## 9 Conclusions and Further Research

The aforementioned OCLC survey collecting information about Linked Data projects in library institutions registered 112 projects, of which “most are primarily experimental in nature” (Smith-Yoshimura, 2016). Some projects have been presented and discussed at conferences and workshops and documented in reports and papers, but few have reported detailed information about the methods used for transformation or have been systematically evaluated. The PhD project contributes knowledge that can inform the road ahead for this experimental field. The two first research questions of the PhD project are formulated as:

1. What are the main challenges in transforming bibliographic metadata following Linked Data principles?
2. What qualities characterize bibliographic metadata published as Linked Data?

The main findings suggest that the challenging factors in the transition to a Linked-Data-based metadata regime are the inconsistencies in the legacy data and the proper instantiation of the chosen target standards. The main challenge is to choose and build target standards that both enable a variety of potential fitness-for-use scenarios and ensure interoperability at a global level.

A recurrent claim is that the standards used in this field lack proper user testing. Accordingly, the third research question is formulated as:

3. How do users conceptualize entities and relationships in the bibliographic universe?

The thesis includes a study that operationalizes this question by examining how the participants conceptualize derivative relationships. The findings from this study contribute knowledge that can be used to further develop emerging and existing standards. The elicited conceptualizations among the participants show great variations but also tendencies toward two main groupings, of which the largest include some form of abstraction in the collocation of related documents.

Shreeves et al. (2005, p. 231) ask whether “quality metadata is shareable data?” and conclude that more “research is needed to understand the trajectory of metadata as it travels from the initial

design of the cataloging workflow to its use in a federated collection.” This PhD project shows that this question and the conclusion have become even more relevant in the context of ongoing Linked Data efforts. This project continues existing work on metadata quality and user conceptualizations and suggests further research on several topics. The project identifies issues regarding inconsistencies in legacy data that present a major challenge to the successful transition to new metadata regimes. More research on how transformations better can handle such inconsistencies is needed. As more and more texts are born digital or being digitized, research on how such digital texts can be interpreted automatically and used as input and evidence in the transformation processes is warranted. One of the most beneficial areas to investigate is the instantiation of FRBR entities to increase both completeness and semantic accuracy. As new standards are developed and implemented, more research to deepen the understanding of reasons for the existing inconsistencies would also be beneficial. Do the new metadata regimes result in better data quality, or do they perpetuate old causes?

The library community traditionally has been concerned with building things correctly and perhaps has paid less attention to continuously evolving fitness-for-use criteria. More research should examine information needs and conceptualizations related to bibliographic metadata structures. The study design in the thesis proved to work well for elicitations and could be used in further studies including more documents. More studies on how users interact with metadata structures in operating bibliographic systems are also needed to gain more knowledge that can help inform the core business of libraries.

The projects relate to both system and user perspectives. The important overall questions concern the relationship between these perspectives: what are real user needs, and how can we build something that assist them? The project contributes knowledge that can support some answers and illustrates the need for more knowledge. These questions, therefore, should be the driving force for both future research and practical efforts of the library community.

## 10 Paper Summaries

The following are brief overviews of each article included in the thesis. For the full papers, see Appendix I–VI.

### A. Quality of Linked Bibliographic Data: The Models, Vocabularies, and Links of Data Sets Published by Four National Libraries

Paper A examines different quality aspects of bibliographic data published on the Web according to Linked Data principles (Berners-Lee, 2006; Hyland et al., 2014). The analysis considers and compares data sets published by four prominent European national libraries, the BNF, BNB, BNE, and DNB. All the data sets were published as RDF graphs applying both locally minted and widely adopted external vocabularies. The most recent data dumps of these graphs were downloaded and ingested into a triple store on a local server. The main analyses were performed with SPARQL queries providing comparable statistical data about RDF composition, vocabulary usage, and interlinking. The metrics used in the analysis were based on previous research on Linked Data quality (Hogan et al., 2012; Schmachtenberg et al., 2014). In addition, a limited sample of data, a subgraph containing triples describing Bob Dylan and his fictional novel *Tarantula* was constructed from each data set and studied to gain more detailed insight into the chosen structures and semantics.

The analyses showed that all data sets conformed to the main Linked Data principles by being published on the Web applying W3C standards and by providing a relative high number of links to external data sets (compared to studies of interlinking in other data sets). It also showed that the data sets applied high-quality existing vocabularies. The publishers, however, also used a high number of locally minted vocabularies. The implementation of vocabularies, abstraction levels of bibliographic entities, such as FRBR entities, and interlinking targets differed widely across the data sets, decreasing the level of interoperability between them. The case study also revealed some inconsistencies and other quality issues at the instance level.

## B. User Conceptualizations of Derivative Relationships in the Bibliographic Universe

Paper B takes the lack of user testing of emerging bibliographic models as the starting point and motivation for a study on user conceptualizations. The study follows up on previous research in this area (especially Pisanski & Žumer, 2010a, 2010b) but uses other methods and a more model independent approach. In this study, 107 library students at the beginning of undergraduate studies were asked to draw concept maps depicting their conceptualizations of the attributes of and relationships between three related documents. The participants (50/50) were presented with two equivalent but different document families containing a play (Ibsen's *Peer Gynt* and Shakespeare's *Romeo and Juliet*), a movie adaptation of the play, and two types of CDs compiling music related to the plays (*Peer Gynt Suites* by Grieg and Halvorsen inspired by Ibsen's play and the soundtrack to the movie adaptation of *Romeo and Juliet*).

Analysis of 98 (interpretable) concept maps resulted in a spreadsheet containing binary data on the presence or absence of the main entities and the relationships between them. The data were further examined statistically with cluster analysis based on standard methods for binary data. The cluster analysis resulted in five groups. Two groups (accounting for 63% of the participants) utilized a central node to collocate all three or two of the documents, one group (9%) applied only arbitrary relationships, and two groups (28%) directly related the document nodes. These groupings were further reinforced by the participants' use of attributes and name labels for the nodes.

The study viewed the central node as a multi-entity model involving a form of abstraction of derivative relationships. Based on the attributes and naming practices, different levels of abstraction levels were identified and discussed, but further research with other document types is needed to capture more knowledge. The results of this current study, however, showed tendencies of both FRBR works as described in LRM (Riva et al., 2017), a superwork as described by Svenonius (2000) and a fictional world as described in the literature on transmedial storytelling (Vukadin, 2014). A main finding of the study was the participants' differing approaches at a more detailed level.

### C. From Many Records to One Graph: Heterogeneity Conflicts in the Linked Data Restructuring Cycle

Paper C is aimed at identifying conflicts and challenges in the process of transforming legacy data organized as records into RDF graphs on the Semantic Web, conforming Linked Data principles. The reported effort is an exploratory literature review providing an overview of the concepts, insights, and discussions found in relevant writings. The first sections of the paper explain and present different views on essential concepts related to the ongoing transformation processes in the library domain, such as RDF and graph-based networks, Semantic Web, Linked Data principles, and semantic interoperability. In the latter sections of the paper, conflicts and challenges are grouped and presented according to phases in the Linked Data restructuring cycle. The different phases of such cycle are identified in the existing literature (e.g., Hyland, 2010) and confirmed by an experimental study transforming MARC data into RDF graphs reported in paper D of this thesis.

Specific challenges are identified for each phase. Some prominent examples are the characteristics of legacy data and the potentially divergent use cases for the transformed data. Existing data are produced in line with standards mandating data structures and data types challenging to adapt to Linked Data principles. Legacy data typically consist of textual descriptions rather than machine-readable data. Regarding the conversion phase, the paper discusses literature that reports that the textual characteristics of source data are problematic in the conversion of formats demanding extensive use of authoritative, machine-readable identifiers.

The crucial part of the modelling phase is the selection of a target model for the data. A particular challenge touched upon is the potentially conflicting use cases that the new model should enable. Another challenge discussed in the literature is to use models that reduce loss from legacy data. Regarding interlinking, it has been claimed (e.g., Halpin, Hayes, McCusker, McGuinness, & Thompson, 2010) that Linked Data publishing suffers an identity crisis due to heterogeneity conflicts, such as discrepancies in the interpretation of flexible ontologies.

#### D. Ordo ad Chaos—Linking Norwegian Black Metal

The paper reports an experimental effort to facilitate the publication of bibliographic data on the Web-conforming Linked Data principles. The main objective of the experiment is to gain knowledge about the challenges arising during the publishing process. The experiment involves and emphasizes the transformation of legacy MARC records and an automated interlinking procedure. The concrete case is to publish data describing a selection of black metal artists and recordings from Norway's national discography and to interlink the artists to their corresponding aliases in the comprehensive MusicBrainz database.

Initially, 99 records in the NORMARC (ISO2709) format were harvested and transformed into MARCXML. A simple target ontology based on metadata elements from the Music Ontology, Dublin Core, and FOAF was developed. Then the MARCXML data were transformed with XSLT into RDF according to this ontology. This process revealed inconsistency issues in the legacy data in both the names found in the headings and the structural features. Due to the extensive use of pseudonyms in the black metal genre, the artists were registered under widely different names throughout the data sample. Tracks were also registered differently; some records left them out, others mentioned them only in notes, and those records containing tracks properly registered at an instance level applied different MARC fields to do so.

The interlinking of the data was inspired by Raimond, Sutton, and Sandler (2008), who proposed taking the surrounding graph of a linking candidate into consideration. A conceptual procedure using the titles of tracks related to an artist as qualifiers to identify and match that given artist was outlined. The flexible procedure could be tuned to the thresholds for string matching (e.g., by considering the edit distances between the titles and names) and the decisive balance between the artist (the linking candidate) and the tracks (the qualifiers found in the surrounding graph).

## E. Mediation Machines: How Principles from Traditional Knowledge Organization Have Evolved into Digital Mediation Systems

The paper discusses digital information systems' ability to mediate cultural resources. Mediation techniques embedded in search and recommendation systems are compared with those activities developed for mediating culture heritage in libraries, archives, and museums (LAM institutions). Mediation is an essential task in all kinds of libraries, archives, and museums holding collections of resources made available to audiences. Library-science programs train librarians to become skilled intermediaries, able to analyze information and recreational needs and to connect these needs to relevant resources. Meanwhile, Google (n.d.) claims that its "mission is to organize the world's information and make it universally accessible and useful." Its vast suite of search systems, databases, and other services has helped it come far toward realizing this goal. Does this mean that Google is a mediator, performing the same kind of mediation as LAM institutions?

This question is investigated through a literature review and a discussion of central topics. Digital mediation systems is found to follow many principles and techniques of traditional knowledge organization, such as those related to classification and metadata. Furthermore, they mimic librarians who know their users, knowledge organization systems, and collections. An important challenge is the mechanical rationality embedded in the computation of recommendations, which may limit users' exposure of materials of interest the system finds irrelevant.

## References

- Abele, A., McCrae, J. P., Buitelaar, P., Jentsch, A., & Cyganiak, R. (2017). The Linking Open Data cloud diagram. Retrieved March 6, 2018, from <http://lod-cloud.net/>
- Alemu, G., Stevens, B., Ross, P., & Chandler, J. (2012). Linked Data for libraries: Benefits of a conceptual shift from library-specific record structures to RDF-based data models. *New Library World*, 113(11/12), 549–570. <https://doi.org/10.1108/03074801211282920>
- Andresen, L. (2004). After MARC: What then? *Library Hi Tech*, 22(1), 41–43. <https://doi.org/https://doi.org/10.1108/07378830410524486>
- Avram, H. D. (1975). *MARC, its history and implications*. Washington: Library of Congress.
- Baker, T., Coyle, K., & Petiya, S. (2014). Multi-entity models of resource description in the Semantic Web. *Library Hi Tech*, 32(4), 562–582. <https://doi.org/10.1108/LHT-08-2014-0081>
- Bates, M. J. (2005). An introduction to metatheories, theories, and models. In K. E. Fisher, S. Erdelez, & L. McKechnie (Eds.), *Theories of information behavior* (pp. 1–24). Medford, NJ: Information Today.
- Bawden, D., & Robinson, L. (2018). Curating the infosphere: Luciano Floridi's Philosophy of Information as the foundation for library and information science. *Journal of Documentation*, 74(1), 2–17. <https://doi.org/10.1108/JD-07-2017-0096>
- Bergman, M. K. (2009). The Open World Assumption: Elephant in the Room. Retrieved from <http://www.mkbergman.com/852/the-open-world-assumption-elephant-in-the-room/>
- Berners-Lee, T. (2006). *Linked data: Design issues*. W3C. Retrieved from <http://www.w3.org/DesignIssues/LinkedData.html>
- Berners-Lee, T. (2007). Giant Global Graph. Retrieved from <http://dig.csail.mit.edu/breadcrumbs/node/215>
- Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American*, 284(5), 34–43. <https://doi.org/10.1038/scientificamerican0501-34>
- Bibliothèque nationale de France. (2018). About data.bnf.fr. Retrieved from <http://data.bnf.fr/en/about>



- Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked Data: The story so far. *International Journal on Semantic Web and Information Systems*, 5(3), 1–22.  
<https://doi.org/10.4018/jswis.2009081901>
- Bliss, H. E. (1929). *Organization of knowledge and the system of the sciences*. New York: Henry Holt and Company.
- Brinxmat. (2015). Of records and RDF. Retrieved from  
<https://brinxmat.wordpress.com/2015/05/01/of-records-and-rdf/>
- Broughton, V., Hansson, J., Hjørland, B., & López-Huertas, M. J. (2005). Knowledge organisation: Report of working group 7. In L. Kajberg & L. L. (Eds.), *European Curriculum Reflections on Education in Library and Information Science*. København: Royal School of Library and Information Science.
- Bruce, T. R., & Hillmann, D. I. (2004). The continuum of metadata quality: Defining, expressing, exploiting. In D. I. Hillmann & E. L. Westbrook (Eds.), *Metadata in practice* (pp. 203–222). Chicago, IL: American Library Association.
- Buckland, M. K. (1997). What is a “document”? *Journal of the American Society for Information Science*, 48(9), 804–809. [https://doi.org/10.1002/\(SICI\)1097-4571\(199709\)48:9<804::AID-ASI5>3.0.CO;2-V](https://doi.org/10.1002/(SICI)1097-4571(199709)48:9<804::AID-ASI5>3.0.CO;2-V)
- Byrne, G., & Goddard, L. (2010). The strongest link: Libraries and Linked Data. *D-Lib Magazine*, 16(11). <https://doi.org/doi:10.1045/november2010-byrne>
- Campbell, D. (2007). The birth of the new web: A foucauldian reading of the semantic web. *Cataloging & Classification Quarterly*, 43(3–4), 9–20.  
[https://doi.org/https://doi.org/10.1300/J104v43n03\\_01](https://doi.org/https://doi.org/10.1300/J104v43n03_01)
- Carlyle, A. (1997). Fulfilling the second objective in the online catalog: Schemes for organizing author and work records into usable displays. *Library Resources & Technical Services*, 41(2), 79–100. <https://doi.org/10.5860/lrts.41n2.79>
- Carlyle, A. (1999). User categorisation of works: Toward improved organisation of online catalogue displays. *Journal of Documentation*, 55(2), 184–208.  
<https://doi.org/10.1108/EUM0000000007143>
- Chan, L. M., & Zeng, M. L. (2006). Metadata interoperability and standardization: A study of methodology, Part I. *D-Lib Magazine*, 12(6). <https://doi.org/doi:10.1045/june2006-chan>

- Chang, S.-N. (2007). Externalising students' mental models through concept maps. *Journal of Biological Education*, 41(3), 107–112. <https://doi.org/10.1080/00219266.2007.9656078>
- Clarke, R. I. (2015). Breaking records: The history of bibliographic records and their influence in conceptualizing bibliographic data. *Cataloging & Classification Quarterly*, 53(3–4), 286–302. <https://doi.org/10.1080/01639374.2014.960988>
- Clarke, S. G. D. (2009). Knowledge organization system standards. In M. J. Bates & M. N. Maack (Eds.), *Encyclopedia of Library and Information Sciences, Third Edition* (pp. 3164–3175). Boca Raton: CRC Press. <https://doi.org/doi:10.1081/E-ELIS3-120044538>
- Coyle, K. (2016). *FRBR, before and after: A look at our bibliographic models*. Chicago: American Library Association.
- Coyle, K. (2017). Two FRBRs, many relationships. Retrieved from <http://kcoyle.blogspot.no/2017/05/two-frbrs-many-relationships.html>
- Coyle, K., & Hillmann, D. I. (2007). Resource Description and Access (RDA). *D-Lib Magazine*, 13(1/2). <https://doi.org/10.1045/january2007-coyle>
- Crotty, M. (1998). *The foundations of social research*. London, UK: Sage.
- Cygniak, R., Wood, D., & Lanthaler, M. (2014). *RDF 1.1 Concepts and abstract Syntax. W3C Recommendation*. Retrieved from <https://www.w3.org/TR/rdf11-concepts/>
- Dahl, T. A., Knutsen, U., & Tallerås, K. (2012). Mellom tradisjonen og weben: Katalogisering, metadata og bibliotekarutdanning. In R. Audunson (Ed.), *Krysspeilinger: Perspektiver på bibliotek- og informasjonsvitenskap* (pp. 141–163). Oslo: ABM-Media.
- Dahlberg, I. (2006). Knowledge organization: A new science? *Knowledge Organization*, 33(1), 11–19.
- Direktoratet for forvaltning og ikt. (2016). Interoperabilitet: Overordna arkitekturprinsipp. Retrieved from <https://www.difi.no/fagomrader-og-tjenester/digitalisering-og-samordning/nasjonal-arkitektur/prinsipper/interoperabilitet-overordna-arkitekturprinsipp>
- Dunsire, G., Hillmann, D. I., & Phipps, J. (2012). Reconsidering universal bibliographic control in light of the Semantic Web. *Journal of Library Metadata*, 12(2–3), 164–176. <https://doi.org/10.1080/19386389.2012.699831>

- Dunsire, G., Hillmann, D. I., Phipps, J., & Coyle, K. (2011). A reconsideration of mapping in a semantic world. In *International Conference on Dublin Core and Metadata applications* (pp. 26–36). Retrieved from <http://dcpapers.dublincore.org/pubs/article/view/3622/1848>
- Dushay, N., & Hillmann, D. I. (2003). Analyzing metadata for effective use and re-use. In *International Conference on Dublin Core and Metadata applications* (pp. 161–170). Retrieved from <http://dcpapers.dublincore.org/pubs/article/view/744>
- Duval, E., Hodgins, W., Sutton, S., & Weibel, S. L. (2002). Metadata principles and practicalities. *D-Lib Magazine*, 8(4).
- Floridi, L. (2005). Is semantic information meaningful data? *Philosophy and Phenomenological Research*, 70(2), 351–370. <https://doi.org/10.1111/j.1933-1592.2005.tb00531.x>
- Floridi, L. (2010). *Information: A very short introduction*. New York, NY: Oxford University Press.
- Foucault, M. (1973). *The birth of the clinic: An archaeology of medical perception*. New York, NY: Pantheon.
- Frohmann, B. (1992). The power of images: A discourse analysis of the cognitive viewpoint. *Journal of Documentation*, 48(4), 365–386.
- García-Castro, R., & Gómez-Pérez, A. (2011). Perspectives in semantic interoperability. In *Proceedings of the International Workshop on Semantic Interoperability (ICAART 2011)* (pp. 13–22). <https://doi.org/10.5220/0003346700130022>
- Gardašević, S. (2013). Semantic web and Linked (Open) Data possibilities and prospects for libraries. *INFOtheca - Journal of Informatics & Librarianship*, 14(1), 26–36.
- Gasser, L., & Stvilia, B. (2001). A new framework for information quality. *Urbana Champaign: University of Illinois at Urbana Champaign*.
- Glushko, R. J. (2013). *The discipline of organizing*. Cambridge, MA: The MIT Press.
- Godby, C. J. (2013). *The relationship between BIBFRAME and OCLC's Linked Data model of bibliographic description: A working Paper*. Dublin, Ohio: OCLC Research. Retrieved from <http://oclc.org/content/dam/research/publications/library/2013/2013-05.pdf>

- Godby, C. J., & Denenberg, R. (2015). *Common ground: Exploring compatibilities between the Linked Data models of the Library of Congress and OCLC*. Dublin, Ohio: Library of Congress & OCLC Research. Retrieved from <https://www.oclc.org/content/dam/research/publications/2015/oclcresearch-loc-linked-data-2015-a4.pdf>
- Gómez-Pérez, A. (2004). Ontology evaluation. In S. Staab & R. Studer (Eds.), *Handbook on ontologies* (pp. 251–273). Berlin: Springer. [https://doi.org/10.1007/978-3-540-24750-0\\_13](https://doi.org/10.1007/978-3-540-24750-0_13)
- Gonzales, B. M. (2014). Linking Libraries to the Web: Linked Data and the Future of the Bibliographic Record. *Information Technology and Libraries*, 33(4), 10–22. <https://doi.org/10.6017/ital.v33i4.5631>
- Google. (n.d.). Our company. Retrieved from <https://www.google.com/about/our-company/>
- Greenberg, J. (2005). Understanding metadata and metadata schemes. *Cataloging & Classification Quarterly*, 40(3–4), 17–36.
- Greenberg, J. (2009). Metadata and digital Information. In M. J. Bates & M. N. Maack (Eds.), *Encyclopedia of Library and Information Sciences, Third Edition* (pp. 3610–3623). Boca Raton: CRC Press. <https://doi.org/doi:10.1081/E-ELIS3-120044415>
- Greenberg, J., Pattuelli, M. C., Parsia, B., & Robertson, W. D. (2001). Author-generated Dublin Core metadata for Web resources: A baseline study in an organization. In *International Conference on Dublin Core and Metadata Applications* (pp. 38–45). Retrieved from <http://dcpapers.dublincore.org/pubs/article/view/647>
- Grimen, H. (2008). Profesjon og kunnskap. In A. Molander & L. I. Terum (Eds.), *Profesjonsstudier* (pp. 71–85). Oslo: Universitetsforlaget.
- Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2), 199–220. <https://doi.org/10.1006/knac.1993.1008>
- Guy, M., Powell, A., & Day, M. (2004). Improving the quality of metadata in eprint archives. *Ariadne*, (38). Retrieved from <http://www.ariadne.ac.uk/issue38/guy/>
- 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/10.1177/01655515155594729>

- Halpin, H., Hayes, P., McCusker, J., McGuinness, D. L., & Thompson, H. (2010). When owl:sameAs isn't the same: An analysis of identity in Linked Data. In P. F. Patel-Schneider, Y. Pan, P. Hitzler, P. Mika, L. Zhang, J. Z. Pan, ... B. Glimm (Eds.), *The Semantic Web – ISWC 2010* (Vol. 6496, pp. 305–320). Berlin: Springer. [https://doi.org/10.1007/978-3-642-17746-0\\_20](https://doi.org/10.1007/978-3-642-17746-0_20)
- Haslhofer, B., & Klas, W. (2010). A survey of techniques for achieving metadata interoperability. *ACM Computing Surveys*, 42(2), 1–37. <https://doi.org/10.1145/1667062.1667064>
- 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](http://www.agi-imc.de/internet.nsf/0/f106435e0fd9ffc1c125699f002ddf31/$FILE/dubin_core.pdf)
- Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *Mis Quarterly*, 28(1), 75–105. Retrieved from <http://dl.acm.org/citation.cfm?id=2017217>
- Hider, P. (2012). *Information resource description: Creating and managing metadata*. London: Facet Publishing.
- Hillmann, D. I. (2008). Metadata quality: From evaluation to augmentation. *Cataloging & Classification Quarterly*, 46(1), 65–80. <https://doi.org/10.1080/01639370802183008>
- Hillmann, D. I., & Phipps, J. (2007). Application profiles: Exposing and enforcing metadata quality. In *International Conference on Dublin Core and Metadata Applications* (pp. 52–62). Retrieved from <http://dcpapers.dublincore.org/pubs/article/view/866>
- Hjørland, B. (2007). Semantics and knowledge organization. *Annual Review of Information Science and Technology*, 41(1), 367–405. <https://doi.org/10.1002/aris.2007.1440410115>
- Hjørland, B. (2008). What is Knowledge Organization (KO)? *Knowledge Organization*, 2/3(35), 86–101.
- Hogan, A., Umbrich, J., Harth, A., Cyganiak, R., Polleres, A., & Decker, S. (2012). An empirical survey of Linked Data conformance. *The Journal of Web Semantics*, 14, 14–44. <https://doi.org/10.1016/j.websem.2012.02.001>

- Huvila, I., Anderson, T. D., Jansen, E. H., McKenzie, P., & Worrall, A. (2017). Boundary objects in information science. *Journal of the Association for Information Science and Technology*, 68(8), 1807–1822. <https://doi.org/10.1002/asi.23817>
- Hyland, B. (2010). Preparing for a Linked data enterprise. In D. Wood (Ed.), *Linking enterprise data* (pp. 51–64). Springer US. [https://doi.org/10.1007/978-1-4419-7665-9\\_3](https://doi.org/10.1007/978-1-4419-7665-9_3)
- Hyland, B., Atemezing, G., & Villazón-Terrazas, B. (2014). *Best practices for publishing Linked Data: W3C working group note*. W3C. Retrieved from <https://www.w3.org/TR/ld-bp/>
- IFLA Cataloguing Section & IFLA Meetings of Experts on an International Cataloguing Code. (2016). *Statement of international cataloguing principles*. Den Haag: IFLA.
- IFLA Study Group on the Functional Requirements for Bibliographic Records. (1998). *Functional requirements for bibliographic records: Final report*. München: K.G. Saur.
- Isidoro, A. (2013). Google’s Knowledge Graph: One step closer to the Semantic Web? Retrieved from <https://econsultancy.com/blog/62241-google-s-knowledge-graph-one-step-closer-to-the-semantic-web>
- Jenkins, H. (2006). *Convergence culture: Where old and new media collide*. New York, NY: New York University Press.
- Kaufman, L., & Rousseeuw, P. (2009). *Finding groups in data: An introduction to cluster analysis*. Hoboken, NJ: Wiley.
- Keet, C. M. (2013). Open World Assumption. In W. Dubitzky, O. Wolkenhauer, H. Yokota, & K.-H. Cho (Eds.), *Encyclopedia of Systems Biology* (pp. 1567–1567). New York, NY: Springer New York. [https://doi.org/10.1007/978-1-4419-9863-7\\_734](https://doi.org/10.1007/978-1-4419-9863-7_734)
- Knoblock, C. A., Szekely, P., Ambite, J., Gupta, S., Goel, A., Muslea, M., ... Mallick, P. (2012). Semi-automatically mapping structured sources into the semantic web. In *Proceedings of the Extended Semantic Web Conference*. Crete, Greece. Retrieved from [http://link.springer.com/chapter/10.1007/978-3-642-30284-8\\_32](http://link.springer.com/chapter/10.1007/978-3-642-30284-8_32)
- 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>
- Lankes, D. R. (2011). *The atlas of new librarianship*. Cambridge, MA: MIT Press.

