Hugo C. Huurdeman

(in) formal classification

connecting formal and informal knowledge organization systems

Abstract

Recent developments in knowledge organization systems (KOS) have added new dimensions to traditional classification practices, as knowledge is now not only organized by official committees and editors, but also by the users of these systems themselves.

The aim of this study is to provide new insights into mapping formally created knowledge organization systems to socially created knowledge organization systems. To this end, a literature review, a theoretical analysis and a practical analysis have been carried out. The theoretical analysis consists of a comparison between the knowledge structure of Freebase, a semantic encyclopedia based on Linked Open Data, and the Dewey Decimal Classification (DDC), a traditional classification system. The practical analysis is based on a statistical mapping of Freebase “domains” to classes of the Dewey Decimal Classification, using information available in the WorldCat database. This analysis is of a quantitative nature.

The findings of this study can be approached on multiple levels. On a broad level, it shows that it is possible to map the structure of a socially created knowledge organization system to the structure of a traditional, formally created system, although the study also indicates that there are some fundamental differences in these systems that cannot be solved in an easy way. When we look at the level of the statistical mapping between Freebase and the DDC, this study shows that the mapping covers most of the available Freebase domains, and that it could feasible to use this approach on a broader scale. On the other hand, the study also shows that there are limits as to which features of Freebase's knowledge structure can be represented using the DDC.

Structure

This thesis starts with an outline of the topic and research questions, and continues with an extensive literature review, that discusses formally and socially created KOS, and theories and practices in mapping. Subsequently, the methodology is discussed, followed by a theoretical analysis of the structure of the DDC and Freebase, and a practical analysis based on a statistical mapping. Finally, the thesis is wrapped up in the conclusion, that also indicates suggestions for future research.

Master's thesis for the International Master in Digital Library Learning
Tallinn University, Oslo and Akershus University College, University of Parma

(in) formal classification: connecting formal and informal knowledge organization systems

by Hugo C. Huurdeman

supervisor: dr. Nils Pharo

version 1 (June 27, 2012)

www.timelessfuture.com/dill-thesis

Acknowledgements

First of all, I would like thank all of the professors and staff involved in the DILL program. In particular I would like to thank Nils, for his guidance in the process of writing this thesis. Furthermore, Elise Conradi, for facilitating my internship at the National Library of Norway, and for supplying many ideas and insights for this thesis.

I also would like to thank the DILL-4 family – in these nearly two years in Norway, Estonia and Italy, fellow students have become friends, and friends have become family. In particular Rasmus, Jenny, Dydimus, Jakaria, and Muharrem: for friendship, encouragement and (in)formal discussions when writing this thesis, and beyond.

Thanks to my friends and family at home, for encouraging me to embark on this wonderful journey, that has given me many new experiences and visions on life.

And finally, Lili, for inspiration and support.

Table of Contents

1 Introduction 11

1.1 Background 11

1.2 Research problem 12

1.3 Motivations 12

1.4 Aims 13

1.5 Research questions 13

1.6 Methodology 14

1.6.1 Theoretical analysis 14

1.6.2 Practical Analysis 14

1.6.3 Research paradigm 15

1.7 Limitations 15

1.8 Summary 16

2 Literature Review 17

2.0.1 Introduction 17

2.0.2 Defining information, knowledge and knowledge organization systems 17

2.1 Formally created knowledge organization systems 19

2.1.1 Formally created knowledge organization systems 19

2.1.2 Classification systems 21

2.1.2.1 Introduction 21

2.1.2.2 Principles 22

2.1.2.3 Types 22

2.1.2.4 Elements 23

2.1.2.5 Advantages & disadvantages 23

2.1.3 Dewey Decimal Classification 24

2.1.3.1 Introduction 24

2.1.3.2 Organization 25

2.1.3.3 Main ideas 26

2.1.3.4 Structure 26

2.1.3.4.1 Elements 26

2.1.3.4.2 Division 27

2.1.3.4.3 Relationships 28

2.1.3.5 Advantages & disadvantages 30

2.2 Socially created knowledge organization systems 31

2.2.1 Socially created knowledge organization systems 31

2.2.2 Semantic community knowledge bases 32

2.2.2.1 Introduction 32

2.2.2.2 Principles 33

2.2.2.3 Types 33

2.2.2.4 Elements 34

2.2.2.5 Advantages and disadvantages ... 34
2.2.3 Freebase ... 35
2.2.3.1 Introduction ... 35
2.2.3.2 Organization ... 36
2.2.3.3 Main ideas ... 36
2.2.3.4 Structure ... 38
2.2.3.4.1 Elements ... 38
2.2.3.4.2 Underlying structure ... 40
2.2.3.4.3 Relationships ... 42
2.2.3.5 Advantages & disadvantages ... 43
2.3 Mappings ... 45
2.3.1 The concept of mapping knowledge organization systems ... 45
2.3.1.1 Definition ... 45
2.3.1.2 Types of mappings ... 46
2.3.1.3 Mapping relationships ... 46
2.3.2 Existing mappings ... 48
2.3.2.1 Dewey Decimal Classification ... 48
2.3.2.1.1 Mapping DDC to thesaurus terms ... 48
2.3.2.1.2 Mapping DDC to subject headings ... 49
2.3.2.1.3 Mapping DDC to other classification systems ... 49
2.3.2.1.4 DDC as a switching system ... 50
2.3.2.2 Freebase ... 50
2.3.2.2.1 Mapping Freebase's concepts to external concepts ... 51
2.3.2.2.2 Mapping different knowledge structures to Freebase's structure ... 51
2.3.2.2.3 Mapping Freebase to other knowledge structures ... 52
2.3.2.2.4 Related mappings ... 52
2.4 Summary ... 53
3 Methodology ... 54
3.0.1 Introduction ... 54
3.1 Justification ... 54
3.1.1 Methodology ... 54
3.1.1.1 Development ... 54
3.1.1.2 Elements ... 54
3.1.1.3 Data sources ... 54
3.1.2 Limitations ... 56
3.2 Procedures ... 57
3.2.1 Introduction ... 57
3.2.2 Tools ... 58
3.2.3 Steps ... 58
3.2.3.1 Method 1 ... 59
3.2.3.1.1 Dewey “class” level ... 59

3.2.3.1.2 Dewey “division” level ... 60
3.2.3.1.3 Dewey “subdivision” level ... 60
3.2.3.2 Method 2 ... 60
3.2.3.2.1 Matching method 1 and 2 ... 61
3.2.3.3 Method 3 ... 61
3.3 Summary ... 62
4 Theoretical analysis ... 63
4.1 Criteria to compare knowledge organization systems ... 63
4.2 Comparing the structure of the DDC and Freebase ... 64
4.2.1 Domain of interest ... 64
4.2.2 Complexity ... 64
4.2.3 Size ... 65
4.2.4 Formality ... 66
4.2.5 Usage ... 66
4.2.6 General modeling principles ... 67
4.3 Conclusion ... 68
5 Practical Analysis ... 69
5.0.1 Introduction ... 69
5.1 Results ... 69
5.1.1 Overview ... 69
5.1.2 Basic mapping (method 1 & method 2) ... 69
5.1.2.1 General mapping statistics ... 69
5.1.2.1.1 Single matches ... 70
5.1.2.1.2 Multiple matches ... 70
5.1.2.2 Mapping statistics based on Freebase domain category ... 71
5.1.3 Mapping refinement (method 3) ... 73
5.1.3.1 Process ... 73
5.1.3.2 General mapping statistics ... 73
5.1.3.2.1 Single matches ... 73
5.1.3.2.2 Multiple matches ... 74
5.1.3.3 Mapping statistics based on Freebase Domain Category ... 74
5.1.3.4 Mapped domains and classes ... 75
5.1.3.4.1 Overview ... 75
5.1.3.4.2 Observations ... 77
5.1.3.4.3 Mapped Freebase domains & DDC classes ... 79
5.1.3.5 Unmapped Freebase domains ... 79
5.2 Analysis ... 82
5.2.1 Results summary ... 82
5.2.1.1 General results ... 82
5.2.1.2 Results for different Freebase domains ... 82
5.2.1.3 Results in different Dewey sections ... 83

5.2.2 Results analysis 83
5.2.2.1 Research question 83
5.2.2.2 Further analysis 84
5.2.3 Methodology improvements 86
5.3 Summary 87
6 Conclusions 88
6.1 Discussion 89
6.1.1 Implications 89
6.1.2 Suggestions for further Research 90
7 Bibliography 92
8 Appendix 97

List of Figures

-Figure 1: Distinguishing different vocabularies, based on their expressiveness (adapted from Weller, 2007)..... 21
-Figure 2: Dewey.info interface..... 24
-Figure 3: Class network around DDC's 782.29 class (Green & Panzer, 2009)..... 29
-Figure 4: Screenshot of Freebase..... 35
-Figure 5: ASK-Ken Visual Knowledge Browser..... 38
-Figure 6: Thinkbase screenshot..... 42
-Figure 7: Mapping relationships (S = source class, T = target class)..... 46
-Figure 8: Freebase Schema Explorer..... 55
-Figure 9: DeweyBrowser interface..... 56
-Figure 10: Main Dewey classes for the title search for "soccer"..... 59
-Figure 11: Dewey results at the division level..... 60
-Figure 12: Dewey results at the section level..... 60
-Figure 13: Knowledge structures arranged by their complexity and extent of the captured knowledge domain (Weller, 2010)..... 65
-Figure 14: Frequency distribution of Freebase Domains Categories, based on the number of instances (topics)..... 71
-Figure 15: Mapped Freebase classes..... 76
-Figure 16: Color codes and abbreviations..... 76
-Figure 17: Mapped Freebase domains to main DDC classes & their subordinate classes (percentage) 79
-Figure 18: Excel: basic information from Freebase (excerpt)..... 97
-Figure 19: Excel: data collection sheet (for the Freebase domains above)..... 98

List of Tables

-Table 1: Main DDC classes..... 27
-Table 2: Dewey tables..... 27
-Table 3: Relative Index entry for Garlic..... 28
-Table 4: Metaschema: higher order relationships in Freebase43..... 43
-Table 5: Basic research steps..... 58
-Table 6: Basic comparison of the Dewey Decimal Classification and Freebase..... 68
-Table 7: Freebase domains and matching DDC classes (one-to-one)..... 70
-Table 8: Freebase domains and matching DDC classes (one-to-many)..... 70
-Table 9: DDC classes assigned per Freebase domain category..... 72
-Table 10: Unique Freebase domains and matching DDC classes..... 73
-Table 11: Freebase domains and matching DDC classes (multiple matches allowed)..... 74
-Table 12: The number of Freebase domains with DDC mapped classes, using method 3..... 75
-Table 13: Freebase: unmapped domains..... 80
-Table 14: Absolute and relative values of unmapped domains..... 80
-Table 15: Grouped unmapped classes..... 81

List of Abbreviations

AAT	Getty's Art & Architecture Thesaurus
ALA	American Library Association
DILL	Digital Library Learning
DDC	Dewey Decimal Classification
DNB	German National Library (Deutsche Nationalbibliothek)
EPC	Dewey Decimal Classification Editorial Policy Committee
FOAF	Friend Of A Friend
ISO	International Organization for Standardization
KOS	Knowledge Organization System
IMDB	Internet Movie Database
LCC	Library of Congress Classification
LCSH	Library of Congress Subject Headings
LOD	Linked Open Data
OCLC	Online Computer Library Center
RDF	Resource Description Framework
SCS	Semantic Classification Search
SKOS	Simple Knowledge Organization System
SWD	Schlagwortnormdatei
UDC	Universal Decimal Classification
URI	Uniform Resource Identifier
YAGO	Yet Another Great Ontology

1 Introduction

1.1 Background

The organization of information and knowledge is a basic drive in humans, and has been with us for ages, in all aspects of life (Taylor & Joudrey, 2009). An important reason to organize is because we want to retrieve. For example, systematically organizing our kitchen utensils means that we know where we can find our items, and 'retrieval' is easier. This also holds true for other, more formal items, like books.

In libraries, classification systems have been used for a long time to organize books. One of the oldest systems that is still being used on a large scale is the Dewey Decimal Classification. It was initiated by Melvil Dewey in 1873 (Dewey, 2011), and has been published in 23 subsequent DDC editions, the latest in 2011.

These days, the internet plays a major role in the organization and dissemination of information. However, there are some limitations in the way information is handled on the internet, and it can be hard to "structure, find and retrieve information precisely and effectively" (Weller, 2007). In the past, several initiatives to use classification systems for information retrieval on the internet have been deployed, but many of them do not exist anymore (e.g. Koch, Neuroth, & Day, 2003). To solve the problem of finding relevant information in a different way, a new kind of 'web' has been proposed: the semantic web (Berners-Lee, 2001).

The premise of this 'web of data' is that access to information can be greatly improved, because machines know the meaning (semantics) of information, and automatic links between information sources can be created. More and more organizations, such as governmental and research institutions, are making their structured data available as Linked Open Data, that can be used for establishing these semantic connections. Freebase, an online encyclopedia in the style of Wikipedia, aggregates much of this data and provides editing functionalities to its users. Freebase offers the resulting datasets, in its turn, as Linked Data. It is also possible to access a subset of the Dewey Decimal Classification via Dewey.info. This creates the opportunity to make a connection between Dewey and Freebase, using this Linked Data.

However, a basic requirement for a connection between Dewey, a formal classification system, and Freebase, a community-driven ontology (Peters, 2009), is to create a mapping between both knowledge structures. This is the topic of this thesis, and it involves different theoretical and practical issues in the field of knowledge organization.

1.2 Research problem

This thesis is focused on the creation of a mapping between a formal Knowledge Organization System and a socially created Knowledge Organization System. Creating connections between different types of KOS can be difficult, as indicated by Koch, Neuroth, & Day (2001); Salah, Gao, Suchecki, & Scharnhorst (2011) there are several issues when mapping classification systems to other knowledge organization systems, due to the inherent structure of these systems.

The focus will be on the structure of the Dewey Decimal Classification, the most widely used classification system in libraries around the world (OCLC, 2003), and the (ontology-based) structure of Freebase[1], a “community knowledge base” (Weller, 2010).

1.3 Motivations

The advent of the internet for retrieving information, and the development of text-based search engines to query this information, has made the role of classification systems less important in the process of indexing, classifying and retrieving information. However, the DDC still has its place in the library world, is being updated on a regular basis and a wealth of (mostly physical) information resources have been classified using Dewey.

The rising use of 'folksonomies' (Morville & Rosenfeld, 2006) indicates that also ordinary users can be involved in classifying resources. The popularity of Wikipedia shows that users can also be motivated to create content collaboratively. Open initiatives like Wikipedia often offer possibilities to reuse their information in different ways. For example DBpedia[2] adds semantic structure to Wikipedia[3] content (see Auer et al., (2007)), and Freebase combines Wikipedia's information with other semantic information resources.

Several opportunities are opened up by combining the formally defined Dewey and the informally defined Freebase, making it worthwhile to assess the feasibility of a mapping between the two. For example, Freebase could be enhanced with information about books in WorldCat about the topic you are currently browsing, using mapped Dewey numbers. Or, library catalogs could be enhanced with contextual information from Freebase about their books' topics, also making use of a mapping between the two knowledge organization systems.

On a higher level, as indicated by Salah, Cheng, Suchecki, & Scharnhorst (2011), the exploration of these kind of mappings can aid us in the organization of knowledge,

1 http://www.freebase.com [retrieved: 19-06-2012]
2 http://www.dbpedia.org [retrieved: 13-06-2012]
3 http://en.wikipedia.org [retrieved: 13-06-2012]

to improve navigation, to bridge systems and to facilitate "overviews and deep insights into available knowledge".

1.4 Aims

The aims of this research are the following:

- to establish a theoretical grounding by analyzing the structure of classification systems and community-driven ontologies, in particular the Dewey Decimal Classification and Freebase, and by comparing these knowledge organization structures
- to carry out a mapping between Dewey and Freebase, based on statistical data
- to analyze and evaluate this mapping, using the collected data
- to summarize and visualize the (dis)similarities of the structure of Freebase and the Dewey Decimal Classification, with the help of the collected data

The research could also shed more light on the possibilities to connect formally and socially created knowledge organization systems in general, and similar to Salah, Cheng, et al. (2011): "even if the outcome is not without ambiguity, the process helps us to better understand the nature of the knowledge generation systems we deal with".

1.5 Research questions

The main research question can be defined as follows:

- To what extent can the structure of a semantic collaborative knowledge base be mapped to the structure of a classification system?

The research will focus in particular on the Dewey Decimal Classification and Freebase, but also aims to shed light on the more general concept of mapping different types of knowledge organization systems.

The literature review will provide a basic foundation for this thesis, based on the following guiding questions:

- What are formally created knowledge organization systems? (section 2.1.1)
- What are classification systems, and how are they organized? (section 2.1.2)
- What is the Dewey Decimal Classification, and how is it organized? (section 2.1.3)
- What are socially created knowledge organization systems? (section 2.2.1)
- What are semantic community knowledge bases, and how are they organized? (section 2.2.2)
- What is Freebase, and how is it organized? (section 2.2.3)
- What is a mapping, in the context of knowledge organization systems? (section 2.3.1)

- Which mappings have been carried out between the Dewey Decimal Classification, Freebase and other knowledge organization systems? (section 2.3.2)

The analysis chapters of this thesis are aimed at helping to answer the main research question, and are based on the following guiding questions, focused on the Dewey Decimal Classification and Freebase:

Theoretical analysis (chapter 4)

- How do the Dewey Decimal Classification and Freebase compare in terms of structure, based on the criteria to compare KOS defined by Weller (2010)?

Practical analysis (chapter 5)

- To what extent can Freebase domains be mapped to classes of the Dewey Decimal Classification?

1.6 Methodology

The methodology consists of two main elements:

1) a *theoretical analysis*: an analysis of the structure of Freebase and the DDC, based on the literature review
2) a *practical analysis*: a quantitative approach to mapping, using statistical data.

This way, the feasibility of a mapping of Freebase domains, which are broad categorizations of information to classes of the Dewey Decimal Classification is assessed. The data collection for 2) will be done using a statistical approach. In this process, the likely candidates for Dewey classes that Freebase domains can be mapped to will be determined, in order to answer the main research question.

1.6.1 Theoretical analysis

The theoretical analysis is based on different criteria to compare KOS defined by Weller (2010, p. 210-216), which are *complexity, domain of interest, size, formality, usage* and *general modeling principles*. Available literature, as gathered for the literature review in chapter 2, will serve as a basis for analyzing the structure of the Dewey Decimal Classification, and Freebase.

1.6.2 Practical Analysis

The first part of the mapping will be done using a statistical analysis with the DeweyBrowser interface[4]. This mapping on a statistical level (Vizine-Goetz, 2001) involves searching for Freebase domains (the main categories of Freebase) in the DeweyBrowser interface, and selecting appropriate classes based on statistical relevance (involving the number of retrieved WorldCat items for a class). In addition to that, the Library of Congress subject headings (LCSH), included in the DeweyBrowser and as-

[4] http://deweybrowser.oclc.org

signed to Dewey classes, will be used for gathering statistical information about which classes the Freebase domains can be mapped to.

A further method that is used in this thesis, in order to retrieve more precise Dewey classes for a Freebase domain, is to perform (fulltext) searches in the DDC's Relative Index; again using the names of Freebase domains. This can be combined with the results from the mappings above, to provide a more precise mapping[5].

Please refer to chapter 3 for more information about the methodology.

1.6.3 Research paradigm

Pickard (2007) defines different research paradigms that can be used for the research process. She distinguishes *positivism, post-positivism* and *interpretivism* as major research paradigms. Positivism is the belief in a tangible, social reality, that exists independently from those that create this reality. Post-positivism beliefs in this social reality as well, but adds the acceptance that the reality "will always be inhibited by imperfections in detecting its nature" (i.e. total independence is not possible). Finally, interpretivism beliefs in "multiple, constructed realities that cannot exist without the social context that create them".

The main research paradigm in this thesis will be post-positivistic. This means that the methodologies are meant to be objective and are of a quantitative nature, but that the context is also taken into account, especially in the theoretical analysis, and in the analysis of the mapping that results from this thesis. In general, the context of knowledge organization systems has a substantial influence on their structure and contents, and also a mapping is influenced by the context and assumptions of a KOS. For example, the structure of the Dewey Decimal Classification, and many classifications in general, is based on the belief that "the expert" knows, while the structure of Freebase, and collaborative knowledge bases in general, is based on collaborative knowledge of the users of the system. This assumption might have implications for the resulting mappings, and the conclusions based on this mapping.

1.7 Limitations

The methodology, as described in the previous sections, consists of two main parts. The first part uses information gathered in the literature review. The literature review aims to cover all major work related to the research questions. However, realistically speaking, it is not possible to cover all possible literature.

The second part of the methodology is a statistical mapping, that is dependent on the DeweyBrowser interface. The performance of the DeweyBrowser and the reliability of its contents influence the results of the statistical mapping. Also the choice for a statistical mapping instead of an intellectual (manual) mapping means that there might

[5] in a similar style as done in Salah, Cheng, et al. (2011)), that also combined different mapping steps.

be false positives (i.e. a Freebase domain assigned to a class it does not belong to) or false negatives (i.e. a Freebase domain that is not assigned to the class that it is supposed to be assigned to). However, the large number of classified WorldCat items available in the DeweyBrowser should weigh up against that. Moreover, an intellectual mapping would involve personal and contextual variables, that could also influence the results.

1.8 Summary

This introductory chapter has introduced the broad topic of this thesis, as well as the research problem, motivations and aims. Furthermore, the research questions, that will be answered in the literature review and analysis chapters, have been defined. The methodology has been introduced, that will be elaborated in chapter 3. Finally, a summary of the limitations of the study was discussed.

2 Literature Review

2.0.1 Introduction

The past years have shown a shift in the authorship of knowledge organization systems: while in the past these schemes were formally defined by a select group of experts, nowadays, also users can (informally) organize knowledge, with various implications. This literature review firstly looks at formally created KOS, which we define as created by (a group of) experts. This includes the Dewey Decimal Classification.

Secondly, the literature review looks at socially created KOS, wholly or partially devised by their users, and Freebase in particular. The research questions in section 2.1 and 2.2 are converging from broad to specific.

The third section of this chapter discusses the concept of mapping, and specific mappings carried out in the past from and to Freebase, and the DDC.

For the literature review in general, over ninety sources from different sources were used, including journal articles, conference papers, academic books and book chapters. For the section with the literature review about Freebase, also forum posts, popular literature and Wiki documentation related to Freebase were consulted, to augment the relatively limited amount of academic literature on Freebase.

2.0.2 Defining information, knowledge and knowledge organization systems

The concept of information can be interpreted in multiple ways. Buckland (1991) distinguishes three meanings of information: *information-as-process* (the act of "becoming informed"), *information-as-knowledge* ("knowledge communicated concerning some particular fact, subject, or event") and *information-as-thing*, described by Buckland in the following way:

> The term "information" is also used attributively for objects, such as data and documents, that are referred to as "information" because they are regarded as being informative, as "having the quality of imparting knowledge or communicating information; instructive." (Oxford English Dictionary, 1989, vol. 7, p. 946).

In this thesis we use information in the sense of Buckland's definition of *information-as-thing*, so as "objects that are regarded as being informative". Buckland notes that information systems can only deal with information-as-thing, since it is tangible, as opposed to the intangible information-as-process and information-as-knowledge.

Knowledge, in a general sense, "exists in the mind of an individual who has studied an object, understands it, and perhaps has added to it through research or other means" (Taylor & Joudrey, 2009). Even though knowledge has this intangible form, it is possible to create a representation of this knowledge (for example a textual representation), which is a tangible, informative object and thus information-as-thing.

If we view knowledge as *information-as-thing*, we can represent it in a system. To organize knowledge representations, we make use of Knowledge Representation Systems. In the field of Library and Information Sciences (LIS), the organization of (representations of) knowledge is often denoted as "knowledge organization systems", abbreaviated as KOS. In this thesis, we use the following definition for KOS, as proposed by Hodge (2000):

> "The term knowledge organization systems is intended to encompass all types of schemes for organizing information and promoting knowledge management. Knowledge organization systems include classification and categorization schemes that organize materials at a general level, subject headings that provide more detailed access, and authority files that control variant versions of key information such as geographic names and personal names. Knowledge organization systems also include highly structured vocabularies, such as thesauri, and less traditional schemes, such as semantic networks and ontologies."

In this chapter, we will give an outline of the different knowledge organization systems that are mentioned by Hodge; in particular classification schemes, subject headings, thesauri and ontologies. Furthermore, we will discuss the newer concept of folksonomies and other socially created knowledge organization systems in the subsequent section.

2.1 Formally created knowledge organization systems

This section starts with an overview of formally defined KOS, then focuses on classification systems, and finally, the Dewey Decimal Classification, a specific classification system. This section aims to answer the following guiding questions, converging from broad to specific:

- What are formally created knowledge organization systems? (section 2.1.1)
- What are classification systems, and how are they organized? (section 2.1.2)
- What is the Dewey Decimal Classification, and how is it organized? (section 2.1.3)

2.1.1 Formally created knowledge organization systems

This section briefly introduces formally created knowledge organization systems, that are usually created by a team of "experts". The section focuses on controlled keyword indexing, classifications, thesauri and ontologies, systems that fit in Hodge's (2000) definition of KOS, as they organize information for the purpose of retrieval.

Examples of classical methods of knowledge representation are *controlled keyword indexing*, *classifications systems* and *thesauri*. *Controlled keyword indexing* captures synonyms, in order to create a controlled vocabulary. There are generally no hierarchical relationships. Subject heading lists are lists of terms and phrases, that can be used to represent the (subject) content of an information resource, an example being the Library of Congress Subject Headings (LCSH), that is often used in library catalogs. Controlled vocabularies are "controlled natural language tools used to facilitate access to information by using pre-defined and pre-coordinated natural language terms" (Chowdhury & Chowdhury, 2007). The main advantage of using a controlled vocabulary is that it standardizes index terms, and can improve search results.

Classification systems (see section 2.1.2) are structured hierarchically, and use a non-verbal notation to "represent concepts and relations between them". The aim of a classification system is to represent knowledge in "a uniform and language-independent way" (Weller, 2007). The relationships are mainly hierarchical, though classifications often also include other (implicit) relationships (Mitchell, 2001). Examples of classification systems are the Library of Congress Classification and the Dewey Decimal Classification.

Thesauri mainly focus on the equivalence relationship, and contain preferred and non-preferred terms, that represent the same concept. So they are "bringing together various representations of terms, along with an indication of a mapping of that term in the universe of knowledge" (Chowdhury & Chowdhury, 2007). Different explicit relationships can be used in thesauri, like superordinate, subordinate and coordinate relationships (broader, narrower and related terms). An example of a thesaurus is Getty's

Art & Architecture Thesaurus (AAT). Compared to controlled keyword indexing, thesauri are "more strictly hierarchical", and generally narrower in scope (Taylor & Joudrey, 2009).

The classical KOS are currently complemented by *ontologies*, a relatively new way to structure information, that "define the types and items of an area of knowledge, encode the knowledge of a domain and make it reusable" (Colillas, 2012). Or, as Taylor & Joudrey (2009) put it:

> "An ontology defines the nature of reality by identifying the concepts, entities, terms, and categories in a particular domain in order to model the relationships among them. It is created to keep conceptual and semantic ambiguity at a minimum in an information and technical environment"

So, by keeping conceptual and semantic ambiguity at a minimum, ontologies can be used to exchange information in a formal way:

> "In order to exchange the semantics of information, one first needs to agree on how to explicitly model it. Ontologies are a mechanism for representing such formal and shared domain descriptions" (Fluit, Sabou, & Van Harmelen, 2003).

It is not only possible to use ontologies for interpersonal communication, but also for "human–computer and inter-computer interactions" (Weller, 2007). By representing shared domain descriptions that enhance interoperability between systems in different knowledge domains (Taylor & Joudrey, 2009), ontologies play an essential role in the semantic Web.

An important characteristic, which differentiates ontologies from controlled vocabularies, classifications and thesauri, is the possibility to add self-defined knowledge relations (for instance relations of equivalence, hierarchical relations and associative relations). It therefore increases the "expressiveness" of ontologies (see figure 1).

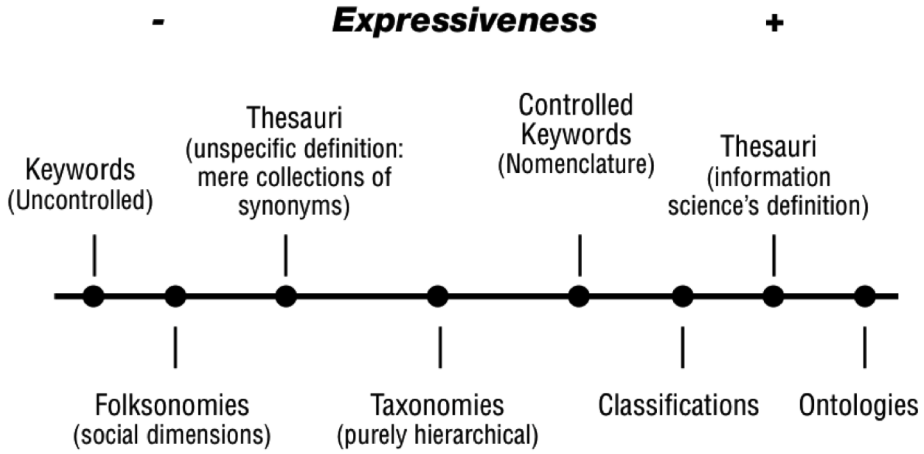

Figure 1: Distinguishing different vocabularies, based on their expressiveness (adapted from Weller, 2007)

2.1.2 Classification systems

In this thesis, we have chosen to focus on classification systems, that are discussed in this section.

2.1.2.1 Introduction

Classification systems have been used for a very long time to organize knowledge, "going back to at least the third century B.C." (Dextre Clarke, 2011). Classification, as defined by Chan (2007), is "the process of organizing knowledge in some systematic order". She elaborates:

> "The essential act of classification is the multistage process of deciding on a property or characteristic of interest, distinguishing things or objects that possess that property from those which lack it, and grouping things or objects that share a common property or characteristic into a class."

So in essence, classification can be seen as a multistage process in which a systematic organization is applied. A common element of this systematic organization is the definition of relationships between classes.

Classification systems have been used for a very long time to organize books, and many classification systems have been geared towards the use in a (physical) library setting. A classical definition of classification systems is the following:

> "the systematic arrangement by subject of books and other material on shelves or of catalogue and index entries in the manner that is most useful to those who read or who seek a definite piece of information" (Maltby, 1975)

We can see that this definition focuses on organizing classical carriers of knowledge, like books, the intended use for classification systems at that time.

2.1.2.2 Principles

The origins of classification systems lie in the philosophical principles of classification. Classic classification theory starts with the "universe of knowledge" as a whole. A "top-down" division in different classes and subclasses is made, based on a certain characteristic. These methods were mainly applied in the end of the 19th century, and biological taxonomy was the underlying model (Chan, 2007).

These types of classification systems use a top-down approach, and the divisions are made from broad (for example "Science") to specific (for example "Geodetic Surveying"). Also, as Chan discusses, the "array of classes on each level [...] form a coordinate relationship to one another and are collocated according to the affinity of their relationships". The classes that are located on a certain level of the hierarchy, are usually mutually exclusive and totally exhaustive; meaning that they do not overlap each other, and that they together represent all aspects of a subject.

Some classification systems that originated in last part of the 19th century and the beginning of the 20th century still exist to of this day, like the Library of Congress Classification (LCC), the Dewey Decimal Classification (DDC) and the Universal Decimal Classification (UDC).

S.R. Ranganathan saw the limitations of the "top-down" approach as used in the DDC and UDC, and that there was a need for a different type of classification, and he made use of the notion that documents, and objects, have multiple dimensions: *facets* (Morville & Rosenfeld, 2006). Ranganathan defined the following facets: Personality, Matter, Energy, Space and Time[6]. In this thesis, we mainly focus on the traditional classification systems, with the DDC as the main example.

2.1.2.3 Types

Based on the principles of classification systems as listed above, we can distinguish two basic types of classification schemes:

- *enumerative* classification schemes
- *faceted* (or *analytico-synthetic*) classification schemes

Traditional library classification systems, usually list "all subjects and their subdivisions, and provide ready-made symbols for them" (Chan, 2007). A prime example of a fully enumerative classification scheme is the LCC. Another approach is to do 'facet analysis and synthesis'. During the analysis phase, a subject is divided into its component parts (the facets), and in the synthesis phase, these component parts are reassembled, based on the properties of the document at hand.

An example of a fully faceted classification scheme is the Colon Classification, as devised by Ranganathan (Ranganathan & Gopinath, 1987), and published in seven editions from 1928 to 1987 (Satija & Singh, 2009).

[6] For examples, see: http://www.iskoi.org/doc/colon.htm [retrieved: 14-06-2012]

There are also examples of mixed classification systems (having both enumerative and faceted properties), of which the DDC is an example.

2.1.2.4 Elements

A characteristic of most classification systems is the "notation", usually a combination of numerals and/or letters, that can be used for both physical arrangement and systematic display (in for example library catalogs). As Chan (2007) indicates, the notation carries the meaning, and this meaning is independent from language. In addition to physical arrangement, the notation can be used for information retrieval functions. Besides the notation, a classification system consists of the following key components, as distinguished by Chan (2007):

- a *schedule*, "the sequence of class numbers and captions, arranged in class number order"
- the *tables*, "additional numbers used in conjunction with numbers from the schedules"
- an *index*, "a list of terms, with the corresponding class numbers", and
- *additional documentation*, for example manuals and instructions.

These components mainly apply to the enumerative or hybrid classification schemes, since fully faceted schemes (like the Colon Classification) often use a different kind of structure.

2.1.2.5 Advantages & disadvantages

Some of the advantages of using classification systems in the traditional sense (to index and retrieve physical items) are that they are language-independent, and have user-friendly possibilities to navigate the hierarchy (Peters, 2009). Koch et al. (2001) also indicate some advantages of using traditional classification systems to organize knowledge on the web: they provide browsing facilities, searches can be broadened and narrowed (using the notation), context is provided and multilingual access is possible.

A general problem with classification systems is that they are, to some degree, subjective. Classification systems are influenced by their context and language: "the act of classification is filled with ambivalence, and is tainted with the equivocal nature of language" and "the cultural and political context" (Salah, Cheng, et al., 2011). Classification decisions made by different people can also vary, especially in the case of books that contain multiple topics: "in practice, the necessary decisions to select categories involve enough choices that they will not be made the same way by two different people" (Lesk, 2005). This can create inconsistencies in the classification of materials by different parties involved.

The universal properties of the notation can also have disadvantages: for some users the notation can be hard to memorize, and hard to comprehend. Furthermore, as Peters (2009) puts it, the "prescribed structure with *n* classes (e.g. ten classes for a dec-

imal classification), makes it hard for classifications to expand breadth-wise". Once all classes at a certain level are full, it is not possible to add new classes to that level, since the decimal notation allows a maximum of *n* classes at a certain level. However, it is possible to expand the hierarchy "depth-wise", i.e. move further down the hierarchy.

Another common problem is that related material can be scattered in different sections of a classification. In addition to that, "some popular schemes do not always subdivide classes in a logical manner", and this can make browsing more difficult. Shirky (2005) discusses the underlying reasons:

> "what's being optimized is number of books on the shelf. That's what the categorization scheme is categorizing. It's tempting to think that the classification schemes that libraries have optimized for in the past can be extended in an un complicated way into the digital world".

According to Shirky (2005) classification can work well for some types of material (for example material with clear boundaries), but it generally does not work well with large corpora, unstable entities, unclear boundaries, and with uncoordinated, amateur users. The next section discusses a particular classification system: the Dewey Decimal Classification.

2.1.3 Dewey Decimal Classification

This section focuses on the organization, main ideas, structure, merits and demerits of the Dewey Decimal Classification.

2.1.3.1 Introduction

The most widely used classification system in the world is the DDC, or Dewey Decimal Classification (OCLC, 2003). It has been initiated by Melvil Dewey in 1873, and has been published in 23 subsequent DDC editions. A printed revision is released every ca. 7 years, and the latest revision is DDC 23, released by OCLC in 2011 (Dewey, 2011).

Figure 2: Dewey.info interface

In addition to the printed edition, there is a web-based version, called WebDewey[7], that includes all contents of the printed edition, plus regular updates. Another initiative, Dewey.info[8] (see figure 2) contains the top levels of the DDC 23 classification (2011), the top levels of the DDC 22 classification (2003) and the full Dewey Abridged Edition 14 (2004). The contents of Dewey.info are released as Linked Data under a Creative Commons license[9], including the Dewey notations and captions, as well as information about the Dewey edition, release date and hierarchy (Huurdeman, 2011). Similarly, the German National Library (DNB) has released the "essential elements" of the DDC Deutsch, the German version of the Dewey Classification (Deutsche Nationalbibliotek, 2011). Compared to Dewey.info, this version includes more classes (51,478) and additional semantic relations, for instance Dewey's see-references.

Dewey organizes information using a hierarchical "notation". The notation, as described in the previous sections, is the system of symbols that represents the classes in a classification system. In the case of Dewey, this is done using Arabic numerals (Dewey, 2011). The class numbers identify the meaning of a class, but also the relationship to other classes: longer numbers generally represent narrower classes than shorter (broader) numbers in the same class hierarchy. In addition to the hierarchical notation, there is also the possibility to use different "facets" in the classification system, since advanced numbers can be "built" using extra tables and other notations[10].

2.1.3.2 Organization

Since 1988, the DDC is maintained by OCLC. Currently, it is being developed by an editor in chief and four assistant editors. According to Mitchell & Vizine-Goetz (2009), "the editors study the distribution of topics in WorldCat to determine literary warrant (the existence of a certain level of literature on a topic) for updates – they also monitor the subject literature, news feeds, and other information resources, plus consult with users" [11].

A second check is being done by the Decimal Classification Editorial Policy Committee (EPC), an international advisory board that advises the DDC editor, and the OCLC on changes, possible innovations, and general development of the Dewey Decimal Classification. It is a joint committee of the ALA (American Library Association) and OCLC, and it also has members from outside the United States. In addition to that, there are several translations of the DDC, maintained by partners in different countries

7 http://www.oclc.org/dewey/versions/webdewey/ [retrieved: 24-06-2012]
8 http://dewey.info [retrieved: 05-05-2012]
9 http://creativecommons.org/licenses/by-nc-nd/3.0 [retrieved: 04-05-2012]
10 So the DDC has both enumerative and analytico-syntactic properties (see 2.1.2.3).
11 As we can see from this citation, "literary warrant" is an important guiding principle, and "the DDC is developed on the basis of literary warrant; that is, literature determines the need for category creation" (Mitchell, 2001).

(for instance Germany and Norway). These translation partners are also "corresponding members" of the EPC.

2.1.3.3 Main ideas

The DDC is "a general classification system which aims to classify documents of all kinds falling in any knowledge domain" (Satija, 2007). The initial division of nine main classes was based on the three divisions of knowledge that Francis Bacon had distinguished before: reason (science), located in the classes 100-600, imagination, in the classes 700-800 and the record of events and conditions in class 900 (Chan, 2007; Maltby, 1975). Satija (2007) indicates that this reflects the educational consensus of the late nineteenth century, when the DDC was developed.

As a result of this division of knowledge, the subjects in Dewey are ordered by discipline. Therefore, a subject can appear in different places. For example "radio" can be seen as as a broadcasting medium (384.54), as a field of engineering (621.384) or as a leisure channel (791.44). There has been criticism on the discipline-based division of classes. An example is Langridge (1989), who indicates that a discipline "does not distinguish between specializations (…) and the fundamental forms of knowledge from which they derive". He proposes "forms of knowledge", ways of looking at the world, and the "topics" they discuss as a more appropriate structure, instead of the discipline-based structure in many classifications systems.

2.1.3.4 Structure

2.1.3.4.1 Elements

The Dewey Decimal Classification contains different types of entries, that serve various purposes. Most Dewey numbers in the DDC are *ordinary entries*, which is a notational number with a corresponding heading. As the Dewey classification is in continuous transition, Dewey classes are sometimes moved and rearranged, resulting in *unusable headings*. These numbers, indicated with square brackets, cannot be used for classifying an object. There are also numbers that are still empty, called *unassigned numbers* These numbers are not used yet, indicated by the [unassigned] heading. *Optional numbers*, indicated with "Option: Class with" can be used to customize the DDC to a particular library's needs. Some numbers in the DDC have no specific meaning (for example the classes starting with "Other ..."). They are called *hooked numbers*, and books will usually only be classified in subclasses of this number. Finally, *centred headings* are used when a subject is spread over a span of numbers and indicated by the ">" symbol (Satija, 2007).