- 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 Linked Data Incubator Group. (2011). *Library Linked Data incubator group: Final report*. W3C. Retrieved from <http://www.w3.org/2005/Incubator/lld/XGR-lld/>
- Library of Congress. (n.d.). Bibliographic Framework initiative. Retrieved from <https://www.loc.gov/bibframe/>
- Library of Congress. (2012). *Bibliographic Framework as a Web of data: Linked data model and supporting services*. Washington DC: Library of Congress. Retrieved from <http://www.loc.gov/marc/transition/pdf/marclld-report-11-21-2012.pdf>
- Library of Congress Working Group on the Future of Bibliographic Control. (2008). *On the record: Report of the Library of Congress Working Group on the Future of Bibliographic Control*. Retrieved from <http://www.loc.gov/bibliographic-future/news/lcwg-ontherecord-jan08-final.pdf>
- Lindquist, T., Dulock, M., Törnroos, J., Hyvönen, E., & Mäkelä, E. (2013). Using Linked Open Data to enhance subject access in online primary sources. *Cataloging & Classification Quarterly*, 51(8), 913–928. <https://doi.org/10.1080/01639374.2013.823583>
- Madison, O. M. A. (2005). The origins of the IFLA Study on Functional Requirements for Bibliographic Records. *Cataloging & Classification Quarterly*, 39(3–4), 15–37. [https://doi.org/10.1300/J104v39n03\\_02](https://doi.org/10.1300/J104v39n03_02)
- Mallea, A., Arenas, M., Hogan, A., & Polleres, A. (2011). On blank nodes. In L. Aroyo, C. Welty, H. Alani, J. Taylor, A. Bernstein, L. Kagal, ... E. Blomqvist (Eds.), *The Semantic Web – ISWC 2011* (pp. 421–437). Berlin: Springer. [https://doi.org/10.1007/978-3-642-25073-6\\_27](https://doi.org/10.1007/978-3-642-25073-6_27)
- McCuinness, D. (2002). Ontologies come of age. In D. Fensel, J. A. Hendler, H. Lieberman, & W. Wahlster (Eds.), *Spinning the semantic web: Bringing the World Wide Web to its full potential* (pp. 171–195). Cambridge, MA: MIT Press.
- McKenna, L. (2017). Engaging librarians in the process of interlinking RDF resources. In H. O. Blomqvist E., Maynard D., Gangemi A., Hoekstra R., Hitzler P. (Ed.), *The Semantic Web – ESWC 2017* (pp. 216–225). Champaign, IL: Springer. [https://doi.org/10.1007/978-3-319-58451-5\\_16](https://doi.org/10.1007/978-3-319-58451-5_16)

- Merčun, T., Žumer, M., & Aalberg, T. (2016). Presenting bibliographic families. *Journal of Documentation*, 72(3), 490–526. <https://doi.org/10.1108/JD-01-2015-0001>
- Merčun, T., Žumer, M., & Aalberg, T. (2017). Presenting bibliographic families using information visualization: Evaluation of FRBR-based prototype and hierarchical visualizations. *Journal of the Association for Information Science and Technology*, 68(2), 392–411. <https://doi.org/10.1002/asi.23659>
- Miller, G. (1997). Building bridges. In D. Silverman (Ed.), *Qualitative research: Theory, method and practice* (pp. 24–44). London: Sage.
- Nilsson, M., Baker, T., & Johnston, P. (2009). Interoperability levels for Dublin Core Metadata. Retrieved from <http://dublincore.org/documents/interoperability-levels/>
- Nilsson, M., Johnston, P., Naeve, A., & Powell, A. (2007). The future of learning object metadata interoperability. In K. Harman & A. Koohang (Eds.), *Learning objects: Standards, metadata, repositories, and LCMS* (pp. 255–313). Santa Rosa, CA: Informing Science Press.
- Norman, D. (2013). *The design of everyday things: Revised and expanded edition*. New York, NY: Basic Books.
- OASIS SOA Technical Committee. (n.d.). OASIS SOA reference model. Retrieved from <https://www.oasis-open.org/committees/soa-rm/faq.php>
- Ochoa, X., & Duval, E. (2009). Automatic evaluation of metadata quality in digital repositories. *International Journal on Digital Libraries*, 10(2–3), 67–91. <https://doi.org/10.1007/s00799-009-0054-4>
- 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>
- Park, J.-R. (2009). Metadata quality in digital repositories: A survey of the current state of the art. *Cataloging & Classification Quarterly*, 47(3–4), 213–228. <https://doi.org/10.1080/01639370902737240>
- Park, T. (2006). Semantic interoperability and metadata quality: An analysis of metadata item records of digital image collections. *Knowledge Organization*, 33(1), 20–34.
- Pattuelli, M. C., & Rubinow, S. (2013). The knowledge organization of DBpedia: A case study. *Journal of Documentation*, 69(6), 762–772. <https://doi.org/10.1108/JD-07-2012-0084>



- Pisanski, J., & Žumer, M. (2010). Mental models of the bibliographic universe. Part 1: Mental models of descriptions. *Journal of Documentation*, 66(5), 643–667.  
<https://doi.org/10.1108/00220411011066772>
- Pisanski, J., & Žumer, M. (2010). Mental models of the bibliographic universe. Part 2: Comparison task and conclusions. *Journal of Documentation*, 66(5), 668–680.  
<https://doi.org/https://doi.org/10.1108/00220411011066781>
- Pisanski, J., & Žumer, M. (2012). User verification of the FRBR conceptual model. *Journal of Documentation*, 68(4), 582–592. <https://doi.org/10.1108/00220411211239129>
- Radford, G. P. (2003). Trapped in our own discursive formations: Toward an archaeology of Library and Information Science. *The Library Quarterly*, 73(1), 1–18.
- Raimond, Y., Sutton, C., & Sandler, M. (2008). Automatic interlinking of music datasets on the semantic web. In *Linked Data on the Web - LDOW 2008*.
- Rajabi, E., Sicilia, M.-A., & Sanchez-Alonso, S. (2014). An empirical study on the evaluation of interlinking tools on the Web of Data. *Journal of Information Science*, 40(5), 637–648.  
<https://doi.org/10.1177/0165551514538151>
- Reitz, J. M. (n.d.). Bibliographic control. In *Online Dictionary for Library and Information Science*. ABC-CLIO. Retrieved from [http://www.abc-clio.com/ODLIS/odlis\\_b.aspx](http://www.abc-clio.com/ODLIS/odlis_b.aspx)
- Riley, J. (2017). Understanding metadata: What is metadata, and what is it for? Baltimore, MD: National Information Standards Organization. Retrieved from [http://www.niso.org/apps/group\\_public/download.php/17446/Understanding Metadata.pdf](http://www.niso.org/apps/group_public/download.php/17446/Understanding%20Metadata.pdf)
- Riva, P., Bœuf, P. Le, & Žumer, M. (2017). IFLA Library Reference Model. Den Haag: IFLA. Retrieved from [https://www.ifla.org/files/assets/cataloguing/frbr-lrm/ifla\\_lrm\\_2017-03.pdf](https://www.ifla.org/files/assets/cataloguing/frbr-lrm/ifla_lrm_2017-03.pdf)
- Rosenfeld, L., Morville, P., & Arango, J. (2015). *Information architecture: For the web and beyond* (4th ed.). Sebastopol, CA: O'Reilly.
- Saracevic, T. (2005). How were digital libraries evaluated. In *Libraries in the Digital Age (LIDA)* (pp. 1–13). Dubrovnik and Mljet.
- Schmachtenberg, M., Bizer, C., & Paulheim, H. (2014). Adoption of the Linked Aata best practices in different topical domains. In P. Mika, T. Tudorache, A. Bernstein, C. Welty, C. Knoblock, D. Vrandečić, ... C. Goble (Eds.), *The Semantic Web – ISWC 2014* (pp. 245–260). Champaign, IL: Springer. [https://doi.org/10.1007/978-3-319-11964-9\\_16](https://doi.org/10.1007/978-3-319-11964-9_16)

- Schreur, P. (2012). The academy unbound: Linked Data as revolution. *Library Resources & Technical Services*, 56(4), 227–237.
- Shreeves, S. L., Knutson, E. M., Stvilia, B., Palmer, C. L., Twidale, M. B., & Cole, T. W. (2005). Is quality metadata shareable metadata? The implications of local metadata practices for federated collections. In *Proceedings of the Association of College and Research Libraries - ACRL 2005*. Association of College and Research Libraries.
- Smiraglia, R. P. (2006). Whither knowledge organization. *Knowledge Organization*, 33(1), 8.
- Smiraglia, R. P. (2007). Bibliographic families and super works. In A. G. Taylor (Ed.), *Understanding FRBR: What it is and How it will Affect our Retrieval Tools* (pp. 73–86). Westport, CT: Libraries Unlimited.
- Smiraglia, R. P. (2014). *The elements of knowledge organization*. Cham: Springer.
- Smiraglia, R. P., & Leazer, G. H. (1999). Derivative bibliographic relationships: The work relationship in a global bibliographic database. *Journal of the American Society for Information Science*, 50(6), 493–504.
- Smith-Yoshimura, K. (2016). Analysis of international Linked Data survey for implementers. *D-Lib Magazine*, 22(7/8). <https://doi.org/10.1045/july2016-smith-yoshimura>
- Smith, B. (2004). Ontology. In L. Floridi (Ed.), *The Blackwell guide to the philosophy of computing and information* (pp. 155–166). MA: Blackwell publishing.
- Sokal, R. R., & Michener, C. D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*. <https://doi.org/citeulike-article-id:1327877>
- Studer, R., Benjamins, V. R., & Fensel, D. (1998). Knowledge engineering: Principles and methods. *Data & Knowledge Engineering*, 25(1–2), 161–197. [https://doi.org/10.1016/S0169-023X\(97\)00056-6](https://doi.org/10.1016/S0169-023X(97)00056-6)
- Stvilia, B., Gasser, L., Twidale, M. B., Shreeves, S. L., & Cole, T. (2004). Metadata quality for federated collections. In I. N. Chengalur-Smith, L. Raschid, J. Long, & C. Seko (Eds.), *Proceedings of the International Conference on Information Quality - ICIQ04* (pp. 111–125).

- Stvilia, B., Gasser, L., Twidale, M. B., & Smith, L. C. (2007). A framework for information quality assessment. *Journal of the American Society for Information Science and Technology*, 58(12), 1720–1733. <https://doi.org/10.1002/asi.20652>
- Suominen, O., & Hyvönen, N. (2017). From MARC silos to Linked Data silos? *O-Bib*, 4(2), 1–14. Retrieved from <http://dublincore.org/resources/training/#2017suominen>
- Svenonius, E. (2000). *The intellectual foundation of information organization*. Cambridge, MA: MIT Press.
- Svensson, L. (2013). Are current bibliographic models suitable for integration with the Web? *Information Standards Quarterly*, 25(4), 6. <https://doi.org/10.3789/isqv25no4.2013.02>
- Szekely, P., Knoblock, C. A., Yang, F., Zhu, X., Fink, E., Allen, R., & Goodlander, G. (2013). Connecting the Smithsonian American Art Museum to the Linked Data cloud. In P. Cimiano, O. Corcho, V. Presutti, S. Rudolph, & L. Hollink (Eds.), *The Semantic Web – ESWC 2013* (Vol. 7882, pp. 593–607). Berlin: Springer. [https://doi.org/10.1007/978-3-642-38288-8\\_40](https://doi.org/10.1007/978-3-642-38288-8_40)
- Tennant, R. (2002). MARC must die. *Library Journal*. Retrieved from <http://www.libraryjournal.com/article/CA250046.html>
- Thomale, J. (2010). Interpreting MARC: Where’s the bibliographic data? *Code4lib*, (11). Retrieved from <http://journal.code4lib.org/articles/3832>
- Tillett, B. (1991). A taxonomy of bibliographic relationships. *Library Resources and Technical Services*, 35, 150–158.
- Tillett, B. B. (2013). RDA and the Semantic Web, Linked Data environment. *Italian Journal of Library and Information Science*, 4(1), 139–145. <https://doi.org/10.4403/jlis.it-6303>
- Tolk, A., & Muguira, J. (2003). The levels of conceptual interoperability model. In *Proceedings of the 2009 Spring Simulation Multiconference - SpringSim 2009* (pp. 1–9). San Diego, CA: Society for Computer Simulation International.
- van Hooland, S. (2009). *Metadata quality in the cultural heritage sector: Stakes, problems and solutions*. Brussels: 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.

- Verborgh, R., Vander Sande, M., Hartig, O., Van Herwegen, J., De Vocht, L., De Meester, B., ... Colpaert, P. (2016). Triple Pattern Fragments: A low-cost knowledge graph interface for the Web. *Journal of Web Semantics*, 37–38, 184–206.  
<https://doi.org/10.1016/J.WEBSEM.2016.03.003>
- 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 M. Hausenblas & E. Simperl (Eds.), *Proceedings of the European Data Forum 2012 - EDF 2012*. Copenhagen: Ceur.
- Vrandečić, D. (2010). *Ontology evaluation*. Karlsruhe: Karlsruher Institut für Technologie. Retrieved from <http://www.aifb.kit.edu/images/b/b5/OntologyEvaluation.pdf>
- Vukadin, A. (2014). Bits and pieces of information: Bibliographic modeling of transmedia. *Cataloging & Classification Quarterly*, 52(3), 285–302.  
<https://doi.org/10.1080/01639374.2013.879976>
- W3C Schema.org Community Group. (n.d.). Schema.org. Retrieved from <http://schema.org/>
- 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>
- Ward, J. (2003). A quantitative analysis of unqualified dublin core metadata element set usage within data providers registered with the open archives initiative. In *Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries* (pp. 315–317). IEEE Computer Society. Retrieved from <http://dl.acm.org/citation.cfm?id=827140.827196>
- Westbrook, L. (2006). Mental models: A theoretical overview and preliminary study. *Journal of Information Science*, 32(6), 563–579. <https://doi.org/10.1177/0165551506068134>
- Westrum, A.-L. (2014). Oslo Public Library chooses RDF Linked Data as core metadata format. Retrieved from <http://digital.deichman.no/blog/2014/06/19/oslo-public-library-chooses-rdf-linked-data-as-core-metadata-format/>
- Westrum, A.-L., Rekkavik, A., & Tallerås, K. (2012). Improving the presentation of library data using FRBR and Linked data. *Code4Lib Journal*, (16).
- Willer, M., & Dunsire, G. (2013). *Bibliographic information organization in the semantic web*. Oxford: Chandos Publishing.

- Wilson, P. (1968). *Two kinds of power: An essay on bibliographical control*. Berkeley, CA: University of California Press.
- Wisser, K. (2014). The errors of our ways: Using metadata quality research to understand common error patterns in the application of name headings. In S. Closs, R. Studer, E. Garoufallou, & M.-A. Sicilia (Eds.), *Metadata and Semantics Research - MTSR 2014* (Vol. 478, pp. 83–94). Berlin: Springer. [https://doi.org/10.1007/978-3-319-13674-5\\_9](https://doi.org/10.1007/978-3-319-13674-5_9)
- Working Group on FRBR/CRM Dialogue. (2016). Definition of FRBRoo: A conceptual model for bibliographic information in object-oriented formalism. Den Haag: IFLA. Retrieved from [https://www.ifla.org/files/assets/cataloguing/FRBRoo/frbroo\\_v\\_2.4.pdf](https://www.ifla.org/files/assets/cataloguing/FRBRoo/frbroo_v_2.4.pdf)
- Yasser, C. M. (2011). An analysis of problems in metadata records. *Journal of Library Metadata*, 11(2), 51–62. <https://doi.org/10.1080/19386389.2011.570654>
- Yasser, C. M. (2012). An experimental study of metadata training effectiveness on errors in metadata records. *Journal of Library Metadata*, 12(4), 372–395. <https://doi.org/10.1080/19386389.2012.735573>
- Yee, M. M. (1994). The concept of work for moving image materials. *Cataloging & Classification Quarterly*, 18(2), 33–40. [https://doi.org/10.1300/J104v18n02\\_04](https://doi.org/10.1300/J104v18n02_04)
- Yee, M. M. (2013). Can bibliographic data be put directly onto the Semantic Web? *Information Technology and Libraries*, 28(2), 55–80. Retrieved from <https://ejournals.bc.edu/ojs/index.php/ital/article/view/3175>
- 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>
- Zeng, M. L., & Qin, J. (2016). *Metadata*. London: Facet publishing.
- Zhang, Y., & Salaba, A. (2009). What is next for functional requirements for bibliographic records? A delphi study. *The Library Quarterly*, 79(2), 233–255. <https://doi.org/10.1086/597078>
- Aalberg, T., & Žumer, M. (2013). The value of MARC data, or, challenges of fibrisation. *Journal of Documentation*, 69(6), 851–872. <https://doi.org/10.1108/JD-05-2012-0053>



Tallerås, K. (2017). Quality of linked bibliographic data: The models, vocabularies, and links of data sets published by four national libraries. *Journal of Library Metadata*, 17(2), 126–155.







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

Kim Tallerås

Department of Archivistics, Library, and Information Science, Oslo and Akershus University College of Applied Sciences, Oslo, Norway

### ABSTRACT

Little effort has been devoted to the systematic examination of published Linked Data in the library community. This article 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 data sets. 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.



### KEYWORDS

linked data; bibliographic data; semantic web; data quality; knowledge organization

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 modeling, 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 article 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 data sets. The study is aimed at answering the following research questions:

---

**CONTACT** Kim Tallerås  [Kim.Talleras@hioa.no](mailto:Kim.Talleras@hioa.no)  Oslo and Akershus University College of Applied Sciences, PO box 4 St. Olavs plass, NO-0130 Oslo, Norway.

Color versions of one or more of the figures in the article can be found online at [www.tandfonline.com/wjlm](http://www.tandfonline.com/wjlm).

© Kim Tallerås

Published with license by Taylor & Francis

- 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 Data sets 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 article 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 identifying 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 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 nonproprietary 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 that 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 data sets.

Despite these detailed guidelines, studies show that Linked Data sets 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 (Villazoñ-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 (LLDI Group, 2011) published its final report, which, 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 endeavoring 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, Machado, & 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 whether these parallel publishing activities merely create new Linked Data silos filled with heterogenic data (Suominen & Hyvönen, 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<sup>1</sup> 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, data set 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 reuse 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 that include LBD.

Hogan et al. (2012) analyze and statistically rank 188 pay level—domains (PLD)<sup>2</sup> 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 reuse 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 Data sets harvested through a Web crawl and find that 56% of the analyzed data sets 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 six external sets (Schmachtenberg et al., 2014). Almost all of the sets (99.9%) use elements from nonproprietary 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 an LBD set.

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 to determine that data sets 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 data sets,

including LBD from the BNE and Library of Congress. The test shows that most errors in the data sets, 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 data sets from several LBD providers (Papadakis et al., 2014). Hallo, Luján-Mora, Maté, and Trujillo (2015) also investigate the quality of data sets that are part of the corpus studied in this article. 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 data sets assessed in the study must contain directly available, comparable, and nonexperimental bibliographic data published by a library institution. Based on these criteria, the following data sets were selected:

- BNB The British National Bibliography was first published as Linked Data in 2011. It includes both books and serial publications made available in separate data sets. In this evaluation, only the book set is considered.
- BNE The Biblioteca Nacional de España has published LBD since 2011. This data set 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 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 has included bibliographic data since 2012. For this evaluation, two data sets are downloaded and combined: the Deutsche Nationalbibliografie (DNBTitel) and the Integrated Authority File (GND).

Other data sets may also fit the selection criteria described here, but an analysis of the chosen data sets 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 subsets. Most of these are complementary and interlinked through common URIs. For example, the DNBTitel data set mainly contains detailed information about documents, including references to URIs from the GND set where authors and other persons related to the

**Table 1.** Download Date, Last Modified Date, License Information, and Set Names of the Four Corpus Sets.

	Download Date	Modified Date	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

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 data sets 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 between late February and early April 2016 and were the most recently updated sets commonly available for download at that time.

### 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 an 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 a 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<sup>3</sup> vocabulary (expressed with the namespace `dct`<sup>4</sup>) 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:<sup>5</sup>

```

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

```

**Table 2.** Number of Triples, Entities, and Data-Level Constants.

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

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 link connects the URIs within an 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 data set. 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 data set (<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  $C = U \cup B \cup L$ . 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.

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 data sets are tailored for particular user tasks, are 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.<sup>6</sup> 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 data sets to which a source data set links, independent of the predicate used in those links (Schmachtenberg et al., 2014). Two data sets 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 in which 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).

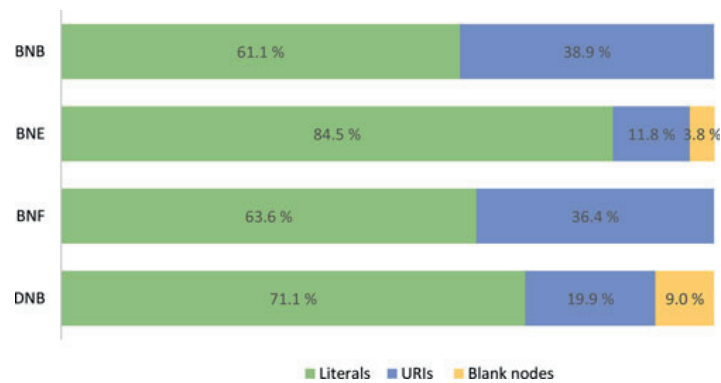
Second, the analysis considers only external data sets providing RDF data. In practice, this means that links to DBpedia but not Wikipedia are counted. This is in line with previous studies (Hogan et al., 2012) and Linked Data principles. Third, for each particular predicate used in external RDF links (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 data set. 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' home pages). A corpus with fewer than 300 links to a particular data set, 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 cloud requires at least 1,000 triples<sup>7</sup>). A reduced amount of external data sets 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



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

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).

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

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 data sets differ, so specific queries mirroring these models could be developed to retrieve the information mentioned. Instead, to treat the data sets as neutrally as possible, all the queries are based

**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.

	Persons		Manifestations		Concepts	
	Class	Percentage of distinct entities with membership	Class	Percentage of distinct entities with membership	Class	Percentage 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%	rdaf:br:Manifestation	27.4%	skos:Concept	9.0%
DNB	gndo:Person*	26.2%	bibo:Documents	30.7%	gndo:SubjectHeading*	0.6%

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 data sets. 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 a 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 data sets, as does the composition of RDF components. This indicates that the full data sets do not contain significant amounts of data not directly related to the bibliographic entities that could skew the comparative perspectives of the following analysis in this study. Moreover, while the modeling 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 data sets *reuse* 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 and at examining reuse, 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 reuse. 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.

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 subclass 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

**Table 4.** Top 10 Property and Class Terms by the Number of Triples and Class Memberships for Each Corpus Set.

No.	Property	Percentage of triples	Class	Percentage of rdf:type triples	Property	Percentage of triples	Class	Percentage of rdf:type triples
			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:PT053	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%	rdafibr:Manifestation	27.4%	owl:sameAs	6.7%	rdf:Seq	25.2%
3	dcterms:created	4.1%	rdafibr:Expression	27.4%	gndo:surname	4.2%	gndo:UndifferentiatedPerson	13.2%
4	rdfs:seeAlso	4.1%	skos:Concept	9.1%	gndo:firstname	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%	rdafibr: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	rdafibr:expression Manifested	3.0%	dcmitt: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 5.** Number of Vocabularies and Vocabulary Terms Used in the Sets.

	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

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 `rdافربر:`). 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 share only three.

The corpus sets can also be distinguished by other characteristics. The BNF uses 24 different vocabularies to describe its bibliographic data, but the BNE uses only 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 MARC 21 relator code vocabulary<sup>8</sup> (for examples, see Appendix IV).

### Vocabulary reuse

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 data sets outside this context. This study, therefore, uses

**Table 6.** Percentage of Local Vocabularies and Vocabulary Terms and the Percentage of the Triples in the Sets Using These Terms.

	Percentage 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%

the more moderate term *local vocabulary*. A local vocabulary is further defined by an institutional connection to the publisher of the particular data set 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 reuse. 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.

The BNE in particular and 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.

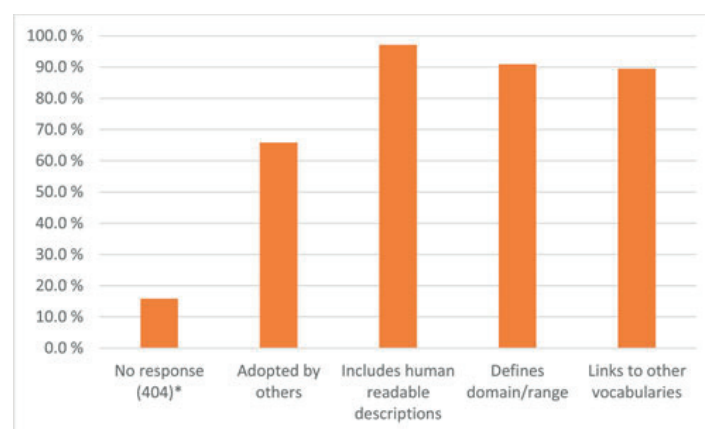
Data providers apply local terms for several reasons—for example, to facilitate logical consistency in a given data set or to express semantic relationships not covered by existing vocabularies. In the case of the BNE, its predominant use of local terms is probably due to intrinsic consistency issues. 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 wherein 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 that could be used instead of the local terms in the corpus sets. That, however, is a substantial task that 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 reuse 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 data sets 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



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



address vocabulary terms correctly. The four vocabularies initially returning a 404 response but later manually identified are therefore included in the examinations.

Whether (other) data set 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 Data sets. Both services provide a search interface for vocabularies and return the number of data sets 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 data sets other than those in the corpus. On average, 65.8% of the vocabularies are used at least by one other data set. 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 data sets for any of these measurements.

### Interlinking

Reusing 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 data sets. 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 data set. The external RDF links in the corpus sets are counted in line with the limitations listed in the methods section.<sup>9</sup> The analysis of linking practices is based on the main components of external RDF links: the properties used and the external target data sets. These components correspond to metrics from earlier Linked Data–quality research. Counting external data sets allows comparing the outdegree of a particular data set 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 and the highest ratio of interlinked entities. Linked Data guidelines tend to favor `owl:sameAs` links (external RDF links using the property `sameAs` from

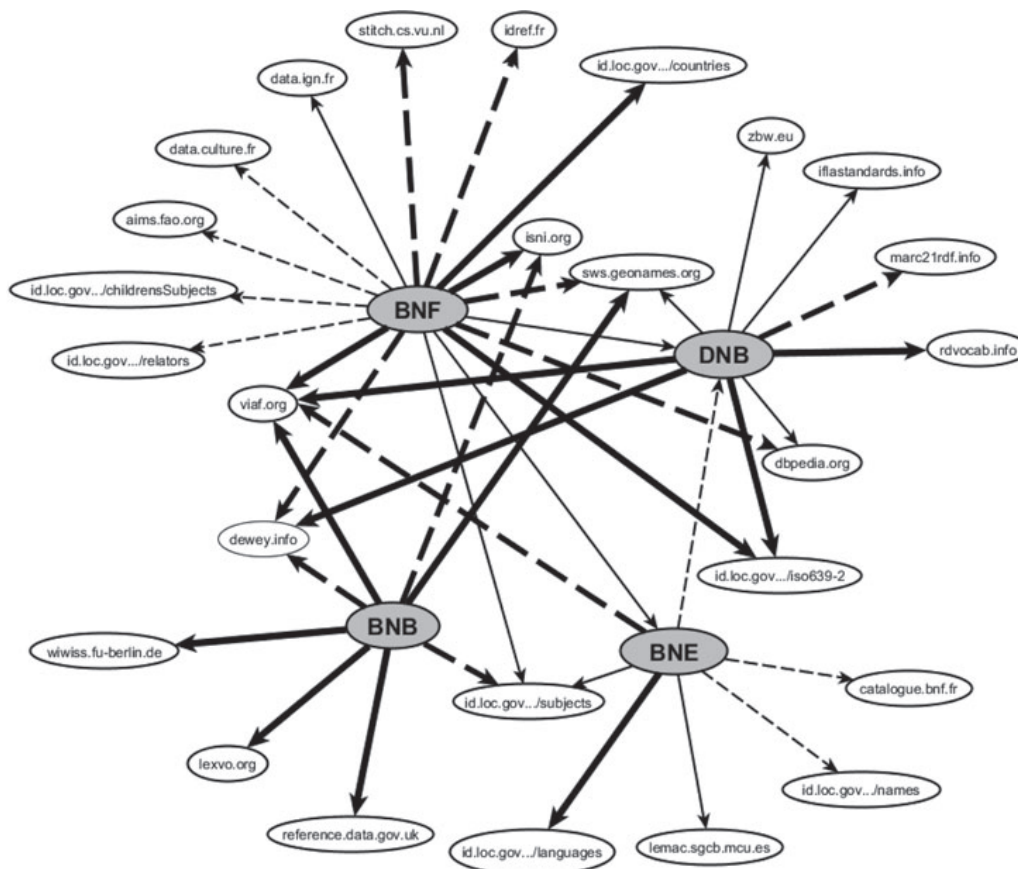
**Table 7.** External RDF Links for All Triples and per Entity.

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

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.

**Outdegree**

The metric outdegree is defined as the number of unique external data sets 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 data sets 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 data sets 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 data sets is analyzed, particularly their origin in the library domain, defined as being hosted by a library institution. Figure 3 shows the full network of



**Figure 3.** The four corpus sets and the external data sets 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.

**Table 8.** Various Aspects of Outdegree in Each Set, Average and Total for the Corpus.

	BNB	BNE	BNF	DNB	Total	Avg.
No. of data sets	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 nonlibrary PLDs linked with predicate owl:sameAs	0	0	3	1	4	1

links between the corpus sets and the external data sets. The thickness of lines indicates the number of RDF links between the data sets. Table 8 lists the outdegree of each set. Table 9 provides an overview of the 10 data sets that are the targets of most RDF links, along with the distribution in each corpus set.

In total, the corpus sets link to 28 unique data sets 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 nonlibrary data sets (the aforementioned three and idref.org). At the data set level, 20 of the 28 data sets are linked to only one corpus set. Three data sets have links to two corpus sets, and four data sets to three corpus sets and only one external data set (viaf.org) have links to all four sets. Of the 22

**Table 9.** Ten External Data Sets 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.

	Percentage of RDF links, total and individual sets	Distinct objects of RDF links, total and individual sets
<a href="http://id.loc.gov/vocabulary/iso639-2/">http://id.loc.gov/vocabulary/iso639-2/</a>	<b>23.1%</b>	<b>486</b>
BNB	59.5%	446
DNB	15.7%	486
<a href="http://rdvocab.info/termList/RDACarrierType/">http://rdvocab.info/termList/RDACarrierType/</a>	<b>20.3%</b>	<b>4</b>
DNB	46.3%	4
<a href="http://viaf.org/viaf/">http://viaf.org/viaf/</a>	<b>19.7%</b>	<b>10,341,459</b>
DNB	31.5%	8,141,903
BNE	22.0%	559,783
BNF	11.3%	1,807,538
BNB	7.2%	1,040,851
<a href="http://www4.wiwiss.fu-berlin.de/bookmashup/books/">http://www4.wiwiss.fu-berlin.de/bookmashup/books/</a>	<b>5.6%</b>	<b>3,262,475</b>
BNB	23.0%	3,262,475
<a href="http://sws.geonames.org/">http://sws.geonames.org/</a>	<b>5.4%</b>	<b>148,845</b>
BNB	20.7%	156
BNF	0.7%	101,629
DNB	0.2%	47,104
<a href="http://lexvo.org/id/iso639-3/">http://lexvo.org/id/iso639-3/</a>	<b>5.2%</b>	<b>272</b>
BNB	21.0%	272
<a href="http://reference.data.gov.uk/id/year/">http://reference.data.gov.uk/id/year/</a>	<b>5.0%</b>	<b>224</b>
BNB	20.6%	224
<a href="http://id.loc.gov/vocabulary/languages/">http://id.loc.gov/vocabulary/languages/</a>	<b>3.3%</b>	<b>256</b>
BNE	75.8%	256
<a href="http://isni.org/">http://isni.org/</a>	<b>3.3%</b>	<b>1,651,998</b>
BNF	7.5%	1,196,185
BNB	5.0%	725,148
<a href="http://dewey.info/class/">http://dewey.info/class/</a>	<b>3.1%</b>	<b>215,059</b>
BNB	1.4%	1733
BNF	2.1%	63
DNB	5.1%	214,005

**Table 10.** Overlapping viaf.org Entities Limited to Person Entities and owl:sameAs Links in Different Set Combinations and Between 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%

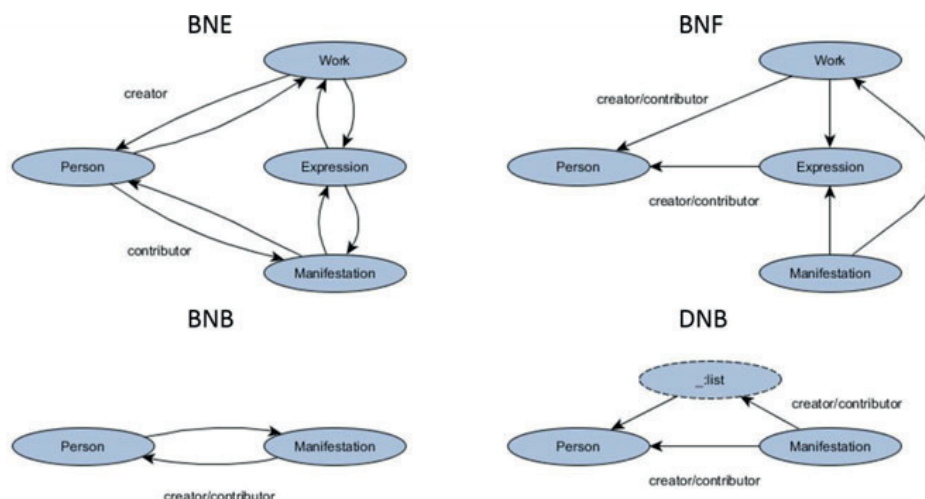
PLDs, 15 datasets are linked to one corpus set, three are linked to two corpus sets, and another two data sets from three corpus sets, and two data sets (id.loc.gov and viaf.org) to all four sets. All the corpus sets provide owl:sameAs links, with an average outdegree of three.

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/>).