2.1.3.4.2 Division

The basic structure of the Dewey Decimal Classification consists of 1,000 classes, divisions and sections, organized from general to specific, that could be visualized using a tree structure. Dewey contains 10 general classes at the top level, that are numbered from 000 to 900:

```
000 Computer science, information & general works
100 Philosophy & psychology
200 Religion
300 Social sciences
400 Language
500 Science
600 Technology
700 Arts & recreation
800 Literature
900 History & geography
```

Table 1: Main DDC classes

The ten general classes are each divided into ten divisions, and each division, in its turn, is divided into ten sections. These top levels, taken together, are known as the *DDC Summaries* (Mitchell, Beall, Matthews, & New, 1996). Furthermore, there can be subdivisions below this level, separated by a decimal point after the first three digits.

In addition to the summaries, there are six auxiliary tables, that make it possible to do *notational synthesis*: the construction of numbers by combining different facets ("number building", see 2.1.2.3). The following auxiliary tables can be used for notational synthesis:

```
Table 1: Standard Subdivisions
Table 2: Geographic Areas, Historical Periods, Persons
Table 3: Subdivisions for the Arts, for Individual Literatures, for
 Specific Literary Forms
Table 4: Subdivisions of Individual Languages and Language Families
Table 5: Ethnic and National Groups
Table 6: Languages
```

Table 2: Dewey tables

The index keywords for the DDC are available in the "Relative Index". This index "shows the relationship between subjects and the disciplines in which they appear." (Mitchell et al., 1996). Index numbers can be used to retrieve a Dewey notational number for a specific subject. An entry in the Relative Index looks like this[12]:

[12] Example taken from Mitchell et al. (1996)

Garlic	641.3526
Garlic—botany	584.33
Garlic—cooking	641.6526
Garlic—food	641.3526
Garlic—garden crop	635.26
Garlic—pharmacology	615.32433

Table 3: Relative Index entry for Garlic

The example above shows the Relative Index entry for Garlic. The first number is the "interdisciplinary number", where the general works about Garlic should be classed. The other numbers show the appearance of Garlic in various areas of the DDC, for example in the context of cooking and botany.

In the list of index terms, the subjects and disciplines (Dewey classes) in which they appear are listed alphabetically. Generally, topics are ordered by discipline in the DDC, and "the Relative Index provides a useful reverse display of topics and the disciplines in which they appear" (Mitchell, 2001). This means that it can aid the user in retrieving suitable DDC numbers.

2.1.3.4.3 Relationships

The structure of Dewey and other classification systems can be "viewed as directed graphs, with classes as nodes and relationships between classes as edges" (Green & Panzer, 2009). These relationships can be explicit (for example in the form of class notes), but also implicit (for instance implied in the hierarchical structure). No special labeling exists for the different kinds of relationships in DDC, but the type can often be determined from the context in which it appears, for example a taxonomic schedule, parts of a discipline or listing of geographic features (Mitchell, 2001).

Relationships in the DDC are expressed through the *notational hierarchy* and the *structural hierarchy.* The *notational hierarchy* is visible in the length of the Dewey notation:

> "Numbers at any given level are usually subordinate to a class whose notation is one digit shorter, coordinate with a class whose notation has the same number of significant digits, and superordinate to a class with numbers one or more digits longer" (Mitchell, 2001).

The *structural hierarchy* can be expressed through different types of headings (for instance centred headings) and class notes (see notes[13], see-also notes[14], class-elsewhere notes[15] etc.). Most relationships that occur in the DDC are embodied in these structural hierarchy (for example equivalence and associative relationships). A disadvantage is

[13] indicating a topic is located elsewhere
[14] showing that a topic is also available somewhere else
[15] indicating a topic should be classed elsewhere

that they are usually not formally defined, so they have to be derived from for instance the see references. Various other relationships can be created when synthesizing numbers (i.e. building numbers using the auxiliary tables).

A related topic is the relationship between topics and classes. Green & Panzer (2009) indicate that classes can be defined as "a semantic space defined by associated topic neighborhoods". A topic that is represented by a class in Dewey is associated with it mainly through captions and notes (like the see, see-also and class-elsewhere notes) . These relationships have been visualized in figure 3 (Green & Panzer, 2009). The see and class-elsewhere notes in the Dewey schedules create additional connections in the Dewey classes. For example comprehensive works (see figure 3) have to be classified under the main class number, which is "Liturgical forms".

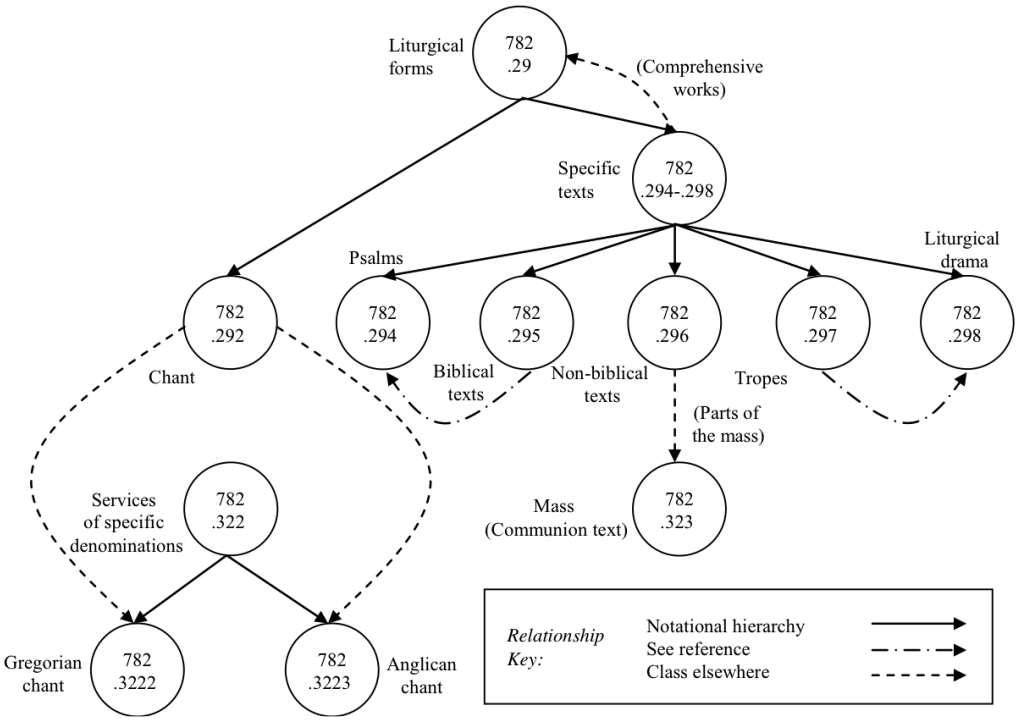

Figure 3: Class network around DDC's 782.29 class (Green & Panzer, 2009)

Green & Panzer (2009) indicate that these different relations create and expand a "neighborhood" around a certain focal topic. Index terms also expand the neighborhood around a topic, since they sometimes name topics that are broader or narrower than the captions and notes of a class. These aspects can make it difficult to derive the exact topics that are included in a particular Dewey class.

2.1.3.5 Advantages & disadvantages

Some of the advantages of using DDC are that it is a practical system that is being updated on a regular base, that it utilizes relative location[16], that the Relative Index conveniently "groups different aspects of the same subject" and that the DDC uses a universally recognizable system with arabic numerals, that is language-independent. The captions and other textual contents of the DDC are also translated in "more than 30 languages"[17], making it possible to be utilized in a wide variety of countries.

The DDC features a *mnemonic notation*, meaning that the same number (ending) is used in different locations of the DDC, helping users to memorize numbers (Chan, 2007).

The Dewey Decimal Classification has often been criticized, and frequent criticism is related to an Anglo-American bias in different classes (Chan, 2007; Lau, 2008; Shirky, 2005). Another drawback of the DDC, similar to other classification systems that originated in the same era, is that related classes are frequently separated (e.g. `400 Languages` and `800 Literature`).

The formal structure of the DDC's organization could have drawbacks in terms of flexibility, a general issue in formally created classification systems: "classification schemes often reveal difficulty in reacting to new areas of study and changing terminology since they are usually updated through formal processes by organised bodies" (Koch et al., 2001). Also, libraries already indexed their physical collections with earlier editions, thus "the user libraries face difficulties and much expense in switching to a new way of doing things, however superior in principle, because it becomes more cumbersome to retrieve items in their heritage collections" (Dextre Clarke, 2011).

The DDC inherited much of the thoughts on structuring knowledge from the late 19th century, though this is being corrected in new editions. However, as Chan (2007) indicates, the growth in certain disciplines has caused an "uneven structure" in the DDC (e.g. `300`, `500` and `600` "have become overcrowded"). The fact that the DDC uses 10 numerals for each level also means that topics with more than 10 subjects have to be accommodated on different levels of the hierarchy. Finally, some specific subjects (especially the synthesized numbers) make use of very lengthy numbers, and therefore are not very user-friendly.

[16] Related topics are shelved close (relative) to each other, and topics do not have a fixed location, which has the advantage that items do not necessarily have to be reshelved when additional items are added

[17] http://www.oclc.org/dewey/about/translations/default.htm [retrieved: 26-06-2012]

2.2 Socially created knowledge organization systems

This section addresses socially created knowledge organization systems, in which users play an essential role in creating the structure of the knowledge organization system. It starts with an overview of socially created KOS, in particular folksonomies, continues with semantic community knowledge bases and finally focuses on Freebase, an example of a semantic knowledge base. This section aims to answer the following guiding questions:

- What are socially created knowledge organization systems? (section 2.2.1)
- What are semantic community knowledge bases, and how are they organized? (section 2.2.2)
- What is Freebase, and how is it organized? (section 2.2.3)

2.2.1 Socially created knowledge organization systems

The previous sections have introduced classification systems which originated in the late 19th century. These days, with the advent of the internet, there are alternatives to the traditional classification systems. Various tools offer opportunities to collaboratively write web content, and to provide novel access methods to content, for instance in the form of "tags". As Salah, Cheng, et al. (2011) put it: "millions of users actively create, share, and classify various digital content and collections on the web". A prime example of a platform in which digital content is shared is the online encyclopedia Wikipedia[18]. Since its release in 2001, Wikipedia has expanded rapidly, and includes over 20 million articles in 283 languages as of 2011[19]. A Wiki has been defined as "a collaborative website whose content can be edited by anyone who has access to it" (Boulos, 2009). The focus of Wikis lies more on their contents than on its knowledge structure, however.

The classification done by users on the internet, involving tags, or user-created descriptors (Peters, 2009) is often done in the form of "free", or "social" tagging (Weller, 2007). This collaborative form of categorization has been dubbed "folksonomy" by Thomas Vander Wal (Morville & Rosenfeld, 2006). Weller (2010) indicates that folksonomies are "the embodiment of social indexing principles".

A folksonomy is part of the Web 2.0 paradigm, which is "a new generation of tools for the retrieval, deployment, representation and production of information" (Peters, 2009). Well-known examples of Web 2.0 applications that use social indexing principles are Flickr[20] and del.icio.us[21].

[18] http://en.wikipedia.org [retrieved: 06-05-2012]
[19] http://en.wikipedia.org/wiki/History_of_Wikipedia [retrieved: 06-03-2012]
[20] http://www.flickr.com [retrieved: 08-06-2012]
[21] http://www.delicious.com [retrieved: 08-06-2012]

The style of adding structure in folksonomies is the opposite of formal classification systems: ordinary users are annotating content with metadata, and are doing this in an "organic, bottom-up fashion" (Hirsch, Grundy, & Hosking, 2008). As discussed section 2.1.2.2, formal classification systems make use of the opposite "top-down" approach. An advantage of social tagging is the speed: "users can create tags quickly in response to new developments and changes in terminologies" (Weller, 2007). Also, universal properties are being added in this process, because folksonomies "include everyone's vocabulary and reflect everyone's needs without cultural, social, or political bias" (Kroski, 2006, as quoted by Weller, 2007). Users can add tags voluntarily, but there have also been experiments with other methods of tag creation, for example in Von Ahns' "Games with a Purpose", in which users can add tags in a playful manner, using (simple) games (Von Ahn, 2006).

A distinction can be made between "soft" semantic structures, as employed by Wikipedia, and "hard" semantic structures, as specified in semantic web modeling languages, like RDF. Tiropanis, Davis, Millard, & Weal (2009) indicate that "hard" semantic structures make it possible to process the meaning of resources and their relationships, and that they allow one to do reasoning (i.e. to derive new knowledge from existing facts).

New hybrid forms of socially created knowledge organization systems are also appearing. The transition from a "soft" semantic structure to a "hard" semantic structure is embodied in semantic Wikis, and Freebase. The next section will focus on the particular kind of tools that embody this transition: semantic knowledge bases.

2.2.2 Semantic community knowledge bases

2.2.2.1 Introduction

This section is discussing two types of semantic community knowledge bases: semantic Wikis, based on open-source technologies, and community knowledge bases that are using proprietary database systems, but that allow users to edit and reuse their data.

Semantic community knowledge bases are defined here as highly structured collaborative tools using community created knowledge, that allow for reasoning with their contents, as opposed low-structured collaborative tools like Wikis (Buffa, Gandon, Ereteo, Sander, & Faron, 2008). Semantic community knowledge bases have a knowledge structure, that is either ontology-based, or ontology-like[22]. These tools use community-created knowledge, and Semantic Wikis are an example of these structured collaborative tools (Breslin, Passant, & Decker, 2009).

Weller (2010) distinguishes between collaborative ontology engineering, and community-based approaches to ontologies. Collaborative ontology engineering makes use

[22] An ontology-like structure resembles an ontology, but is not necessarily formally defined as one

of a fixed team and is formal in nature, while community-based ontology engineering is "a specific form of collaborative ontology engineering based on the contribution of an open community (in contrast to a fixed team)" (Weller, 2010).

2.2.2.2 Principles

As indicated in the previous section, ontology-based structures form the foundation of semantic community knowledge bases like semantic Wikis, and of other platforms, that are very similar to semantic Wikis, of which Freebase is an example. A semantic Wiki can be defined as "a Wiki engine that uses technologies from the semantic Web to embed formalized knowledge, content, structures and links in the wiki pages" (Buffa et al., 2008). Hirsch et al. (2008) define a semantic Wiki as "a collaborative knowledge repository which provides semantically enriched contents".

The community aspects of these semantic knowledge bases can be valuable for the engineering and optimization of their underlying structure: "as ontologies are supposed to be both complex and highly domain-specific, their production and implementation is currently costly and laborious" (Weller, 2007). However, involving communities for the production and implementation might make it less costly, and Weller points out that "ontology engineering may profit from the growing communities that are tagging web documents and are becoming aware of the use of metadata". It would even be possible to use "entirely collaborative approaches in which users may actively contribute to the construction of ontologies from the very beginning".

2.2.2.3 Types

As stated in the introduction, this section focuses on semantic community knowledge bases that are based on open-source technology (semantic Wikis), and those based on proprietary database technology.

Examples of semantic Wikis are Semantic MediaWiki[23], OntoWiki[24] and SweetWiki[25], initiated by non-profit organizations. Several companies have tried to build a business model around similar, but proprietary systems. Examples of these are Twine[26] (discontinued), Powerset[27] (acquired by Microsoft in 2008[28]), OpenLink DataSpaces[29] and Freebase. Many of the past initiatives to build semantic community knowledge bases have been discontinued, both in the case of (semantic) Wikis based on open-source technologies, and proprietary systems.

[23] http://semantic-mediawiki.org/ [retrieved: 14-06-2012]
[24] http://ontowiki.net/Projects/OntoWiki [retrieved: 14-06-2012]
[25] http://www-sop.inria.fr/teams/edelweiss/wiki/wakka.php?wiki=SweetWiki [retrieved: 14-06-2012]
[26] http://web.archive.org/web/20100116015513/http://www.twine.com/ [retrieved: 14-06-2012]
[27] http://web.archive.org/web/20091110152736/http://powerset.com/ [retrieved: 14-06-2012]
[28] http://venturebeat.com/2008/06/26/microsoft-to-buy-semantic-search-engine-powerset-for-100m-plus/ [retrieved 04-05-2012]
[29] http://ods.openlinksw.com/wiki/ODS [retrieved: 14-06-2012]

2.2.2.4 Elements

We can differentiate the main elements that semantic community knowledge bases consist of, which are the semantic backbone, providing the structure for the content, and the content itself. Buffa et al. (2008) make a difference between "the use of Wikis for ontologies" and "the use of ontologies for Wikis". "The use of Wikis for ontologies" means that Wiki pages are considered as concepts, and the links in a wiki as object properties, so the wiki forms the basis of the semantic Wiki. This concept is also dubbed "Wikitology", in which "the wiki becomes the front-end of the ontology maintenance system" (Buffa et al., 2008). An example of software that uses this paradigm is Semantic MediaWiki. On the other hand, ontologies can be considered the main and essential element of a semantic Wiki, therefore dubbed "the use of ontologies for wikis". In this case you usually have to create the ontology before creating the pages. Examples of these types of Wikis are IkeWiki / KiWi[30] and SweetWiki.

2.2.2.5 Advantages and disadvantages

Using semantic collaborative knowledge bases instead of non-semantic tools could yield advantages. The possibilities to connect them to other semantic knowledge sources, to import semantic data sources and to add reasoning engines can improve the accessibility and usability of the data. Furthermore, semantic knowledge bases can make use of a wealth of material by integrating available Linked Data, for example Wikipedia and MusicBrainz[31].

As Buffa et al. (2008) indicate in their paper, semantic Wikis are still in development, and there are no clear standards yet. There are many various semantic Wiki engines, that use different approaches to integrate semantic elements, but a substantial number of them is not updated anymore[32], and none of them have emerged as the universal standard.

Commercial semantic community knowledge bases have a similar disadvantage, since they do not use a common standard either. The history has shown that many of these tools have been discontinued, or have been taken over by other companies and discontinued as independent tools. Powerset is an example, since it was integrated into Bing Search by Microsoft and is not available as a standalone tool anymore. A semantic community knowledge base that is still available, is Freebase. Freebase will be discussed in the next section.

[30] http://www.kiwi-project.eu/ [retrieved: 14-06-2012]
[31] MusicBrainz is an open music encyclopedia, see http://musicbrainz.org/ [retrieved 06-05-2012]
[32] See http://semanticweb.org/wiki/Semantic_wiki_projects [retrieved: 24-06-2012] for a list of semantic Wiki projects and their current status

2.2.3 Freebase

2.2.3.1 Introduction

Freebase (figure 4) is an "open repository of structured data" in the style of Wikipedia (Bollacker, Evans, Paritosh, Sturge, & Taylor, 2008), or, as Weller (2010) defines it, "an open database for factual information about people, places and things". The initial goals of Freebase have been quite ambitious: "Freebase is a database system designed to be a public repository of the world's knowledge" (Bollacker et al., 2008). As Danny Hillis, one of the founders, puts it, they have been "trying to create the world's database, with all of the world's information" (Markoff, 2007).

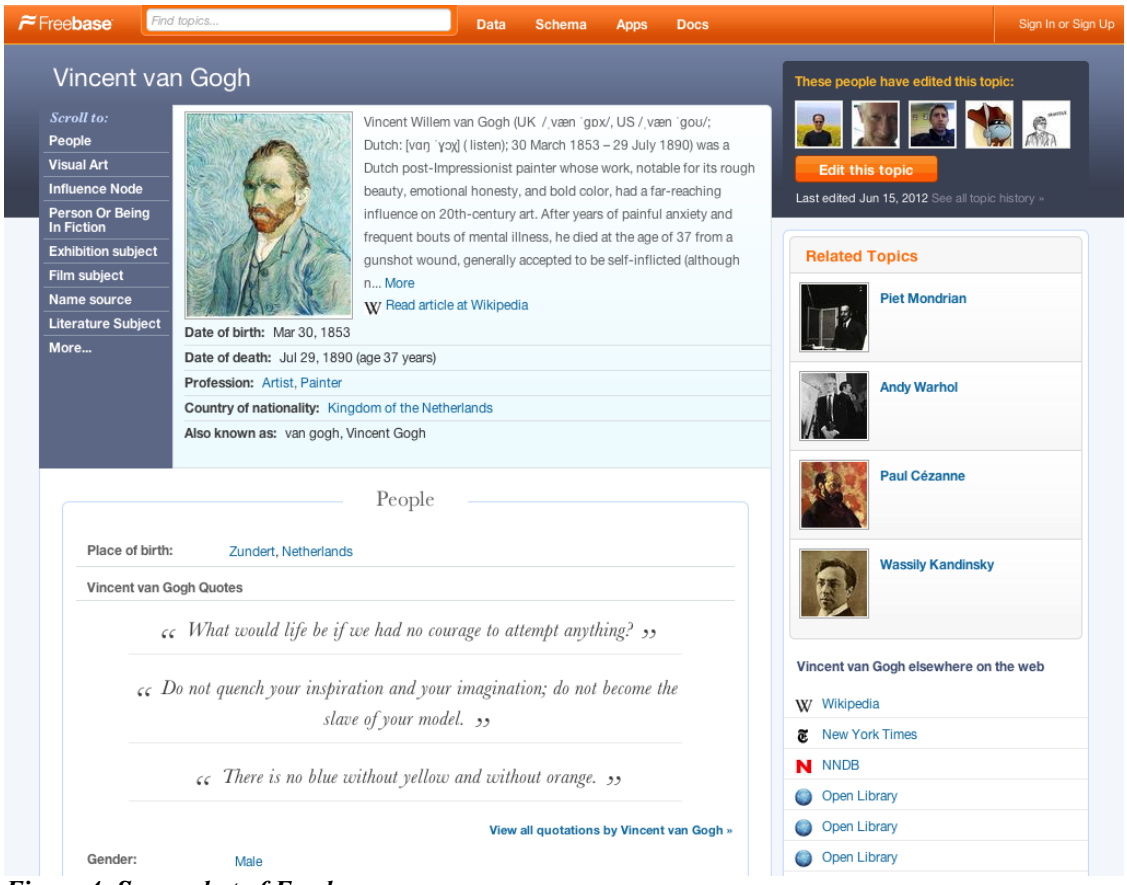

Figure 4: Screenshot of Freebase

Freebase has been categorized by Weller (2010) as a community knowledge base, that has "taken on the Wikipedia principle of letting a community collect world knowledge". Breslin et al. (2009) consider Freebase a "knowledge service leveraging semantics" and, similar to Weller, an "open collaborative Knowledge Database". Other authors call Freebase a semantic Wiki, like Buffa et al. (2008); Hirsch et al. (2008); Mika & Greaves (2008).

We can derive from the community aspects described above, that Freebase is a website made possible by the Web 2.0 paradigm. It contains elements of both folk-

sonomies and ontologies. Tim O'Reilly also referred to this combination in 2007: “in many ways, Freebase is the bridge between the bottom up vision of Web 2.0 collective intelligence and the more structured world of the semantic Web” (O’Reilly, 2007). An interesting aspect is that the users of Freebase might not be aware of its semantic data properties: “the users of knowledge base projects are not told to build an ontology but to contribute little pieces of information to a system” (Weller, 2010).

2.2.3.2 Organization

Freebase was initiated by the San Franciso-based company Metaweb Technologies, in 2007 (Breslin et al., 2009). Danny Hillis and Robert Cook, managed to secure $60 million in funding, and setup an alpha version of Freebase (Bollacker et al., 2008). In the first phase, Freebase combined community-created data with imported data, for example Wikipedia and MusicBrainz. The functionalities of Freebase have gradually been extended, and in 2009, Freebase started to offer its contents in the RDF format[33], a standardized ontology language, in addition to its existing API services.

On July 16th, 2010, Google announced that it had acquired Freebase, and also stated its intentions to “maintain Freebase as a free and open database for the world”, and to develop it further (Menzel, 2010). On March, 14, 2012, Google announced that they will integrate their “database of structured data” into their search results, possibly providing direct answers to search queries in their interface; partially made possible by the use of the acquired Metaweb / Freebase database (Efrati, 2012).

2.2.3.3 Main ideas

As indicated in the introduction, Freebase offers information about a wide variety of topics, and it aggregates open data from various data sources, like Wikipedia, the Open Library Project[34] and Wordnet[35]. Freebase, like Dewey.info, uses a Creative Commons Attribution license, albeit a more open one, since it “allows anybody to use the data for any purpose, as long as they give attribution to the contributor” (OpenBusiness, 2007). Freebase has been growing very rapidly, and contains “more than 10 million topics, more than 3,000 types, and more than 30,000 properties”[36].

The initial steps in creating Freebase were to seed it with initial data sets, that are available as open data. The basic idea was that these data sets should be of interest to general population, as opposed to “those that are highly esoteric and specialized” (Bollacker, Cook, & Tufts, 2007). The initial data sets included general knowledge, scholarly information, location information and popular culture. Freebase has been called a “melting pot” of datasets, as it integrates many different datasets[37]. References to ex-

[33] http://rdf.freebase.com [retrieved: 09-05-2012]
[34] http://openlibrary.org [retrieved: 05-05-2012]
[35] http://wordnet.princeton.edu [retrieved: 05-05-2012]
[36] http://www.freebase.com/docs/data [retrieved: 04-05-2012]

ternal datasets are also included in the topics in Freebase (as "internal identifiers"), so it is possible to jump to external datasets (such as Wikipedia).

Freebase has been inspired by Wikipedia, firstly in the sense that it uses the "post-hoc" moderation model, which means that corrections by users are immediately visible, and reviewed by other users, and secondly in that it aims to have exactly one article per concept (OpenBusiness, 2007). These articles are called "topics" in Freebase.

As Freebase is an example of a "Web 2.0" application, users can collaborate, edit the gathered data, add new information and create new connections between different topics. It is also possible to integrate Freebase into one's own website, and to develop new web applications and visualizations using Freebase data, with the provided Application Programming Interface (API) (Huurdeman, 2012). Several applications have been developed using Freebase's API[38]. They have different features, for example visualizing its structure (e.g. Thinkbase[39], see also Hirsch et al. (2008)), providing new browsing possibilities (e.g. Parallax[40]), inferring knowledge (e.g. the Genealogy Viewer[41]) or focusing on certain properties of the relationships of its data (e.g. Free-influencer[42]). An example of a standalone application that uses Freebase's data is Small Demons, that allows users to check people, places and things that feature in books, and browse using those properties[43]. An advanced visualization of Freebase is ASK-Ken, a "Visual Knowledge Browser", that uses Freebase's content and structure to present "a Node-Link diagram that allows to visually navigate through interconnected topics"[44] (see figure 5).

To edit data in Freebase, it is possible to register as a user, or to make use of a Google account. Users are free to model their own structures, and for guidance a "Data Modeling Style Guide" is available on the Freebase's Wiki[45]. An incentive for users to add and edit contents is created by "promoting" users to "Top Contributor", after a number of edits and additions to Freebase (Mattison, 2008). A higher level than the Top Contributor user level is the "Expert" level. Experts[46] have more rights for editing the schema and supporting users. They are invited by Freebase's staff from time to time, based on different criteria, for example familiarity with Freebase and the semantic

37 In a presentation by Jamie Taylor (graph data model engineer at Google / Freebase), available via http://www.google.com/events/io/2011/sessions/querying-freebase-get-more-from-mql.html [retrieved: 15-04-2012]

38 A list of Freebase apps can be found at: http://www.freebase.com/view/freebase/featured_application [retrieved: 04-05-2012]

39 http://thinkbase.cs.auckland.ac.nz [retrieved: 04-05-2012]

40 http://www.freebase.com/labs/parallax/ [retrieved: 04-05-2012]

41 http://genealogy.alexander.user.dev.freebaseapps.com/ [retrieved: 04-05-2012]

42 http://www.we-love-the.net/FreeInfluencer [retrieved: 04-05-2012]

43 http://www.smalldemons.com [retrieved: 04-05-2012]

44 http://askken.heroku.com/ [retrieved: 14-06-2012]

45 http://wiki.freebase.com/wiki/Main_Page [retrieved: 14-06-2012]

46 http://www.freebase.com/view/freebase/badges/freebaseexpert [retrieved: 19-03-2012]

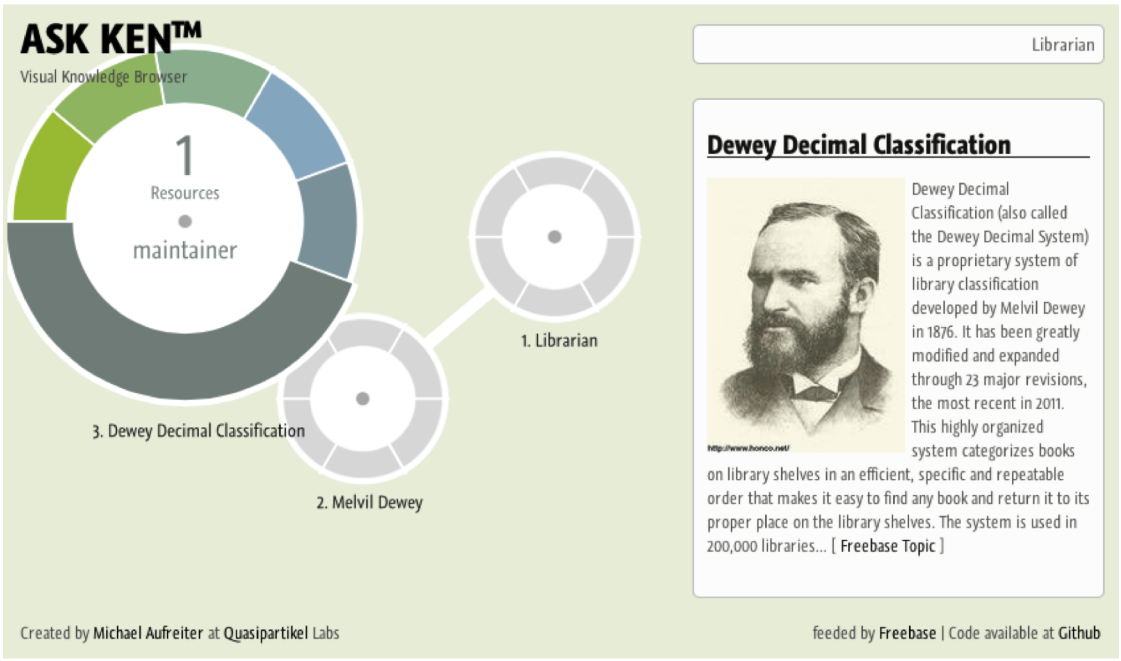

Figure 5: ASK-Ken Visual Knowledge Browser

Web, activities on Freebase and forum posts[47]. This way users are encouraged to add and edit content, and given an incentive to help others, for instance on the Freebase forums.

There have also been experiments with other types of data collection from users, for example using the "Freebase Typewriter"[48]. In this game-like application, users have to judge what the type is of Freebase topics that do not have a type yet, and answer "yes" or "no" to the question whether a topic belongs to a certain type or not.

Freebase contains an elaborate history system, that allows one to find out who created and deleted topics, what changes have been made, and so forth. It also makes it possible to query the underlying graph of Freebase and to receive its properties at a certain date, by including a timestamp with the query.

2.2.3.4 Structure

2.2.3.4.1 Elements

The main knowledge structure of Freebase consists of several elements, that will be explained in the following sections. These elements are: *topics*, *types*, *domains*, *bases*, *domain categories* and *properties*.

[47] http://wiki.freebase.com/wiki/Freebase_Experts [retrieved: 19-03-2012]
[48] http://typewriter.freebaseapps.com/ [retrieved: 14-03-2012]. Note: the game was not working at the moment of writing.

Topics

Topics form the foundation of Freebase. A topic can be compared with an article on Wikipedia, and the word "topic" is vague on purpose, since it can represent a lot of different things, like physical entities, artistic creations and abstract concepts (Huurdeman, 2012). On every page on Freebase, the 'edit and show details' option makes it possible to view the (ontological) details of a topic. It is then possible to edit it, to 'fill in the blanks' and to make new ontological connections between topics.

A topic is "an object representing a discrete entity" (Bollacker et al., 2007), comparable with an article on Wikipedia. It can represent specific objects (for example "George W. Bush" or "Obama"), but also broad concepts ("Biology"). As "discrete entity" indicates, it is important that a topic in Freebase only represents one concept or entity. It has one globally unique identifier (GUID), that refers only to that topic (Bollacker et al., 2007). A topic always possesses at least one "type" (Mattison, 2008), described below.

Types

A type is described as "an object that is used to semantically group topics", and "a topic associated with a type is considered to be an instance of that type" (Bollacker et al., 2007). For example, "George W. Bush" could be typed as "Person" and "US president".

Bollacker et al (2007) describe types as a "loose collection of structuring mechanisms and conventions, rather than a rigid system of ontologies and representations." Nevertheless, types are related to ontologies, and Weller (2010) states that "Freebase types loosely correspond to ontology classes". She also indicates, however, that there is no explicit hierarchical view of the Freebase types.

Domains and bases

A Freebase domain is a "broad category of information" (Mattison, 2008). A domain groups a number of types that belong to a logical category. An example of a domain is "Biology", that contains for example the types "animal-breed" and "organism-classification".

The domains in Freebase have been approved by Freebase's staff, and are called "Freebase Commons". However, it is also possible for users to create their own domains; these are called "bases", and can be edited by one or more authors. They were described by Mattison (2008) as "personal domains, essentially places where you create your own Types and Domains". He also states that bases of sufficient quality can be promoted by the Freebase staff to be an "official", public domain.

An example of a base is "pet breeds", that contains extensive information about the properties of different dog breeds, and this base is maintained by an active community.

Domain categories

Domain categories are "groupings of related domains". They are not ontologically relevant[49], but only meant for organizing the "types" page, and for showing a teaser of Freebase content on the homepage. There are ten different domain categories[50]: *Sports, Arts & Entertainment, Time & Space, System, Society, Science & Technology, Special interests, Products & Services, Transportation* and *Commons.*

Properties

A property is described as a particular "flavor" of a type, and Freebase's properties "can be compared to ontology properties" (Weller, 2010). A type can contain its own custom properties, for example "Date of Birth" for the "Person" type. A property can be a *literal* (e.g. "17-05-1975" for a "Date of Birth"). It can also be a *relationship* (e.g. "Amsterdam" for "Place of Birth"), which means that the property is connected to another node (the topic "Amsterdam", in this case). When the property is a relationship, Freebase uses an "expectedType", a type that is expected as the other end of the relationship (for example a "Location" for "Place of Birth"), as the example above showed. This is "the equivalent of the range in ontology properties" (Weller, 2010).

A type can also inherit properties from a parent domain, for example the "Person" type inherits "Parents" from the superordinate "People" domain. The list of associated and inherited properties is called "Schema" in Freebase.

2.2.3.4.2 Underlying structure

Freebase has an "ontology-like" (Breslin et al., 2009), graph-based structure, based on the elements described above, which "is designed to scale to a large number and diversity of users and data" (Bollacker et al., 2007). So the structure has to facilitate the integration of a lot of data, but also be easy to understand: "its features are designed to be simple and easily mapped onto the ontological structures of other systems" (Bollacker et al., 2007).

The ontological validity of Freebase's structure is not the most important goal: "Rather than ontological correctness or logical consistency, Freebase's type system is designed for collaborative creation of structure." This means that it is even possible for users to add contradictory types and properties, "in order to reflect users' differing opinions and understanding" (Bollacker et al., 2008). This is similar to folksonomies (see section 2.2.1), that "acknowledge local and situated knowledges by including the

[49] As mentioned in the description of Freebase's Domain Category page: "Domain Categories have no real meaning and they are not hierarchical. They really are used only for organization [sic] the "Types" page.", http://www.freebase.com/view/freebase/domain_category [retrieved: 04-05-2012]

[50] The decision to have 10 domain categories is interesting, since it is the same number as the number of main classes of the DDC. It is not clear if this is influenced by classical (decimal) classification systems or just for presentation purposes.

voice of multiple ontologies, rather than prescribing how information should be organized" (Lau, 2008).
The idea of using self-defined knowledge relations is "one major characteristic of ontologies" (Weller, 2007). The difference with classification systems is that these relations are made explicit, while in classification systems, the relations (excluding the hierarchical relationships as expressed in the notation) are mostly implicit. This leads us to the question whether Freebase's structure can actually be defined as an ontology.

There are different views on how to define ontologies. Weller (2010) describes 6 ways to distinguish an ontology from traditional KOS:

a. machine-readability (the use of representation languages), which enables automatic reasoning over their contents, and implicit information might be inferred
b. the possibility to freely define various types of semantic relations (properties)
c. the inclusion of datatype properties that may also link values to concepts
d. the specification of additional attributes for properties, such as transitivity or reflexivity
e. the ability to distinguish individual concepts from general concepts (via instances)
f. the possibility to design an ontology for other purposes other than pure document indexing

We can observe that Freebase meets most of the requirements to be qualified as an ontology in Weller's way of distinguishing ontologies. However, d) is not valid for Freebase: advanced properties cannot be added in a formal way. We can compare this with a different method of distinguishing an ontology, as proposed by Peters (2009):

a. use of standardized ontology languages (e.g. OWL)
b. possibility to do automatic inference using terminological logic
c. occurrence of and differentiation by common terms and instances
d. use of specific relations (as well as hierarchical relations)

In this definition, a) is not valid for Freebase's structuring mechanism: advanced properties cannot be added in a formal way, and the representation language that Freebase uses for its structure is not officially standardized (for instance by the ISO or W3C).

The structure of Freebase has been visualized in ThinkBase, which is described by its authors as a 'visual Semantic Wiki'. This web application is a "visual navigation and exploration tool for Freebase" (Hirsch et al., 2008). Figure 6 shows a screenshot of the tool, that visualizes the graph-based structure of Freebase, with various connected nodes. In the Java-based Web application, it is possible to click on the different elements to navigate through the graph.

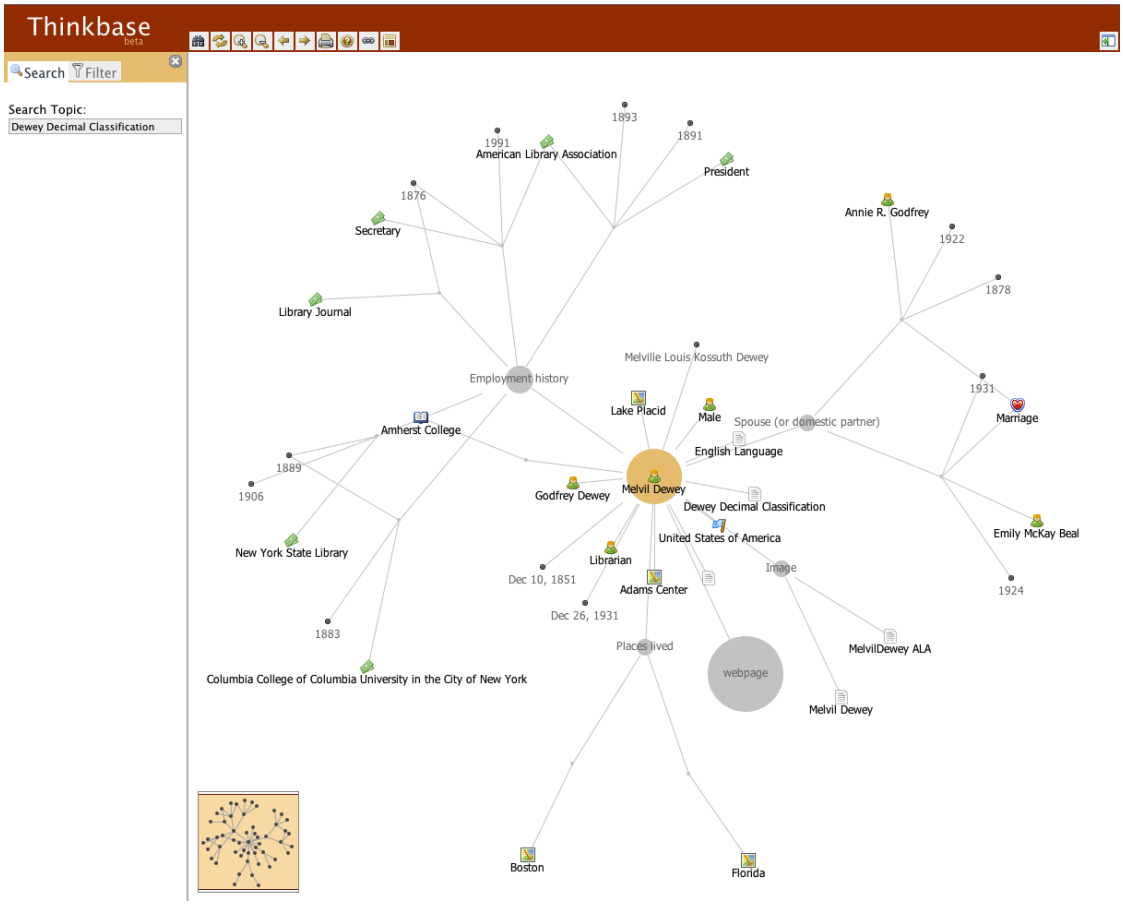

Figure 6: Thinkbase screenshot

2.2.3.4.3 Relationships

Some of the relationships that are included in Freebase have been described in section 2.2.3.4.1. There is also an initiative "to describe properties (or paths formed through the use of multiple properties) in terms of more general relationships". This is called Metaschema. An example from the Metaschema documentation is a Film Director, that has "directed" a film, and a Film Actor has "acted" in a film. If we describe these two properties in a more generic way, we can say that both "contributedTo" the film[51].