Table 9 shows the 10 most popular external data sets 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.

### 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 data sets. In addition, Dylan does not come from any of the four countries that published the data sets studied. It, therefore, is less likely that the data describing him and his book are given special treatment as might be the



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

case for bibliographic data describing famous writers sharing the same nationality as the data set 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.

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 (`bn eo:OP5001/OP1001` is the creator of/is created by). Other contributors such as translators (as in the sample) are related to manifestations (`bn eo:OP3006/OP3005` has a contributor/contributes to).<sup>10</sup> 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 subproperties 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<sup>11</sup> 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 data sets 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 connected to their main creators only 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 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 and 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 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 in which 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 have been proven 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 Data set. 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 Data set.

The test of subsets based on Nobel laureates 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 reuse 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 data set), 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 than 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 data sets 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 data sets represents a varied collection of potential linkage candidates for bibliographic data. The BNF, in particular, provides links to an impressive number of data sets. 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 data sets not hosted by library institutions. The analysis also reveals that a high proportion of external data sets, 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 data sets 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



FRBRizations. Based on the examined versions of the data sets, however, this study shows that these FRBRizations 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 data sets 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 article 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 reuse of widely adopted vocabularies. At the same time, the study reveals quality issues. The data sets 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 data sets 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 & Hyvönen, 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.

## Notes

1. Blank nodes are nodes in an RDF graph that indicate 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).
2. A PLD is a subdomain of the public, top-level domain that users usually pay to access.
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/>
6. <https://www.w3.org/TR/sparql11-query/>
7. <http://lod-cloud.net/>
8. <https://www.loc.gov/marc/relators/relacode.html>
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 provides nearly 150,000 links to filmportal.de. These two sites do not offer RDF data and, therefore, are not included in the analysis.
10. The BNE ontology contains a property for expressing a relationship between manifestations and creators (`bne: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](https://www.w3.org/TR/rdf-schema/#ch_seq)

## 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, Germany: Springer. [https://doi.org/10.1007/978-3-642-33876-2\\_31](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 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. W3C. 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.
- 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/fl106435e0fd9ffc1c125699f002ddf31/\\$FILE/dubin\\_core.pdf](http://www.agi-imc.de/internet.nsf/0/fl106435e0fd9ffc1c125699f002ddf31/$FILE/dubin_core.pdf)
- Hentschke, J. (2017). Linked data service of the German national library. Retrieved from [http://www.dnb.de/EN/Service/DigitaleDienste/LinkedData/linkedata\\_node.html](http://www.dnb.de/EN/Service/DigitaleDienste/LinkedData/linkedata_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., & Villazoñ-Terrazas, B. (2014). *Best practices for publishing Linked Data: W3C Working Group Note 09 January 2014*. W3C. 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*. Munich, Germany: K. G. Saur.
- Janowicz, K., Hitzler, P., Adams, B., Kolas, D., & Vardeman, C., II. (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, NY: ACM Press. <https://doi.org/10.1145/2566486.2568002>
- LeBoeuf, P. (2012). A strange model named FRBRoo. *Cataloging and 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: Library of Congress. Retrieved from <http://www.loc.gov/marc/transition/pdf/marclid-report-11-21-2012.pdf>
- LLDI Group. (2011). *Library Linked data incubator group: Final report*. W3C. Retrieved from <https://www.w3.org/2005/Incubator/llid/XGR-llid-20111025/>
- Mallea, A., Arenas, M., Hogan, A., & Polleres, A. (2011). On blank nodes. In *The Semantic Web-ISWC 2011: 10th International Semantic Web Conference Proceedings, Part I* (pp. 421–437). Berlin, Heidelberg: Springer. [https://doi.org/10.1007/978-3-642-25073-6\\_27](https://doi.org/10.1007/978-3-642-25073-6_27)
- Malmsten, M. (2009). Exposing library data as linked data. Presented at the IFLA satellite pre-conference sponsored by the Information Technology Section “Emerging trends in technology: Libraries between Web 2.0, the Semantic Web and search technology.” <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.181.860&rep=rep1&type=pdf> [Google Scholar].
- 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). IFLA. 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.), *The Semantic Web-ISWC 2014: 13th International Semantic Web Conference, Riva del Garda, Italy. Proceedings, Part I, LNCS 8796* (pp. 245–260). Cham: Springer. [https://doi.org/10.1007/978-3-319-11964-9\\_16](https://doi.org/10.1007/978-3-319-11964-9_16)
- Simon, A., Wenz, R., Michel, V., & Mascio, A. Di. (2013, May). 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*. (Vol. 7882, pp. 563–577). Berlin, Germany: Springer. <https://doi.org/10.1007/978-3-642-38288-8>

- Suominen, O. & Hyvönen, N. (2017). From MARC silos to Linked Data silos? *o-bib* 4(2), 1–13. <http://doi.org/10.5282/o-bib/2017H2S1-13>
- Van Hooland, S. (2009). *Metadata quality in the cultural heritage sector: Stakes, problems and solutions*. Bruxelles: 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*. Great Britain, London: Facet publishing. Retrieved from <http://difusion.ulb.ac.be/vufind/Record/ULB-DIPOT:oai:dipot.ulb.ac.be:2013/156413/TOC>
- Villazoñ-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, DK: European Data Forum 2012, Vol: Vol-877, published on CEUR-WS.org.
- 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>

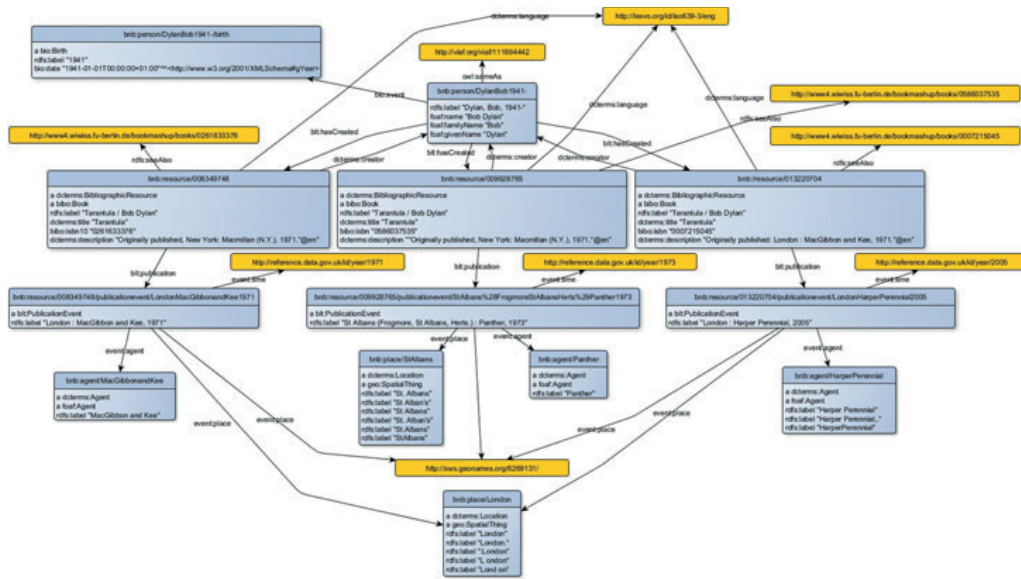
## Appendix I: Namespaces

---

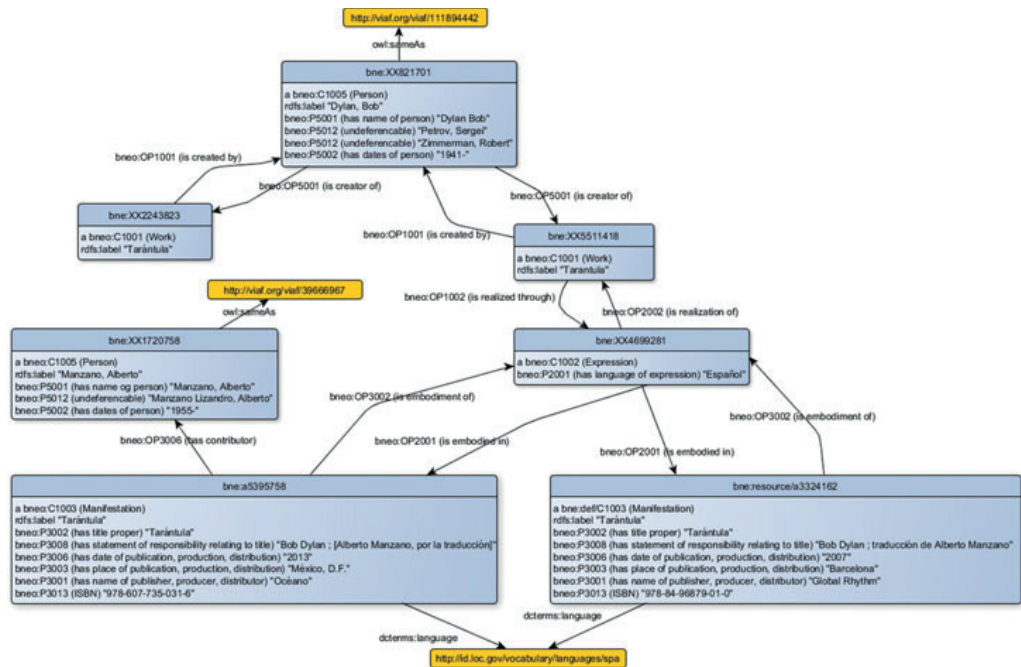
bio:	< <a href="http://purl.org/vocab/bio/0.1/">http://purl.org/vocab/bio/0.1/</a> >
blt:	< <a href="http://www.bl.uk/schemas/bibliographic/blterms#">http://www.bl.uk/schemas/bibliographic/blterms#</a> >
bnb:	< <a href="http://bnb.data.bl.uk/id/">http://bnb.data.bl.uk/id/</a> >
bne:	< <a href="http://datos.bne.es/resource/">http://datos.bne.es/resource/</a> >
bneo:	< <a href="http://datos.bne.es/def/">http://datos.bne.es/def/</a> >
bnf:	< <a href="http://data.bnf.fr/ark:/12148/">http://data.bnf.fr/ark:/12148/</a> >
bnfo:	< <a href="http://data.bnf.fr/ontology/bnf-onto/">http://data.bnf.fr/ontology/bnf-onto/</a> >
bnfrel:	< <a href="http://data.bnf.fr/vocabulary/roles/">http://data.bnf.fr/vocabulary/roles/</a> >
dce:	< <a href="http://purl.org/dc/elements/1.1/">http://purl.org/dc/elements/1.1/</a> >
dcmit:	< <a href="http://purl.org/dc/dcmitype/">http://purl.org/dc/dcmitype/</a> >
dcterms:	< <a href="http://purl.org/dc/terms/">http://purl.org/dc/terms/</a> >
dnb:	< <a href="http://d-nb.info/">http://d-nb.info/</a> >
event:	< <a href="http://purl.org/NET/c4dm/event.owl#">http://purl.org/NET/c4dm/event.owl#</a> >
foaf:	< <a href="http://xmlns.com/foaf/0.1/">http://xmlns.com/foaf/0.1/</a> >
frgeo:	< <a href="http://rdf.insee.fr/geo/">http://rdf.insee.fr/geo/</a> >
geo:	< <a href="http://www.w3.org/2003/01/geo/wgs84_pos#SpatialThing">http://www.w3.org/2003/01/geo/wgs84_pos#SpatialThing</a> >
geonames:	< <a href="http://www.geonames.org/ontology/ontology_v3.1.rdf/">http://www.geonames.org/ontology/ontology_v3.1.rdf/</a> >
geosparql:	< <a href="http://www.opengis.net/ont/geosparql#">http://www.opengis.net/ont/geosparql#</a> >
gnd:	< <a href="http://d-nb.info/gnd/">http://d-nb.info/gnd/</a> >
gndo:	< <a href="http://d-nb.info/standards/elementset/gnd#">http://d-nb.info/standards/elementset/gnd#</a> >
igno:	< <a href="http://data.ign.fr/ontology/topo.owl/">http://data.ign.fr/ontology/topo.owl/</a> >
interval:	< <a href="http://reference.data.gov.uk/def/intervals/">http://reference.data.gov.uk/def/intervals/</a> >
isbd:	< <a href="http://iflastandards.info/ns/isbd/elements/">http://iflastandards.info/ns/isbd/elements/</a> >
library:	< <a href="http://purl.org/library/">http://purl.org/library/</a> >
madsrdf:	< <a href="http://www.loc.gov/mads/rdf/v1#">http://www.loc.gov/mads/rdf/v1#</a> >
mo:	< <a href="http://musicontology.com/">http://musicontology.com/</a> >
ore:	< <a href="http://www.openarchives.org/ore/terms/">http://www.openarchives.org/ore/terms/</a> >
org:	< <a href="http://www.w3.org/ns/org#">http://www.w3.org/ns/org#</a> >
owl:	< <a href="http://www.w3.org/2002/07/owl#">http://www.w3.org/2002/07/owl#</a> >
rdacarrier:	< <a href="http://rdvocab.info/termList/">http://rdvocab.info/termList/</a> >
rdafibr:	< <a href="http://rdvocab.info/uri/schema/FRBRentitiesRDA/">http://rdvocab.info/uri/schema/FRBRentitiesRDA/</a> >
rdag1:	< <a href="http://rdvocab.info/Elements/">http://rdvocab.info/Elements/</a> >
rdag2:	< <a href="http://rdvocab.info/ElementsGr2/">http://rdvocab.info/ElementsGr2/</a> >
rdau:	< <a href="http://rdaregistry.info/Elements/u/">http://rdaregistry.info/Elements/u/</a> >
rdfs:	< <a href="http://www.w3.org/2000/01/rdf-schema#">http://www.w3.org/2000/01/rdf-schema#</a> >
schema:	< <a href="http://schema.org/">http://schema.org/</a> >
sf:	< <a href="http://www.opengis.net/ont/sf#">http://www.opengis.net/ont/sf#</a> >
skos:	< <a href="http://www.w3.org/2004/02/skos/core#">http://www.w3.org/2004/02/skos/core#</a> >
umbel:	< <a href="http://umbel.org/umbel#">http://umbel.org/umbel#</a> >

---

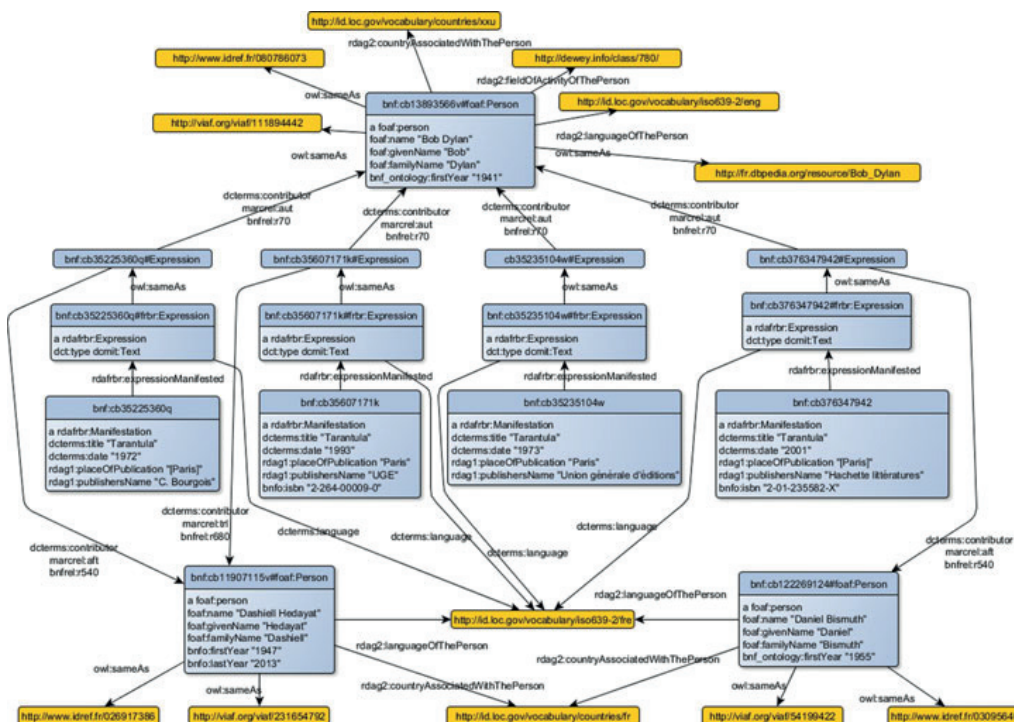
## Appendix II: Case Study of Dylan in BNB



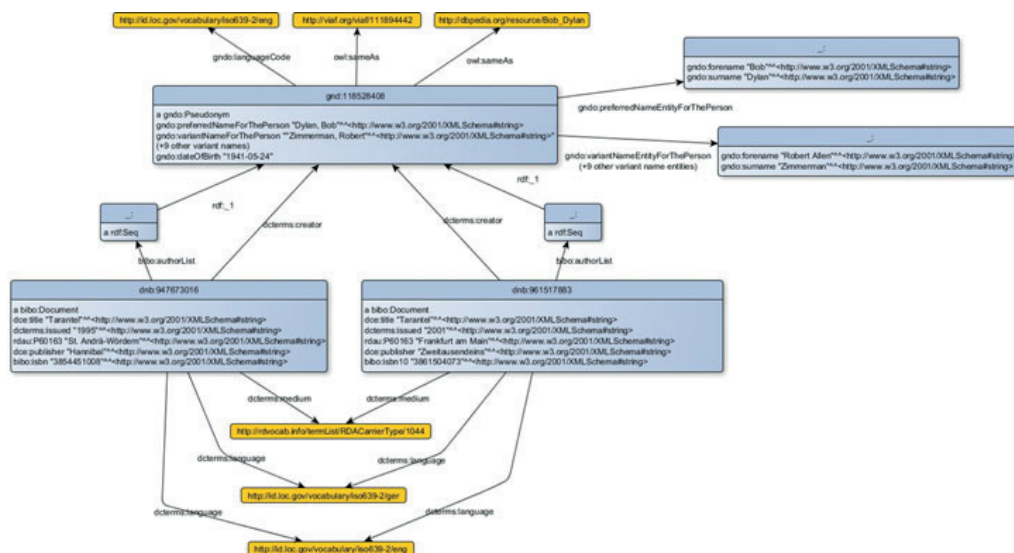
## Appendix III: Case Study of Dylan in BNE



## Appendix IV: Case Study of Dylan in BNF



## Appendix V: Case Study of Dylan in DNB



Tallerås, K., Dahl, J. H. B., & Pharo, N. (2018). User conceptualizations of derivative relationships in the bibliographic universe. Revised and accepted version in the process of being published in *Journal of Documentation*.





# User Conceptualizations of Derivative Relationships in the Bibliographic Universe

Kim Tallerås, Jørn-Helge B. Dahl, Nils Pharo

OsloMet – Oslo Metropolitan University

This is a revised and accepted version of a paper that is in the process of being published in *Journal of Documentation*.

## ABSTRACT

### Purpose

Considerable effort is devoted to developing new models for organizing bibliographic metadata. However, such models have been repeatedly criticized for their lack of proper user testing. This paper presents a study on how non-experts in bibliographic systems map the bibliographic universe and, in particular, how they conceptualize relationships between independent but strongly related entities.

### Methodology

The study is based on an open concept-mapping task performed to externalize the conceptualizations of 98 novice students. The conceptualizations of the resulting concept maps are identified and analyzed statistically.

### Findings

The study shows that the participants' conceptualizations have great variety, differing in detail and granularity. These conceptualizations can be categorized into two main groups according to derivative relationships: those that apply a single-entity model directly relating document entities and those (the majority) that apply a multi-entity model relating documents through a high-level collocating node. These high-level nodes seem to be most adequately interpreted either as superwork devices collocating documents belonging to the same bibliographic family or as devices collocating documents belonging to a shared fictional world.

### Value

The findings can guide the work to develop bibliographic standards. Based on the diversity of the conceptualizations, the findings also emphasize the need for more user testing of both conceptual models and the bibliographic end-user systems implementing those models.

### Keywords

Bibliographic systems, User studies, Cataloguing, Linked data, Information modelling, Conceptualizations, Ontologies, FRBR, Mental models, Metadata

## 1. INTRODUCTION

At the time of writing, science fictions fans battle in heated debates over whether the new *Alien* and *Blade Runner* movies are part of the same fictional universe. The final season of the television series *Game of Thrones* has been launched well ahead of the remaining volumes in the book series that initially inspired it. Another anticipated television series is an adaptation of the book *Pride and Prejudice and Zombies*, which itself is a mash-up of Jane Austen's classic novel with the zombie craze in contemporary pop culture. Such entities seem to orbit each other in a *bibliographic universe*, "just as the physical universe reels with gravity and physical forces that propel, impel, and propel planets, stars, asteroids, and other bodies to exist in relation to each other" (the ideas of Wilson, 1968, as interpreted by Smiraglia, 2014, p. 10). In practice, these entities can cover the same topics or even transmedial storylines, share author and fictional characters, and belong to families of works related through various types of derivations.

When organizing bibliographic data in information systems, it is crucial to control the forces of the bibliographic universe in a way that increases the fitness for use. One particular challenge to controlling such a universe is the application of complex derivative relationships. This paper presents a study on how non-experts in bibliographic systems map the bibliographic universe and, in particular, how they *conceptualize* relationships between independent but strongly related works. The study is based on an open concept mapping task performed to externalize the bibliographic conceptualizations of novice students in library and information science. The resulting conceptualizations are analyzed statistically to reveal typical structures.

The paper has the following organization. Section 2 provides background information on bibliographic modelling and the research question, while section 3 describes the theoretical framework. Section 4 provides an overview of previous research, and section 5 presents the research methodology. Sections 6, 7 and 8 convey the results, discussion, and concluding remarks.

## 2. BACKGROUND

In bibliographic systems, relationships are indirectly applied based on descriptive metadata expressing *shared characteristics* (Tillett, 2001) about responsibility, topicality, and publishing events. Other bibliographic relationships, such as adaptations and non-trivial derivations, cannot be applied in as a straightforward way but are included as elements in existing bibliographic models. These include the *Library Reference Model* (LRM) (Riva *et al.*, 2017), the latest formalization of models belonging to the so-called FRBR-family<sup>1</sup>. The LRM includes the original FRBR entities for *works*, *expression* (of works), and *manifestations* (of expressions). Together, these *W/E/M*<sup>2</sup> entities enable representing both successive derivative relationships, such as new marginally changed editions (enforcing a new manifestation entity), and more significant modifications, such as a translation (enforcing a new expression entity). In addition, the LRM provides

---

<sup>1</sup> <https://www.ifla.org/node/2016>.

<sup>2</sup> The entity Item is also one of the so-called Group 1 entities in the original FRBR model and represents a concrete exemplar of a manifestation. However, in this study this physical level (item) is not considered.

derivative relationships directly between works, for example, in cases when one work has served as inspiration for another.

The LRM specifications state that the model was developed based on what are believed to be important entities and relationships for users of bibliographic systems. The users are represented through a set of specific user tasks (e.g., *to explore*), which should be facilitated by “the support of discovery by making relationships explicit, by providing contextual information and navigation functionality” (Riva *et al.*, 2017, p. 10). In models like the LRM, the included elements and, not least, their structural organization represent a *conceptualization* of the bibliographic universe, a simplified, abstract model of what exists in that particular universe of discourse. According to theories on mental models (Norman, 2013), users interacting with information systems depend heavily on their own conceptualizations when solving tasks. Thus, to facilitate, for example, the exploration task as defined by the LRM, the conceptualizations facilitating “contextual information and navigation functionality” (Riva *et al.*, 2017, p. 10) should reflect the conceptualizations of the users.

A repeated claim is that bibliographic models lack proper user testing (see, e.g., Coyle, 2016; Pisanski and Žumer, 2010a; Zhang and Salaba, 2009). The models typically reflect experts’ accumulated ideas about important user tasks (for instance, the LRM builds on tasks that can be traced back to the bibliographic pioneer Charles Cutter (1904)). Pisanski and Zumer (2010a, 2010b, 2012) examined users’ mental models of W/E/M entities but mostly evaluated the resemblance between mental models and W/E/M structures as they are mandated by the FRBR model. Although this approach has provided valuable insights into users’ verification of that particular model, it could be beneficial to complement this research by testing users independent of an already-given structure. Another motivation for the present study is found in the bibliographic universe characterized by intertextuality and transmedia franchises generating immense numbers of complex derivative relationships, as exemplified in the introduction. Studies focusing on user conceptualizations in that particular context have not been found.

Thus, this paper is motivated by both the dearth of user testing in the domain of bibliographic modelling in general and the lack of knowledge on how users conceptualize derivative relationships in particular. These gaps lead to the following research question: how do users conceptualize derivative relationships between entities in the bibliographic universe?

### **3. THEORETICAL FRAMEWORK**

The vision of the semantic web and the Linked data principles introduced to guide its realization have become the driving theoretical framework of recent developments in bibliographic metadata (van Hooland and Verborgh, 2014; Willer and Dunsire, 2013). This framework promotes interoperability through the establishment of a global network of metadata, facilitated by the use of standards (Berners-Lee, 2006; Hyland *et al.*, 2014). Such standards may be more or less technical and tailored to support the network structure, such as the Resource Description Framework (RDF), or they may be ontologies that reflect the conceptual structures of the entities and relationships constituting a particular domain. Although an RDF-like network is part of the present research design,

as described in section 5, the main concerns of this study are ontologies and their conceptual building blocks<sup>3</sup>.

### **3.1 Conceptualizations**

From an information science perspective, Gruber (1993, p. 199) defined an ontology as an “explicit specification of a conceptualization”<sup>4</sup>. Smith (2004, p. 161) elaborated on the implications of such conceptualizations: “As we engage with the world from day to day, we participate in rituals, and we tell stories. We use information systems, databases, specialized languages, and scientific instruments. [...] Each of these ways of behaving involves, we can say, a certain conceptualization. What this means is that it involves a system of concepts in terms of which the corresponding universe of discourse is divided up into objects, processes, and relations in different sorts of ways. [...] Tools can be developed to specify and to clarify the concepts involved and to establish their logical structure”.

This notion of conceptualizations corresponds to what is often referred to as *mental models* in human–computer interaction, usability, and other related research fields. Norman (2013, p. 25), a leading proponent of this tradition, defined mental models as “the conceptual models in people’s minds that represent their understanding of how things work. [...] People create mental models of themselves, others, the environment, and the things with which they interact”. Theories on mental models derive from psychology, but since the 1940s, they have gradually been subjected to different interpretations in a number of fields (Westbrook, 2006). Like previous studies on users’ internal models of bibliographic structures (e.g. Pisanski and Žumer, 2010a), this present study is based on Norman’s (2013) perspective on mental models. This perspective is related to ontologies and principles underlying the development of modern bibliographic standards.

In this perspective, to improve usability for, say, data consumers who need to understand and use a particular Linked data set in a local system, it is imperative to model the Linked data set in a way that reflects the common conceptualizations shared among the potential data consumers. Ontologies can limit or enable the information architecture of end-user interfaces by providing rich, granular, simple, or shallow data structures. In such cases, ontologies should be based on an idea of how a generic user of those interfaces conceptualizes the entities in the given universe of discourse.

In bibliographic model development, designers often model users as stereotypes by defining user tasks or use cases<sup>5</sup>. These show the commitment of the development process to facilitating the user experience but often assume specific structures. Take, for example, a use case relevant to the research question of this study from the development

---

<sup>3</sup> Regarding the nuances between a conceptual model and a data model, see, for example, Coyle (2017).

<sup>4</sup> Studer *et al.* (1998) developed Gruber’s (1993) ontology definition by stating that the conceptualizations should be *shared*.

<sup>5</sup> An example of the bibliographic extension for the schema.org vocabulary: [https://www.w3.org/community/schemabibex/wiki/Use\\_Cases](https://www.w3.org/community/schemabibex/wiki/Use_Cases); an example of the Linked Data for Libraries model: <https://wiki.duraspace.org/display/ld4l/LD4L+Use+Cases>.

of the BIBFRAME model, defined as “Broadening Search (Discover Adaptations of Work)” (Library of Congress, 2013):

“Sue’s research paper for Classics 201 must identify the themes in Homer’s *Odyssey* as they relate to present day. She has a paperback copy of the book, but thinks that various adaptations of the Work (Movies, Plays, translations, etc.) would help with her research. In order to do this, she first finds the Work associated with the specific Instance she has in hand. From there, she explores the various relationships to other related Works.”

The entity types *instance* and *work*, which represent a certain conceptualization of the bibliographic universe, are considered part of Sue’s mental model. From the use case, it is clear that the implementation of these entities are preconditions to solve the task. The literature on bibliographic organization (see next section) discuss vast numbers of specific entities. However, this study examines mental models and users’ conceptualizations as independent as possible from such constructs.

### ***3.2 Derivative Constellations in the Bibliographic Universe***

The universe of discourse examined in this paper is the bibliographic universe, defined as a concept space containing all recorded knowledge (Smiraglia, 2014, p. 10). Bibliographic entities refer to instances of that recorded knowledge (e.g., a novel, movie, or piece of music). Such instances can be grouped in multi-entity models reflecting their (dis)similarities (Baker *et al.*, 2014). This implies that in addition to the single entities representing a novel or a movie, there are more complex entities bringing their variations together. The mentioned W/E/M entities, for example, bring together different expressions or manifestations of a particular work. Svenonius (2000, p. 35) provided a similar grouping based on sets, including

- “The set of all documents sharing essentially the same information (work),
- The set of all documents sharing the same information (edition),
- The set of all documents descended from a common origin (superwork)” (Svenonius, 2000, p. 35)

Elaborating on superworks, Svenonius (2000, p. 38) explained that they *collocate* (a term adopted in the following analysis) works that are similar “by virtue of emanating from the same ur-work”. As a concrete example of a *Hamlet* superwork, she listed collocated works such as the “original text, motion pictures, sound recordings of readings, analyses of the play, commentaries, playbills, derivative works like *Rosencrantz and Guildenstern Are Dead*” (Svenonius, 2000, p. 38). Svenonius (2000) also commented that a superwork can serve as an interesting tool for effective navigation. Smiraglia (2007) discussed *the bibliographic family*, a similar concept introduced by Wilson (1968). A bibliographic family collocates kindred works. The family structures are all “unique in the relationship the members bear to the originating work [...] yet distinct patterns occur among the members” (Smiraglia, 2007, p. 74). Smiraglia (2007) described such patterns as different types of derivative relationships that create a network of *instantiations*. An instantiation is “a concrete exemplar of a work as it has appeared at a specific point in the lifetime of the work” (Smiraglia, 2007, p. 83). Others, including Carlyle (1999) and Yee (1994), have also touched on the idea of a high-level collocating device. What all these approaches have in common is the shared premise of a specific starting point: the existence of a first

instantiation of a work that serves as the prime mover or the common ancestor of all the other works that form a family.

An instantiation is a generic term for different types of derivative relationships (Smiraglia, 2007), which were investigated extensively by both Smiraglia and Tillett in the 1990s. Tillett (1991) studied bibliographic relationships in general, including derivative relationships. Smiraglia and Leazer (1999) elaborated on Tillett's examples and definitions and listed seven types of common derivations:

- Simultaneous derivations
- Successive derivations
- Translations
- Amplifications
- Extractions
- Adaptations
- Performances

As mentioned in the background section, the W/E/M structure of the FRBR model encompasses some of these relationships. A successive derivation (e.g., a revised "second edition") can, if the intellectual or artistic content is unaffected, be represented by a new manifestation entity. Changes to the content result in a new expression or even a new work if "a significant degree of independent intellectual or artistic effort is involved" (Riva *et al.*, 2017, p. 20). A translation is widely understood as a new expression entity, whereas an adaptation is considered a new work. Other derivative relationships between works are defined with varying levels of granularity in FRBR-based models. For example, the RDA vocabulary <sup>6</sup> contains 14 specified sub-attributes representing various forms of adaptations, such as "is adapted as a motion picture" (P10085) and "is adapted as a television program" (P10085). The FRBRoo ontology, which harmonizes the original FRBR model with the museum-oriented CIDOC CRM model (LeBoeuf, 2012), includes a complex work concept (F15) that is quite similar to the notion of a superwork or bibliographic family. According to the FRBRoo specification it covers the notion that "The conceptual unity observed across a number of complete sets of signs, which makes it possible to organise publications into 'bibliographic families.' This is modelled as: F15 Complex Work is a F1 Work, and F15 Complex Work R10 has member (is member of) F1 Work" (Working Group on FRBR/CRM Dialogue, 2016, p. 26).

Vukadin (2014) points out that in addition to providing a practical means for collocating bibliographic entities in a superwork set, the FRBRoo F15 complex work concept can be used in cases when it is difficult to identify a common ancestor of the entities. This is common in so-called transmedia works that contain stories taking place in a shared fictional world but are often instantiated simultaneously across multiple media platforms. Such fictional worlds typically are developed through stories referencing the same characters, places, or events within or across authorships. In particular, they are studied in literary and media science as intertexts (in the tradition of Genette, 1997) or as transmedia storytelling (Jenkins, 2006).

---

<sup>6</sup> Resource Description Access (RDA) is the cataloging code developed to replace the existing AACR2 code. RDA has been described as a Linked data vocabulary (<http://www.rdaregistry.info/>).

## 4. PREVIOUS RESEARCH

As pointed out, the existing bibliographic models are based on theoretical constructs, not empirical data reflecting end users' understanding of bibliographic entities. Some researchers, though, have matched users' preferences with the FRBR model or tested interfaces for systems built upon the model. Yee (2005) evaluated the search facilities of four FRBRized catalogs and found that they were designed neither to take into account how existing bibliographic records can exploit the FRBR model nor to understand the original purpose of these records.

Carlyle and Becker (2008) conducted a survey asking if users would accept substitutes of FRBR manifestations, expressions, and related works when searching for known items. Their results showed that differences in manifestation types (e.g., a website or a printed copy) were as important as differences in expressions (e.g., different languages) when evaluating substitutability. Most surprisingly, Carlyle and Becker (2008) found that the participants accepted 30% of the related works suggested as substitutes.

Pisanski and Žumer (2010a, 2010b, 2012) compared how users' mental models of the bibliographic universes matched the W/E/M/I entities of the FRBR model. In their first study (Pisanski and Žumer, 2010a), 30 participants were given cards representing W/E/M/I entities of two books. In the first task, the participants were asked to sort the cards "into at least three groups based on the criterion of concrete/abstract (physical/non-physical) nature" and to name the groups (Pisanski and Žumer, 2010a, p. 649). Card co-occurrence was used to perform cluster analysis, which showed that no "constantly similar mental models" could be found (Pisanski and Žumer, 2010a, p. 655).

In the second task, the participants were asked to create a concept map describing the inter-relation of the cards and, specifically, "what comes out of what" (Pisanski and Žumer, 2010a, p. 655). The authors found that 14 of the 30 participants formed at least one work-expression-manifestation-item (four lengths) chain, and another 10 participants formed at least one chain of three lengths. Only two maps, however, corresponded exactly to the FRBR model.

In the third task (Pisanski and Žumer, 2010b), the participants were shown 11 pairs of items whose members differed in one W/E/M/I entity. The participants ranked the pairs according to their substitutability, and the analysis showed that the rankings matched the pairs' FRBR level. In other words, *items* were considered to be easily substitutable, whereas pairs that differed on the *work* level could not be substituted for one another (Pisanski and Žumer, 2010b).

Pisanski and Žumer (2012) followed up with a study in which the participants (120 students) were asked to select among six graphs representing potential relationships between W/E/M/I entities. The majority of the participants chose the graph representing the FRBR view, which indicates that it was the preferred way of coupling W/E/M/I entities.

A few user studies of library systems with FRBR-inspired interfaces have been conducted. Zhang and Salaba (2009) examined how users succeeded in performing different tasks in three FRBR-inspired catalogs. The users most successfully accomplished tasks that had the target of finding a work. The participants had problems with (in order of increasing difficulty) finding manifestations, identifying manifestations, and obtaining items (Zhang

and Salaba, 2009). Based on these findings, the authors developed a new prototype catalog, which they evaluated against a non-FRBRized catalog. Zhang and Salaba (2009) reported that 85% of the users preferred the FRBR prototype. Users, not surprisingly, performed tasks tailored toward works, expressions, and manifestations better in the FRBR system than the regular catalog.

Merčun and colleagues developed the FrbrVis prototype system (with FRBRized records) and assessed it against a traditional system (without FRBRized records) in two usability studies. In the first study (Merčun et al., 2016), 120 participants were asked to perform specific tasks interacting with bibliographic families representing different levels of complexity; in the second (Merčun et al., 2017), they were free to explore the system. The controlled study found that the FrbrVis prototype performed better than the traditional system, both in general and when taking into account the complexity level of tasks.

To summarize, research investigating how users understand bibliographic universes have mostly used the FRBR model as their point of departure. Conceptually, users generally find different items, manifestations, and expressions of the same work to be substitutable and, to a certain degree, allow related works to be substituted for one another. When asked to map how different FRBR entities are related, users are less consistent but tend to prefer the FRBR model from among the alternatives presented. Some attempts to FRBRize existing records have been made, but evaluations indicate that these projects have been only partially successful. Promising FRBR prototype displays have been developed, and it will be interesting to see whether these can be implemented in future catalogs.

In contrast to previous research, the users in this study are not presented with existing solutions or bibliographic records but, rather, conceptualize bibliographic families based on their own understandings of what the documents' important characteristics are.

## **5 METHOD**

### ***5.1 Concept Mapping***

Concept mapping serves as a method to reveal the bibliographic conceptualizations held by the participants in a study. The literature describes two forms of concept maps: hierarchical concept maps and network concept maps (Ruiz-Primo and Shavelson, 1996). Novak and Cañas (2006) deemed the hierarchical model to be not flexible enough for the purposes of studies such as the present one. Chang (2007, p. 107), who studied novice students' modelling of the homeostasis of blood sugar, concluded that the network concept map "is suitable for knowledge encompassing complex processes or interrelationships". As well, networks, or *graphs*, represent both flexible and sophisticated tools for organizing entities in a cultural heritage context (Murray and Tillett, 2012). Accordingly, the participants in this study were asked to draw a network representing how they view the documents, their essential attributes, and the (derivative) relationships between them. This method does not favor any hierarchical understandings of the bibliographic universe, leaving the participants free to draw any kinds of concepts and relationships.

The instructions for handling entity identification and organization in the concept mapping process may still affect the outcome. The task model should not limit the



elicitation of the conceptualizations by either its complexity or difficulty of application, but neither should it provide a means to represent the conceptualizations too abstractly. In this study, an RDF-like network was used as a guide for the concept maps. Since RDF is based on a graph model with named nodes and directed edges it is a relatively intuitive and expressive guide for a concept mapping task. To provide the participants with a concrete guideline before they drew their concept maps, they were shown an example network conceptualizing an alternative universe of discourse (ships and persons related to the ships). They also received a short introduction explaining the task in detail. Finally, the participants were presented with a contextual purpose: their finalized conceptualizations should mirror what they believe would be a reasonable organization and selection of information for a general multiuser information system “like the ones used in libraries”.

## **5.2 Participants**

The study participants were all first-year bachelor students in library and information science. The task was given in the students’ first lecture on bibliographic metadata. The participants’ competency in this field of research was expected to be low and comparable to that of ordinary users of information systems. The participants also completed a post-task questionnaire, with questions on their gender, age, and previous experience with metadata, cataloguing, and programming related to their education, work, and hobbies. In addition, the participants could comment on the task in a text box.

## **5.3 Documents**

Each participant was given three pieces of paper depicting three documents representing a book, a movie, and a music record. Before the main experiment, a pilot study was performed with five participants. The pilot testers performed the same tasks that we planned to use. Based on the pilot study some adjustments to the introduction were made. Apart from that, the study design remained unchanged. Documents from two different bibliographic families were used. Family *PG* contained:

- 1) The title page (recto and verso) of *Peer Gynt* by Henrik Ibsen, a Norwegian edition from 1962, published by Gyldendal
- 2) The DVD cover (front and back) of the 2006 television adaption of *Peer Gynt* directed by Bentein Baardson and produced by the Norwegian Broadcasting Corporation
- 3) The CD and the liner notes of *Music from the Mountains*, a collection of Peer Gynt suites composed by Edvard Grieg and Harald Sæverud, conducted by Ari Rasilainen, performed by the Norwegian Radio Orchestra and published by Finlandia Records in 1997

Family *RJ* contained:

- 1) The title page (recto and verso) of *Romeo and Juliet* by William Shakespeare translated to Norwegian by André Bjerke and published by Aschehoug in 2000
- 2) The DVD cover (front and back) of the 1996 movie *Romeo + Juliet* directed by Baz Luhrmann and published by Twentieth Century Fox

- 3) The CD and the CD cover (backside) of the *Romeo + Juliet* soundtrack published by Capitol Records in 1996

Documents that represent works belonging to large bibliographic families were selected, increasing the probability that the participants were familiar with them. The documents also represent typical entities that can be found in a library. In order to avoid constraining the tasks for the informants, no representations of the original manuscripts of the plays were included. Smiraglia and Leazer (1999) found that the size of a bibliographic family grows with the age of the progenitor work; therefore, universes that contain relatively old items were selected. *Peer Gynt* was written in 1867, while *Romeo and Juliet* was first printed in 1597. These two works have given rise to a great variety of creative inspirations and interpretations, so their accumulated bibliographic entities realized in a variety of media platforms have contributed to shared fictional worlds where stories unfold based on (or at least referencing) a set of given characters, places, and events.

The two families contained similar but not identical relationships between the documents. They both contained a play presenting a version of the original work. They also contained a movie and a musical record. In the PG family, the movie and the music represented independent adaptations. In the RJ family, the movie was an adaptation of the play, but the music contained already-published songs by different artists collected as a soundtrack for the movie. It was therefore less connected to the original play. The differences in the document families and relationships were incorporated into the research design to control for these variables in the experiment. In the following, *movie* represents the DVDs containing the movies, *book* represents the books containing the plays, and *music* represents the CDs containing the musical recordings.

## 6 ANALYSIS

### 6.1 Cluster Analysis

A total of 107 participants was recruited for the experiments. Their concept maps were interpreted and encoded by two researchers in two iterations. The first iteration provided an overview of the maps' common characteristics, such as the main nodes and the relationships between them. Eight concept maps could not be further analyzed due to a lack of identifiable or interpretable attributes. The remaining 99 concept maps were drawn according to the task instructions. They all contained a minimum of three nodes that could be identified as representations of the three documents from the handouts and the relationships connecting them directly or indirectly. The nodes were depicted as named circles or boxes, relationships as arrows or lines. Many relationships were named. Document nodes were identified as those being related to a minimum number of attributes, such as title, publisher, publication year or carrier/expression type (see Section 6.5 and 6.6 for details). In addition, indirect relationships between such document nodes were often formalized through a *central node*, as in the example concept map shown in Figure 1.

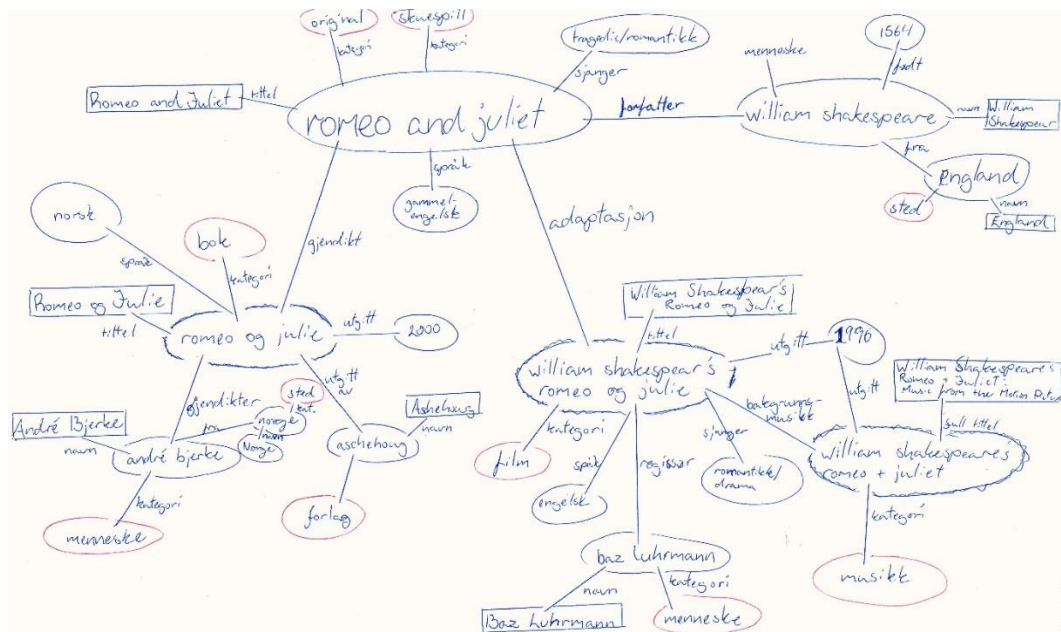
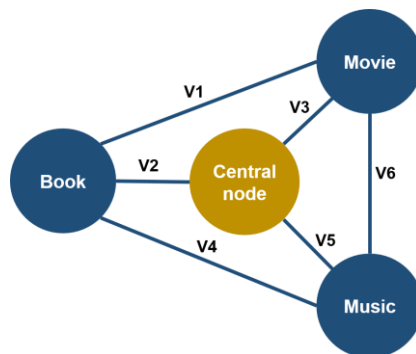


Figure 1 A representative concept map from the RJ family.

In the second iteration, the relationships between the document nodes were encoded in a spreadsheet as present or absent based on the criteria described above. To identify and group common relationship models, a hierarchical cluster analysis of six binary variables representing the identified relationships between the main nodes (Figure 2) was performed. In addition to the document nodes, the cluster analysis included the central node among the main nodes.

Cluster analysis offers a set of methods for grouping objects based on their characteristics and structures already present in data (Kaufman and Rousseeuw, 2009). Specific methods are chosen based on the types of variables (e.g., interval scaled, nominal, or binary). In order to perform the cluster analysis one need an operation to calculate the dissimilarities between objects and one to cluster the results. The well-known, simple matching coefficient (Sokal and Michener, 1958) was used for the (symmetric) binary data to develop a distance matrix, and the average linkage method (from the hclust package in R<sup>7</sup>) was utilized to build hierarchies.



<sup>7</sup> <https://stat.ethz.ch/R-manual/R-devel/library/stats/html/hclust.html>.

Figure 2 The main nodes identified, the concept maps, and the relationships between them treated as binary variables (present/not present) in the cluster analysis.

The results of the cluster analysis are visualized as a dendrogram, shown in Figure 3.

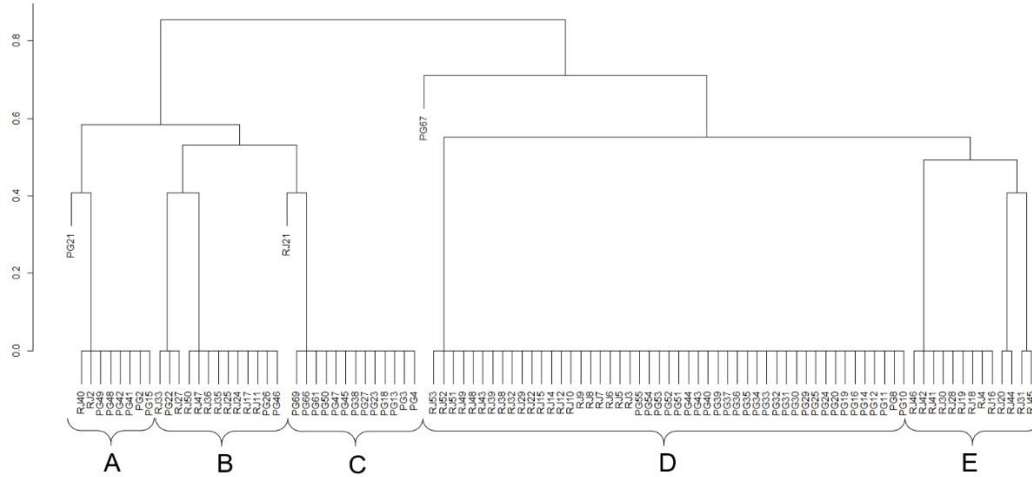


Figure 3 Dendrogram showing the results of the cluster analysis, with two main clusters and five sub-clusters identified.

The results of the cluster analysis reveal that the concept maps mostly belong to two clusters with five sub-clusters (A-E). Table 1 shows the distribution of the concept maps in each sub-cluster, the two main clusters, and each document universe. The most notable difference between the two main clusters is whether they include a central node for handling relationships. The 62 concept maps clustered in clusters D and E all include such a node; the universes belonging to the three other clusters (A, B, and C) do not. One concept map, placed between the C and D clusters in the dendrogram, is an outlier with a unique combination of relationships. In the following analysis, the attributes characterizing the five sub-clusters are examined. The outlier conceptualization is considered so atypical that it is removed from the statistics. Thus, 98 concept maps are included in the examinations. The analysis of the common properties in the various sub-clusters examines the directions of the relationships, primarily based on explicit naming (e.g., “adaptation of”, “version”, and “belongs to”) but also other expressed features indicating direction (e.g., arrows).

Cluster	PG family	RJ family	Total	Main clusters in %
A	7	2	9	37%
B	3	10	13	
C	13	1	14	
D	28	21	49	63%
E	0	13	13	
<b>Total</b>	<b>51</b>	<b>47</b>	<b>98</b>	<b>100%</b>

Table 1 Distribution of concept maps by cluster. Clusters A, B, and C contain central nodes; clusters D and E do not.



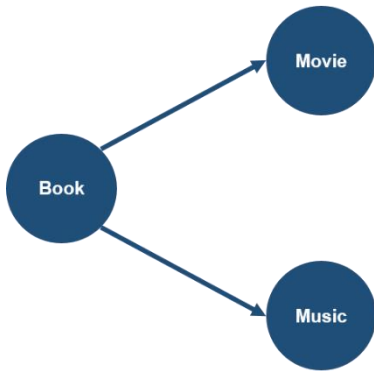


Figure 5 Cluster C with an example of a concept map. The document nodes are directly related from the book to the movie and the music.

**6.3 Cluster A: Document Nodes with Relationships Based on Shared Characteristics**

The document orientation of the concept maps in cluster A are similar to those in clusters B and C. What distinguishes the maps in cluster A is the lack of derivative relationships between the documents. The documents are instead linked indirectly via shared characteristics, such as authors, dates, genres, or topics. The relationships between the documents seem more arbitrary, as illustrated with dotted lines in Figure 6.

Seven concept maps in this cluster describe the PG family, while three describe the RJ family. No particular characteristics that can explain the skewed distribution are identified.

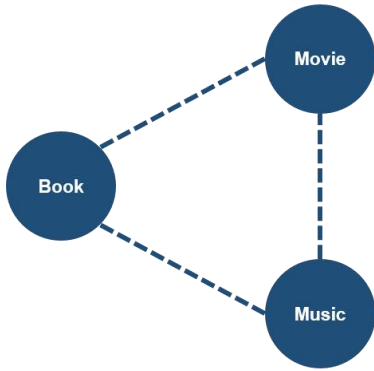


Figure 6 Cluster A with an example of a concept map. The document nodes are related via shared characteristics.



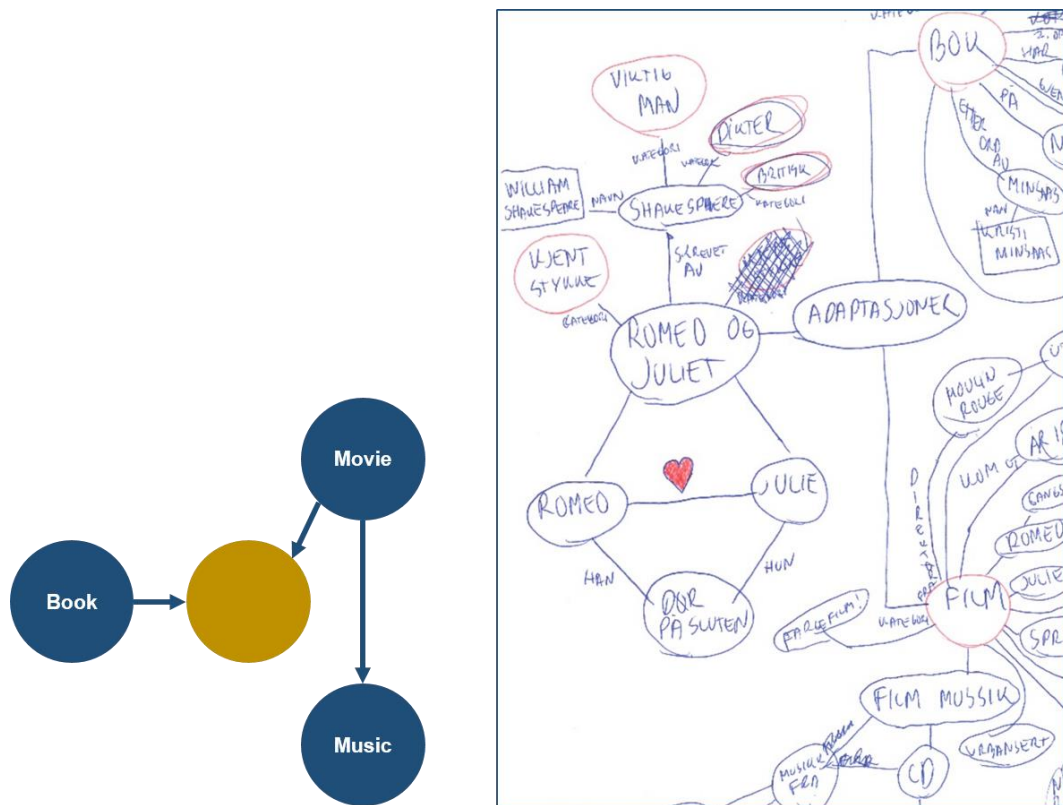


Figure 8 Cluster B with an example of a concept map. The document nodes are partly related indirectly via the central node (the book and the movie) and partly related directly (the movie and the music).

### 6.5 Attributes

Altogether, 72 different attributes, or descriptive characteristics of the documents, were identified. The concept maps each contained 18 attributes on average. The nature of the supplied material likely was a contributing factor to which attributes the informants included. For example, visually clear attributes (e.g., a publisher presented in a large font) were included in the concept maps more frequently than visually weaker ones. Although this study was more concerned with the overarching structures than the details of the attributes, the representations of the three most common attribute types were examined. Table 2 shows the distribution of the attributes across the clusters. Due to the different genres of the two music documents, the responsible composer for the music document in the PG family and the artists in the music document in the RJ family were included.

		A (n=9)	B (n=13)	C (n=14)	D (n=49)	E (n=13)	All models
Title	Book	89%	85%	79%	37%	46%	55%
	Movie	100%	92%	57%	41%	54%	57%
	Music	100%	77%	64%	59%	54%	66%
	Central node				100%	100%	63%
Responsibility	Book (author)	100%	100%	100%	41%	46%	64%
	Movie (director)	89%	77%	50%	55%	77%	63%
	Music (componist/artist)	78%	69%	79%	67%	38%	67%



	Central node (author)				69%	85%	46%
Date of publication	Book	78%	77%	71%	73%	85%	77%
	Movie	56%	92%	64%	69%	85%	73%
	Music	78%	54%	64%	76%	69%	71%
	Central node				4%	0%	2%

Table 2 Distribution of key attributes across the clusters.

Table 2 shows that many concept maps in clusters D and E have title and responsibility attributes directly linked to the central node; this is not the case for the date of publication. The date of publication is a typical manifestation attribute (in FRBR terminology), and the analysis reveals that the concept maps in the central-node clusters mostly attach these attributes to the document node.

Table 3 presents the distribution of different attribute types for the central node. Of the concept maps in clusters D and E, 73% have an author related to the central node, whereas 31% have a genre related to it. Only a few concept maps have a date of origin or an original language (the latter applies solely to concept maps in the RJ family) related to the central node. Of the concept maps, 40% have central nodes related to various fictional characters (e.g., “Mor Åse” and “Juliet”), while 15% have other attributes from the fictional world of the relevant documents, such as places (“Verona”) and events (“The death of Romeo by poison”).

	% of central node conceptualizations
Author	73%
Date of origin	7%
Original language	7%
Genre	31%
Related fictional characters	40%
Related fictional places or events	15%

Table 3 Attributes related to the central node.

## 6.6 Naming

Beyond a general request to make the nodes interpretable, the task instructions gave the participants no specific guidance on how to name the nodes. Examining the concept maps found that this creative freedom yielded additional insights into the conceptualizations. The central nodes are exclusively named “Peer Gynt” or “Romeo and Juliet”. Such a naming practice was interpreted to indicate, or at least to originate from, a title. The naming of the document nodes is somewhat more complex. In addition to the use of document titles, two other sources of names are identified: carrier and expression types. The carrier category includes names that specify a carrier device, such as a CD or DVD. The expression category contains content or media types including names such as “text”, “music” and “video”. Table 4 shows the distribution of the naming categories across models and document types. For all document types and models, on average, 58% of the participants name their document nodes with a title, 34% a expression type, and 6% a

carrier type. A closer look at the distribution across the different models reveals a dominant tendency: concept maps belonging to the central-node-only cluster (cluster D) include fewer titles and more carrier and expression types than the other clusters of concept maps. The concept maps in cluster A have the highest frequency of titles, while the other non-central-node-clusters (B and C) also include more titles than the central-node clusters.

Models	Title	Carrier	Expression
A (n=9)	82%	0%	19%
B (n=13)	77%	8%	10%
C (n=14)	64%	2%	33%
D (n=49)	20%	16%	57%
E (n=13)	46%	5%	49%
Average (n=98)	58%	6%	34%

Table 4 Types of the names of document nodes across models.

## 6.7 Participants

The results from the post-task questionnaire show no significant differences in the gender or average age of the participants creating the concept maps across the clusters. Overall, 25% of the participants reported that they had some prior experience with metadata, which seems to have influenced their conceptualizations. In cluster A, 60% of the participants reported that they had previous metadata experience, whereas only 17% of the participants with concept maps in cluster D did so. The other clusters had 20%–30% participants with prior experience, similar to the total average. An interesting possible explanation may be found in the cataloguing tradition of the Norwegian library sector, where the participants most likely gained their experience. In Norway, cataloguers are trained to catalogue documents according to standards (AACR2 and MARC) that mandate few relationships representing derivations between documents. This document orientation may have influenced the arbitrary relationships in the conceptualizations found in cluster A. Moreover, the central nodes found in cluster D concept maps created by participants with at least some experience are very different from the conceptualizations mandated by the current standards.

## 6.8 Main Findings

This study was intended to examine conceptualizations of derivative relationships. Cluster analysis of the relationships between the main nodes in the concept maps resulted in five clusters. Two clusters (D and E, representing 63% of the concept maps) include a central node used to relate all or some of the document nodes. In the other clusters, the document nodes are related directly (clusters B and C) or indirectly via shared characteristics (cluster A). Cluster A thus represents a significant document orientation which does not include the derivative relationships between the documents.

Statistical analysis of the attributes and naming of the nodes confirmed the identified clusters. The concept maps without a central node tend to have titles as the names of document nodes, whereas clusters with a central node tend to use the names of document nodes to explicitly identify the type of expression or carrier the documents represent (e.g., “video” or “music”). Clusters with a central node tend to have persons of

responsibility related to the central node but provide other attributes at the document level, such as the date of publication. Many concept maps belonging to the central-node clusters also relate to the central node information from the fictional world to which the documents belong, such as related fictional characters, places, and events.

If the concept maps are considered expressions of the participants' conceptualizations, the findings suggest that the participants hold conceptualizations that:

- relate documents solely via shared characteristics (cluster A)
- relate documents directly (clusters B and C)
- relate documents through a central node (cluster D)
- combine a central node with direct relationships between the documents (cluster E)

Regarding bibliographic modelling, two different approaches to conceptualizing the entities and relationships of the bibliographic universe are identified. The document-oriented nodes and relationships in clusters A, B, and C can be generalized into a single-entity model, with the documents themselves at the center; the book is "a book". Clusters D and E, in contrast, introduce a level of abstraction with their central nodes and indicate a multi-entity model; the book can be differentiated into several entities reflecting its meaning, expression, and physicality (Baker *et al.*, 2014).

Based on these groups of concept maps, a spectrum can be established (Figure 9), ranging from document-oriented conceptualizations constituting a single-entity model to conceptualizations with relationships handled by an entity representing an abstraction of the documents, constituting a multi-entity model.