The Metaschema effort has mapped 3,500 properties to 46 "higher-order" Metaschema patterns (see table 4)[52]. The advantage is that applications built with Freebase data can make use of this more generic data, and can create more generic interfaces, that do not have to respond to all 3,500 properties, but only to the 46 higher order relationships.

[51] Note: the Metaschema elements do not seem to have originated from an existing (ontological) standard

[52] Taken from http://wiki.freebase.com/wiki/Metaschema [retrieved: 12-05-2012]

Abstract/Concrete	Genre	Place of Origin
Adaptation	Identifier	Practitioner
Administration	Leadership	Production
Broader/Narrower	Location	Publication
Categorical	Means of Demise	Series
Certification	Means of Expression	Service Area
Character Appearance	Measurement	Status
Character Portrayal	Membership	Subject
Composition	Name	Succession
Contribution	Ownership	Superclass/Subclass
Creation	Organizational Center	Symbol
Discovery	Parent/Child	Time Point
Distribution	Participation	Title
Event/Location	Peer	Whole/Part
Exhibition	Permitted Use	
Fictional	Place of Occurrence	

Table 4: Metaschema: higher order relationships in Freebase

2.2.3.5 Advantages & disadvantages

Freebase has a number of advantages. First of all, it provides a freely available platform, and free data sources in various (semantic) formats, using a relatively open Creative Commons Attribution license. Freebase has a user base that actively contributes to improve its structure and contents, providing "free" labour. Freebase uses many expansive and popular data sources to get its data, which include Wikipedia, MusicBrainz, census data and location information (Markoff, 2007). The available data can be accessed in multiple ways, for example via a JSON-based API and using the Metaweb Query Language (Breslin et al., 2009). The contents of Freebase are also downloadable as weekly "data dumps"[53], that can be converted to for example RDF or XML[54].

A disadvantage of Freebase could be that it is owned by a commercial company. As indicated before, Freebase was founded in 2007 by MetaWeb Technologies, and acquired in 2010 by Google. This might mean some risks for the continuity in the future. For example, Microsoft acquired Powerset in 2008, and subsequently integrated the technology into their Bing search engine, closing down the original site. Also Google has introduced services that it later shut down, for example the "Google Labs" services[55]. However, it has been stated by Google[56] that they will continue to offer Freebase as an open platform.

Another demerit of Freebase is that it has a peculiar structure, that might not be immediately clear to the general audience, as compared to other Web 2.0 initiatives like Wikipedia. Also, it uses a proprietary database technology; even though it is possible to download all contents of Freebase, it is not possible to install its technology on your

[53] Via http://download.freebase.com/datadumps/ [retrieved: 14-03-2012]
[54] As done in for example BaseKB, that offers an RDF-version of Freebase data, see http://basekb.-com/ [retrieved: 24-06-2012]
[55] http://googleblog.blogspot.nl/2011/07/more-wood-behind-fewer-arrows.html [retrieved 14-06-2012]
[56] http://googleblog.blogspot.nl/2010/07/deeper-understanding-with-metaweb.html [retrieved 14-06-2012]

own server, although it is possible at the moment to view the source code, and create your own applications using their platform and API. Finally, the website does not make use of W3C's SPARQL standard[57] for making queries in Freebase's semantic knowledge base, but of their own MQL (the “Metaweb Query Language”) format, which has a different syntax than SPARQL.

[57] http://www.w3.org/TR/rdf-sparql-query [retrieved: 06-05-2012]

2.3 Mappings

Mappings, in the context of KOS, play an essential role in integrating different systems. This section focuses on mappings of the DDC and Freebase. It aims to answer the following guiding questions:

- What is a mapping, in the context of knowledge organization systems? (section 2.3.1)
- Which mappings have been carried out between the Dewey Decimal Classification, Freebase and other knowledge organization systems? (section 2.3.2)

2.3.1 The concept of mapping knowledge organization systems

2.3.1.1 Definition

A wide range of knowledge organization systems has been developed in the past, and many of these KOS are still in use. Sometimes it might be valuable to combine systems, for example for information retrieval purposes. This "interoperability" of KOS can be achieved in different ways, as described by Weller (2010). She distinguishes the following ways to connect knowledge organization systems:

1. *reuse* and *upgrades*
2. *matching* and *mapping*
3. *merging* and *integration*

Firstly, a KOS, or elements of a KOS, can be used as a starting point for creating new ontologies, which is called *reuse* (or *upgrade*).

A second way to combine systems is by *matching* them, which is described as taking two schemas[58], a source and target, and "finding relations and correspondences between single elements located in different KOS" (Rahm and Bernstein (2001), as cited by Bellahsene, Bonifati, Duchateau, & Velegrakis (2011)). According to Weller (2010), *mapping* is "one of the most common kinds of interaction between different KOS". When two ontologies are mapped, the corresponding (matched) elements are being linked together, also called "ontology alignment". As Bellahsene et al. (2011) indicate, this mapping is "a relationship, i.e. a constraint, that must hold between their respective instances". Mappings have been carried out frequently in the context of ontologies, but also in the context of classification systems. In some cases, more than two ontologies are mapped, usually to a single "master" ontology.

A third way to combine KOS is to *merge* them. This is "the creation of one new KOS from two (or more) source models". This way, the knowledge in the original ontologies is combined into one, new ontology: a union.

[58] Rahm and Bernstein use a high level of the concept of "schema", so it can mean for example database schemas, ontologies or generic models.

2.3.1.2 Types of mappings

Mappings can be carried out manually and automatically. In the context of classification systems, Vizine-Goetz (2001) makes a distinction between *statistical mappings*, for example with the help of associations in WorldCat[59] (Hickey & Vizine-Goetz, 2001) and *editorial mappings*[60], which are manual mappings.

Statistical mappings are often based on co-occurrences, which is establishing mappings based on "the co-occurrence of terms from different schemes in the same metadata or catalog record" (Zhang, Peng, Huang, & Li, 2011). Hubrich (2010) elaborates on this concept: "Two concepts are regarded as similar if they are assigned to the same information resource, regardless of the peculiarities of the respective subject indexing method." Hubrich adds that a "critical mass of indexed data" is required to be able to use this technique. This type of mapping, based on WorldCat data, was for instance performed when Dewey terms were linked to LCSH headings by the OCLC (see also 2.3.2.1).

The disadvantage of statistical mappings, according to Vizine-Goetz (2001) is that they "are essentially reactive, since they depend on pre-established terminology as the source vocabulary". Statistical mappings make use of information that has been gathered in the past, for instance by using terms in existing catalog records.

2.3.1.3 Mapping relationships

According to Koch et al. (2003), the structures and levels of detail, the vocabularies, languages and cultural contexts of different classifications can vary widely, and therefore it is assumed that "a simple equivalence between the content of two classes will be rare". Therefore, they established five mapping relationships:

- Fully equivalent
- Narrower equivalent
- Broader equivalent
- Minor overlap
- Major overlap

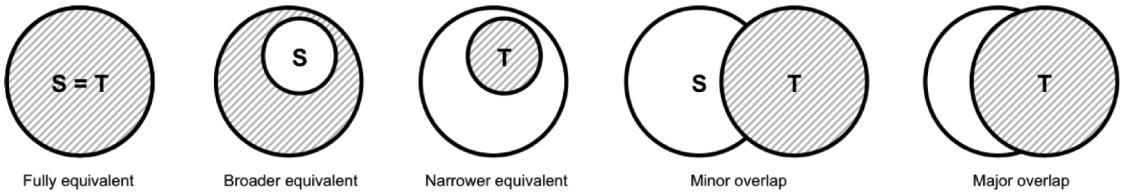

Figure 7: Mapping relationships (S = source class, T = target class)

The *fully equivalent* relationship means that the two mapped classes are generally the same. The *narrower equivalent* indicates that the mapped target class is a subset of the

[59] WorldCat
[60] also called "intellectual mapping"

source class, and the *broader equivalent* indicates that the target class contains more than the source class (a superset). *Minor overlap* involves a target class that has some equivalence to the source class, but also other topics, and *major overlap* indicates that there is a large equivalence to the source class, and other related topics in addition to that.

The CrissCross project (Jacobs, Mengel, & Müller, 2010), used a different system to indicate the mapping relationships between the German Schlagwortnormdatei[61] (SWD) and the DDC. This relation was shown with the "Degrees of Determinacy", having the following gradations:

- D4: "fully congruent to the scope of the class"
- D3: "a slight degradation of D4"
- D2: "less congruency than D3"[62]
- D1: "only slight conceptual congruency between heading and class"

We can derive from this list that D4 is comparable to the *fully equivalent* relationship of Koch et al. (2003), while D1-D3 have a different approach (only indicating the level of equivalence, but not if this equivalence is broader or narrower).

The mapping relationships proposed by Koch et al. (2003) are mainly aimed at the concepts that are represented by classes, as they "reconcile concepts rather than the terms used to represent those concepts" (McCulloch & Macgregor, 2008). The same is also true for the CrissCross project's mapping relationships. So the terms of two mapped classes might differ (e.g. on a grammatical or lexical level), but they represent the same concept.

On a more general level, we can also look at existing standards for mapping between vocabularies, being the British BS 8723, "Structured vocabularies for information retrieval – Guide" (part 4), and SKOS, the Simple Knowledge Organization System, which is a W3C recommendation. Dextre Clarke (2010) has summarized their characteristics, and introduced the upcoming ISO 25964 standard[63]. Both BS 8723 and SKOS involve hierarchical mapping types, being either "broader" or "narrower" matches. In BS 8723, equivalence mappings can be simple or compound[64], while in SKOS there can be an "exactMatch", "closeMatch", or "relatedMatch", but no compound match.

In general, the direction of mappings is usually regarded as being unidirectional (for example in the CrissCross project, or in BS 8723). However, in SKOS, mapping relationships are regarded as two-way relationships, involving symmetry or reciprocity. Dextre Clarke (2010) indicates that this might be caused by the lack of compound

[61] The SWD is a subject headings authority file
[62] e.g. if one SWD heading is one concept of many in one DDC class)
[63] ISO 25964: *Information and documentation – Thesauri and interoperability with other vocabularies*
[64] A compound match is common in thesauri, in which a combination of concepts can be used to express the idea of another, specific subject

matches in SKOS, that involve additional complexity and prevent changing the direction of a relation.

2.3.2 Existing mappings

2.3.2.1 Dewey Decimal Classification

In the past, mappings have been made from the DDC to several other systems, and other systems have been mapped to the Dewey Decimal Classification. This can have several advantages, as indicated by Mitchell & Vizine-Goetz (2009):

> "mappings between Dewey and thesauri, subject heading lists, and other classification schemes enrich the vocabulary associated with DDC numbers and enable the use of the DDC as a switching system"

So an important advantage of these mappings is the enrichment of the DDC's vocabulary, but they also add the possibility to use the DDC as an intermediary system. This section firstly discusses mappings from the DDC to thesauri, subject heading lists, and other classification schemes, and subsequently discusses the use of the DDC as a switching system.

2.3.2.1.1 Mapping DDC to thesaurus terms

Saeed & Chaudhry (2002) describe a project in which the Dewey Decimal Classification is used to build taxonomies for knowledge organization. Relevant hierarchies from the DDC Schedules are combined with terms from the DDC Relative Index, to develop the knowledge structure. Finally, a controlled vocabulary is added to the hierarchy, by combining the DDC terms and captions with similar terms from the IEEE Web Thesaurus[65]. A manual mapping was carried out, in which a distinction was made between *direct mapping* and *indirect mapping*. A direct mapping occurs when a thesaurus term matches a DDC caption exactly, and an indirect mapping is done when a thesaurus term falls under the same concept as a DDC entry, but does not match the caption exactly.

The project showed that the "DDC provides rich vocabulary in terms of its captions, notes, and indexing terminology" (Saeed & Chaudhry, 2002). It also indicated that many captions at the higher level of the DDC hierarchy could not be found in the thesaurus due to their general nature, and only very rarely the structure of thesaurus terms could be added to the existing hierarchical structure of a DDC class.

[65] http://www.ieee.org/publications_standards/publications/services/thesaurus_access_page.html [retrieved: 2012-05-20], "a hypertext interface for browsing terms that are arranged alphabetically" (Saeed & Chaudhry, 2002)

2.3.2.1.2 Mapping DDC to subject headings

Nowadays, the electronic version of DDC, WebDewey, already includes mappings from Dewey numbers to a number of subject heading systems: Library of Congress Subject Headings (LCSH), Medical Subject Headings (MeSH), and H.W. Wilson's Sears List of Subject Headings (Mitchell & Vizine-Goetz, 2009). The mappings of the DDC to LCSH are also visible in the *DeweyBrowser* prototype, which provides "access to several million records from the OCLC WorldCat database and to a collection of records derived from the abridged edition of DDC" (Vizine-Goetz, 2006). Users can perform text-based searches for books, and the DDC and LCSH subject headings provide context to these searches, and can also be used to retrieve items[66].

CrissCross is a German project that involves a mapping of the German subject headings authority file, Schlagwortnormdatei (SWD), to DDC notations (Jacobs et al., 2010). This mapping used three guidelines. The first was a *one-to-many strategy*: because of the discipline-based structure of the DDC, "one SWD heading could be mapped to several classes"[67]. Secondly, *deep level mapping* aimed at representing meanings as specific as possible, necessitating also notational synthesis. Thirdly, their mapping used *Degrees of Determinacy* (D), which indicated how accurate a match between the SWD and DDC was (see section 2.3.1.3 for an explanation of Degrees of Determinacy).

2.3.2.1.3 Mapping DDC to other classification systems

Several national projects, for example in Germany, Italy and Sweden, involve mappings from Dewey to other important classification systems that are used in these countries. Another, more recent example is a mapping from the DDC to the Chinese Library Classification, in which a combination between a statistical and manual (editorial) mapping was being used, with the statistical mapping being the first choice, and the manual mapping being complementary (Zhang et al., 2011), because the statistical mapping proved to be "more reliable".

Mappings from the DDC to other classification systems are also available via the *Classification Web* system of the Library of Congress, a subscription-based service, that incorporates statistical correlations among the LCC (Library of Congress Classification) and the DDC. This is done based on an analysis of the co-occurrence of those three in the bibliographic records of the Library of Congress (Mitchell & Vizine-Goetz, 2009).

[66] The DeweyBrowser interface is used as a data source for the mapping done in this thesis

[67] The opposite of the one-to-many (1:n) strategy is a one-to-one (1:1) strategy, in which a heading is mapped to only one class

2.3.2.1.4 DDC as a switching system

A “switching system” is a (classification) system that serves as a “hub”, i.e. all other systems are mapped to this classification, so this system works as an intermediary (McCulloch & Macgregor, 2008). In the Renardus project[68], the DDC was used as a switching language. The Renardus service was an attempt to facilitate subject browsing across different gateways, located in Denmark, Finland, Germany, the Netherlands, Sweden, and the United Kingdom (Koch et al., 2001). The DDC was used in the cross-browsing service to “mediate between the different classification systems in use (..) as a common switching language and browsing structure” (Koch et al., 2003).

Koch et al. (2003) also discuss some issues that occur when mapping from the DDC to other KOS in the context of the Renardus project, issues that were related to the depth of the mapping (in terms of hierarchy), to the treatment of classes with both generalities and specialities, to the in- or exclusion of non-topical classes, and so forth. Other issues include the “shattering” of disciplines (for instance engineering) in universal systems (like the DDC), as compared to specialized subject classifications.

2.3.2.2 Freebase

Freebase makes use of an ontology-like structure, and also provides URIs[69] to their concepts. This way it is possible to connect it to other knowledge structures[70], and to perform “ontology mapping”, or “ontology alignment” (see also section 2.3.1.1). This is a research topic in the domain of ontology engineering and the semantic Web that receives a lot of attention, since it is “a critical operation for information exchange on the semantic web” (Falconer & Storey, 2007). Various tools are available to do ontology mapping. As opposed to other ontologies like DBpedia, there is not much scientific literature about mapping Freebase's ontology-like structure to other knowledge organization systems.

This section introduces connections from Freebase's concepts to the same concepts in other knowledge databases, mapping to and from other knowledge structures, and, because of the limited amount of literature on mappings to and from Freebase, a discussion of a mapping of related knowledge structures.

2.3.2.2.1 Mapping Freebase's concepts to external concepts

Freebase has some built-in connections to other knowledge structures, as it provides “foreign keys”, which are keys to access information in systems outside of Freebase (done using URIs on remote sites). The Freebase URI template system can discover URIs from URLs automatically, and create links to other systems using the foreign

[68] http://renardus.sub.uni-goettingen.de/ [retrieved: 06-05-2012]
[69] A URI (Uniform Resource Identifier) is a text string that identifies a name or a resource
[70] http://wiki.freebase.com/wiki/URI_Templates [retrieved: 16-03-2012]

keys. For example, the topic (the concept page) of the film "Taxi Driver" on Freebase[71] is linked to the title page of the film on Wikipedia, IMDB, Netflix, and Metacritic. The key used to access the same topic on IMDB is "tt0075314", and stored in Freebase's system together with the namespace ("/authority/imdb/title"). This way, connections to other descriptions of the same topic on Freebase are available via its associated topic page.

2.3.2.2.2 Mapping different knowledge structures to Freebase's structure

The idea of mapping different ontologies to Freebase has been realized in the "Web Ontologies" base on Freebase, the purpose of which "is to be able to model the relationships between prominent web ontologies and map them onto equivalent Freebase types and topics"[72]. As as result, it is possible to improve the search functions in Freebase, since URIs and concepts from other ontologies can be retrieved, that apply to specific Freebase topics. Included ontologies are for example OpenCyc[73] and OpenCalais[74]. In a topic in Freebase, for instance "Person"[75], the equivalent classes with their URIs are listed, in this case FOAF, OpenCyc, YAGO, etcetera. However, not much documentation about the mappings in the Web Ontologies base seems to exist[76], and applications that use this data do not seem to be available yet.

During the development of Freebase and beyond, many existing databases have been mapped to Freebase. For example, the MusicBrainz music database has been mapped to Freebase[77], which means that MusicBrainz's "Next Generation Schema"[78] was connected to Freebase's Music domain[79]. The experience gained during the creation of this mapping lead to changes in the structure of the Music domain in Freebase, discussed on one of Freebase's Wiki pages[80]. The idea behind this change was to make it easier to synchronize data with the different music databases that can provide data to Freebase. Therefore it could be seen as a practical approach to schema modeling, since the theory of formal knowledge organization systems is not the main influence on Freebase's structure, but instead the practical usage of Freebase for representing datasets.

---

71 http://www.freebase.com/view/en/taxi_driver [retrieved: 19-05-2012]
72 http://ontologies.freebase.com/ [retrieved: 23-05-2012]
73 OpenCyc is "the open source version of the Cyc technology, the world's largest and most complete general knowledge base and commonsense reasoning engine", http://www.cyc.com/opencyc [retrieved: 14-06-2012]
74 OpenCalais extracts semantic information from unstructured text (for example on websites), in a semantic web format, http://www.opencalais.com/ [retrieved: 14-06-2012]
75 http://www.freebase.com/view/en/person/-/base/ontologies [retrieved: 23-05-2012]
76 The accompanying discussion forum for the Web Ontologies base (http://www.freebase.com/discuss/threads/base/ontologies) only contains five topics (as of 23-05-2012), the latest being from 2011
77 http://wiki.freebase.com/wiki/MusicBrainz_data_load [retrieved: 19-05-2012]
78 http://wiki.musicbrainz.org/NGS [retrieved: 19-05-2012]
79 http://www.freebase.com/view/music [retrieved: 19-05-2012]
80 http://wiki.freebase.com/wiki/Music_schema_open_questions [retrieved: 19-05-2012]

2.3.2.2.3 Mapping Freebase to other knowledge structures

Freebase's structure has not been mapped to many other knowledge structures. In one project, an informal mapping from Freebase domains to Dewey classes was performed in the context of "Semantic Classification Search", an application that enhances the DDC with contextual descriptions from Freebase (Huurdeman, 2011). The data for this application was Linked Data from the DDC and Freebase. Some issues, inherent in the structure and contents of Dewey and Freebase, influenced the reliability of this mapping:

1. Ambiguous or unclear Dewey captions
2. Very broad (for example *Business*), or very narrow Freebase domains (for example *Ice Hockey*)
3. Vague Freebase domains and domain descriptions (for example *Influence*)
4. Small Freebase domains, i.e. containing few instances (for example *Zoos*), and broad Freebase domains, i.e. containing many instances (for example *Books*)

The experience gained during this project have been used as a starting point for the statistical mapping in the other direction (from Freebase to the DDC), that is described in chapter 5 of this thesis.

2.3.2.2.4 Related mappings

An example of a mapping between a formal and informal knowledge organization system, which could be relevant for this thesis, has been carried out by Salah, Cheng, et al. (2011). The Universal Decimal Classification (UDC), that has a similar structure as the DDC, since it was based on the same principles, was mapped to Wikipedia categories[81]. In their research Salah et al. indicate that a simple mapping between the Wikipedia and UDC category structure is "problematic", because of the act of classification itself, and the differences in the structure and distribution of both systems. To solve some of these issues, they performed a multistage mapping between UDC and Wikipedia categories (instead of a single mapping). Therefore they applied four different levels of mappings in their research:

- naïve mapping by users
- term matching
- manual reading of ambiguous categories (by checking their occurrences in UDC)
- search for UDC terms in Wikipedia category page names

The results of these four mappings were combined in order to achieve a basic mapping between the UDC and Wikipedia, and to visualize the differences in the structure of both knowledge structures.

[81] Freebase uses a large portion of Wikipedia's data and concepts

2.4 Summary

This chapter has started with a discussion of formally created knowledge organization systems, classification systems and the Dewey Decimal Classification. The second section described socially created knowledge organization systems, semantic community knowledge bases and Freebase. The final section discussed the concept of mapping in the context of the two main KOS discussed in this thesis: the DDC and Freebase. This section also contained examples of mappings in the context of both systems.

The next chapter will elaborate on the methodology used for the theoretical and practical evaluation in this thesis.

3 Methodology

3.0.1 Introduction

This thesis aims to assess the feasibility of creating a mapping between formally and informally created knowledge organization systems, in particular the Dewey Decimal Classification and Freebase. In order to create this mapping, a theoretical analysis (see chapter 4) and a statistical analysis[82] of Freebase domains and Dewey classes (chapter 5) are carried out. The first section contains a justification of the methodology, while the second section describes the procedures for carrying out the mapping.

3.1 Justification

3.1.1 Methodology

3.1.1.1 Development

Different methodologies regarding data collection have been considered during the preparation phase for this thesis. This included doing user research with a custom developed application ("Semantic Classification Search", see section 2.3.2.2.3). However, the use of a custom application could introduce more unknown variables to the research, for instance because of usability issues in the application. Doing user research would also mean that the focus would be moved away from comparing the structure of the DDC and Freebase, and move more to the usability and functionalities of the custom-built application.

For this reason it was decided to use data that is available via the OCLC's Dewey-Browser as the basis for this research.

3.1.1.2 Elements

As discussed in chapter 1, the methodology consists of two elements:

1. a theoretical analysis of the structure of Freebase and the DDC, based on the literature review
2. a quantitative approach to mapping, using statistical data.

The theoretical analysis is guided by a set of criteria that have been defined by Weller (2010): *complexity*, *domain of interest*, *size*, *formality*, *usage* and *general modeling principles*. The statistical mapping is based on data from Freebase and the Dewey-Browser. It maps Freebase domains (broad categories of information) to DDC classes.

3.1.1.3 Data sources

We can differentiate data sources for Freebase domains, and data sources for the contents of the Dewey Decimal Classification.

[82] also dubbed "statistical mapping" (Vizine-Goetz, 2001)

The information about Freebase domains and associated properties is retrieved from the "Freebase Schema Explorer" (figure 8)[83]. Data that could be retrieved from this Freebase application includes:

- the *name* of the Freebase domains
- the *ID* of the Freebase domains
- the number of *associated types and instances*
- the *creation date*
- a *textual description*.

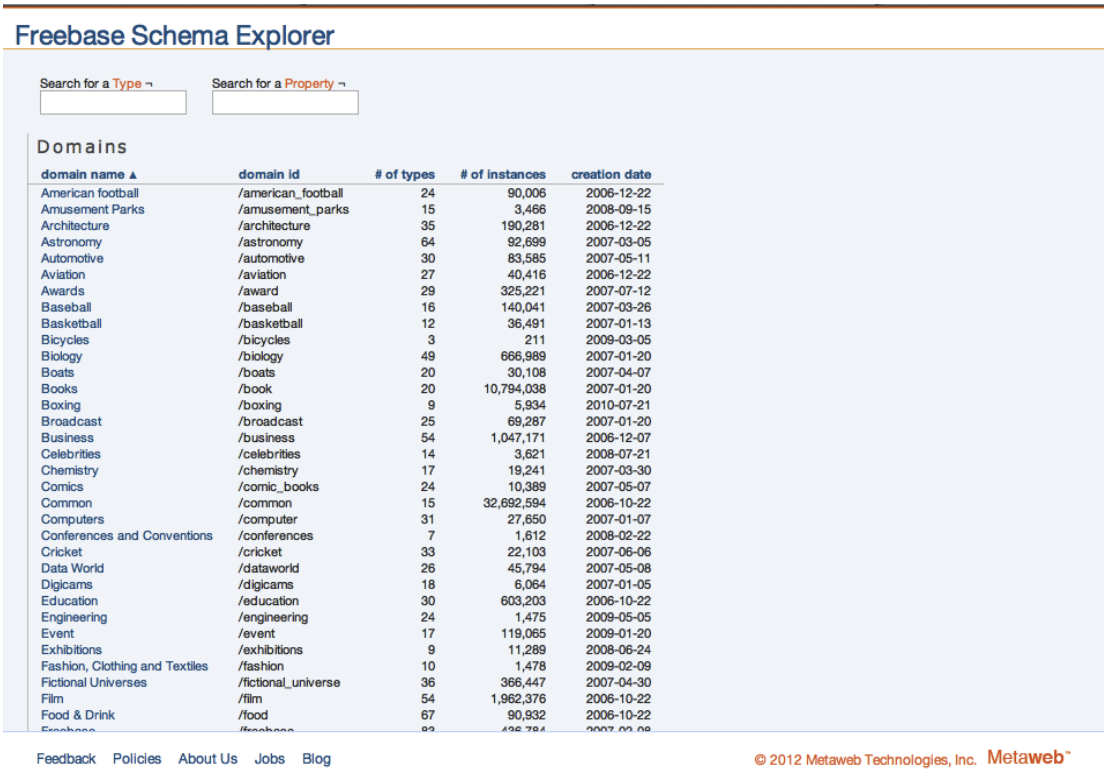

Figure 8: Freebase Schema Explorer

Another source of data is the DeweyBrowser (figure 9)[84]. This application, available online via OCLC ResearchWorks[85], the OCLC's "research laboratory", contains DDC classes, LCSH subject headings and WorldCat book titles as fields that can be used to search for books. It returns the number of books per Dewey class, division and section. It "provides access to approximately 2.5 million records from the OCLC Worldcat database"[86], and is based on the Apache Solr system. To obtain quantitative data for the mapping, it was decided to focus on specific search indexes for WorldCat items available in the DeweyBrowser:

83 http://schemas.freebaseapps.com/ [retrieved: 29-04-2012]
84 http://deweybrowser.oclc.org [retrieved: 29-04-2012]
85 http://www.oclc.org/research/activities/researchworks.htm [retrieved: 29-04-2012]
86 http://www.oclc.org/research/activities/browser/browser.htm [retrieved: 28-05-2012]

- the “title” index (the titles of WorldCat items)
- the “keyword” index (the subject keywords assigned to WorldCat items)

These elements were chosen because they exist for every WorldCat item, so in theory all relevant items can be retrieved; this is not true for some other fields, like the “description” field, that is not available for all available WorldCat items.

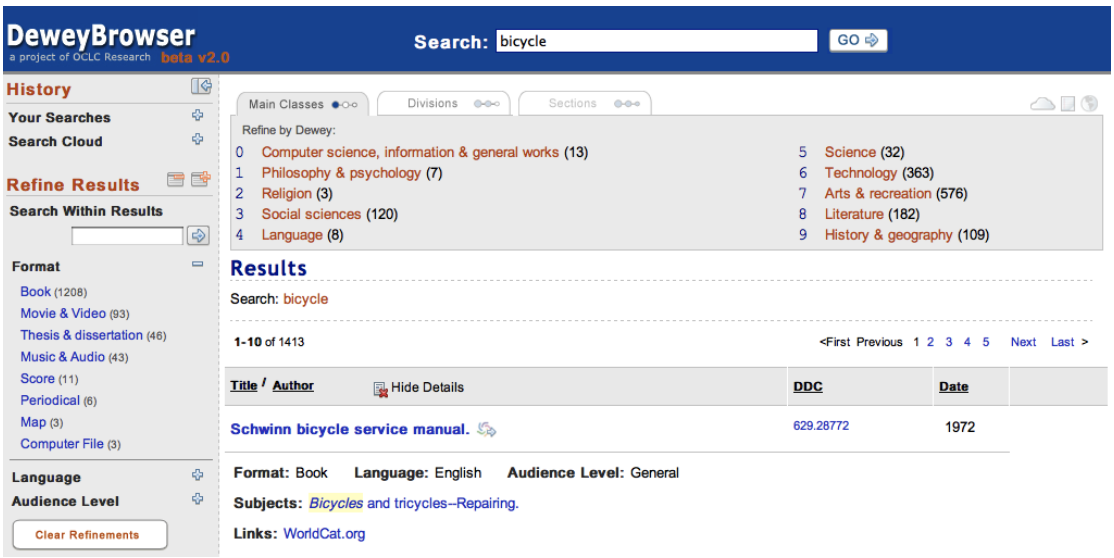

Figure 9: DeweyBrowser interface

The search in WorldCat titles and descriptions provides an indirect way of searching for corresponding DDC elements for Freebase domains. By using these fields, it is possible to avoid direct term matching, i.e. only matching the name of a Freebase domain with the caption of a DDC number on a textual level. The advantage is that in this way also synonyms and spelling variations will be retrieved.

3.1.2 Limitations

Some limitations of the used quantitative approach to mapping are the following:

- It is essentially a “reactive” approach, as discussed in section 2.3.1.2: it depends on the existing classifications done using the DDC in the WorldCat database. If Freebase domains are too “new”, they might not give substantial results in the WorldCat database (via the DeweyBrowser)
  - this limitation is inherent in the chosen approach, so it cannot be avoided; however, changes in classifications of books in the WorldCat database will be visible in the DeweyBrowser interface when its database is updated.
- The WorldCat database can only be queried up to the third, or *section* level, most likely due to copyright restrictions. In addition to that, the DeweyBrowser

prototype makes use of a snapshot of WorldCat items, so it might not reflect all changes in WorldCat directly.
    - in order to alleviate this problem, it was decided to do an additional step of mapping, using Dewey's Relative Index (see section 3.2.3.3), and to combine this with the results from the first step of analysis.
- The Freebase domains can be vague, or unclear, thereby reducing the number of valid results from Dewey.
    - this limitation is partially overcome by combining a search in the titles of WorldCat items in the DeweyBrowser interface, with a search in the LCSH subject headings that are included in the DeweyBrowser interface, that provide another access point to the WorldCat material (meaning that also synonyms of the domains can be found). The relevant and matching results are subsequently combined in the Excel sheet that is used in the project.
- The decision to map Freebase *domains*, and not *types*, *topics*, or *domain categories* to Dewey means that the results of the mapping could be different than when using other elements of Freebase's structure.
    - previously, some experience has been gained in the development of Semantic Classification Search, and the results seemed to indicate that using Freebase domains would be the preferred option for mapping. The Freebase Commons consists of a set of domains in Freebase that is generally stable, since they have been approved by the Freebase staff[87]. Also, the total number of domains (80) is more workable than the approximately 3,000 types, 30,000 properties and 10 million topics that are included in Freebase.

3.2 Procedures

3.2.1 Introduction

As described in the previous sections, the methodology for creating a basic mapping between the Dewey Decimal Classification and Freebase uses a quantitative approach to mapping, and makes use of statistical data.

[87] Refer to 2.2.3.4.1 for more information about Freebase domains and the Freebase Commons

3.2.2 Tools

The following tools are used for the data collection:

- *Freebase Schema Explorer* (online)
  - The Freebase Schema Explorer, as described before, is used to retrieve information about Freebase domains
- *DeweyBrowser* (online)
  - The DeweyBrowser is used to retrieve the Dewey classes with the highest number of WorldCat results
- *Dewey Relative Index* (book) and *WebDewey* (online)
  - The DDC Relative Index is used to refine the results of the first mapping step.

The following tool is used to capture the data:

- *Microsoft Excel*
  - Microsoft Excel is being used to register the collected data from the sources above. The Excel sheet makes us of advanced techniques to record, verify and analyze the data (see 8. Appendix).

3.2.3 Steps

The different steps involved in the creation of this mapping are summarized in table 5 below, and consist of three methods, that are used for a basic mapping and a mapping refinement.

	Basic mapping
Method 1	Search for Freebase domains (main categories) in WorldCat book titles - retrieve number of WorldCat book results: - results for a particular Dewey class - total number of results - select DDC class covering the highest percentage of WorldCat results
Method 2	Search for Freebase domains (main categories) in WorldCat book subject headings - retrieve number of WorldCat book results: - results for a particular Dewey class - total number of results - select DDC class covering the highest percentage of WorldCat results
	Mapping refinement
Method 3	Search for Freebase domains (main categories) in the DDC Relative Index - retrieve matching Dewey classes - combine with results part 1

Table 5: Basic research steps

3.2.3.1 Method 1

The first step of performing the mapping is to search for Freebase domains (their labels) in WorldCat book titles, using the following query structure:

```
ti:(<searchterm>)
```

The prefix "ti:" indicates that we search for the title of an item in WorldCat, and "<searchterm>" stands for the search term that is used for the query, in this case the Freebase domain identifier. A list of available search prefixes for the WorldCat database is available via the OCLC website[88], some of which are also applicable to the DeweyBrowser interface.

3.2.3.1.1 Dewey "class" level

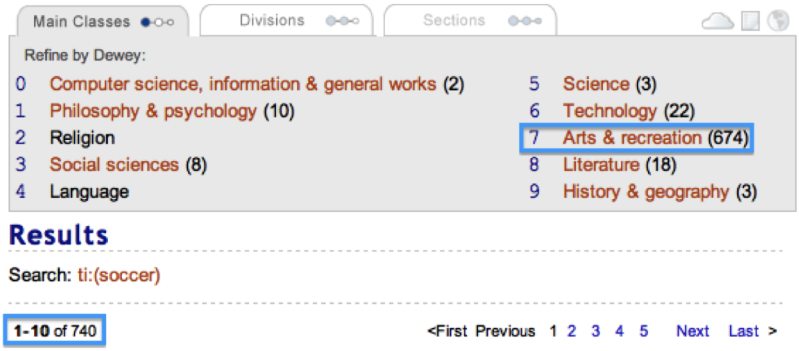

Figure 10: Main Dewey classes for the title search for "soccer"

As can be seen in figure 10, the DeweyBrowser interface shows the number of matching results for WorldCat items, and indicates in which Dewey classes the results appear. Figure 10 contains a query for the Freebase domain *Soccer* in the title index. The results for the different Dewey classes are shown, and the marked 7 class *Arts & recreation* has the highest number of results (674), out of the total of 740 results, which is 95%.

The class with the highest number of relevant results is subsequently selected for inclusion in the Excel sheet, and the number of results and total results are registered.

[88] http://www.oclc.org/support/documentation/worldcat/searching/searchworldcatindexes [retrieved: 28-05-2012]

3.2.3.1.2 Dewey "division" level

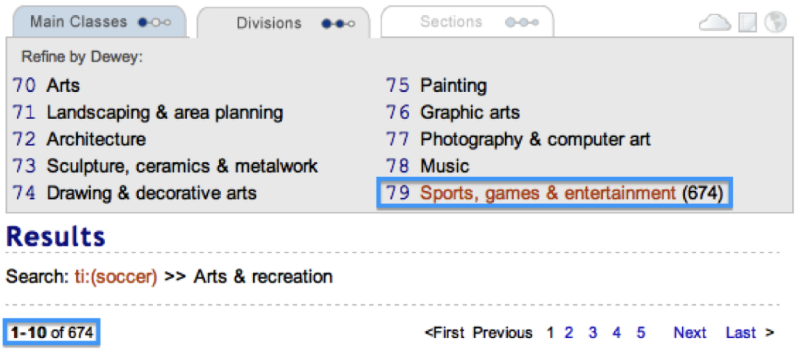

Figure 11: Dewey results at the division level

Subsequently, it is possible to see the number of results at the Dewey division level for the search term, see figure 11. In this example, we see that the `79` DDC class contains all (674 out of 674) results.

3.2.3.1.3 Dewey "section" level

Finally, we can go down to the section level of Dewey, and record the number of results. In this case, the `796` class (`Athletic & outdoor sports & games`) has 665 out of 674 results, which is 84%, so this class is added to the Excel sheet with the project's data.

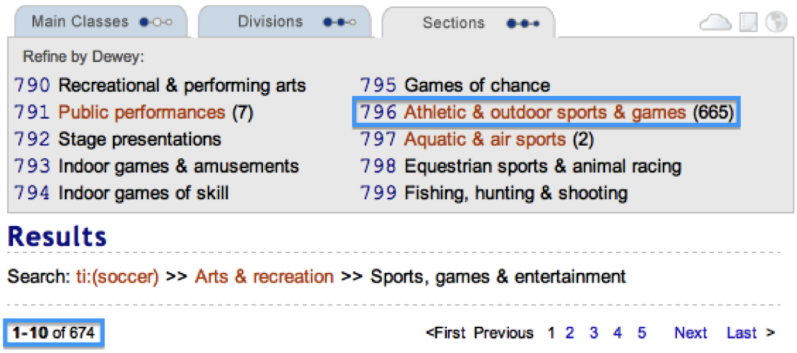

Figure 12: Dewey results at the section level

These steps are repeated for all Dewey classes, in case they have more than 20% of all results (the "threshold").

3.2.3.2 Method 2

The second step in performing the mapping is to perform queries for the Freebase domains in the subject index of the DeweyBrowser. This process is the same as described for Method 1 (see previous section), but the searches are performed using the following search string:

```
su:(<searchterm>)
```

The prefix “su:” indicates that we search for a subject heading, and “<searchterm>” stands for the search term that is used for the query. For each Freebase domain, these steps are repeated, in order to retrieve the matching Dewey classes, divisions and subdivisions.

3.2.3.2.1 Matching method 1 and 2

An additional step in the mapping process, after getting the results for both method 1 and method 2, is to compare the results, and to find the matching elements. A matching element is a DDC class that occurs in method 1 as well as method 2.

Using the “IF” function of Microsoft Excel, the results (i.e. the DDC number) for the title and subject heading searches are compared. If both values match, they are saved in an additional column of the worksheet. It is possible to set a threshold value in the worksheet, that will check if the featured Dewey partial number (at the class, division or section level) has a percentage of relevant items that is high enough to be included (for example 20% of all results). The value above this threshold is subsequently displayed.

3.2.3.3 Method 3

The methods described in the previous sections are refined using the Relative Index of the Dewey Decimal Classification, that is available in book form (Dewey, 2011) and via an online tool, called WebDewey. The following steps are taken in this refinement process:

1. The matches between mapping 1 and 2 are used to search for Freebase domains in Dewey classes, by using Dewey's Relative Index. If a more narrow class is found, using the name of the Freebase domain or a synonym, it is added to the Excel sheet. For example, DDC class `796` could be refined using the Relative Index to the more precise class `796.332: American Football` that matches the Freebase domain identifier.
   Exceptions:
   - The resulting classes from the matches between mapping 1 and 2 that are too broad (i.e. appearing at the class level) are discarded, unless there is a match between the Freebase domain name and the caption of the Dewey class.
2. Subsequently, for the Freebase domains with a result that appeared in either method 1, or in method 2, but not in both, the resulting Dewey class with the highest percentage of relevant results will be chosen. Like in step 1), Dewey's

Relative Index is searched for this class, and the class is added to the list of mapped classes if a valid result is found.