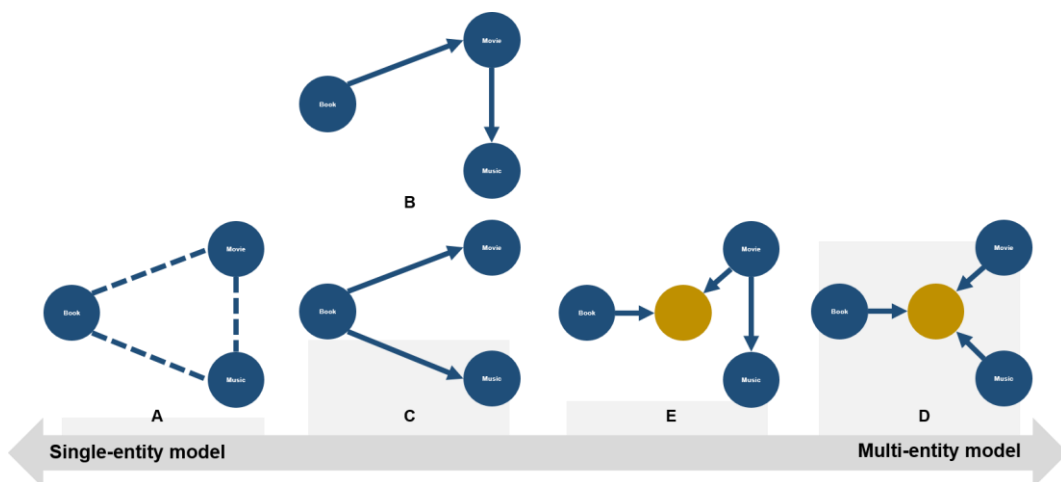


Figure 9 The five clusters along a spectrum from single-entity to multi-entity conceptualizations. The bars in the background indicate the amount of conceptualizations distributed across the spectrum

Conceptualizations with (single) entities that contain all the attributes of the documents describing both the formal characteristics and the aspects of their content and functionality can be found on the left side of the spectrum. "Nothing" appears to exist outside those entities. On the right side of the spectrum, conceptualizations with multiple entities differentiated by their varying views of the document attributes can be found. In

these conceptualizations, the attributes appear to float more freely in a concept space that can be organized according to different views.

As explained in Section 5.3, slightly different document families were used to detect any influences on the conceptualizations. Some of the resulting clusters clearly relate to a particular document family and illustrate that the families, to some extent, do influence characteristics, such as the direction of relationships (clusters B and C) and the semantics of the central node (cluster E). The analysis, however, reveals that the variations are equally distributed among the clusters. Clusters B and C are nearly the same size and are considered to represent the same conceptualization. Cluster E, which is interpreted to represent an independent conceptualization, contains only RJ concept maps. The concept maps in cluster E handle the music node (the soundtrack of the movie adaptation) in a way (excluding it from the central node entity) that would have been unlikely in the context of the PG family documents, where the music node represents an independent adaptation of the Ibsen play. It, therefore, is assumed that some participants in the PG group with concept maps in the D cluster would have made concept maps belonging to the E cluster if they were in the RJ group.

Considering the number of attributes included, relatively large differences exist between the families. Whereas, for instance, 87% of the participants include a genre in the PG group, only 32% include a genre in the RJ group. In the case of actors, the distribution is 48% and 72%, respectively. The differences are most likely due to the graphical presentations in the supplied document representations (e.g., font size and color) and the participants' greater familiarity with some actors than others.

These figures show quite clearly that the document universes are interpreted differently when it comes to specific information but similarly in terms of the expressed high-level relationships between the document nodes.

## 7 DISCUSSION

As described in section 4, previous research has been concerned with verifying FRBR structures. Although the present study was designed to avoid an initial influence from particular bibliographic models, it is interesting to reflect on the findings in light of bibliographic entities, as they are outlined in section 3.2. In particular, the significant presence of a central node that signals a form of abstraction leads toward multi-entity structures like the W/E/M entities. In cluster E, two documents (the movie and the book on which the movie is based) are related through a collocating central node. The third document node (the music document not directly based on the book but strongly related to the movie) is related not to the central node but directly to the movie node. Hence, in the conceptualizations of cluster E, the central node organizes two cultural artifacts with strongly related content but excludes a third that is obviously related but also has quite different content. This recalls the logic behind W/E/M entities that conceptually collocates and separates varieties of cultural products based on similarities in content. Although cluster D does not directly contain ordinary W/E/M-entities, several of its factors are also reminiscent of FRBR logic. According to FRBR (e.g., in the form of LRM relationships between an agent and its W/E/M entities; Riva *et al.*, 2017, pp. 66–67),

information about primary responsibility “for the creation of the intellectual or artistic content” is linked to the work entity, while the translators or the record company is linked to the expression or the manifestation. Information about responsibility for elements other than the content mostly is linked directly to the document nodes of cluster D. Thus, in the conceptualizations of cluster D, an FRBR-like chain of entities which includes almost half of the concept maps can be sensed.

Of the concept maps in clusters D and E, 73% have Shakespeare or Ibsen (the main persons of responsibility in the bibliographic families) related to the central node. However, it is quite apparent from the document representations that neither Ibsen nor Shakespeare has any responsibility for the content of the music node, which is part of the entity represented by a central node in cluster D. Thus, the central node cannot entirely be interpreted as an FRBR work. Some participants may have used the central node as a representation of the original work, although only a few participants included attributes like the original language and the original year of publication (see Section 6.5).

Perhaps it is also plausible to interpret the central node as an even more abstract collocating device. Instead of being responsible for all the documents attached to the central node, Ibsen and Shakespeare are linked to the central node because they are responsible for the originator works of the bibliographic families. From this perspective, the central node more resembles a superwork entity. This may also be the case for cluster E conceptualizations. A movie adaptation or a dramatization of a text is usually interpreted as a new work within FRBR, so the collocation of the book and the movie node is perhaps more accurately understood as a superwork function, collocating works belonging to a larger bibliographic family.

In addition, the study demonstrates that the central nodes are attached to much information belonging to an even more abstracted level: the fictional world related to the content of the documents. Thus, at least three types of conceptual abstractions can be drawn from the analysis of the central nodes:

- FRBR work
- Superwork
- Fictional world

The FRBR work has only a weak presence, though. Nevertheless, this study shows that for the purpose of expressing derivative relationships, users seem to prefer multi-entity conceptualizations including a superwork entity or characteristics of the relevant fictional world. It is also worth mentioning that 27% of the participants include direct derivative relationships between document nodes in their concept maps.

Another interesting question for this research field is whether any existing models reflect the present findings. Clarke (2015) claims that the current framework for bibliographic development based on semantic-web and Linked-data principles is progressing towards a

new way of conceptualizing data. In this perspective, metadata are not necessarily exchanged as “units that include all the bibliographic information about a resource together in one place [...] like a MARC record”, as in traditional cataloging, but also as limited individual statements (RDF-triples) “from multiple locations” (Clarke, 2015, p. 300). This requires bibliographic models that contribute to and enable flexible, complex semantics on different levels of abstraction. Several prominent libraries have attempted to facilitate such exchange and interoperability by publishing their bibliographies online as Linked data based on application profiles that include FRBR entities (Tallerås, 2017). Although the present findings show that some participants hold document-oriented conceptualizations reminiscent of traditional cataloging, the majority applies models dividing the universe into different levels of abstraction.

Emerging models, such as RDA, and Linked-data-based bibliographic models, such as BIBFRAME, also enable the modelling of FRBR works. The superwork level may be inferred from the explicitly expressed derivative relationships between works, but no models dedicate a conceptual entity at the superwork level, with the prominent exception of FRBRoo. Regarding the fictional-world level of abstraction, FRBRoo certainly can be used to collocate relevant works but does not provide any sophisticated semantics for expressing such relationships.

The identified conceptualization thus, to some extent, can be realized through existing models, but it appears that few utilize the available opportunities in current systems. Previous research has also mostly been concerned with the established FRBR levels, both in terms of user verification and the information architecture in user interfaces. Based on the present findings, more attention should be paid to the more high-level abstractions of superworks and fictional worlds.

## **8 LIMITATIONS AND FURTHER RESEARCH**

Although the distinction between the two main clusters of drawings is evident, some of the five subclusters are rather small. Further research with more participants is therefore necessary to provide deeper insight.

The participants were not observed during or interviewed after the concept mapping task. The knowledge about their motivations and strategic decisions during development of their concept maps is therefore limited. An alternative research design would have been to observe and interview a smaller group of participants.

The two bibliographic families used in the study have similar origins, in that both stem from a play. Thus, they represent a specific selection of documents. It is necessary with additional studies including more and other types of documents to improve the understanding of derivative relationships in bibliographic universes in general. Other documents, such as a translated text presented together with a representation of the text in its original language, probably would provide more insight into user conceptualizations of bibliographic structures. To compare the results with those of previous FRBR-oriented research, documents with such relationships should be prominently incorporated in the design of future research.

The informants were all library and information science students, and therefore do not represent a general population. The students, however, were in their second week of the first semester in a bachelor degree and had no previous formal training in bibliographic metadata standards. One fourth of the students, on the other hand, did report some degree of previous experience with metadata (as discussed in section 6.7).

## 9 CONCLUDING REMARKS

A bibliographic universe comprises “the totality of things over which bibliographical control is or might be exercised” (Wilson, 1968, p. 6). Wilson (1968) further described two distinct kinds of control. Descriptive control provides the means, traditionally by cataloging, to create (arbitrary) lists that enable retrieval of all the entities characterized by certain attributes (“All plays by Ibsen”). Exploitative control, in contrast, is the ability to procure the best entities available serving a specific purpose. The first kind of power is evaluative neutral; the second involves appraisal by the user (Wilson, 1968, p. 22). According to Wilson (1968), exploitative control is more important, but descriptive control is a precondition for achieving exploitative control; to identify the best entities, these entities must be known, and to be known, they must be described. The same is true for the gravitational forces in the bibliographic universe manifested by relationships between entities. The present study shows that users conceptualize such relationships quite differently. Some utilize attributes that describe the shared characteristics of the documents—the traditional apparatus of descriptive control. Others directly relate documents by applying accurate derivative links. The majority of participants, however, applies a multi-entity model in which document entities are related through nodes at a higher level of abstraction describing the characteristics of a bibliographic family or a shared fictional world. Such information is essential to exercise exploitative control in an ever-expanding bibliographic universe containing the storylines of transmedia franchises and the derivative accumulations of popular bibliographic families.

Today, one perhaps could argue that the problem of descriptive control has more or less been solved, especially given the provision of digitized content enabling the automated generation of descriptions. Existing bibliographic ontologies provide means to describe complex derivative relationships. However, the ability to exploit these descriptions—to exercise exploitative control—is still an open problem and a holy grail for the world’s leading libraries, search engines, and recommender systems.

## REFERENCES

- Baker, T., Coyle, K. and Petiya, S. (2014), “Multi-entity models of resource description in the Semantic Web”, *Library Hi Tech*, Vol. 32 No. 4, pp. 562–582.
- Berners-Lee, T. (2006), *Linked Data: Design Issues*, W3C, available at: <http://www.w3.org/DesignIssues/LinkedData.html> (accessed 7 November 2012).
- Carlyle, A. (1999), “User categorisation of works: toward improved organisation of online catalogue displays”, *Journal of Documentation*, Vol. 55 No. 2, pp. 184–208.
- Carlyle, A. and Becker, S.R. (2008), “ASIS&T 2008 annual meeting poster ‘FRBR and the “known-item” search.’”, *Proceeding of the American Society for Information Science and Technology*, Vol. 45 No. 1, pp. 1–9.

- Chang, S.-N. (2007), "Externalising students' mental models through concept maps", *Journal of Biological Education*, Vol. 41 No. 3, pp. 107–112.
- Clarke, R. (2015), "Breaking records: The history of bibliographic records and their influence in conceptualizing bibliographic data", *Cataloging & Classification Quarterly*, Vol. 53 No. 3–4, pp. 286–302.
- Coyle, K. (2016), *FRBR, before and after: a look at our bibliographic models*, American Library Association, Chicago.
- Coyle, K. (2017), "Two FRBRs, many relationships", *Coyle's InFormation*, available at: <http://kcoyle.blogspot.no/2017/05/two-frbrs-many-relationships.html> (accessed 5 October 2017).
- Cutter, C. (1904), "Rules for a dictionary catalog", *The US Bureau of Education Special Report on Public Libraries-Part II*.
- Genette, G. (1997), *Paratexts: Thresholds of interpretation*, Cambridge University Press, Cambridge.
- Gruber, T.R. (1993), "A translation approach to portable ontology specifications", *Knowledge Acquisition*, Vol. 5 No. 2, pp. 199–220.
- van Hooland, S. and Verborgh, R. (2014), *Linked Data for Libraries, Archives and Museums: How to clean, link and publish your metadata*, Facet publishing, London.
- Hyland, B., Ateazing, G. and Villazón-Terrazas, B. (2014), "Best Practices for Publishing Linked Data", *W3C Working Group Note*, available at: <https://www.w3.org/TR/ld-bp/> (accessed 5 October 2017).
- Jenkins, H. (2006), *Convergence culture : where old and new media collide*, New York University Press, New York.
- Kaufman, L. and Rousseeuw, P. (2009), *Finding groups in data: an introduction to cluster analysis*, Wiley, Hoboken, NJ.
- LeBoeuf, P. (2012), "A strange model named FRBROO", *Cataloging & Classification Quarterly*, Vol. 50 No. 5–7, pp. 422–438.
- Library of Congress. (2013), "BIBFRAME Use Cases and Requirements", available at: <http://bibframe.org/documentation/bibframe-usecases/> (accessed 22 September 2017).
- Merčun, T., Žumer, M. and Aalberg, T. (2016), "Presenting bibliographic families", *Journal of Documentation*, Vol. 72 No. 3, pp. 490–526.
- Merčun, T., Žumer, M. and Aalberg, T. (2017), "Presenting bibliographic families using information visualization: Evaluation of FRBR-based prototype and hierarchical visualizations", *Journal of the Association for Information Science and Technology*, Vol. 68 No. 2, pp. 392–411.
- Murray, R.J. and Tillett, B.B. (2012), "Cataloging theory in search of graph theory and other ivory towers", *Information Technology and Libraries*, Vol. 30 No. 4, pp. 170–184.
- Norman, D. (2013), *The design of everyday things: Revised and expanded edition*, Basic Books, New York, NY.



- Novak, J.D. and Cañas, A.J. (2006), "The origins of the concept mapping tool and the continuing evolution of the tool", *Information Visualization*, Vol. 5 No. 3, pp. 175–184.
- Pisanski, J. and Žumer, M. (2010a), "Mental models of the bibliographic universe. Part 1: mental models of descriptions", *Journal of Documentation*, Vol. 66 No. 5, pp. 643–667.
- Pisanski, J. and Žumer, M. (2010b), "Mental models of the bibliographic universe. Part 2: comparison task and conclusions", *Journal of Documentation*, Vol. 66 No. 5, pp. 668–680.
- Pisanski, J. and Žumer, M. (2012), "User verification of the FRBR conceptual model", *Journal of Documentation*, Vol. 68 No. 4, pp. 582–592.
- Riva, P., Bœuf, P. Le and Žumer, M. (2017), "IFLA Library Reference Model", The International Federation of Library Associations and Institutions, Den Haag, available at: [https://www.ifla.org/files/assets/cataloguing/frbr-lrm/ifla\\_lrm\\_2017-03.pdf](https://www.ifla.org/files/assets/cataloguing/frbr-lrm/ifla_lrm_2017-03.pdf).
- Ruiz-Primo, M. and Shavelson, R. (1996), "Problems and issues in the use of concept maps in science assessment", *Journal of Research in Science*, Vol. 33 No. 6, pp. 569–600.
- Smiraglia, R.P. (2007), "Bibliographic families and super works", in Taylor, A.G. (Ed.), *Understanding FRBR: What It Is and How It Will Affect Our Retrieval Tools*, Libraries Unlimited, Westport, CT, pp. 73–86.
- Smiraglia, R.P. (2014), *The elements of knowledge organization*, Springer, Cham.
- Smiraglia, R.P. and Leazer, G.H. (1999), "Derivative Bibliographic Relationships: The Work Relationship in a Global Bibliographic Database.", *Journal of the American Society for Information Science*, Vol. 50 No. 6, pp. 493–504.
- Smith, B. (2004), "Ontology", in Floridi, L. (Ed.), *The Blackwell Guide to the Philosophy of Computing and Information*, Blackwell publishing, MA.
- Sokal, R.R. and Michener, C.D. (1958), "A statistical method for evaluating systematic relationships", *University of Kansas Science Bulletin*, available at: <https://doi.org/citeulike-article-id:1327877>.
- Svenonius, E. (2000), *The Intellectual Foundation of Information Organization*, MIT Press, Cambridge, Massachusetts.
- Tallerås, K. (2017), "Quality of linked bibliographic data: the models, vocabularies, and links of data sets published by four national libraries", *Journal of Library Metadata*, Vol. 17 No. 2, pp. 126–155.
- Tillett, B. (1991), "A taxonomy of bibliographic relationships", *Library Resources and Technical Services*, Vol. 35, pp. 150–158.
- Tillett, B.B. (2001), "Bibliographic relationships", in Bean, C.A. and Green, R. (Eds.), *Relationships in the Organization of Knowledge*, Springer Netherlands, pp. 19–35.
- Vukadin, A. (2014), "Bits and pieces of information: bibliographic modeling of transmedia", *Cataloging & Classification Quarterly*, Vol. 52 No. 3, pp. 285–302.
- Westbrook, L. (2006), "Mental models: a theoretical overview and preliminary study",

- Journal of Information Science*, Vol. 32 No. 6, pp. 563–579.
- Willer, M. and Dunsire, G. (2013), *Bibliographic information organization in the semantic web*, Chandos Publishing, Oxford.
- Wilson, P. (1968), *Two kinds of power: An essay on bibliographical control*, University of California Press, Berkeley, CA.
- Working Group on FRBR/CRM Dialogue. (2016), “Definition of FRBROO: A conceptual model for bibliographic information in object-oriented formalism”, The International Federation of Library Associations and Institutions, Den Haag, available at: [https://www.ifla.org/files/assets/cataloguing/FRBROO/frbroo\\_v\\_2.4.pdf](https://www.ifla.org/files/assets/cataloguing/FRBROO/frbroo_v_2.4.pdf) (accessed 5 October 2017).
- Yee, M.M. (1994), “The Concept of Work for Moving Image Materials”, *Cataloging & Classification Quarterly*, Vol. 18 No. 2, pp. 33–40.
- Yee, M.M. (2005), “FRBRization: a method for turning online public finding lists into online public catalogs”, *Information Technology and Libraries*, Vol. 24 No. 3, pp. 77–95.
- Zhang, Y. and Salaba, A. (2009), “What Is next for functional requirements for bibliographic records? A delphi study”, *The Library Quarterly*, Vol. 79 No. 2, pp. 233–255.

## Paper C

Tallerås, K. (2013). From many records to one graph: Heterogeneity conflicts in the Linked Data restructuring cycle. *Information Research*, 18(3).



Proceedings of the Eighth International Conference on  
Conceptions of Library and Information Science,  
Copenhagen, Denmark, 19-22 August, 2013

**From many records to one graph: heterogeneity conflicts  
in the linked data restructuring cycle**

[Kim Tallerås,](#)

Oslo and Akershus University College,  
Department of Archive Studies, Library and  
Information Science, Lilleborggatan. 3, 04480  
OSLO, Norway

Abstract

**Introduction.** During the last couple of years the library community has developed a number of comprehensive metadata standardization projects inspired by the idea of linked data, such as the BIBFRAME model. Linked data is a set of best practice principles of publishing and exposing data on the Web utilizing a graph based data model powered with semantics and cross-domain relationships. In the light of traditional metadata practices of libraries the best practices of linked data imply a restructuring process from a collection of semi-structured bibliographic records to a semantic graph of unambiguously defined entities. A successful interlinking of entities in this graph to entities in external data sets requires a minimum level of semantic interoperability.

**Method** The examination is carried out through a review of the relevant research within the field and of the essential documents that describe the key concepts.

**Analysis** A high level examination of the concepts of the semantic Web and linked data is provided with a particular focus on the challenges they entail for libraries and their meta-data practices in the perspective of the extensive

*restructuring process that has already started.*

**Conclusion** *We demonstrate that a set of heterogeneity conflicts, threatening the level of semantic interoperability, can be associated with various phases of this restructuring process from analysis and modelling to conversion and external interlinking. It also claims that these conflicts and their potential solutions are mutually dependent across the phases.*

CHANGE FONT

## Introduction

The report *On the record* ([Library of Congress Working Group on the Future of Bibliographic Control, 2008](#)) states that the '*library community's data carrier, MARC, is based on forty-year old techniques for data management and is out of step with programming styles of today*'. The report recommends future library standards to be integrated into a Web environment. Three years later Library of Congress followed up the conclusions from the report and announced that a '*new bibliographic framework project will be focused on the Web environment, Linked Data principles and mechanisms, and the Resource Description Framework (resource description framework)*' ([Library of Congress, 2011](#)). In November 2012 the primer *Bibliographic Framework as a Web of Data: Linked Data Model and Supporting Services* ([Library of Congress, 2012](#)) was released providing an initial draft of a dedicated linked data model for bibliographic metadata (BIBFRAME in short).

Knowledge organizational approaches in the library community are increasingly characterized by a desire to harmonize with the Web architecture ([Coyle, 2010](#); [Hodge, 2000](#)). During the last couple of years the community has developed a number of comprehensive metadata standardization projects inspired by the idea of linked data such as the BIBFRAME model. Linked data is a set of best practice principles of publishing and exposing data on the web utilizing a graph based data model powered with semantics and cross-domain relationships: '*the semantic Web done right*' according to the web pioneer Sir Tim Berners-Lee ([Heath, 2009](#)).

In the light of traditional and current metadata practices of libraries the best practices of linked data imply a restructuring process from a collection of semi-structured bibliographic records to a semantic graph of unambiguously defined entities. Graphs are not new, neither as applied technology for knowledge organization (e.g. *The Network model*, a database model from the late 1960s) nor as a field of study (Graph theory as a mathematical field dates back Leonard Euler`s experimentations in the 1700s). Nevertheless, as a model for metadata structuring in libraries graphs introduce a new and challenging model for describing and organizing collections. This article demonstrates that the challenges can be associated with various phases of the restructuring process mentioned above - from analysis and modeling to conversion and external interlinking. Further it claims that these challenges and their potential solutions are mutually dependent across the phases. A poor initial analysis of the original model and the metadata that are designed according to this model could for example influence the design of a new (linked data model) and the final interlinking to external resources.

The concept of semantics is neither new in this context. Hjørland ([2007](#)) argues that semantic issues '*underlie all research questions*' in Library and Information Science and especially in the subfield Knowledge organization (KO). He also remarks that many consider the Semantic Web as one of the "important frontiers". The semantic Web is essentially an ambition to link data across different domains and to enable machines to act upon the links. The ambition requires that machines *understand* external data, or in other words that a minimum level of *semantic interoperability* is provided. Semantic interoperability is a key concept in this analysis of semantic Web orientated restructuring.

The article provides a high level examination of the concepts of the semantic Web and linked data, such as semantic interoperability. It has a particular focus on the challenges they entail for libraries and their metadata practices in the perspective of the extensive restructuring process that has already started. The examination is carried out through a review of relevant research within the field and of the essential documents that describe the key concepts.

The initial sections introduce and discuss the notions of graphs and semantic Web. The latter sections deal with the various phases of the restructuring process.

## The giant global graph

Formally a graph  $G$  is a structure which consists of a set of nodes  $N$  and a set of edges  $E$  expressed as a pair,  $G = (N, E)$ . The nodes represent objects, and the edges are relationships (or properties) connecting the nodes. An example of an applied graph is the World Wide Web which can be regarded as a set of interlinked documents where each document is a node and the links are edges connecting the documents. This Web graph is used in Google's PageRank algorithm to assign (relative) weighting to documents based on their incoming links (utilizing the Eigenvector centrality measure as described in [Page, Brin, Motwani, and Winograd, 1999](#)). Another example of a graph is a set of bibliographic metadata, where entities like authors, titles and year of publication is represented as nodes  $N = \{\text{Henrik Ibsen, A dolls's house, 1879}\}$ , and the edges are properties relating the authors to the correct titles, and the titles to the year of publication  $E = \{\text{Henrik Ibsen-A doll's house, A doll's house-1879}\}$ .

After inventing the essential components of today's Web architecture Sir Tim Berners-Lee later introduced the idea of an extension of the Web enabling not only relationships between documents but also between the things that the documents were about: In practice, a graph of interlinked data objects published and exposed on the Web. The idea was first presented as a *Semantic Web* ([Berners-Lee, Hendler, and Lassila, 2001](#)), then connected to a concrete technological infrastructure and a set of best practice publishing guidelines and revitalized as *linked data* ([Berners-Lee, 2006](#)). Berners-Lee has later used the terms Giant global graph ([2007](#)) and the *Web of data* ([Bizer, Heath, and Berners-Lee, 2009](#)) to express more or less the same concept. There are some discussion about the meaning of these terms, but a common interpretation is that the Semantic Web, the Giant Global Graph and the Web of data signify a high-level *vision*, whereas linked data represents the methods for realizing the vision ([Heath, 2009](#)).

## Linked data



One of the main challenges in realizing a semantic Web is the heterogeneous nature of the metadata in various communities. An essential principle in the numerous guidelines for publishing linked data ([Berners-Lee, 2006](#); [Heath and Bizer, 2011](#); [W3C, 2012](#)) is therefore to use established standards like resource description framework. According to its suite of specifications [1] resource description framework provides a framework for representing resources as a set of statements based on a Graph data model. The statements consist of two nodes, a *subject* and an *object*, and a *predicate* that connects them. The statement '*Henrik Ibsen wrote A doll`s house*' can be outlined as a resource description framework statement where Henrik Ibsen is the subject, A doll`s house the object and the property wrote is the predicate. The three components form a *triple*, and a single resource description framework graph is the totality of such triples in a given universe of statements. There are some discussion on how well the resource description framework specifications are founded in the established mathematical concept of graphs (e.g. [Hayes and Gutierrez, 2004](#)), nevertheless the resource description framework graph is often characterized formally as a *directed labelled graph* since the edges always point from a subject towards an object and explicitly denote the property of the subjects. In order to make the resource description framework graphs machine processable and to integrate them with the Web architecture Uniform Resource Identifiers (URI) [2] are used to identify subjects, predicates and in some cases objects. Borrowing a predicate from the Dublin Core Metadata Terms [3] to label the edge in the example above, a triple based on URI`s can be expressed in triple notation as:

```
http://example.org/A_dolls_house  
http://purl.org/dc/terms/creator http://example.org/henrik_ibsen .
```

Objects are also allowed to be literal values as "1979" in the following triple:

```
http://example.org/A_dolls_house  
http://purl.org/dc/terms/issued "1879" .
```

And the objects can be URIs created outside the local resource description framework graph as the URI from DBpedia [4] in the triple:

[http://example.org/henrik\\_ibsen](http://example.org/henrik_ibsen)  
<http://www.w3.org/2002/07/owl#sameAs>  
[http://dbpedia.org/resource/Henrik\\_Ibsen](http://dbpedia.org/resource/Henrik_Ibsen) .

The three examples form a resource description framework graph as visualized in Figure 1.

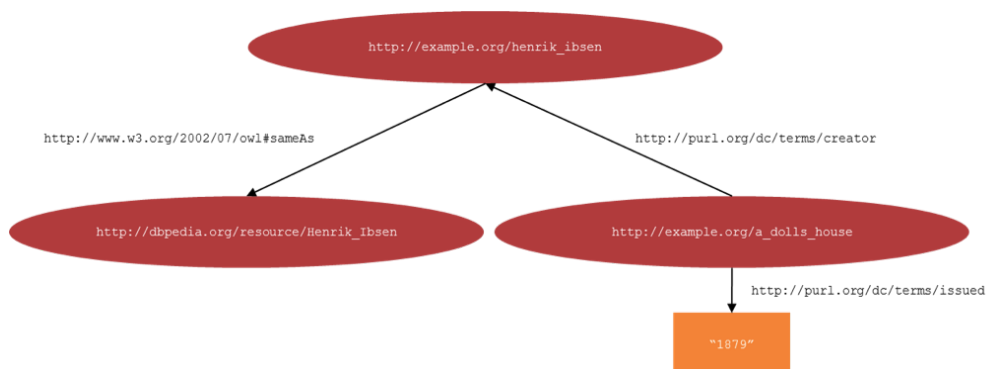


Figure 1 - A simple resource description framework graph of three triples

## Interlinking of data

The basic resource description framework graph in Figure 1 also exemplifies some of the other essential principles of best practice linked data, using (HTTP) URIs as names for things being one of them [5]. In order to achieve a Giant global graph of truly interlinked data it is fundamental to provide links to URIs in external data sets. This is achieved in the example resource description framework graph by the link to a representation of the author Henrik Ibsen in the DBpedia data set which contains resource description framework structured information derived from Wikipedia. The property `sameAs` is taken from the Web Ontology Language (OWL) [6] and used as a predicate to denote the concurrence of the two representations of the author. In the vision of a Semantic Web such links based on HTTP URIs pointing to standardized data representations provides a platform for computational reasoning across institutions and communities. Reusing properties and classes from established and widely adopted vocabularies and ontologies, like Dublin Core and OWL, is considered a good practice which makes it easier to interpret and process the data for client applications. However linked data sources often mix self-defined and existing properties and classes.

## 2.3 Conceptualization

In the literature of linked data it is difficult to find a definite division between the terms vocabularies and ontologies. Gruber famously defined an ontology as a '*specification of a conceptualization*' (1993). The same broad definition could be used to describe a vocabulary (or a metadata schema). However ontologies tend to be used frequently to describe complex systems that provide a set of inference rules and description logic enabling computational reasoning, while vocabularies are used quite consistently to describe less complex collections of conceptual terms like the aforementioned Dublin Core Metadata Terms. In the field of Knowledge representation TBox and ABox are often used to separate between a high level representation system and the actual instance data generated in accordance with such systems (Bergman, 2009; Ferrara, Lorusso, Montanelli, and Varese, 2008). TBox (T for *Terminological*) constitutes a set of concepts, properties and constraints on their usage. ABox (A for *Assertions*) constitutes a set of assertions which are structured according to the TBox, for example, a collection of resource description framework-triples. Within a slightly broad definition, both ontologies, vocabularies, metadata schemas and standards providing some sort of concepts, properties and constraints could be defined as a TBox.

## Semantic interoperability

At the core semantic Web and linked data are about connecting data across heterogeneous domains enabling computers to understand the data and their relations. '[...] *Information is given well-defined meaning*' (Berners-Lee, Hendler, and Lassila, 2001) and this meaning is enhanced with machine-interpretability by the use of standards like resource description framework, unique identifiers and referenced ontologies (as described above in section 2).

Some have questioned such a definition of semantics (Uschold, 2003), and others have discussed whether it's in accordance with established approaches in computer science and linguistics (Almeida, Souza, and Fonseca, 2011; Sheth, Ramakrishnan, and Thomas, 2005). Regardless of these objections and discussions, it is natural to associate an operational understanding of semantics in the context of linked data with the overall

goal to provide *interoperability* across heterogeneous domains.

While interoperability in general can be defined as the ability of two or more systems to exchange information and to use this information, *semantic interoperability* specifies a certain requirement to achieve this goal: The ability of two or more systems to exchange and share intended meaning ([Kalfoglou, 2010](#); [Nilsson, 2010](#); [Park, 2006](#)). Semantic interoperability often constitutes one level in a conceptual model which distinguishes it from other levels of interoperability such as syntactic interoperability concerning exchange formats and technical interoperability concerning exchange protocols (see e.g. [Nilsson, Baker, and Johnston, 2009](#); [Tolk and Muguira, 2003](#); [Tolk, 2006](#)).

For a system to understand the intended meaning of information in other systems, the information being exchanged needs to be equipped with a minimum of disambiguous machine-interpretable description. In a linked data conformant resource description framework graph the interpretable description is to be found in the referenced ontologies defining the meaning of certain properties and classes, as described and exemplified above in Figure 1. García-Castro and Gomez-Perez ([2011](#)) provide a definition of semantic interoperability where this functionality is outlined explicitly as '*the ability that semantic systems have to interchange ontologies and use them*'.

The challenges to achieve semantic interoperability can also be defined and explained negatively by the existence of a certain degree of *semantic heterogeneity* between two systems. Pluempitiwiriyaewj and Hammer ([2000](#)) have classified occurrences of semantic heterogeneities in XML data sources. Some of their main conflict classes can be related to conflicts arising in the process of interlinking instances described with disparate ontologies:

- *structural conflicts*, when the same (or overlapping) classes or properties are represented differently in two ontologies due to discrepancies in the level of generalization/specialization
- *data conflicts*, when the same concept is represented differently due to incorrect spelling and different identification systems

Ferrara, Lorusso, Montanelli and Varese ([2008](#)) highlights three sources of heterogeneity challenging the matching of instances across populated ontologies: structural heterogeneity, data value differences, and logical heterogeneity. The first two equals the structural conflicts and data conflicts mentioned above. The latter is concerning differences in the way ontologies are implementing rules for reasoning. In addition to these conflicts Ferrara ([2005](#)) has described semantic heterogeneity scenarios related to flexible schemas providing semi-structured data, where conflicts arises from the inconsistencies in usage and interpretation of the schema rules.

Semantic heterogeneity conflicts are potential obstacles to achieving the degree of semantic interoperability necessary for a successful realization of the Semantic web. Bizer, Heath, and Berners-Lee ([2009](#)) have announced data fusion and schema mapping to be one of the main research challenges related to linked data. The next section discusses the potential obstacles in relation to the restructuring of library data.

## **The linked data restructuring cycle**

Cultural heritage institutions like libraries possess huge amounts of metadata already catalogued and stored according to the principles of established community standards. Representing these data as resource description framework graphs and linking them '*to other people`s data*' ([Berners-Lee, 2006](#)) leads into a cycle of restructuring. This cycle can be derived from the best practice guidelines and is analysed and described in detail by the LOD2 project [[7](#)] and by other parties (e.g. [Hyland, 2010](#); [W3C, 2012](#)). In the context of a concrete case study of restructuring library data Tallerås, Massey, Dahl, and Pharo ([2013](#)) have synthesised (and simplified) existing efforts in describing a linked data restructuring cycle (see [Figure 2](#)). With an exception of the evaluation aspects each of the phases in the cycle will be discussed in separate subsections, with a special focus on the case of restructuring library data. The cycle can be viewed as an iterative process with the starting point in an analysis of a certain domain. A new ontology is developed; the data is converted in accordance with this ontology and interlinked with data in other datasets published on the web. The

latter phase can be considered as an on-going evaluation with the potential to restart the process initiating a deeper analysis, remodelling of the ontology and a tuning of the conversion algorithm and interlinking technique.

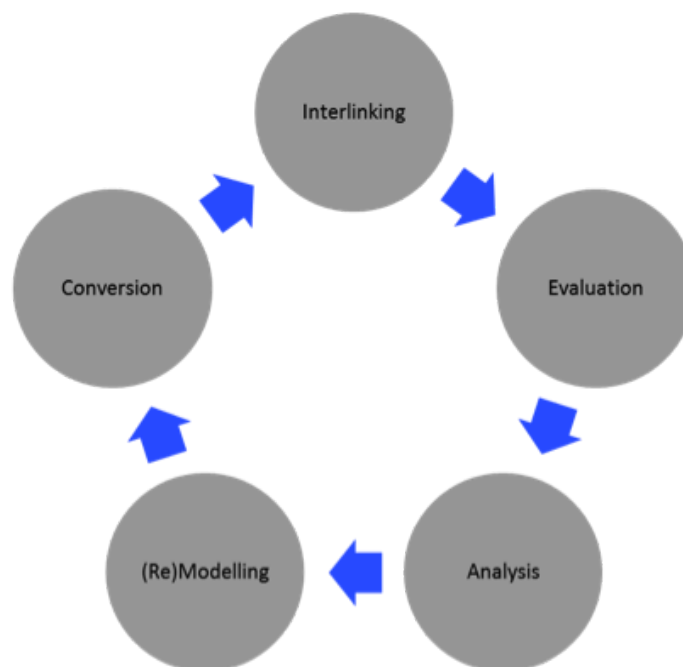


Figure 2 - Linked data restructuring cycle

### Analysis: Library data

Parallel to the developments at the Library of Congress, described in the introduction, the library community has witnessed a great number of 'bottom up' linked data initiatives. The national libraries of Sweden ([Malmsten, 2009](#)), France ([Bibliothèque nationale de France, 2012](#)), Germany ([Hauser, 2012](#)) and Great Britain ([The British Library, 2013](#)) have all carried out major projects involving a conversion of their catalogue data into a variety of resource description framework implementations. OCLC have made a data set of the three top levels of the Dewey Decimal Classification System in 23 languages available as linked data [8]. They also host the Virtual International Authority File (VIAF) project connecting authority records from several national and other libraries, also made available as resource description framework[9]. See Dunsire and Willer (2011) for an extensive overview of other linked data projects in the library community.

These projects reveal a desire for change, and a belief in the vision of the Giant global graph. What are then the dissatisfactory aspects of the existing traditions of metadata production motivating such desires and beliefs?

And what challenges concerning restructuring are to be found in traditional library metadata?

### The bibliographic record

Since the middle of the 1800s *universal bibliographic control* (universal bibliographic control) has been an expressed objective in the library community ([Svenonius, 2000](#)). universal bibliographic control is the vision of a shared worldwide bibliography of every book ever published. To support this vision library history has offered different bibliographic systems based on available technology. These systems have undergone two major revolutions, respectively, the transition from the book catalog to the card catalog and from the card catalog to the automated systems that characterize current practices. The first revolution was the origin of the independent bibliographic record in form of a card containing description of a certain edition of a book. The second revolution automated this record and made it '*machine readable*' ([Avram, 1975](#)). The struggle to achieve universal bibliographic control has emphasized standards in order to support interoperability and exchange of bibliographic records between the contributing libraries; the ideal has been to catalogue a book only once. The standards have also changed in accordance with the bibliographic systems they were developed to support. Today the most widespread standards are the cataloging rule *Anglo-American Cataloging Rule* (AACR) [[10](#)] and the metadata schema Machine Readable Cataloguing (MARC) [[11](#)]. Both standards were developed during the 1960s.

These standards have increasingly been criticized for several reasons. The general critique concerns their age and that they are out of step with the '*programming styles of today*' (as stated in the report issued by Library of congress cited in section 1). The standards were developed prior to relational databases ([Codd, 1970](#)), and the Web, and lack important knowledge organizational innovation from those infrastructures; for instance the idea of using unique and computable identifiers like database keys and URIs. Instead they are tightly intervened with some of the knowledge organizational principles implied in the card catalogue, the leading technology of the time they were developed ([Coyle and Hillmann, 2007](#); [Thomale, 2010](#)).

This includes carrying on the principle of bibliographic records being geared for human reading and interpretation, resulting in semi-structured MARC records containing mostly text strings. These strings are machine readable, but harder to reason upon for machines than well-structured data in accordance with relational database theory or description logic in ontologies ([Styles, Ayers, and Shabir, 2008](#)). The restructuring phases described below all concerns successful identification of entities. Inconsistent cataloging due to heterogeneity conflicts in terms of data values and human interpretation of standards may lead to both data loss, where the text is not understood by the machine, and redundancy, where two or more text strings in a given data set are representing the same entity (described in more detail in section 4.3).

The bibliographic MARC record also continues the principle of describing a certain edition of a book (a *manifestation* of a work in terms of the FRBR model ([IFLA Study Group on the Functional Requirements for Bibliographic Records, 1998](#))). The lack of consistent references to the platonic idea of a work entity that connects the potential manifold of editions, and the lack of unique identifiers of authors, publisher etc., constitutes a data model of rather unconnected and segregated records ('*islands of data*'), representing the opposite of the idea of a unified data set based on directed and (semantic) labeled graph connecting data objects.

Some have faced this criticism and argue that many of the problems are to be found in the lack of complexity within the systems that manage the MARC records, for example in utilizing the sophistication of relationships expressed in the card catalog (e.g. [Murray and Tillett, 2012](#)). Library of Congress and other major stakeholders have nevertheless, as outlined above, regarded the problems to be too extensive to make mere adjustments.

### **(Re)modelling library data: universal bibliographic control vs. the giant global graph**

The primary method to achieve universal bibliographic control has been standardization. Groups of experts from leading institutions like the Library of Congress have developed and maintained the standards, and consistency



has been secured by the principle of everyone using these standards. In this perspective universal bibliographic control can be described as a '*top down*' approach to interoperability. Linked Data represent a more pragmatic and '*bottom-up*' approach. Berners-Lee, Hendler, and Lassila ([2001](#)) states that the Semantic Web '*will be as decentralized as possible*'. When the new ontologies within different domains and communities are to be designed, metadata managers are free to choose and mix classes and predicates from existing ontologies and vocabularies with their own terms.

Broad definitions of TBox and ABox (as given in Section 2.3) may be useful in a comparative analysis of e.g. OWL-based ontologies and domain specific standards from the library world. Ontologies should according to guidelines for Linked Data and Semantic Web facilitate automated reasoning. This requires that the ontologies describe concepts, properties and rules for their usage in a processable way. Library standards do indeed provide concepts, properties and rules for using them, but they tend to - especially the rules (such as AACR2) - to be oriented towards human consumption and not automated reasoning. To make a good linked data model, it is important that it not only ensures a successful conversion of the instance data in the ABox, but also a machine readable TBox.

Svenonius ([2000](#)) has remarked that the bibliographic records were made to support a fixed set of functions and objectives, such as inventory and the objectives formulated by prominent catalogue innovators like Cutter and Lubetzky, and that technological advance and new media formats have challenged these functions: '*It is hard surprising that using one device to serve several functions should lead to trouble in times of technological change*'. In Rust and Bide ([2000](#)) such conflicts related to intended function of the data (retrieval aspects, cataloguing aspects etc.) is outlined as potential interoperability obstacles.

Through a bottom up approach libraries get the opportunity to handle such obstacles. Different ontologies can be designed according to the needs at the time they occur. This possibility is also utilized in the different linked data projects described above. If the community however wishes to maintain the idea of universal bibliographic

control, as in a worldwide bibliography based on distributed contributions, they also need to balance this flexibility with the actual potential for operational interoperability between the ontologies in use [12]. This also involves technical challenges related the long term archiving of resource description framework data as discussed in Seadle (2013). Lately there has been a lively and interesting discussion about '*Reuse (or not) of existing ontologies*' at the BIBFRAME email list, where supporters of a flexible '*bottom up*' approach arguing for reuse opposites supporters of a new and local model arguing for control and long term sustainability ('*BIBFRAME archives*,' 2013).

## Conversion

Case studies of mapping library records to resource description framework based ontologies have confirmed all of the potential semantic heterogeneity conflicts mentioned in section 3.1., such as inconsistencies and structural discrepancies (Tallerås, Massey, Dahl, and Pharo, 2013; Westrum, Rekkavik, and Tallerås, 2012). They have also shown that such conflicts have a serious impact on the conversion of data. Without unique identifiers for the various entities the conversion is dependent on a computational interpretation of the strings. The entities, for instance represented by the string Ibsen, Henrik from the field for main entries in the MARC record, are reduced to a set of characters to be matched with other sets of characters. Then a decision is made, based on a chosen similarity threshold, as to whether the characters represent the same entity or not. If the similarity measure satisfies the threshold a URI can be assigned as a unique and single identifier for this entity. The URI is further assigned into a series of triples of the kind exemplified in section 2.1. All forms of inconsistencies due to misspellings, cultural or linguistic contexts or different interpretations of the rules on how to describe things, affect such a process negatively, and will make the conversion algorithm fail to assign correct identifiers.

To improve the result of the conversion process some argue that one should use ontologies based on terms exclusively from local schemas, such as a MARC based ontology[13], in order to overcome structural

heterogeneity, secure semantic coherence and reduce the potential *lossiness* in the conversion process ([Dunsire, 2012](#)). This argument is more in line with the traditional *top down* universal bibliographic control approach emphasizing domain specific standards. Others have worked with pre-coordination of existing MARC collections in order to harmonize them to other prominent data models in the community such as the FRBR model, and in order to *clean* the records and reduce inconsistencies prior to the conversion (and interlinking) process ([Aalberg, 2006](#); [Westrum, Rekkavik, Tallerås, 2012](#)). Nilsson ([2010](#)) have described the latter approach as a vertical harmonization within a certain domain.

## Interlinking

When the data is converted to an resource description framework format, they should be linked to similar data in existing resource description framework graphs. Many of the data sets that already have been published as Linked Data describe cultural objects and entities related to them. These data sets are largely overlapping with library data, which constitutes a great potential for an extensive interlinking. The main challenge in this part of the restructuring process is once again related to semantic interoperability and the question of how to decide which URIs that are representing the same concept or the same entity in two different data sets that are structured according to different ontologies. Although this is an area under development, there are already a number of automated methods for approaching the problem. They range from simple string recognition techniques (often referred to as naïve interlinking) to utilizing the graph structures in the resource description framework graphs, machine learning techniques and more probabilistic oriented methods (for some examples see [Doan, Madhavan, Dhamankar, Domingos, and Halevy, 2003](#); [Melnik, Garcia-Molina, and Rahm, 2002](#); [Raimond, Sutton, and Sandler, 2008](#)). In practice the interlinking of resource description framework graphs is a semi-automatic discovery phase, both dependent on manual and automatic approaches. The manual efforts can be related to supervision of automatic systems, but also to direct assignments of links, for instance in the cataloguing process.

Similarity and identity are not fixed categories, albeit the extensive use of the rather unambiguous predicate owl:sameAs to express concurrences in the Linked Data context (defined in OWL as: '*an statement [that] indicates that two URI references actually refer to the same thing*') [14]). Halpin, Hayes, McCusker, McGuinness, and Thompson (2010) claims that linked data experience an '*identity crisis*': '*Just because a construct in a knowledge representation language is explicitly and formally defined does not necessarily mean that people will follow that definition when actually using that construct 'in the wild*". Based on a logical perspective on identity ('*Leibnitz `s law*') they identify a variety of inconsistent usage cases of the owl:sameAs predicate and a number of reasons for them. Some of these can be related directly to heterogeneity conflicts such as discrepancies in the interpretation of flexible ontologies. Bizer, Heath, and Berners-Lee (2009) also address the problem of structural heterogeneity claiming that existing correspondences often are too '*coarse-grained*' to support effective computational reasoning.

## Concluding remarks

A proper analysis of existing data, the standards used to generate them and the domain specific needs and objectives forms the basis for the development of a new data model. This data model must maintain the basic semantics from the existing standards, and at the same time aim to innovate and renew old traditions. The quality of the conversion from the old to the new model depends on how well the model is able to handle heterogeneity conflicts in order to maintain granularity and semantic attributes, and eventually prevent significant loss of data (and semantics). The semantic expressiveness in the new model is also vital for providing precise links to other dataset.

Through references to research, standards and best practice-documents the article have outlined a restructuring process from a record-based data model to best practice linked data. Library data is used as a case to discuss challenges in the various phases of the process. Library data is an interesting case because the library community is already in an active process of restructuring. Each of the phases represents specific challenges regarding

semantic heterogeneity conflicts, but these challenges also connect the phases and make them mutual dependent. The quality of the implementation of each phase will influence on the ability to gain quality in the other phases.

In a future research project it would be interesting to conduct a more thorough examination of concepts such as semantic interoperability and heterogeneity conflicts. In the research literature there exist a manifold of definitions and interpretations, other than those outlined in this article. A classification of these definitions, based on context and specific technological challenges, could for instance be useful in order to establish a fruitful theoretical perspective on the semantic Web.

[1] In particular resource description framework Primer (<http://www.w3.org/TR/rdf-mt/>), resource description framework Concepts and Abstract Syntax (<http://www.w3.org/TR/rdf-concepts/>) and resource description framework Semantics (<http://www.w3.org/TR/rdf-mt/>)

[2] <http://www.w3.org/TR/rdf-primer/#identifiers>

[3] <http://dublincore.org/documents/dcmi-terms/>

[4] <http://dbpedia.org/About>

[5] To gain a seamless Web integration the guidelines recommend HTTP based URIs.

[6] <http://www.w3.org/TR/owl-ref/>

[7] <http://stack.lod2.eu/>

[8] <http://dewey.info/>

[9] <http://viaf.org>

[10] <http://www.aacr2.org/>

[11] <http://www.loc.gov/marc/>

[12] Lately there has been a lively and interesting discussion about Reuse (or not) of existing ontologies' at the BIBFRAME email list: <http://listserv.loc.gov/cgi-bin/wa?A1=ind1303&L=bibframe>

[13] See <http://marc21rdf.info/> for a resource description framework based Vocabulary representing MARC elements

## References

- Aalberg, T. (2006). A Tool for Converting from MARC to FRBR. *ERICIM News*
- Almeida, M., Souza, R., and Fonseca, F. (2011). Semantics in the Semantic Web: a critical evaluation. *Knowledge Organization*, **38**(3), 187-203
- Avram, H. D. (1975). MARC, its history and implications. *Washington, DC: Library of Congress*
- Bergman, M. (2009). The Fundamental Importance of Keeping an ABox and TBox Split. AI3. Retrieved from <http://www.mkbergman.com/489/ontology-best-practices-for-data-driven-applications-part-2/>
- Berners-Lee, T. (2006). Linked data: design issues. W3C
- Berners-Lee, T. (2007). Giant Global Graph. Decentralized Information Group Breadcrumbs. Retrieved from <http://dig.csail.mit.edu/breadcrumbs/node/215>
- Berners-Lee, T., Hendler, J. & Lassila, O. (2001). *The Semantic Web. Scientific American*, **284**(5), 34-43
- BIBFRAME archives. (2013). Retrieved from <http://listserv.loc.gov/cgi-bin/wa?A1=ind1303&L=bibframe>
- Bibliothèque nationale de France. (2012). data.bnf.fr. Retrieved from <http://data.bnf.fr/>
- Bizer, C., Heath, T. & Berners-Lee, T. (2009). Linked Data: The story so far. *International Journal on Semantic Web and Information Systems*, **5**(3), 1-22
- The British Library. (2013). Free Data Services. The British Library. Retrieved March 1, 2013, from <http://www.bl.uk/bibliographic/datafree.html>
- Codd, E. F. (1970). A relational model of data for large shared data banks. *i*, **13**(6), 377-387
- Coyle, K. (2010). Library Data in a Modern Context. *Library Technology Reports*, **46**(1), 5-13
- Coyle, K. & Hillmann, D. (2007). Resource Description and Access (RDA). *D-Lib Magazine*, **13**(1/2)
- Doan, A., Madhavan, J., Dhamankar, R., Domingos, P., and Halevy, A. Y. (2003). Learning to match ontologies on the Semantic Web. *The VLDB Journal*, **12**(4)
- Dunsire, G. (2012). An introduction to open linked data for librarians. Powerpoint presentation at the National Library of Finland
- Dunsire, G. & Willer, M. (2011). Standard library metadata models and structures for the Semantic

- Web. *Library Hi Tech News*, **28**(3), 1-12
- Ferrara, A., Lorusso, D., Montanelli, S. & Varese, G. (2008). Towards a benchmark for instance matching. *Proceedings of the 3rd International Workshop on Ontology Matching (OM-2008)*
- García-Castro, R. & Gomez-Perez, A. (2011). Perspectives in semantic interoperability. *Proceedings of the International Workshop on Semantic Interoperability IWSI 2011 In conjunction with ICAART 2011* (pp. 13-22). SciTePress
- Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, **5**(2), 199-220
- Halevy, A. (2005). Why Your Data Won't Mix - New tools and techniques can help ease the pain of reconciling schemas. *ACM Queue*, **3**(8)
- Halpin, H., Hayes, P., McCusker, J., McGuinness, D. & Thompson, H. (2010). When owl:sameAs isn't the same: An analysis of identity in Linked Data. *The Semantic Web - ISWC 2010* (Vol. 6496, pp. 305-320). Berlin: Springer
- Hauser, J. (2012). Dokumentation des Linked Data Services der DNB. Retrieved from <https://wiki.dnb.de/display/LDS/Dokumentation+des+Linked+Data+Services+der+DNB>
- Hayes, J. & Gutierrez, C. (2004). Bipartite graphs as intermediate model for resource description framework. *The Semantic Web - ISWC 2004* (Vol. 3298, pp. 47-61). Berlin: Springer
- Heath, T. (2009). Linked Data? Web of Data? Semantic Web? WTF? Tom Heath's Displacement Activities. Retrieved from <http://tomheath.com/blog/2009/03/linked-data-web-of-data-semantic-web-wtf/>
- Heath, T. & Bizer, C. (2011). Linked Data: Evolving the Web into a global data space. *Morgan & Claypool*
- Hjørland, B. (2007). Semantics and knowledge organization. *Annual review of information science and technology*, **41**(1), 367-405
- Hodge, G. (2000). Systems of knowledge organization for digital libraries: Beyond traditional authority files. *Washington: The Digital Library Federation Council on Library and Information Resources*
- Hyland, B. (2010). Preparing for a linked data enterprise. Linking enterprise data (pp. 51-64). *Springer US*
- IFLA Study Group on the Functional Requirements for Bibliographic Records. (1998). Functional requirements for bibliographic records: Final report. *München: K.G. Saur*

- Kalfoglou, Y. (Ed.). (2010). Cases on semantic interoperability for information systems integration: Practices and applications. *New York: Information science reference*
- Page, L., Brin, S., Motwani, R. & Winograd, T. (1999). The PageRank citation ranking: Bringing order to the Web. *Stanford InfoLab*
- Library of Congress. (2011). A Bibliographic Framework for the digital age. Retrieved from <http://www.loc.gov/marc/transition/news/framework-103111.html#ftn1>
- Library of Congress. (2012). Bibliographic Framework as a Web of data: linked data model and supporting services. *Washington DC*
- Library of Congress Working Group on the Future of Bibliographic Control. (2008). On the record: Report of the Library of Congress Working Group on the Future of Bibliographic Control
- Malmsten, M. (2009). Exposing library data as linked data. *IFLA satellite preconference*
- Melnik, S., Garcia-Molina, H. & Rahm, E. (2002). Similarity flooding: A versatile graph matching algorithm and its application to schema matching. *Proceedings of the 18th International Conference on Data Engineering*
- Murray, R. J. & Tillett, B. B. (2012). Cataloging theory in search of graph theory and other ivory towers. *Information Technology and Libraries*, **30**(4), 17-184
- Nilsson, M. (2010). From interoperability to harmonization in metadata standardization - Designing an evolvable framework for metadata harmonization. *Royal Institute of Technology, Stockholm*
- Nilsson, M., Baker, T. & Johnston, P. (2009). Interoperability levels for Dublin Core Metadata. Dublin Core Metadata Initiative. Retrieved from <http://dublincore.org/documents/interoperability-levels/>
- Park, T. (2006). Semantic interoperability and metadata quality: An analysis of metadata item records of digital image collections. *Knowledge organization*, **33**(1), 20-34
- Pluempitiwiriwawej, C. & Hammer, J. (2000). A classification scheme for semantic and schematic heterogeneities in XML data sources. *Technical report TROO-004*
- Raimond, Y., Sutton, C. & Sandler, M. (2008). Automatic interlinking of music datasets on the semantic web. *Linked Data on the Web - LDOW 2008*



- Rust, G., & Bide, M. (2000). The indecs metadata framework: Principles, model and data dictionary. *Indecs Framework*
- Seadle, M. (2013). Archiving in the networked world: resource description framework. *Library Hi Tech*, **31**(1), 182-188
- Sheth, A., Ramakrishnan, C. & Thomas, C. (2005). Semantics for the semantic web: The implicit, the formal and the powerful. *International journal on Semantic Web and information systems*, **1**(1)
- Styles, R., Ayers, D. & Shabir, N. (2008). Semantic MARC, MARC21 and the Semantic Web. *Linked Data on the Web - LDOW 2008*
- Svenonius, E. (2000). The Intellectual Foundation of Information Organization. *Cambridge, Massachusetts: The MIT Press*
- Tallerås, K., Massey, D., Dahl, J.H.B. & Pharo, N. (2013). Ordo ad chaos - Linking Norwegian black metal. Libraries, black metal and corporate finance: Current research in Nordic Library and Information Science (pp. 136-150). *Borås: Univeristy of Borås*
- Thomale, J. (2010). Interpreting MARC: Where`s the bibliographic data? *Code4lib*, (11)
- Tolk, A. (2006). What comes after the Semantic Web - PADS implications for the dynamic Web. *PADS*, 55-62
- Tolk, A. & Muguira, J. (2003). The levels of conceptual interoperability model. *Proceedings of the 2009 Spring Simulation Multiconference*
- Uschold, M. (2003). Where are the Semantics in the Semantic Web? *AI Magazine*, **24**(3), 25-36
- W3C. (2012). Linked data cookbook. Retrieved from [http://www.w3.org/2011/gld/wiki/Linked\\_Data\\_Cookbook](http://www.w3.org/2011/gld/wiki/Linked_Data_Cookbook)
- Westrum, A., Rekkavik, A. & Tallerås, K. (2012). Improving the presentation of library data using FRBR and linked data. *Code4Lib Journal*, (16)

#### How to cite this paper

Tallerås, K. (2013). From many records to one graph: heterogeneity conflicts in the linked data restructuring cycle. *Information Research*, **18**(3) paper C18. [Available at <http://InformationR.net/ir/18-3/colis/paperC18.html>]

Find other papers on this subject

Scholar Search

Google Search

Bing

Check for citations, [using Google Scholar](#)

 Liker 0

 Tweet

 Del

11

---

**898**

© the author, 2013.

Last updated: 10 August, 2013

---

[Contents](#) | [Author index](#) | [Subject index](#) | [Search](#) | [Home](#)

---

## Paper D

Tallerås, K., Massey, D., Dahl, J. H. B., & Pharo, N. (2013). Ordo ad chaos: Linking Norwegian black metal. In *Libraries, black metal and corporate finance: Current research in Nordic Library and Information Science* (pp. 136–150). Borås: University of Borås.



# Ordo ad chaos – Linking Norwegian black metal

KIM TALLERÅS, DAVID MASSEY, JØRN HELGE B. DAHL *and* NILS PHARO

## I INTRODUCTION

Despite the fact that collections of metadata often represent knowledge about the same entities and phenomena, they are created using disparate methods. The two collections of music metadata explored in this article are illustrative examples of heterogeneous regimes in terms of handling overlapping data: the national discography of Norway (NORDSKO)<sup>1</sup>, produced in the context of library standards, and the user generated database Musicbrainz<sup>2</sup>. In this article we describe a case study that experiments with automatic matching and interlinking of metadata from the two sources. This is done in order to investigate the process of producing a Linked data-compliant set of data describing recordings in the musical genre of Norwegian black metal.

We try to answer the following research question:

- RQ1: What kind of interoperability issues emerge from a modelling of the metadata of musical recordings, using best practice Linked data guidelines?
- RQ2: What are the necessary steps in an automatic conversion process of musical recordings' metadata, following best practices?

### 1.1 *Linked data*

Linked data (Berners-Lee 2011) is a set of best practice methods for publishing interlinked data on the web which computers can understand and reason on. Based on the objectives of such interlinking, the best practice approaches contain an inherent promise of overcoming challenges regarding heterogeneous metadata regimes. The project of publishing a small corpus of black metal metadata on the web, and then to try and interlink instances in the data set with data from another collection of metadata represents experimental investigations of the validity of such promises.

<sup>1</sup> <http://www.nb.no/baser/nordisko/english.html>

<sup>2</sup> <http://musicbrainz.org>

## 1.2 Methods

The initial step of the experimental case study was to identify a natural candidate to publish data on the web according to the best practices of Linked data. We found the National Library of Norway and their NORDISKO collection to be a reasonable seed candidate. To make the experiment feasible we limited the corpus to metadata describing Norwegian black metal records. This limitation was also motivated by the complex relationships between recordings, people, their pseudonyms and performing groups characterizing the genre.

NORDISKO is a traditional library catalogue, based on metadata records structured with (NOR)MARC<sup>3</sup> and AACR2, the most prevalent metadata standards in the Norwegian library community. The NORDISKO collection has recently been moved from its customary database environment to the federated BIBSYS catalogue. Our corpus is nevertheless based on data from the original database. Our target data set, MusicBrainz, already exists as Linked data, based on a conversion from their relational database schema, and made accessible via SPARQL-endpoints on a variety of servers<sup>4</sup>. Dahl, Knutsen & Tallerås have written more about the differences between traditional library data and metadata produced in the Musicbrainz community (2012).

## 2 THE LINKED DATA PRODUCTION CYCLE

The different steps in the process (and in our practical approach described later in the article) are based on guidelines for best practices of publishing Linked data such as Heath & Bizer (2009), Berners-Lee (2011) and W3C (2011). These guidelines have in common that they are highly cited in both the research literature and in documentation from the professional practice field. We have summarized the process of gaining best practice Linked data as a production cycle visualized in Figure 1<sup>5</sup>.