Finally, these refined DDC numbers are the basis for the further analysis and visualization of the mapping from Freebase domains to the DDC.

3.3 Summary

This chapter has described the process of data collection and analysis, that is used in this thesis. This study makes use of a theoretical analysis using Weller's criteria to compare KOS, and a practical analysis, based on quantative methods. The study uses three basic methods, in order to create a mapping between Freebase and the Dewey Decimal Classification.

4 Theoretical analysis

This is the first of the two chapters of the analysis section of this thesis. In this chapter, multiple criteria to compare knowledge organization systems are applied, aiming to answer the following question:

- How do the Dewey Decimal Classification and Freebase compare in terms of structure, based on the criteria to compare KOS defined by Weller (2010)?

Section 4.1 discusses the general concepts of the criteria to compare KOS that are used in this thesis, and section 4.2 features a comparison between the DDC and Freebase.

4.1 Criteria to compare knowledge organization systems

Six criteria to compare KOS, which have been discussed by Weller (2010), were used as a general guideline for the comparison between the Dewey Decimal Classification and Freebase in this chapter. These criteria are the following:

1. *domain of interest:* this is the domain that the knowledge organization system intends to represent.
2. *complexity:* Weller indicates that the complexity is "based on the use of semantic relations as a method of vocabulary control[89] and an expression of meaning". So this criterion mainly looks at types of relations that can exist in a knowledge organization system, that are influencing the complexity of such a system.
3. *size:* this can be measured by the number of concepts, relations or levels in the hierarchy.
4. *formality:* Uschold and Jasper (1999), as adapted by Weller (2010) define the following formality levels of KOS:
   - *highly informal*: "expressed loosely in natural language"
   - *structured-informal*: "expressed in a restricted and structured form of natural language"
   - *semi-formal*: KOS making "use of standardized structuring principles (mainly semantic relations) without applying a formal language, e.g. classifications and thesauri"
   - *formal*: making use of a "formally defined ontology language"
   - *rigorously formal*: "meticulously defined terms with formal semantics, theorems, and proofs of such properties as soundness and completeness"

   We can apply these formality levels to Freebase and the Dewey Decimal Classification, to see whether the DDC and Freebase are similar in this regard.

[89] Vocabulary control is the control over for instance synonyms, homonyms and related terms

5. *usage:* in this section, we differentiate between the *purpose* of a KOS, for example as an indexing language, the *type of resources* that are classified using the system, and the *application field* of the system.
6. *general modeling principles*: these are general aspects related to the modeling of the concepts in a KOS. In addition to that, we will look at the following modeling structures that can be applied to the DDC and Freebase:
    - *mono- vs. polyhierarchy*: in a monohierarchy, a class is assigned to one parent, while in a polyhierarchy, a class can be assigned to multiple superclasses
    - *mono- vs. multidimensionality*: this can be classification based on one aspect, for example discipline, or multiple aspects, for example type and format.

In the next section, the criteria to compare KOS listed above are applied to the DDC and Freebase.

4.2 Comparing the structure of the DDC and Freebase

4.2.1 Domain of interest

Both Freebase and the DDC have the same general goal, as we have seen in the literature review: Freebase is “a database designed to be a public repository of the world's knowledge” (Bollacker et al., 2008), and the classes of the DDC are “meant to store all the world's knowledge” (Peters, 2009). So we can say, based on this information, that the scope of both systems is very broad, as they intend to capture “the world's knowledge”.

4.2.2 Complexity

As mentioned in the previous section, Weller (2010) has defined complexity as “most fundamentally based on the use of semantic relations as a method of vocabulary control and an expression of meaning”. Based on this definition, KOS can be arranged from simple to complex.

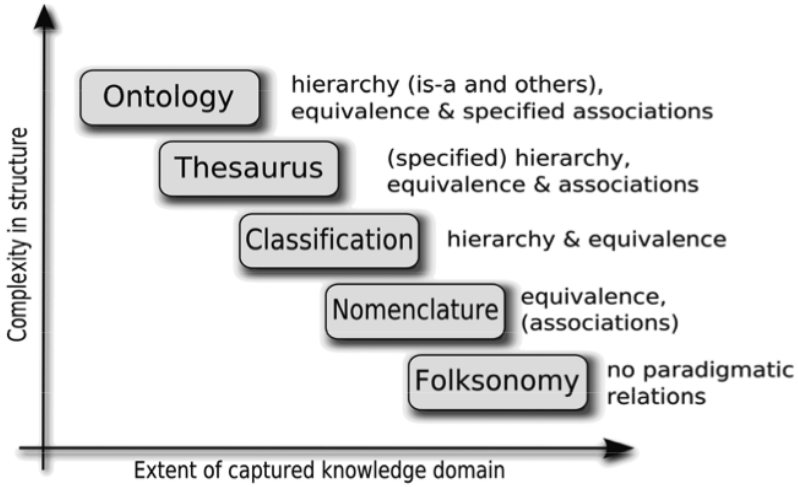

Figure 13: Knowledge structures arranged by their complexity and extent of the captured knowledge domain (Weller, 2010)

Folksonomies are seen as the least complex KOS by Weller (see figure 13, and also section 2.1.1), because they do not apply formal vocabulary control. Implicit relations might be found by doing statistical analyses, though. In Weller's diagram, classifications (like the DDC) are located in the middle, being less complex than thesauri and ontologies, and more complex than folksonomies. According to Weller, the DDC is only having hierarchical and equivalence relations. Following Mitchell (2001), we can add that there are also implicit relations, i.e. relations which exist because of "notes in the schedules, tables, and Manual; and in entries in the Relative Index", but that are not formally defined, as in for example thesauri. The implicit relations are not applied consistently, and can be hard to validate, since they are only available in the DDC's class notes.

Ontologies are the most complex KOS in Weller's list of systems, since it is possible to define custom semantic relations. The structure of Freebase is most similar to an ontology, so we define it as an ontology (for a discussion on whether Freebase's structure is an ontology or not, see section 2.2.3.4.2). Different relationship types are possible in Freebase, by using the associated types and properties. Compared to the DDC, Freebase's structure contains more elaborate hierarchical relationships, specified associations, and also contains equivalence relations.

4.2.3 Size

We can look at different aspects of the size of the DDC and Freebase. Firstly, we can look at the size of the main knowledge structure, and secondly, we can look at the size of different parts of the knowledge structure based on the relative number of instances.

The 22nd edition of the Dewey Decimal Classification contains approximately 50,000 classes (Gödert (2002), as quoted by Jacobs et al., 2010)), but additional classes can be added by "number building" using the auxiliary tables of the DDC.

The knowledge structure of Freebase[90] roughly contains 10 domain categories, 80 domains, 3,000 types and 30,000 properties. This structure is used for approximately 10,000,000 instances, Freebase's topics. In Freebase, there is no exact equivalent of a "class" as used in the DDC, but parts of the DDC's structure can be replicated using Freebase's domain categories, domains, types and properties.

4.2.4 Formality

Weller (2010) adapts the approach of Uschold and Jasper (1999) to define formality. She argues that the DDC can be seen as a *semi-formal system*, since it makes use of "standardized structuring principles", "without applying a formal language". These properties of DDC indicate that there is a fundamental difference between formal knowledge representation languages (using first-order logic) and the more associative way of structuring of classification systems (Zeng, Panzer, & Salaba, 2009).

Arguably, the structure of Freebase can be seen as semi-formal or formal. On the one hand, it could be seen as semi-formal, since it does not make use of a formal ontology language. On the other hand, the contents of Freebase are partially represented in for example the RDF ontology language[91], so Freebase's structure might also be classified as a "formal" system.

4.2.5 Usage

The usage of the knowledge structures of the DDC and Freebase differs. The knowledge structure of the DDC has mainly been invented for indexing and retrieval of items, usually books. The structure of Freebase, however, has been created to be able to do reasoning in addition to indexing and retrieval.

This is related to the kind of the materials that are indexed with both systems; the DDC is mainly aimed at classifying physical items, aiming to "describe documents with unified vocabularies, to support the reformulation of queries and navigating through document collections" (Weller, 2010). Freebase, on the other hand, has recently been created, in a time when documents in a digital form are readily available. Therefore Freebase's structure, created with digital items in mind, and having clear relations between information concepts (entities), might be better suited to be used with internet resources. The DDC has many implicit and relationships that are invisible at

[90] as of 16/03/2012

[91] Freebase's contents are also available in the RDF format at http://rdf.freebase.com/ [retrieved: 26-06-2012], although this service generates a "rather raw view" of Freebase's structure in RDF, see http://web.archive.org/web/20110723121918/http://blog.freebase.com/2008/10/30/introducing_the_rdf_service/ [retrieved: 26-06-2012]

first sight (see for instance Mitchell (2001)), making it for instance harder to reuse its structure, and to automatically reason with its contents.

The general application field of both systems also differs: the DDC has mainly been used by libraries and related knowledge institutions, while Freebase's structure has been created to be used as the basis for an online collaborative knowledge base in the context of the semantic Web.

4.2.6 General modeling principles

The general modeling principles of Dewey and Freebase are different. The DDC has a long history of revisions by an "expert" committee, and mostly standardized principles. The long history also means that there are some legacy and practical issues, as they for instance have to keep current classes intact to ensure compatibility with older editions.

Freebase has rough modeling guidelines, and a generic structure of topics, types and properties that should be used when doing data modeling. The guidelines and principles are not as ironed out as in the DDC, and mostly provided in the form of Wiki documentation.

Other aspects of the general modeling principles that have been described by Weller (2010), are *mono-* versus *polyhierarchies*. The DDC is an example of a monohierarchy, in which a class is assigned to one parent, and it is not possible to assign a class to multiple parents. Freebase, on the other hand, has a much more open structures that resembles a graph, and has more polyhierarchical features: as discussed in the literature review, it is even possible to add contradictory types and properties to the structure, "to reflect users' differing opinions and understanding" (Bollacker et al., 2008).

Also in terms of *mono-* versus *multidimensionality* Freebase and the DDC differ. The classification of materials in the DDC is done based on one aspect (e.g. based on discipline), and therefore an item theoretically[92] can only be assigned to one class. This can give some problems, as Satija (2007) indicates using the example of the book "Cataloguing in Academic Libraries", that can be classified under `025.3` *Cataloguing*, and `027.7` *Academic libraries*. To select which DDC class to use for the classification of an item there are even guidelines to decide the assignment of classes in case of doubt, in the form of "tables of precedence" and general instructions. In Freebase, this problem is not occurring, since it is possible to assign multiple types to a topic, for example "Amsterdam" is typed as a "Location", a "Travel Destination" and a "Literature Subject"[93].

[92] Note: in practice library might assign more than one class number to a book, but physical items can only be shelved in one place.

[93] http://www.freebase.com/view/en/amsterdam [retrieved: 28-05-2012]

4.3 Conclusion

In this chapter, a theoretical comparison between Freebase and the Dewey Decimal Classification has been made. Table 6 provides a summary of the previous sections, and some basic indicators of the differences and similarities of the structure of the DDC and Freebase:

		DDC	Freebase
1	**domain of interest**	"the world's knowledge"	"the world's knowledge"
2	**complexity**	relations: hierarchy & equivalence	relations: hierarchy ("is-a"), equivalence & specified associations
3	**size**	50,000+ classes, extendable via auxiliary tables	10 domain categories, 80 domains, 3,000 types, 30,000 properties, 10,000,000 instances
4	**formality**	semi-formal	Semi-formal / formal
5	**usage**	purpose: indexing, retrieval type of resources: mainly physical application field: libraries	purpose: indexing ("representing facts"), retrieval, reasoning type of resources: digital application field: online knowledge base
6	**general modeling principles**	formal guidelines and documentation -monohierarchy -monodimensional	rough guidelines and Wiki documentation -polyhierarchy -polydimensional

Table 6: Basic comparison of the Dewey Decimal Classification and Freebase

We can see that there are some similarities between the DDC and Freebase, especially in terms of the domain of interest, which is very broad in both cases. Also regarding formality: both can be classified as being "semi-formal", though Freebase has more options to do reasoning with its more structured data, and has more different relationship types, and therefore could also be considered a "formal" system.

Even though the data of Freebase is more structured than in the DDC, the modeling principles are somewhat less structured, as the DDC has formal guidelines and extensive documentation, while Freebase has rough guidelines and Wiki-based documentation. The size of the Dewey Classification and Freebase is hard to compare, since the nature of their knowledge structure varies widely. The original purpose of their data structure is also much different, as the DDC was intended to be used with physical materials, and Freebase has originated in the digital web age.

The differences between Freebase and the DDC could have some influence on the ability to map Freebase domains to DDC classes, since this also depends on the structure of both knowledge organization systems. We will see in the next chapter whether this influences the actual mapping that has been carried out for this thesis.

5 Practical Analysis

5.0.1 Introduction

After the theoretical analysis in the previous chapter, this chapter carries out a practical analysis of the feasibility of mapping Freebase domains to DDC categories.
This chapter aims to answer the following research question:

- To what extent can Freebase domains be mapped to classes of the Dewey Decimal Classification?

Section 5.1 discusses the results of the mapping that is carried out for this thesis, and subsequently, in section 5.2, this mapping is analyzed, the research question is answered, and some possible changes to the methodology are discussed, in order to improve the statistical mapping process.

5.1 Results

5.1.1 Overview

The initial mapping for this study was done between 80 Freebase domains and the DDC classes from the 23rd DDC edition. Section 5.1.2 discusses the basic mapping that was done using the title and subject indexes of the DeweyBrowser (mapping method 1 and 2), and section 5.1.3 discusses the refined mapping, created using Dewey's Relative Index (mapping method 3).

Different levels of equivalence[94] between source and target elements of the mapping influence the number of Dewey classes that is needed to represent one Freebase domain. When a Freebase domain is more or less *fully equivalent* to a DDC class, it would be enough to assign one Dewey class to one Freebase domain (a 1:1, or one-to-one mapping). However, when a Freebase domain is not fully equivalent to a DDC class, it might be necessary to assign multiple DDC classes to one Freebase domain; for example if the Dewey class is a narrower equivalent of a Freebase domain (a 1:*n*, or one-to-many mapping, see also section 2.3.2.1).

5.1.2 Basic mapping (method 1 & method 2)

5.1.2.1 General mapping statistics

The first steps in the mapping process involved searching in the DeweyBrowser's title and subject indexes using the names of Freebase domains, and selecting the most relevant results (based on the relative number of assigned WorldCat items for those domains). The resulting DDC classes, divisions and subdivisions were entered in an Excel-worksheet (see Appendix), culminating in a set of initial mappings.

[94] following Koch et al. (2001)

5.1.2.1.1 Single matches

In the first run of the mapping, Freebase domains were connected to single Dewey classes. The first step (method 1) involved searching in the title index of the DeweyBrowser, following the methodology described in chapter 3, and analyzing the results. Using this method, 65 DDC classes matched the 80 Freebase domains (the Freebase domains without matching DDC classes are discussed in section 5.1.3.5). The second method, searching in the subject headings of the DeweyBrowser, resulted in 69 matches.

Freebase domain	*Matching DDC classes (1:1)*		
	Method 1 (title)	**Method 2 (subject)**	**Combination 1+2**
80 domains	65 matches (81%)	69 matches (86%)	59 matches (73,8%)

Table 7: Freebase domains and matching DDC classes (one-to-one)

Finally, the matches between method 1 and method 2 were combined, by comparing them and selecting only the matching elements. A matching element is a DDC class that occurs in both method 1 and 2. This step resulted in a list of 59 matching DDC classes for all Freebase domains.

5.1.2.1.2 Multiple matches

For some Freebase domains, a mapping that has one target class is not enough; they can be associated with multiple Dewey classes. One of the underlying reasons is that subjects are often located in different parts of the Dewey hierarchy (the DDC orders subjects by discipline, as discussed in section 2.1.3.3).

In a second run of the mapping, also multiple DDC matches for each Freebase domain were allowed. A basic criterion for the inclusion of a class was that it should cover at least 20% of all WorldCat results for a query at the class, division or subdivision level. So only classes, divisions and subdivisions above a threshold of 20% were recorded when using method 1 and 2. This resulted in the DDC classes summarized in table 8:

Freebase domain	*Matching DDC classes (1:n)*		
	Method 1 (title)	**Method 2 (subject)**	**Combination 1+2**
80 domains	76 matches	85 matches	66 matches
	1,17 DDC classes / Freebase domain	1,23 DDC classes / Freebase domain	1,12 DDC classes / Freebase domain

Table 8: Freebase domains and matching DDC classes (one-to-many)

Table 8 shows that there are 1,17 matching DDC classes per Freebase domain using method 1 and allowing multiple classes per domain, and that there are 1,23 matching

DDC classes per Freebase domain when using method 2 (with a selected threshold of 20%). Most likely, if a lower threshold would be chosen, more DDC classes would be returned per Freebase domain. In this mapping, the highest number number of Dewey classes that is assigned to one domain is 3.

If we look at both table 7 and 8, it can be derived that the search for Freebase domains using subject headings generally gave more results.

5.1.2.2 Mapping statistics based on Freebase domain category

We can also look at the results of the mapping by focusing on the Freebase domain categories. A domain category (see 2.2.3.4.1) is a "grouping of related domains". In this thesis, we will look at 9 of the 10 existing domain categories. We omit the "Commons" domain category, because all its domains are also assigned to other domain categories (as all current Freebase domains are part of the Freebase Commons, that have been approved by Freebase's staff).

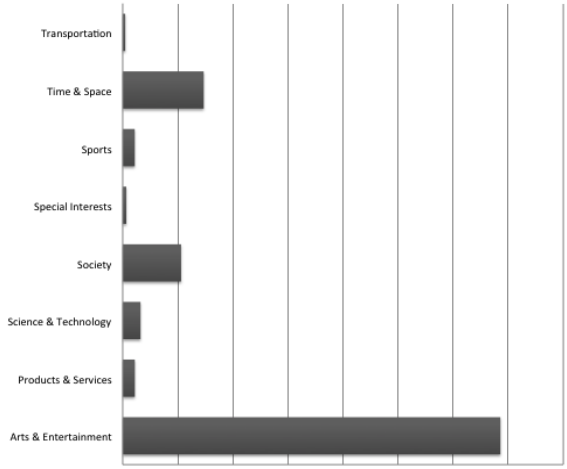

Figure 14: Frequency distribution of Freebase Domains Categories, based on the number of instances (topics)

Figure 14 shows the distribution of Freebase instances in the different domain categories of Freebase. Of the nine main categories of domains in Freebase, the internal "System" domain is not included in the figure. As we can see, the "Arts & Entertainment" category contains the majority of items. In the figure, we can see some general tendencies, mostly related to the availability of datasets that have been imported into Freebase; for example, in 2007, the imported dataset with song information from MusicBrainz contained 4 million entries (Markoff, 2007).

Table 9 shows a breakdown of the results of the statistical mapping using methods 1 and 2, grouped by the domain category assigned to the different domains. Two domains are not assigned to any domain category ("Metaweb System Types" and

"Projects")[95] and are included here under "Unknown". Two other Freebase domains belong to multiple domain categories: Automotive (Transportation / Products & Services), and Spaceflight (Transportation / Science & Technology).

Domain Category	Freebase domains						
	Total	Method 1: DDC numbers assigned		Method 2: DDC numbers assigned		Combination M.1 & M.2: DDC numbers assigned	
Arts & Entertainment	15	14	93,3%	13	86,7%	11	73,3%
Products & Services	3	3	100,0%	3	100,0%	3	100,0%
Science & Technology	10	10	100,0%	10	100,0%	10	100,0%
Society	14	9	64,3%	13	92,9%	8	57,1%
Special Interests	8	7	87,5%	7	87,5%	7	87,5%
Sports	12	12	100,0%	12	100,0%	12	100,0%
System	3	1	33,3%	0	0,0%	0	0,0%
Time & Space	6	1	16,7%	3	50,0%	1	16,7%
Transportation	5	5	100,0%	5	100,0%	5	100,0%
Transportation, Science & Technology	1	1	100,0%	1	100,0%	1	100,0%
Transportation, Products & Services	1	1	100,0%	1	100,0%	0	0,0%
Unknown	2	1	50,0%	1	50,0%	1	50,0%
Total (mean)	**80**	**65**	**78,8%**	**69**	**80,6%**	**59**	**65,4%**
Total (excl. System, Unknown domain category)	75	63	86,2%	68	91,7%	57	73,5%

Table 9: DDC classes assigned per Freebase domain category

The second column of table 9 shows the number of Freebase domains in the domain categories; the third and fourth column show the number of Freebase domains that have one or more DDC numbers assigned to them, using method 1 and method 2. The final column consists of the matches between method 1 and method 2: Freebase domains that were assigned to Dewey numbers, occurring using both method 1 and method 2.