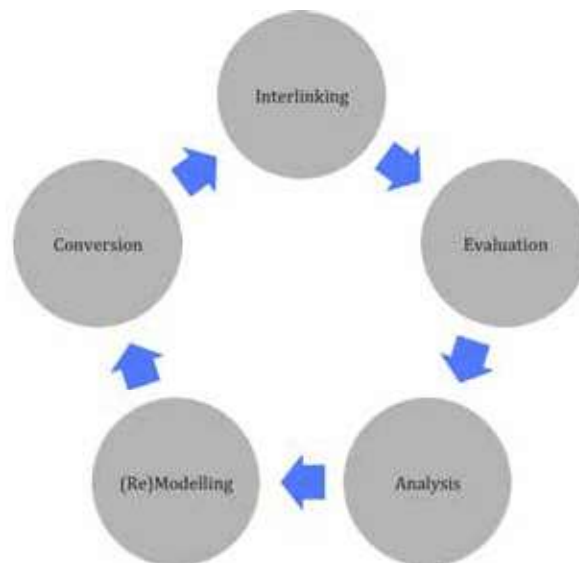
### 2.1 Best practices summarized

A typical production process starts out with an analysis and exploration of existing data and standards in a certain domain. Based on such an analysis an appropriate ontology describing the domain is created. The ontology should utilize established vocabularies and provide a system of unique

<sup>3</sup> <http://www.nb.no/normarc/normarc.html>

<sup>4</sup> Database schema: [http://wiki.musicbrainz.org/MusicBrainz\\_Database/Schema](http://wiki.musicbrainz.org/MusicBrainz_Database/Schema). We have chosen to use a data set and a SPARQL-endpoint hosted by dbtune.org: <http://dbtune.org/musicbrainz/>

<sup>5</sup> There are other and similar visualizations of Linked data cycles see, e.g. the Linked Open Data Lifecycle according to the research programme LOD2, founded by the European Union: [http://www.w3.org/2011/gld/wiki/GLD\\_Life\\_cycle](http://www.w3.org/2011/gld/wiki/GLD_Life_cycle)



**Figure 1. Linked data production cycle**

identification of the resources by assigning URIs. When the ontology is described in a dedicated ontology language like OWL<sup>6</sup>, the data can be converted into a new form of representation. The representation system should be based on RDF, the Resource Description Framework. When the data is successfully converted it is stored in a database environment that provides access through a SPARQL-endpoint<sup>7</sup>, and interlinked with corresponding data already published in other data sets. This can be done by using the property `sameAS` from the ontology language OWL, stating that a given URI in the seed data set represents the same resource as a given URI in a target data set. Finally the data quality and interlinking should be evaluated, and the evaluation can lead to a remodelling of the ontology or changes to the conversion and interlinking procedures. The need for continuous maintenance of the data set implies a recurrent and iterative process.

### *2.2 Semantic interoperability*

Our small experimental corpus allows both manual conversion and interlinking, however, in real life publishing scenarios, when large amounts of data are involved, some sort of automation is required for the Linked data

<sup>6</sup> <http://www.w3.org/TR/owl-overview/>

<sup>7</sup> <http://www.w3.org/TR/rdf-sparql-query/>

production cycle. Automation in itself, will only work if the data sets *understand* each other's data, i.e, if there is a minimum degree of *semantic interoperability* (Kalfoglou, 2010; Ma, Mei, Chung, & Amor, 2006; Mao, 2008; Tolk & Muguira, 2003).

Most of the Linked data collections are based on already existing data extracted from a variety of database environments where they have been produced in conformity with domain and community-specific registration standards. Linked data is therefore a product that ultimately relies on semantic interoperability between inherently heterogeneous data. This is also something that must be considered in areas outside the Linked data production cycle, when data needs to be integrated and exchanged. Our perspective is therefore directed towards the question of whether the different and specific phases of the Linked data production cycle themselves contribute in boosting the degree of interoperability. In the case of library data, for instance, to what extent can, the conversion from a MARC based record model to an RDF-based graph model contribute to methods for automatic interlinking? Or to what extent does the reuse of existing vocabularies contribute to semantic coherence across the data sets?

### 3 PRACTICAL APPROACHES

Publishing Linked data, based on existing library data, implies a process of three main steps:

- a) modelling an RDF-structure of properties and classes from existing vocabularies
- b) converting the existing MARC/AACR2 RDF-data to the new RDF-structure
- c) linking entities in the new data set to entities in another Linked data collection

#### 3.1 *The seed data*

We used the Z39.50-protocol to retrieve a corpus from NORDISKO of 99 records based on a selected list of influential Norwegian black metal bands<sup>8</sup>. The records were converted from the ISO2709 exchange format to MARCXML. We then created a simple ontology consisting of classes and properties taken from three existing vocabularies: Music Ontology, Dublin Core and FOAF<sup>9</sup>. The ontology covered representations of artists, their bands, records, tracks, and the relationships between these entities. The ontology was described in the OWL ontology language, and the se-

<sup>8</sup> <http://www.nb.no/baser/z3950>

<sup>9</sup> <http://purl.org/ontology/mof/>, <http://purl.org/dc/elements/1.1/>, <http://xmlns.com/foaf/o.i/>



lected pieces of data from the MARCXML records were transformed into RDF/XML using XSLT. The development of the XSLT transformation implied, in practice, a development of a crosswalk connecting semantics in the MARC format to semantics in the vocabularies we used in our ontology (Figure 2).

The valid RDF-file was uploaded into a Virtuoso triple store providing a SPARQL-endpoint. For interlinking instances in our new data set with instances in the MusicBrainz collection we analyzed the MusicBrainz ontology and made a graph-matching algorithm based on similarity measures inspired by Raimond, Sutton, & Sandler (2008). The algorithm was based on SPARQL queries and a PHP library calculating the string match (in our case Levenshtein distance<sup>10</sup>) between two instances. The interlinking process was described in a flow chart (Figure 7).

#### 4 MODELLING NORWEGIAN BLACK METAL

As indicated in the introduction Norwegian black metal can be characterized by complex structures between bands and musicians, and between musicians caused by the common practice of using pseudonyms, and different pseudonyms in different contexts. Who is who? And to which records, created by which band, do they contribute musically or otherwise? It is also

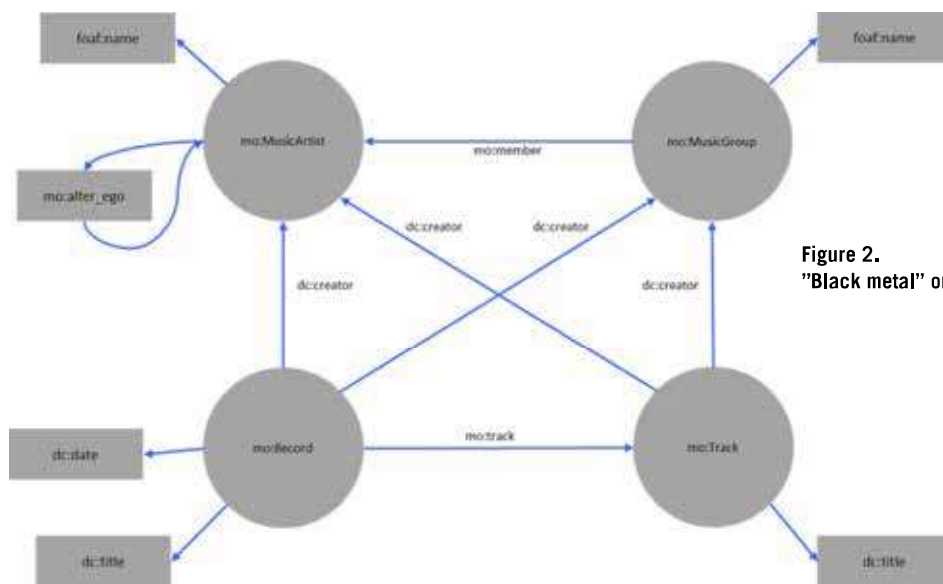


Figure 2.  
"Black metal" ontology

<sup>10</sup> <http://php.net/manual/en/function levenshtein.php>

common to contribute to each other's records, and form different bands with members from other bands in the same genre. This complexity was a major motivation for selecting the genre as an experimental corpus. In practice, the corpus satisfied our expectations concerning the challenges of identifying, relating and separating people by their names only, especially in the modelling and the conversion process.

#### *4.1 The ontology*

For our ontology (see Figure 2) we used only the most basic information in the MARC records, like the titles of records, titles of tracks and the names of the artists contributing to the record.

Since the bands do not necessarily have the same members on all records, and some people, in some cases, only contribute to a single track on a record, it was important to make a direct relationship from `mo:MusicArtist` to both `mo:Record` and `mo:Track`. This way we avoided the situation where all the artists appear to be the composers of all the tracks and records. Some artists use different pseudonyms on different records and tracks. We were not immediately able to cluster the variety of names used for the same person to one authoritative URI. This was due to the manifestation-oriented nature of the MARC record with its unpredictable connections between variant names. We therefore kept all pseudonyms and variant names as they appeared in the seed data set for further refining or clustering via string and graph-matching techniques later.

#### *4.2 Previous knowledge*

It was challenging, and in some cases impossible, to separate the development of a domain specific ontology based on our knowledge of the structures in the MARC records and in existing data sets already published on the web. To some degree, this leads in practice to a collapse between what we have described as separate phases in our production cycle, especially between (re)modelling, conversion and the interlinking (to come). The choices of existing vocabularies were also inevitably influenced by our previous knowledge about the domain and earlier efforts to overcome challenges regarding redundancy, inconsistency, ambiguity etc. The best practice guidelines are not very specific concerning the dangers of inheriting errors and flaws by tuning our own ontology towards known structures. However, Berners-Lee (2011) suggests, for example, a pragmatic approach encouraging common sense and usefulness.

## 5 THE CONVERSION

The conversion process already started as we developed the ontology, as problematized above. According to W3C guidelines (2011) we developed an ontology based on our domain knowledge followed up by an automated conversion. Our process can be outlined as an iterative development of an XSLT-based crosswalk between two metadata structures, (NOR)MARC and our Black metal ontology. The structures differ significantly as (NOR)MARC is a record oriented structure and our ontology is a RDF-graph of triple statements. They also differ in the way that the content of a MARC record is dictated by the principles of descriptive cataloguing of manifestations, while our ontology has no intentions for the data other than containing disambiguated representations of things in the world. The MARC records have been produced by many indexers over a period of twenty years. Figure 3 shows three (NOR)MARC records from the corpus characterized by inconsistencies between many of the metadata elements that may be explained by changes in practices over time, but even by different interpretations of the standards by the indexers. However, even if inconsistencies can be explained, we were surprised by the large number of variations we found in what could be considered as a random sample of only 99 records.

Record A	Record B	Record C
=110\$aMayhem	=110\$aMayhem	=110\$aMayhem
=24510\$aMediolanum capta est	=24510\$aGrand declaration of war	=24510\$aDeathcrush
=700 0\$aManiac	=700 1\$aManiac	=700 1\$aManiac
=700 0\$aBlasphemer	=700 1\$aBlasphemer	=700 1\$aMessiah
=700 0\$aHellhammer	=700 1\$aHellhammer	=700 1\$aAvnskog, Erik
=700 0\$aNecrobutcher	=700 1\$aNecrobutcher	=700 1\$aButcher, Necro
=710 0\$aMayhem	=700 1\$aFinstad, Børge	=700 0\$aAarseth, Øystein
=7400 \$aCarnage	=700 0\$aManiac\$tTo Daimonion	=700 0\$aSchnitzler, Conrad\$tSilvester Anfang
=7400 \$aNecrolust	=700 0\$aManiac\$tA time to die	=710 0\$aMayhem
=7400 \$aDeathcrush	=700 0\$aManiac\$tView from Nihil	=710 0\$aMayhem\$tNecrolust
=7400 \$aAncient skin	=700 0\$aBlasphemer\$tTo Daimonion	=710 0\$aMayhem\$tDeathcrush
=7400 \$aFreezing moon	=700 0\$aBlasphemer\$tA time to die	=710 0\$aVenom\$tWitching hour
=7400 \$aFall of seraphs	=700 0\$aBlasphemer\$tView from Nihil	=710 0\$aMayhem\$tWeird Manheim
=7400 \$aSilvester Anfang	=700 0\$aManiac\$tA grand declaration of war	=710 0\$aMayhem\$tChainsaw gutsfuck
=7400 \$aChainsaw gutsfuck	=700 0\$aManiac\$tA bloodsword and a colder sun	=710 0\$aMayhem\$tPure fucking armageddon
=7400 \$aFrom the dark past	=700 0\$aBlasphemer\$tA grand declaration of war	=900 1\$aEuronymous\$tAarseth, Øystein
=7400 \$aI am thy labyrinth	=700 0\$aManiac\$tIn the lies where upon you lay	
=7400 \$aSymbols of bloodswords	=700 0\$aManiac\$tCompletion in science of agony	
=7400 \$aPure fucking Armageddon	=700 0\$aBlasphemer\$tA bloodsword and a colder sun	
=900 0\$aEriksen, Rune\$tBlasphemer	=700 0\$aBlasphemer\$tIn the lies where upon you lay	
=900 0\$aStubberud, Jørn\$tNecrobutcher	=700 0\$aManiac\$tCrystallized pain in deconstruction	
=900 0\$aKristiansen, Sven-Erik\$tManiac	=700 0\$aBlasphemer\$tCompletion in science of agony	
=900 0\$aBlomberg, Jan Axel\$tHellhammer	=700 0\$aBlasphemer\$tCrystallized pain indeconstruction	
	=710 0\$aMayhem	
	=900 1\$aNecro\$tNecrobutcher	

Figure 3. Records from NORDISKO

### 5.1 Inconsistencies

A manual inspection of the MARC data showed inconsistencies in several respects: The most serious was a considerable difference in the registration of tracks. Some records used MARC field 740 \$a for track titles. The

tracks registered in this way cannot directly reveal the connection between the track title and the composer/responsible person or group. Other tracks were registered by using 700 \$a (person) or 710 \$a (group) for the track composer combined with \$t for the track title. This, according to AACR2, is the preferred way of registering tracks in albums and their respective composers, yet only a small selection of the records were registered in this way. Some of the MARC records had no tracks registered at all.

There were also differences in how the names of people were registered. Some records had pseudonyms in 700 \$a while others had real names. There is no way of knowing if the name in 700 \$a is a real name or a pseudonym just by looking at the MARC record, therefore this had to be resolved later using the aforementioned matching techniques.

The inconsistencies affect the conversion process negatively. It forces the conversion algorithms to include the complexity of potential variations, and in some cases the algorithm also has to ignore data. The different usage of indicators (e.g. 700 0 and 700 1) had a minor effect in our context.

### 5.2 Pseudonyms

Figure 4 shows a list of names from the seed data set with redundant listings of similar names representing the same person. The redundancy can have many possible causes. Some are described above and concern the lack of predictable connections between variant names in the MARC record.

As mentioned above, we had a special interest in the extensive usage of pseudonyms in the black metal genre. To convert pseudonyms we used the NORMARC field 900, which is used for expressing see and see-also-references. Sometimes this is used for expressing relationships between names and pseudonyms as in MARC record A and C in Figure 3. On other occasions, it references variant names, alternative spellings etc. as in MARC record B in the figure. The algorithm has no means of differentiating between these functionalities.

The scale of inconsistencies in a given library catalogue like NORDISKO, exposed through our experiment, gives some interesting indications

Butcher, Necro  
 Erickson, Rune  
 Eriksen, Rune  
 Fenris  
 Fenriz  
 Goat  
 Goatpervortor  
 Greifi Grishnack  
 Greishnackh, Greifi  
 Grishnackh, Greifi  
 H.M. Daiomonion  
 H.M.P.D.K.  
 Haraldstad, Kjell Vidar  
 Haraldstad, Kjetil Vidar  
 Necro  
 Necrobutcher  
 Stubberud, Jørn

**Figure 4. Unique artists from seed data set after conversion**

of potential sources of semantic interoperability flaws which must be taken into consideration in the Linked data production cycle at large.

## 6 INTERLINKING NORWEGIAN BLACK METAL

According to the premise of automation we need computers to understand and match data from our seed collection with corresponding data in the target collection. Many approaches have been suggested for achieving computational interoperability and interlinking (e.g. Correndo, Salvadores, Millard, Glaser, & Shadbolt, 2010; Melnik, Garcia-Molina, & Rahm, 2002; Wache & Voegelé, 2001; Zheng & Madnick, 2012). In this experiment, we have been inspired by Raimond, Sutton, & Sandler (2008) and their basic techniques for matching graphs structures in data sets describing music. Their techniques utilize algorithms calculating the probability that two instances are actually referring to the same thing by comparing triples connecting literal representations of different entities. When the probability measure is significant we want the machine to carry out the interlinking by using the OWL property `sameAs`.

Figure 5 represents the band Mayhem in our data set and by the corresponding Mayhem in the MusicBrainz dataset. The question mark represents the unstable foundation of a potential semantic interoperability and the efficiency of the matching algorithm.

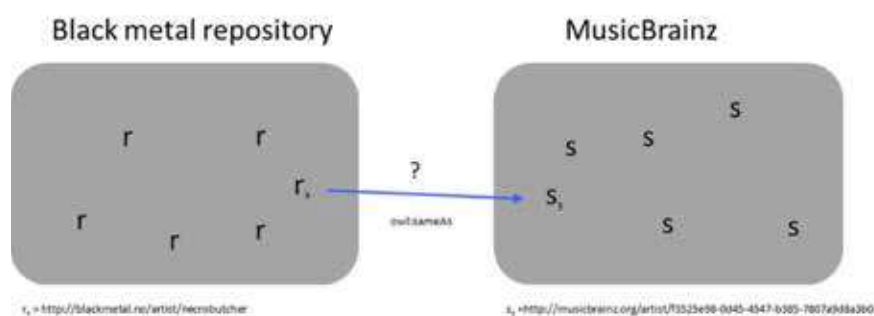


Figure 5. Potential relations between data sets

### 6.1 Matching literals

At the most basic level we could instruct the machine to compare pure literals. This is often referred to as a naïve interlinking. This will only succeed when the instances are relatively unique. A naïve approach would, for instance, fail, if we wanted to match the black metal band Mayhem

which is a literal name shared by 10 different groups or artists in the MusicBrainz collection. In our interlinking algorithm, we eventually match literals, but in order to overcome the obvious problems of the naïve approach, we included some more information from the graph structure into the equation.

### 6.2 Simple graph matching

We based the development of the matching algorithm on a limited case where we wanted to assign links between artists (MusicArtist) in both data sets. The entity artist was chosen because the data had the highest degree of inconsistency, due to the aforementioned challenges regarding the use of pseudonyms and redundant representations of names. By choosing the MusicArtist entity we aimed for a deduplication of these redundant data as a positive side effect of the interlinking process.

In order to strengthen the matching, we chose to include the tracks as a qualifier. There were several reasons for this choice. Tracks gave us an interesting and extra challenge due to the fact that the relationships between artists and their tracks were structured differently in the two data sets. In our RDF graph, we assigned triples directly between the artist and the track on the basis of connections in the MARC records

(<http://blackmetal.no/artist/necrobutcher> dc:creator

[http://blackmetal.no/track/freezing\\_moon](http://blackmetal.no/track/freezing_moon)). Such triples are also to be found in the MusicBrainz collection, but just for some artists and tracks, and seldom in the case of Norwegian black metal. Therefore we had to make a SPARQL request that indirectly related artists and tracks via their memberships of groups, as shown in Figure 6. This indirect relationship may in some cases be misleading because all artists will be credited all the songs of the group, independently of their varying periods of membership. We chose, however, this solution in our experiment as it potentially would provide some interesting outcomes that we could use as a benchmark for future tuning of the algorithm.

Another argument for including tracks was simply that they represented the largest quantity of instances. This made it a robust candidate as a qualifier. We could, in fact, also have chosen to include additional qualifiers such as groups and records. We opted for this solution as a pragmatic and workable strategy. It would nevertheless be interesting to make such inclusions in the future tuning of the algorithm, especially in order to generalize it.

```

prefix foaf: <http://xmlns.com/foaf/0.1/>
prefix mo: <http://purl.org/ontology/mo/>
prefix dc: <http://purl.org/dc/elements/1.1/>

```

```

SELECT DISTINCT ?artist_name
WHERE {
  ?track dc:creator ?group ;
  a mo:Track ;
  dc:title "Freezing moon" .
  OPTIONAL {?group mo:member ?artist .
  ?artist foaf:name ?artist_name . } }

```

Figure 6. SPARQL query matching a given track from seed with tracks in target

Figure 7 shows the relevant section of the graphs, respectively in our seed collection, The black metal repository, and in the target collection, MusicBrainz.  $\text{Artist}_s$  represents a given artist in the seed collection, while  $\text{Artist}_t$  represents a given artist in the target collection. In the example, the artist *Necrobutcher* is set as a starting point for the algorithm. *Freezing moon* and *Funeral Fog* are two of the tracks that Necrobutcher has partly been responsible for through his membership in the band *Mayhem*. In our seed collection this relationship is expressed directly, but in MusicBrainz we have to relate Necrobutcher and these songs via his membership in Mayhem. What we in practice want the algorithm to do in this case is to perform a positive match on the basis of the similarities between all literals involved. In another hypothetical example where the artist literal were dissimilar, but the tracks still represented an exact match (what would be the case if we started out with Necrobutcher's real name, *Jørn Stubberud*, and matched it against Necrobutcher) the total similarity measure would still be quite high, but lower than the measure from the first example. In future experiments we will investigate if certain levels of similarity measures could be used to determine the probability of two different names being an original name and a pseudonym.

### 6.3 ORDO AB CHAO

Figure 8 outlines the algorithm as a flow chart. The initial step is to construct a SPARQL query which retrieves data from our seed collection as a list of artist names ( $\text{Artist}_s$ ) and their track titles ( $\text{Track}_s$ ). The artist names are used as input to the second SPARQL query matching artist in

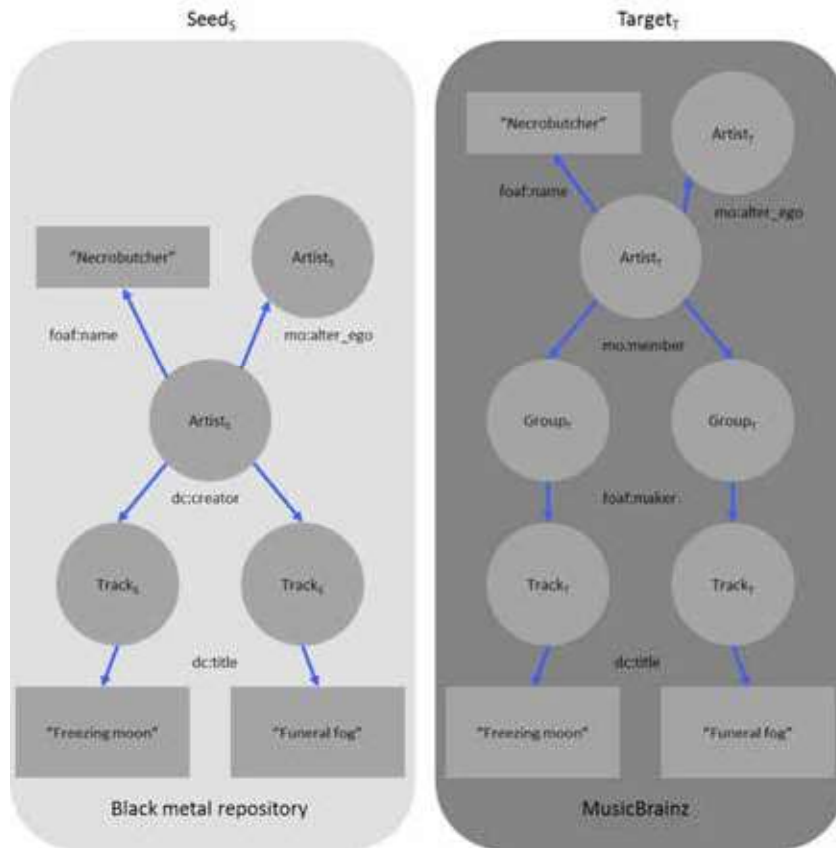


Figure 7. Seed and target graphs

the target data set. If an exact match is found, a fixed score weight ( $\lambda$ ) between 0 and 1 is added. SPARQL lacks fuzzy matching which means slight differences in name forms result in no match. The third SPARQL query matches track related to a given artist in the seed data set with tracks in the target set. In the case of a match, a PHP script calculates a similarity measure between the track titles returned from SPARQL #1 (seed) and SPARQL #2 (target). If the similarity is above a given threshold, the owl:sameAs link is created between  $Artist_s$  and  $Artist_t$ .

The match between  $Artist_s$  and  $Artist_t$  is calculated with the following similarity measure:

$$\text{similarity}(Artist_s, Artist_t) = \lambda + \left( \frac{\text{number of exact matchings}(Track_s, Track_t)}{\text{total number of } Track_s} \right) * (1 - \lambda)$$

Figure 9. The Ordo ab chao similarity measure



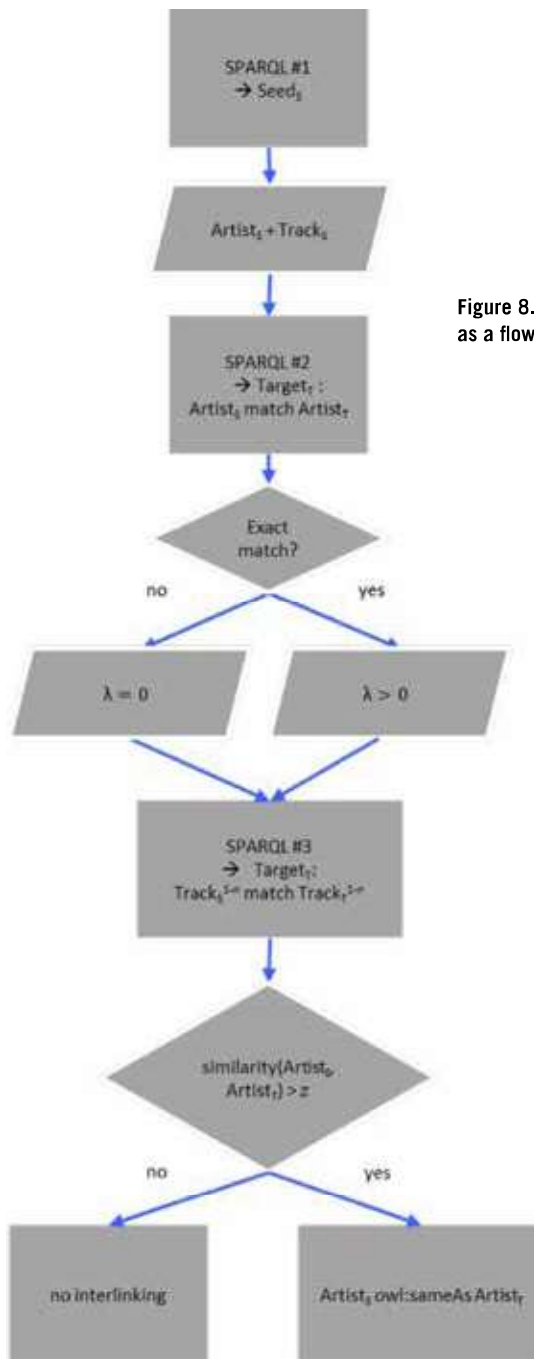


Figure 8. The matching algorithm outlined as a flow chart

## 7 EVALUATION AND CONCLUDING REMARKS

We have investigated two research questions in this paper:

- RQ1: What kind of interoperability issues emerge from a modelling of the metadata of musical recordings, using best practice Linked data guidelines?
- RQ2: What are the necessary steps in an automatic conversion process of musical recordings' metadata, following best practices?

With respect to RQ1 we have shown potential interoperability flaws originating from specific contexts, including indexing inconsistency and inconsistencies in the source material (e.g. in the mixed use of names and pseudonyms) (Section 5.1). Conversion procedures need to take into consideration that similar interoperability issues may occur in both seed and target sets.

For RQ2 we performed an experiment based on a procedure documented in Section 6. In this experiment we have experienced difficulties in both the conversion process from the original format to the recommended RDF graph structure and in the process of automated interlinking. We also found, in our efforts to convert and interlink data, some positive side effects regarding the potential for metadata clean up in the seed collection. This is due to the RDF graph structure which makes it possible to match entities by taking the graph structure into account. However, we have not fully evaluated these effects, although our suggested algorithm of matching artists in two data sets gives some indications that it could be fruitful to investigate further.

One way to evaluate our method would be to use naïve interlinking as a baseline matching method. With the help of a controlled test collection we could compare our approach with the baseline to see how much better our method would be (logically it must be improved since the naïve interlinking is an integrated part of it). Another approach would be to evaluate the matching using typical IR measures like precision and recall, and compare the degree of relevant matchings after adjusting different variables in our laboratory structure, for instance, similarity calculations or data input.

## REFERENCES

- Berners-Lee, T. (2006). *Linked data: design issues*. W3C.
- Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked Data: The story so far. (T. Heath, M. Hepp, & C. Bizer, Eds.) *International Journal on Semantic Web and Information Systems*, 5(3), 1–22. doi:10.4018/jswis.2009081901
- Correndo, G., Salvadores, M., Millard, I., Glaser, H., & Shadbolt, N. (2010). SPARQL query rewriting for implementing data integration over Linked data. *Proceedings of the 1st International Workshop on Data Semantics*.
- Dahl, T. A., Knutsen, U., & Tallerås, K. (2012). Mellom tradisjonen og weben: katalogisering, metadata og bibliotekarutdanning. In R. Audunson (Ed.), *Krysspeilinger: Perspektiver på bibliotek- og informasjonsvitenskap* (pp. 141–163). Oslo: ABM-Media.
- Kalfoglou, Y. (Ed.). (2010). *Cases on Semantic Interoperability for Information Systems Integration: Practices and Applications*. New York: Information science reference.
- Ma, H., Mei, K., Chung, C. K. J., & Amor, R. (2006). Testing Semantic Interoperability. *Joint International Conference on Computing and Decision Making in Civil and Building Engineering* (pp. 1–10).
- Mao, M. (2008). *Ontology mapping: Towards semantic interoperability in distributed and heterogeneous environments*. University of Pittsburgh.
- Melnik, S., Garcia-Molina, H., & Rahm, E. (2002). Similarity flooding: A versatile graph matching algorithm and its application to schema matching. *Proceedings of the 18th International Conference on Data Engineering*.
- Raimond, Y., Sutton, C., & Sandler, M. (2008). Automatic interlinking of music datasets on the semantic web. *Linked Data on the Web*.
- Tolk, A., & Muguiru, J. (2003). The levels of conceptual interoperability model. *Proceedings of the 2009 Spring Simulation Multiconference*.
- W3C. (2012). Linked data cookbook. Retrieved from [http://www.w3.org/2011/gld/wiki/Linked\\_Data\\_Cookbook](http://www.w3.org/2011/gld/wiki/Linked_Data_Cookbook)
- Wache, H., & Voegele, T. (2001). Ontology-based integration of information: a survey of existing approaches. *Ontologies and information sharing*.
- Zheng, X., & Madnick, S. E. (2012). SPARQL query mediation over RDF data sources with disparate contexts. *Linked Data on the Web*.



## Paper E

Tallerås, K., & Pharo, N. (2017). Mediation machines: How principles from traditional knowledge organization have evolved into digital mediation systems. *Information Research*, 22(1).



Proceedings of the Ninth International Conference  
on Conceptions of Library and Information  
Science, Uppsala, Sweden, June 27-29, 2016

## Mediation machines: how principles from traditional knowledge organization have evolved into digital mediation systems

Kim Tallerås and Nils Pharo

**Introduction.** We discuss digital information systems' ability to mediate cultural resources.

Mediation techniques embedded in search and recommendation systems are compared with those activities developed for mediating culture heritage in libraries, archives and museums, or so-called LAM-institutions.

**Method.** Digital mediation systems are examined in light of theories and techniques from knowledge organization, exemplified with implementations of such theories and techniques in public libraries.

**Analysis.** Our analysis sheds light on similarities between the digital mediation in recommendation systems and libraries' mediation of culture, but also reveals some important differences.

**Results.** We find that the digital mediation systems follow many principles and techniques of traditional knowledge organization such as those related to classification and metadata. Further they mimic the librarian who knows her users, knowledge organization systems and collection. An important challenge is the mechanical rationality embedded in the computation of recommendations, which may limit the exposure of material of interest to the user that the system finds irrelevant.

**Conclusion.** Digital mediation systems have implemented traditional theories and techniques of knowledge organization, and they can be interpreted as "mediators" in a LAM context. However, their mechanical approach to information behaviour risk to be inconsistently adaptive to users emotional needs and to not facilitate serendipitous discoveries very well.

Mediation is an essential task in all kinds of libraries, archives and museums, so-called LAM-institutions, holding a collection of resources to be made available to an audience. In library schools librarians are trained to become skilled intermediaries, being able to analyse information or recreational needs, and to connect those needs to relevant resources. Meanwhile, Google claims that their "mission is to organize the world's information and make it universally accessible and useful." Their vast suite of search systems, databases and other services have helped them come far in realizing this goal. Does this mean that Google is a mediator, performing the same kind of mediation as LAM-institutions? Although the users' interaction with Google is "faceless" and based on algorithms, these algorithms use knowledge about human information behavior and needs as a starting point. At the same time companies like Amazon and Netflix use sophisticated algorithms to create tailored recommendations of books and movies for their customers. They can thus be considered as digital intermediaries, in a communication of filtered and targeted information, from machines to users.

However, systems for "faceless" mediation are not a new phenomenon, neither in LAM-institutions. Different systems and techniques have been used to guide users to the relevant documents – from analogue list of books in antiquity and card files in the twentieth century to contemporary online catalogues. This is the tradition of knowledge organization which the digital systems mentioned above partly build upon. In this paper, we aim to discuss knowledge organizational aspects of digital mediation of cultural products. We will start by showing how analogue principles of knowledge organization are implemented

in physical libraries. Thereafter we discuss the concept mediation in light of system-based processes. In the third part, we present digital retrieval and recommendation systems before we discuss some limitations of “mediation machines”.

In the paper, we take mediation of culture in the public library as a starting point and provide some initial examples from this domain of practice. The term “kulturformidling” is used in the Scandinavian library and information science (LIS) literature to denote what we have translated into “mediation of culture” ([Grøn, 2010](#); [Tveit, 2004](#)). The concepts are under continuous debate and no English terms directly cover the activities described by the Scandinavian term. Central in common definitions is the intermediary as a person helping users to find books or other culture products that satisfy their informational or recreational needs.

## Mediation in libraries

In a traditional public library, fiction is organized alphabetically on the shelves according to the authors' last name. Fiction in Norwegian is placed separately from fiction in other languages. The scientific literature is organized according to subject using Dewey's Decimal Classification (DDC). Thus books that the librarians have assessed as being about birds are placed under the DDC number 598, whereas “To kill a mockingbird” by Harper Lee is placed alphabetically on “Lee” in the section of English fiction.

The book on birds will be surrounded by other books on birds, animals and natural history whereas Lee's books will all (or both) be placed together. These examples of providing order by shelving reflect how principles of mediation are embedded in the knowledge organization principles used in physical libraries. Such principles can be general, such as the alphabetical systems, or they may be based on well-established classification standards like the DDC.

In addition to offering direct access to the physical documents, libraries provide searchable access points to their collections via metadata, “data about data”, collected in catalogues. The catalogue has had many forms; organized chronologically in book form; as cardboard cards filed under titles, subjects and author names; and currently in digital form available via the Internet, facilitating access via all metadata recorded for the documents.

Within literary science different kinds of “meta” information is often referred to as “paratexts”. The Danish library and information scientist Jack Andersen ([2002](#)) discusses how metadata in the form of bibliographic records can be interpreted in light of Genette's ([1997](#)) paratext concept and the consequences of such an interpretation on information retrieval and reading:

*For instance, the initial relevance judgments happen when a user is confronted with the bibliographic record. What decisions a user makes as to its relevance are based on the paratextual elements present. That way the bibliographic record affects the reading activity of the user. ([Andersen, 2002, p. 59](#))*

The bibliographic record, continues Andersen, viewed as a text, does not provide mere “access” to the document, but is “rather, a matter of indicating what kind of intellectual content is to be expected” ([Andersen, 2002, p. 59](#)). Reading starts in the bibliographic record, which guides the document selection and subsequent the reading activity. Thus, metadata do not only facilitate documents retrieval, but represents an adjustment and an initial mediation of the documents' content.

Cataloguing principles as formulated by Cutter ([1876](#)) and the International Federation of Library Associations and Institutions ([2009](#)) (IFLA) concretely express such adjustments, stating e.g. that the catalogue shall enable the user:

- 4.1. to find bibliographic resources in a collection as the result of a search using attributes or relationships of the resources:



- 4.1.1. to find a single resource
- 4.1.2. to find sets of resources representing
  - all resources belonging to the same work
  - all resources embodying the same expression
  - all resources exemplifying the same manifestation
  - all resources associated with a given person, family, or corporate body
  - all resources on a given subject
  - all resources defined by other criteria (language, place of publication, publication date, content type, carrier type, etc.), usually as a secondary limiting of a search result;
- 4.2. to identify a bibliographic resource or agent (that is, to confirm that the described entity corresponds to the entity sought or to distinguish between two or more entities with similar characteristics);
- 4.3. to select a bibliographic resource that is appropriate to the user's needs (that is, to choose a resource that meets the user's requirements with respect to medium, content, carrier, etc., or to reject a resource as being inappropriate to the user's needs);
- 4.4. to acquire or obtain access to an item described (that is, to provide information that will enable the user to acquire an item through purchase, loan, etc., or to access an item electronically through an online connection to a remote source); or to access, acquire, or obtain authority data or bibliographic data;
- 4.5. to navigate within a catalogue and beyond (that is, through the logical arrangement of bibliographic and authority data and presentation of clear ways to move about, including presentation of relationships among works, expressions, manifestations, items, persons, families, corporate bodies, concepts, objects, events, and places). ([International Federation of Library Associations and Institutions, 2009, pp. 3–4](#))

The principles are developed to ease library users' access to the documents, e.g. letting the users choose books based on their format (point 4.3). In addition, the principles recommend concrete adaption of the content stating metadata should couple all resources that belong to a particular "work" (point 4.1.2).

Point 4.5 describes a type of navigation across documents, which was difficult in a card-based catalogue, but has become much easier with the documents being digitized and made available with the help of web technology. An experiment at Oslo public library illustrates how literature mediation has gained from digital technology. Metadata representing works by a selection of important Norwegian authors were analyzed in order to see how well they fit the functional requirements for bibliographic records, the so-called FRBR-model, specified by IFLA ([International Federation of Library Associations and Institutions Study Group on the Functional Requirements for Bibliographic Records, 1998](#)). A central part of this model, the so-called group 1-entities, represents different document "conditions". A "work" represents the intellectual or artistic creation (e.g. Shakespeare's *Macbeth*), an "expression" is the form the work takes when it is realized (e.g. the newest Norwegian translation of *Macbeth*), the "manifestation" is the physical embodiment of the expression (e.g. a pocket book edition), and an "item" being one single copy of the manifestation. The analyzed metadata, created following old cataloguing rules developed for the card catalogue, had a manifestation focus. Queries in the metadata resulted in very complex result lists from the OPAC (Online Public Access Catalog). A query on the author Knut Hamsun returned 585 hits, separately listing, e.g., all editions of the same works, all parts of collected works and all translations. In a physical shelf-based system, this way of knowledge organization was a necessity. With some restructuring of the data querying the same data set in an experimental system reduced the result list to 40 genuine Hamsun works ([Westrum, Rekkavik, and Tøllerås, 2012](#)). The interface facilitated user navigation between different translations and editions of the works. Some studies have indicated that the FRBR-model reflects the users' mental models of the bibliographic universe (e.g. [Pisanski and Žumer, 2010](#)).

After these experiments, Oslo public library has decided to transfer their bibliographic metadata to a system based on new standards and technologies since the currently used standards do not solve the needs for literature mediation in a large public library. The standards, the Anglo American Cataloguing Rules and the Machine-Readable Cataloging (MARC) were developed at the very beginning of digitization in the late 1960s. Both standards were, however, constrained by the leading knowledge organization technology: the card catalogue.

In addition to having technologically fallen behind, the knowledge organization systems have been criticized from philosophical and societal perspectives, including questions such as: How did classification systems end up with their particular classes? In what worldview have classification systems been developed? How do the classes in the system influence the use? Although DDC has been revised several times since it was first released, the basic classes developed by Melville Dewey in the late 19th century are still used.

Radford (2003) uses library classification as an example of what Michel Foucault's calls "discursive formation":

*Consider the choices made by a cataloger when allocating books to a subject heading, a call number, and a particular place on the library shelf. How does the cataloger do this task? What is the nature of the preexisting subjects (discursive formations) to which a new book can be assigned a place? What are the rules by which a book is assigned to Philosophy and not to History or Language? (Radford, 2003, p. 4)*

According to Radford the classification system, considered as an intermediary in a mediation situation, is based on rules with a discursive potential. Similarly, to librarians in a physical library the classification system conveys one out of several potential world-views. Such discursive formations not only characterize analogue systems for knowledge organization, but all kinds of mediation systems based on rules and principles of categorization. Systems developed in a digital context included.

### **"There is no shelf"**

In a digital library there are no shelves. The straightjacket requiring that a book physically can only stand in one place is off. Files that contain texts, images, sound and video are retrieved directly or via metadata describing them. A user searching for the author Neil Gaiman will also include works Gaiman co-created with other artists, e.g. Terry Pratchett, as well as documents mentioning Gaiman. Books about Norway and World War II are found when both terms are combined in a query. Some systems will know that this is an interest area of yours and will "mediate" it as a result of the simple query "Norway". The limitations of shelving is replaced by an infinite number of orders of succession.

In principle there are no limitations on what kind of (meta)data that can be used to retrieve documents and information. A user may be interested in audio books in Swahili recorded with female voices, and if the information system stores and indexes metadata representing such characteristics, it will be simple to retrieve matching documents. The same user may also be interested in books that are liked by students in sociology. This is another type of information that systems have started to collect and which can be used in a mediation process. Automatic indexing of whole documents can make all terms in a document potential retrieval endpoints. This has an enormous potential for retrieval, but at the same time raise a lot of challenges for literature mediation. Will a user be interested in being presented all texts containing the term "Gaiman"?

The flexibility and new possibilities offered by digitization are overwhelming. To retrieve and mediate digital collections a whole new set of techniques have been developed that partly build upon the analogue techniques described above and partly are based on analysis of context and user preferences.

## Quality assessment in knowledge organization and information retrieval

The purpose of systems for knowledge organization and information retrieval (IR) is to secure that their users find “documents” (including books, images, music, video, archival records and other media used for representing ideas and knowledge) that may help them solve a task, satisfy an information need or satisfy a need for recreation.

In order to evaluate how good IR systems work, the measures recall and precision are commonly used ([Baeza-Yates, 2011](#)). Recall is defined as the number of relevant documents in a retrieved set of documents divided by the number of relevant documents in the collection. Precision is the number of relevant documents in the retrieved set divided by the total number of retrieved documents. Typically, precision and recall measures are used in experimental evaluation processes following the procedure of the so-called Cranfield experiments. The goal of these experiments was to measure the efficiency of indexing systems ([Sparck Jones and van Rijsbergen, 1976](#)). Originally, the indexing systems were different types of classification systems or other manual systems. Today the same evaluation model is used for measuring the efficiency of algorithms used for search engines.

Criticism raised against the Cranfield model is based on it having as point of departure an “objective” assessment of relevance. The critics claim that relevance is individual and context dependent, making it a “fluent” measure of quality which changes over time. Defendants of the Cranfield model, on the other hand, claims that it is a good tool for securing consistent comparison of different systems since they are compared under equal conditions with controlled variables.

In the late 1960s a counter movement to the system oriented paradigm that Cranfield represented emerged. American pioneer of information science, Robert S. Taylor ([1968](#)) pointed out that the information seeker does not necessarily choose optimal strategies when trying to solve his or her needs. This is an important condition for the recall/precision based evaluation methods. Taylor refers to an empirical study conducted by Victor Rosenberg ([1966](#)) to support the claim that “‘ease of access’ to an information system is more significant than ‘amount or quality of information’ retrievable” ([Taylor, 1968, p. 181](#)). In other words, it is not necessary for the information seeker to invest lots of time and effort to find the “perfect document” as long as she is able to find “good enough” answers. Some years later Nick Belkin developed a “cognitive viewpoint” pointing out it being unreasonable to equalize information needs with document content:

*“[t]he assumption that expression of information need and document text are functionally equivalent also seems unlikely, except in the special case in which the user is able to specify that which is needed as a coherent or defined information structure. A document, after all, is supposed to be a statement of what its author knows about a topic, and is thus assumed to be a coherent statement of a particular state of knowledge. The expression of an information need, on the other hand, is in general a statement of what the user does not know” ([Oddy, Belkin, and Brooks, 1982, p. 64](#))*

Marcia Bates was, with her “berrypicking” model ([Bates, 1989](#)), among the first to develop an alternative model of user-system interaction. In her model she emphasizes that several types of search behavior may satisfy the user’s information needs. Not all of these can be evaluated using precision and recall. A user may, e.g., browse different potential sources, pick a little bit of information from each source, look at reference lists, and get some ideas from colleagues while continuously reformulating her information need, dependent on what is found. The “berrypicking” metaphor is based on such shifts between “berry patches”. It is a good model to explain how users construct their information needs through iterative processes.

When evaluating how well a system is for mediating cultural resources, user-centered approaches such as Bates' model, are invaluable. As we shall see, technological development has made it possible to develop more sophisticated IR systems that take into account user models. One example of such systems is recommendation systems. We shall discuss how recommendation systems build upon user knowledge and problematize the challenges of such systems.

## Recommendation systems

In full-text IR systems, such as Google, the distribution of terms in the documents has played the most important part in indexing algorithms. The term frequency-inverse document frequency (tf-idf) weight ([Sparck Jones, 1972](#)) was developed to reflect how important a term was for representing a document. Other components have been included in the retrieval algorithms, but most of these have been document centric. This also includes Google's PageRank ([Brin and Page, 1998](#)), which was inspired by citation networks and, put very simply, gives weight to web pages depending on their number of ingoing links. Relevance feedback ([Salton and Buckley, 1990](#)) represents an attempt at implementing user preferences in the retrieval process. Users assess the relevance, explicitly or implicitly, of the retrieved documents and the system uses content in the relevant documents to retrieve documents similar to these. Pseudo-relevance feedback ([Efthimiadis, 1996](#)) is particularly interesting, since it is based on the assumption that the highest ranked documents in the retrieved list are relevant and thus the system automatically uses these documents' content to retrieve the final result set. Web based IR systems also often use "cookies" to collect data about the user in order to build profiles to tailor and personalize query results.

Recommendation systems try to predict what documents the users are interested in. Companies like Amazon and Netflix have been in the forefront in developing recommendation techniques, but such techniques are also used by non-commercial services. At Oslo public library, the service [Aktive hyller](#) (Active shelves) use elements based on this technology when it recommends related books based on a patron's current selection. The service collects rating data from three different sources ([Goodreads](#), [Bokelskere](#), and [NovelList](#)) and suggest books that are assessed as similar in topic and genre. The concept "recommendation system" is used to describe everything from simple top 10-lists based on general consumption frequency ("the 10 most read news articles") to personalized recommendations based on complex forms of social profiling and network analysis.

Basically there are two types of recommendation systems, *content-based filtering* and *collaborative filtering systems* ([Ricci, Rokach, and Shapira, 2011](#)). Content-based filtering systems base their recommendations on comparing the characteristics of the documents' content, e.g., their genre, topicality and format. A user that previously has liked crime books in audio format where the action is located in Oslo will probably be interested in other books with the same characteristics. Content-based filtering may be based on traditional knowledge organization techniques such as cataloguing and classification. The content must be described in a precise, consistent and exhaustive way to facilitate the best possible filtering. These description may be created by experts, but in some cases users will be co-creators of the metadata, e.g. when adding un-controlled keyword ('tags'). Oslo public library's experiments with "FRBRizing" their collection, which we described above, makes it possible to accumulate recommendations on the work level and reduce the so-called cold-start problem ([Schein, Popescul, Ungar, and Pennock, 2002](#)) which is caused by having to few recommendations per item.

Collaborative filtering uses characteristics of the user and the user's digital "neighborhood" with other users. Data used in collaborative filtering systems can be "self-exposure" in the form of purchases, library loans, ratings, reviews, wish lists and other forms of assessments. In addition systems may register demographic data, such as age and gender, and implicit data from systems

logs that register clicks, navigation patterns and consumption techniques (e.g. from e-book readers or streaming services). These data are used for user profiling and the idea is that users that have similar profiles have an overlapping taste in books, music, films etc.

Often the different techniques are combined in *inhybrid* recommendation systems. User A, who has a certain profile, may give a book with specific formal and literary characteristics a high rating. In this way, the book increases its coupling with other books with similar characteristics that other users, with profiles that are similar to User A, have rated highly. The combined approach is another way of reducing cold-start problems for new items.

Recommendation systems may be evaluated using techniques similar to those used in experimental IR, i.e. the Cranfield model. When Netflix [organized a competition](#) to improve their recommendation system the goal was to improve the accuracy of their own "Cinematch" algorithm with more than 10 %. Suggested algorithms used a training set of Netflix data consisting of 100 million user ratings given by 480 000 users on 17 700 films. In the competition data set (2.8 million films) ratings were removed and the goal was to use the training set to recreate or predict the ratings of the films in the competition set. This is parallel to the Cranfield-test collection method where queries are matched against relevance-assessed documents. The "best" algorithms are those best at retrieving the documents assessed relevant for the queries ([Sparck Jones and van Rijsbergen, 1976](#)). There are, however, also attempts at involving users more directly in evaluating recommendation systems ([Shani and Gunawardana, 2011](#)).

Digital mediation of literature with recommendation systems is a promising idea. In many ways, the recommendation systems mimic the librarian who knows her user and collection. Interviewing the user and knowing their book borrowing history and how other patron have used the library, the librarian come up with suggestions. The automatization of such processes and the digital library's lack of shelves raise some issues. Of these, the meeting between the rationalizations embedded in computer algorithms and the users' various forms of needs, is among the most important. Related to this we will also discuss how recommendation systems probably decrease the chances of serendipity.

As mentioned above, users often may be satisfied with answers that are good enough when searching for information. Denise Agosto ([2002](#)) discussed the concept "satisficing", originally coined by Herbert Simon, and how it relates to information searching and Web-based decision making. Of particular interest is Agosto's reference to Kuhlthau's ([1991](#)) work on the "information seeking process" and how this is not purely a cognitive process, but also has an affective dimension. Nick Belkin points out that "there has been almost no serious research effort in understanding the role of affect in the information seeking situation in general and the IR situation in particular, nor in IR system design." ([Belkin, 2008, pp. 50-51](#)). IR algorithms are less capable of implementing emotional than cognitive aspects. In particular, this is evident when the algorithms are not only used for solving informational needs, but also to satisfy users' needs for affection and recreation. We do not think that *emotion retrieval* (ER) will take place separately from IR systems, since users will also express their emotional needs with terms that can be matched with an IR algorithm. Thus, ER may be performed with IR algorithms in combination with other techniques. Probably it is possible to adjust weights that better take into account user preferences, e.g. in the form of "likes", in recommendation systems. In his doctoral thesis Moshfeghi ([2012](#)) tested out "emotion information" in two collaborative filtering systems and found that they perform better when taking emotion features into account compared to when only rating information is considered. Considerable work is necessary to meet user's affectional and recreational needs.

The "rationality" of IR systems and recommendation system also may affect

serendipity, i.e. finding something by chance. In a physical library, the user is exposed to shelves of books, magazines, posters on the wall and many other “irrelevant” information sources that may influence him or her. Björneborn (2008) identified ten factors in the library that may be a source of serendipity, including “explorability” and “browsing”. Elaine Toms suggests four approaches to research in order to facilitate serendipity in IR:

1. Role of chance or ‘blind luck’: implemented via a random information node generator.
2. Pasteur principle („chance favours the prepared mind“): implemented via a user profile.
3. Anomalies and exceptions: partially implemented via poor similarity measures.
4. Reasoning by analogy: implementation is unknown at the moment. (Toms, 2000)

André, Schraefel, Teevan and Dumais (2009) points out that serendipity consists of two different aspects, the first being “its accidental nature and the delight and surprise of something unexpected (e.g., the synthesis of copper phthalocyanine)”, whereas the second is “the breakthrough or discovery made by drawing an unexpected connection – the sagacity (e.g, using copper phthalocyanine as dye)”. The focus of system designers, they claim, has been to try to facilitate the former whereas the latter has been ignored. Therefore they argue that a more holistic picture of serendipity and have several suggestion on paths to follow, including the support of domain expertise, creation of common language models and facilitation of networks.

## Summary and conclusions

We have shown how mediation of culture has been embedded in knowledge organization systems since when they were analogue up until the rather sophisticated recommendation systems of today. In her book on mediation of literature (“litteraturformidling”) Åse Kristine Tveit states that to “index is to mediate” (Tveit, 2004, p. 17) (our translation). However, she draws a distinction between this kind of “technical” mediation and a more personal mediation, which requires a direct initiative from the intermediary. The distinction is seemingly in contrast with our description of analogue and digital systems for mediation. One could perhaps argue that knowledge organization represents a *second-order mediation* (inspired by the terminology of Weinberger (2007)), extending mere (first-order) accessibility of material with systematized metadata. However, interpreted as paratexts one could also argue that metadata facilitates direct (third-order) mediation, by guiding cultural consumption. Today, when search and recommendation systems have connected typical LAM-metadata to user data and “mined” them algorithmically, with customized recommendations as a result, they are definitively close to adapt the mediation performed by flesh-and-blood-librarians. Thus, we argue that such “mediation machines” do facilitate direct interaction between cultural products and their potential users, and that they can be interpreted as a mediator of culture in line with the modern practices of LAM-institutions.

This adaptation is not free of challenges. Modern information systems, or “mediation machines”, have the capability to accurately match users’ information need. Such systems, however, face the challenge of becoming too “rational” and not facilitate serendipitous discoveries. Although attempts have been made to address these problems, the ideas are mainly theoretical. Implementing emotion retrieval and serendipity-sensitive retrieval has proven to be difficult.

It should also be noted that the motivation of Google and other commercial vendors of digital mediation services differs a lot from the purposes served by LAM-institutions. The latter have specific social responsibilities, are often funded by public money and regulated by laws. This stand in contrast to the commercial business models of the former. The two types of services we have

compared thus may have very different understanding of mediation as a concept, and further in the realization of mediation techniques. This would be an interesting topic for further investigation.

## About the authors

**Kim Tallerås** is a research fellow at the Department of Archivistis, Library and Information Science at Oslo and Akershus University College of Applied Sciences. He can be contacted at: [kim.talleras@hioa.no](mailto:kim.talleras@hioa.no).

**Nils Pharo** is Professor at the Department of Archivistis, Library and Information Science at Oslo and Akershus University College of Applied Sciences and can be contacted at [nils.pharo@hioa.no](mailto:nils.pharo@hioa.no).

## References

- Agosto, D. E. (2002). Bounded rationality and satisficing in young people's Web-based decision making. *Journal of the American Society for Information Science and Technology*, 53(1), 16–27.
- Andersen, J. (2002). Materiality of works: the bibliographic record as text. *Cataloging & Classification Quarterly*, 33(3–4), 39–65.
- André, P., Schraefel, M. C., Teevan, J., & Dumais, S. T. (2009). Discovery is never by chance: designing for (un)serendipity. In *Proceedings of the Seventh ACM Conference on Creativity and Cognition* (pp. 305–314). New York, NY: ACM.
- Baeza-Yates, R. (2011). *Modern information retrieval: the concepts and technology behind search* (2nd ed.). Addison Wesley.
- Bates, M. J. (1989). The design of browsing and berrypicking techniques for the online search interface. *Online Information Review*, 13(5), 407–424.
- Belkin, N. J. (2008). Some(What) Grand challenges for information retrieval. *SIGIR Forum*, 42(1), 47–54.
- Björneborn, L. (2008). Serendipity dimensions and users' information behaviour in the physical library interface. *Information Research*, 13(4), paper 370. Retrieved from <http://www.informationr.net/ir/13-4/paper370.html> (Archived by WebCite® at <http://www.webcitation.org/6REoRyS5E>)
- Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and Isdn Systems*, 30(1–7), 107–117.
- Cutter, C. A. (1876). *Rules for a printed dictionary catalogue*. Washington.
- Efthimiadis, E. N. (1996). Query expansion. *Annual Review of Information Science and Technology*, 31, 121–87.
- Genette, G. (1997). *Paratexts: thresholds of interpretation* (Vol. 20). Cambridge University Press.
- Grøn, R. (2010). Oplevelsens rammer : former og rationaler i den aktuelle formidling af skønlitteratur for voksne på danske folkebiblioteker : ph.d. afhandling fra forskningsprogrammet Videnskultur og vidensmedier på Danmarks Biblioteksskole. Det Informationsvidenskabelige Akademi.
- International Federation of Library Associations and Institutions. (2009). *IFLA cataloguing principles: the Statement of International Cataloguing Principles (ICP) and its glossary: in 20 languages* (Vol. v. 37). KGSaur.
- International Federation of Library Associations and Institutions Study Group on the Functional Requirements for Bibliographic Records. (1998). *Functional requirements for bibliographic records: final report* (Vol. 19). Saur.
- Kuhlthau, C. C. (1991). Inside the search process: information seeking from the user's perspective. *Journal of the American Society for Information Science*, 42(5), 361–371.
- Moshfeghi, Y. (2012). *Role of emotion in information retrieval* (Ph.D.). University of Glasgow. Retrieved from <http://theses.gla.ac.uk/31118/> (Archived by WebCite® at <http://www.webcitation.org/6kz5q7Xp4>)
- Oddy, R., Belkin, N., & Brooks, H. (1982). ASK for information retrieval: part I. background and theory. *Journal of Documentation*, 61–71.
- Pisanski, J., & Žumer, M. (2010). Mental models of the bibliographic universe. part 1: mental models of descriptions. *Journal of Documentation*, 66(5), 643–667.
- Radford, G. P. (2003). Trapped in our own discursive formations: toward an archaeology of library and information science. *Library Quarterly*, 73(1), 1–18.

- Ricci, F., Rokach, L., & Shapira, B. (2011). Introduction to recommender systems handbook. In F. Ricci, L. Rokach, B. Shapira, & P. B. Kantor (Eds.), *Recommender Systems Handbook* (pp. 1–35). Springer.
- Rosenberg, V. (1966). *The application of psychometric techniques to determine the attitudes of individuals toward information seeking and the effect of the individual's organizational status on these attitudes*. Bethlehem, PA: Center for Information Science, Lehigh Univ.
- Salton, G., & Buckley, C. (1990). Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41(4), 288–297.
- Schein, A. I., Popescul, A., Ungar, L. H., & Pennock, D. M. (2002). Methods and Metrics for Cold-Start Recommendations. In *Proceedings of the 25th Annual International Acm Sigir Conference on Research and Development in Information Retrieval* (pp. 253–260).
- Shani, G., & Gunawardana, A. (2011). Evaluating Recommendation Systems. In F. Ricci, L. Rokach, B. Shapira, & P. B. Kantor (Eds.), *Recommender Systems Handbook* (pp. 257–297). Springer.
- Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1), 11–21.
- Sparck Jones, K., & van Rijsbergen, C. J. (1976). Information retrieval test collections. *Journal of Documentation*, 32(1), 59–75.
- Taylor, R. S. (1968). Question-negotiation and information seeking in libraries. *College & Research Libraries*, 29(3), 178–194. Retrieved from <http://crl.acrl.org/content/29/3/178.full.pdf+html> (Archived by WebCite® at <http://www.webcitation.org/6kz6RGynq>)
- Toms, E. (2000). Serendipitous information retrieval. In *Proceedings of the First DELOS Network of Excellence Workshop on Information Seeking, Searching and Querying in Digital Libraries* (pp. 11–12).
- Tveit, Å. K. (2004). *Innganger : om lesing og litteraturformidling*. Fagbokforl.
- Weinberger, D. (2007). *Everything is miscellaneous: the power of the new digital disorder*. Times Books.
- Westrum, A.-L., Rekkavik, A., & Tallerås, K. (2012). Improving the presentation of library data using FRBR and Linked data. *Code4Lib Journal*, (16). Retrieved from <http://journal.code4lib.org/articles/6424/comment-page-1> (Archived by WebCite® at <http://www.webcitation.org/6kz6hUB8F>)

### How to cite this paper

Tallerås, K. & Pharo, N. (2017). Mediation machines: how principles of traditional knowledge organization have evolved into digital mediation systems. *Information Research*, 22(1), CoLIS paper 1654. Retrieved from <http://InformationR.net/ir/22-1/colis/colis1654.html> (Archived by WebCite® at <http://www.webcitation.org/6oTPowyn2>)

Find other papers on this subject

Scholar Search

Google Search

Bing

Check for citations, [using Google Scholar](#)

© the authors, 2017.

267

Last updated: 8 February, 2017

[Contents](#) | [Author index](#) | [Subject index](#) | [Search](#) | [Home](#)



## **Errata-liste**

Paper B was not yet published when the dissertation was submitted. The paper included in the appendixes is therefore the revised and accepted postprint version.

The published article has the following reference:

Tallerås, K., Dahl, J. H. B., & Pharo, N. (2018). User conceptualizations of derivative relationships in the bibliographic universe. *Journal of Documentation*, 74(4), 894–916.

DOI: <https://doi.org/10.1108/JD-10-2017-0139>