The data in the table indicates that some of the domain categories seem to be more suitable to be mapped than others: *Arts & Entertainment*, *Products and Services*, *Science & Technology*, *Special Interests*, *Sports and Transportation* have a mapping percentage of 73,3% or higher, using all methods.

On the other hand, some classes have very few mappings, for example *System* and *Time & Space*. The combined domain categories "*Transportation / Science and Technology*", "*Transportation / Products and Services*", and the *Unknown* categories have only one domain assigned to them, and therefore are less relevant for the results.

[95] They do not show up on Freebase's website under any of the domain categories

5.1.3 Mapping refinement (method 3)

5.1.3.1 Process

The next step in the creation of the mapping consisted of a refinement, based on the combined results from mapping method 1 and 2 and Dewey's Relative Index. This resulted in a new, adapted mapping. The following steps were taken in this mapping refinement process (also described in section 3.2.3.3):

1. Initially, the matches between mapping 1 and 2 were used to search for Freebase domains in Dewey classes, by using Dewey's Relative Index. If a more precise class was found, using the name of the Freebase domain or a synonym, it was added to the Excel sheet. For example, in this step, the DDC class `796` that results from the matched mapping of step 1 and 2, is refined, using the Relative Index, to the more precise class `796.332: American Football` that matches the Freebase domain identifier.
   Exceptions:
   - The resulting classes from the matches between mapping 1 and 2 that were too broad (the top level classes, and divisions, like `3: Social Sciences`) were discarded, unless there was a match between the Freebase domain name and the caption of the Dewey class (in the case of `2: Religion`).
2. Subsequently, for the Freebase domains with a result that appeared in either method 1, or in method 2, but not in both, the resulting Dewey class with the highest percentage of relevant results was chosen. Like in step 1), Dewey's Relative Index was searched for this class, and the class was added to the list of mapped classes if a valid result was found. For example, the DDC class `320` was refined to `320.351` for the Freebase domain “Government”.

5.1.3.2 General mapping statistics

5.1.3.2.1 Single matches

The refined mapping created as a result of the steps above is summarized in table 10. This table lists the number of unique matches from the first run of the mapping, regarding the instances of the mapping as having 1:1 relationships, and compares this with the matches of method 1 & 2.

Freebase domain	*Matching DDC classes (1:1)*	
	Combination Method 1+2	**Method 3**
80 domains	59 unique matches (73,8%)	60 unique matches (75%)

Table 10: Unique Freebase domains and matching DDC classes

Using the refinement method, one unique match was added to the existing combined list from method 1 and 2, and several other mappings were refined. In the next table, we look at the results of the mapping with the 1:*n* relationship perspective, i.e. multiple DDC classes that can match one Freebase domain.

5.1.3.2.2 Multiple matches

The following table lists the total number of DDC matches resulting of the next mapping run, allowing multiple DDC classes to match one Freebase domain:

Freebase domains	*Matching DDC classes (1:n)*	
	Combination Method 1+2	**Method 3**
80 domains	66 matches	70 matches
	1,12 DDC classes / Freebase domain	1,17 DDC classes / Freebase domain

Table 11: Freebase domains and matching DDC classes (multiple matches allowed)

The summary in table 11 shows that 4 additional classes have been mapped using the third mapping method, if we compare it with the matches between method 1 and 2. This results in a slightly higher number of DDC classes per Freebase domain.

5.1.3.3 Mapping statistics based on Freebase Domain Category

Table 12, like table 9 in the preceding section, shows the results grouped by domain category. It looks at the number of Freebase domains with one or more assigned DDC classes using the combined matches of method 1 and 2, and using method 3.

Domain Category	*Freebase domains*				
	Total	**Matches 1 & 2: DDC numbers assigned**		**Method 3: DDC numbers assigned**	
Arts & Entertainment	15	11	73,3%	12	80,00%
Products & Services	3	3	100,0%	3	100,00%
Science & Technology	10	10	100,0%	10	100,00%
Society	14	8	57,1%	7	50,00%
Special Interests	8	7	87,5%	7	87,50%
Sports	12	12	100,0%	12	100,00%
System	3	0	0,0%	0	0,00%
Time & Space	6	1	16,7%	2	33,33%
Transportation	5	5	100,0%	5	100,00%
Transportation, Products & Services	1	0	0,0%	1	100,00%
Transportation, Science & Technology	1	1	100,0%	1	100,00%
Unknown	2	1	50,0%	0	0,00%
Total	80	59	65,4%	60	70,9%
Total (excl. system, unknown domain category)	75	58	73,5%	60	85,1%

Table 12: The number of Freebase domains with DDC mapped classes, using method 3

This table indicates that there is a slight difference in the results of the two methods, and that method 3 culminated in a slightly higher average percentage of mapped target classes, with the exception of the *Unknown* domain category.

The table does not include the actual domains and classes that have been assigned in the mapping, something which will be covered in the next section.

5.1.3.4 Mapped domains and classes

5.1.3.4.1 Overview

This section contains the concrete source and target elements of the mapping: Freebase domains and DDC classes. The results of method 3 have been used for this diagram (figure 15, see next page).

The first column contains the Freebase domain name (red column) and the DDC caption (blue column). Using a color coding, the mapped Freebase domains and classes are visualized on the right, and ordered based on their DDC number (from class 0, Computer science, knowledge & systems to class 9, History & geography).

Larger bars represent broader Dewey classes (i.e. having a position high in the hierarchy), and shorter bars represent narrower Dewey classes (i.e. at a lower level in the hierarchy). The color of the bars represents the Freebase domain category of each domain.

Freebase domain name	DDC Caption
Computers	Computer science, knowledge & systems
Internet	Internet
Library	Library & information sciences
Internet	World Wide Web
Religion	Religion
Media	Media (Means of communication)
Government	political science (politics and government),public administration
Protected Places	Conservation and protection
Business	Production
Law	Law
Military	Military
Education	Education
Transportation	Commerce, communications & transportation
Radio	Radiobroadcasting
TV	Television
Rail	Railroad transportation
Transportation	Transportation; ground transportation
Fashion, Clothing and Textiles	Costume and personal appearance
Language	Language
Astronomy	Astronomy
Physics	Physics
Chemistry	Chemistry
Geology	Earth sciences & geology
Meteorology	Meteorology
Biology	Life sciences; biology
Zoos and Aquariums	Collections and exhibits of living animals
Medicine	Medicine & health
Engineering	Engineering
Radio	Radio and radar
Boats	Nautical engineering and seamanship
Aviation	Aeronautics
Automotive	Motor land vehicles, cycles
Bicycles	Bicycles
Spaceflight	Space flight
Zoos and Aquariums	Fish culture in aquariums
Food & Drink	Food & drink
Business	General Management
Visual Art	The arts
Architecture	Architecture
Comics	Comic books, graphic novels, ..
Fashion, Clothing and Textiles	Costume
Digicams	Cameras and accessories
Music	Music
Opera	Operas and related dramatic vocal forms
Hobbies and Interests	Activities generally engaged in by individuals
Amusement Parks	Amusement parks
Film	Motion pictures
Radio	Radio
TV	Television
Theater	Stage presentations
Games	Indoor games and amusements
Games	Indoor games of skill
Video Games	Electronic games
Sports	Athletic and outdoor sports and games
Basketball	Basketball
American football	American football
Soccer	Soccer (Association football)
Tennis	Tennis (Lawn tennis)
Baseball	Baseball
Cricket	Cricket
Olympics	Olympic games
Bicycles	Cycling and related activities
Martial Arts	Combat Sports
Boxing	Boxing
Skiing	Skiing and snowboarding
Ice Hockey	Ice hockey
Books	Literature, rhetoric & criticism
Travel	Geography and travel
Physical Geography	Miscellany
Royalty and Nobility	Royal houses, peerage, orders of knighthood

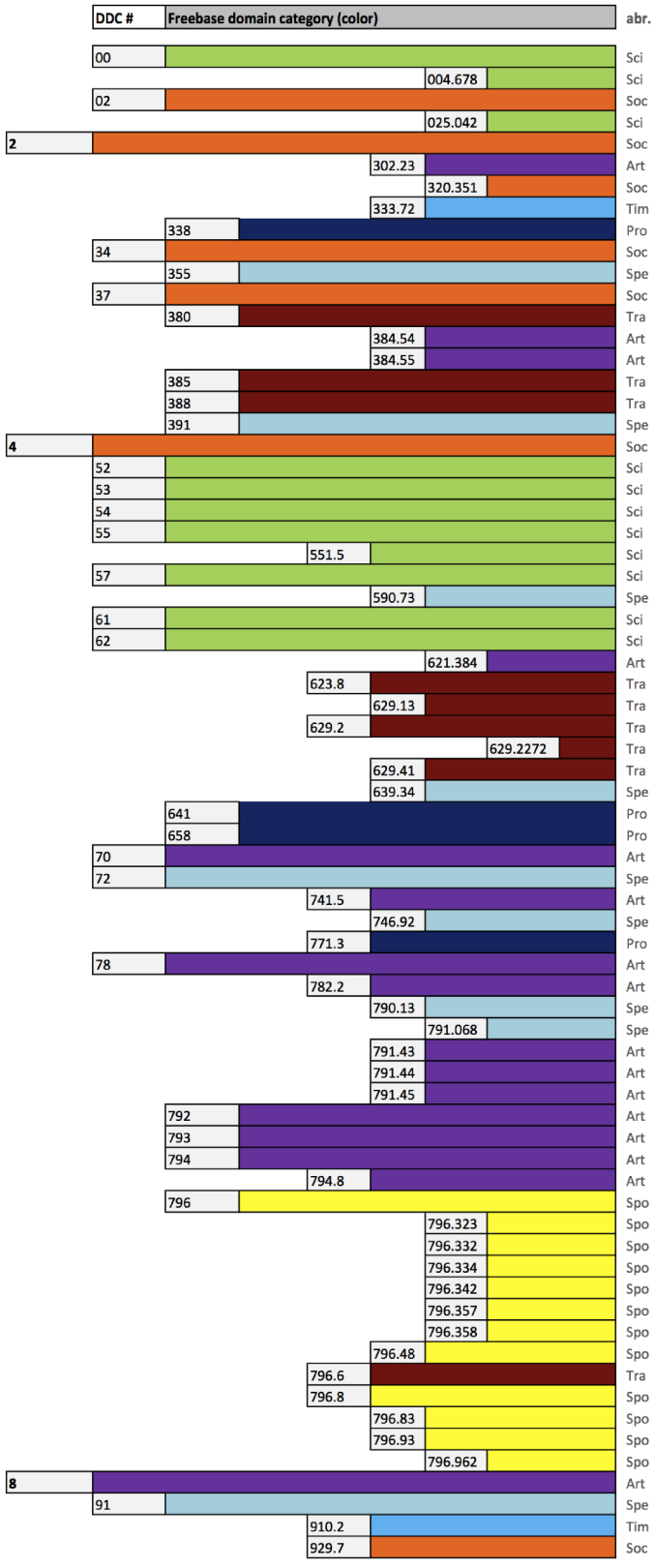

Figure 15: Mapped Freebase classes

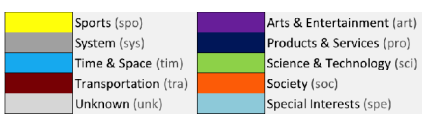

Figure 16: Color codes and abbreviations

5.1.3.4.2 Observations

Using the color coding in figure 15, it is possible to distinguish basic clusters of Freebase domains, and domains that do not belong to these clusters. In many cases, Freebase domains under particular domain categories are mapped to particular ranges of the Dewey classification. For example, the Freebase domains under the Science & Technology domain category are mainly located under the `5` class of the DDC (*Science*). Similarly, the domains under Arts & Entertainment are mainly located under the `70` to `79` divisions of the DDC (*Arts & Recreation*). The Sports domain category exclusively appears under the `796` section of Dewey (*Sports, Games and Entertainment*).

From the color coding we can derive that some of the Freebase domains are probably not in the right location, if we compare the placement of the domain with the domain category that is assigned to it. For example, Radio is located under `384.54`, `621.384` and `791.44`. The color coding reveals, however, that the 7 class of Dewey might be the most suitable place for *Radio* in this diagram, considering the domain category of the surrounding classes (*Arts & Entertainment*). This can be explained by the fact that the *Radio* domain is part of the *Arts & Entertainment* domain category of Freebase. Similarly, *Television* is located at `384.55`, but also `791.45`. It should probably only be located in the `7` class, since parent category of this domain is also *Arts & Entertainment*, and since the surrounding domains assigned to the 7 class have the same category.

We can also focus on a particular Freebase domain category, see what kind of elements it contains, and how they fit into the hierarchy. The sports-related domains of Freebase are mainly found in the 796 subdivision of the DDC (*Athletic & outdoor sports & games*).

Freebase has a number of domains that are used for sports, all part of the Sports domain category: *American Football, Baseball, Basketball, Boxing, Cricket, Ice Hockey, Martial Arts, Skiing, Soccer* and *Tennis*. In addition to that, the Sports domain category has an *Olympics* domain. In general, the mapping of these domains to DDC classes seems to be straightforward, since the mapped domains have only one associated DDC number (i.e. they are fully equivalent, and have a 1:1 relationship). Additionally, the *Sports* domain category also contains a specific domain called "*Sports*" (having the same name as its parent), that includes all sports that do not have their own domain.

Based on the domains assigned to the *Sports* domain category, we can see that the domains that are part of it have various properties: it is possible to differentiate a

broad domain (*Sports*), *specific domains* that represent one sport each (*American Football, Baseball*, ..), and one domain that contains Freebase topics related to an actual sports *event* (*Olympics*). Representing these different kinds of domains in the DDC properly might be hard, since the aim in the DDC (and classification systems in general) is to have "mutually exclusive" classes and "totally exhaustive" classes in the array of classes in a level (Chan, 2007), which means that they should not overlap, and should contain all possible variations.

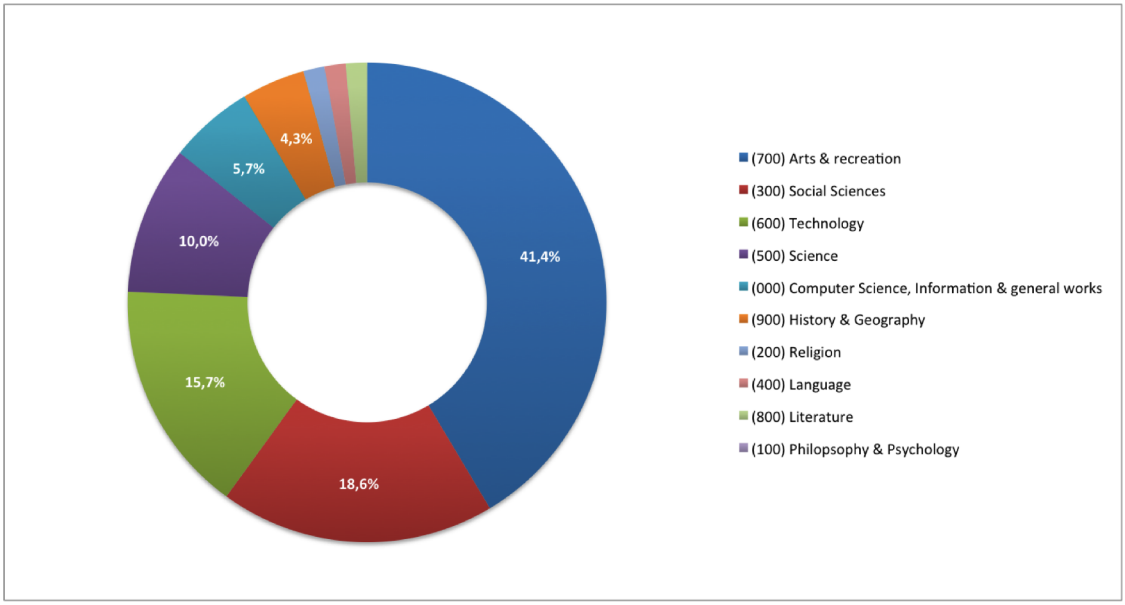

Figure 17: Mapped Freebase domains to main DDC classes & their subordinate classes (percentage)

5.1.3.4.3 Mapped Freebase domains & DDC classes

Figure 17 has been generated using the relative numbers of Freebase domains mapped to different DDC classes, divisions, sections and subdivisions, in order to show which classes are the target elements of which percentage of Freebase domains. Of all the domains, 75% is mapped to the `7` (*Arts & Recreation*), `3` (*Social Sciences*) and `6` classes (*Technology*) of the DDC. The largest percentage of Freebase domains, 41.4%, is mapped to the *Arts & Recreation* DDC class.

On the other hand, some DDC classes contain very few Freebase domains, like the 2 (*Religion*), 4 (*Language*), and 8 (*Literature*) classes, that are all target classes of just one broad Freebase source domain, respectively *Religion, Language* and *Books*.

From this diagram we can derive which class of the Dewey Classification is not covered by the mapping at all: the DDC class `100`, *Philosophy and Psychology*. In the next section, we will take a look at the Freebase domains that could not be mapped with the used methodology.

5.1.3.5 Unmapped Freebase domains

As indicated in the previous sections, not all Freebase domains could be mapped to a DDC class. To get more insights in the properties of the domains that could not be mapped and the underlying reasons, a list of unmapped classes has been generated. Table 13 shows the variety of classes that could not be mapped at all using the statistical methods:

	Domain	Domain Category
1	Awards	Society
2.	Broadcast	Arts & entertainment
3.	Celebrities	Society
4.	Common	System
5.	Conferences and conventions	Society
6.	Data world	System
7.	Event	Time & space
8.	Exhibitions	Society
9.	Fictional universes	Arts & entertainment
10.	Freebase	System
11.	Influence	Society
12.	Location	Time & space
13.	Measurement unit	Time & space
14.	Metaweb system types	Unknown
15.	Organization	Society
16.	People	Society
17.	Periodicals	Society
18.	Projects	Unknown
19.	Symbols	Special interests
20.	Time	Time & space

Table 13: Freebase: unmapped domains

From the list of 20 domains, we can derive that some Freebase domain categories contain many unmapped classes. To analyze this further, we grouped the unmapped classes by Freebase domain category. Table 14 shows that the highest number of unmapped domains is part of the *Society* domain category. If we look at the unmapped domains relative to the total number of domains in a domain category, the *System* and *Unknown* domain have the highest degree of unmapped domains (though it has to be noted that they only have few domains assigned to them).

Domain category	unmapped domains	percentage of total in domain category
Society	7	50,00%
Time & Space	4	66,67%
Arts & Entertainment	3	20,00%
System	3	100,00%
Unknown	2	100,00%
Special Interests	1	12,50%

Table 14: Absolute and relative values of unmapped domains

There might be a few reasons why the *System* and *Unknown* categories could not be mapped. The *System* domain category contains the "*Freebase*", "*Common*", and "*Data World*" domains, and these domains are meant for internal knowledge representations in Freebase; i.e. not really used for adding topics (information) to Freebase. Therefore it is not possible to map the domains under this category to the Dewey Decimal Classification. The *Unknown* category seems to contain similar domains as the *System* domain category ("*Metaweb System Type*", and "*Projects*").

If we remove these two domain categories from our analysis, the remaining list contains 15 entries. In table 15 they are grouped by their domain category:

Domain Category	**Domain**
a) Society	1. Awards 2. Celebrities 3. Conferences and Conventions 4. Exhibitions 5. Influence 6. Organization 7. People
b) Arts & Entertainment	1. Broadcast 2. Fictional Universes 3. Periodicals
c) Time & Space	1. Event 2. Location 3. Measurement Unit 4. Time
d) Special Interests	1. Symbols

Table 15: Grouped unmapped classes

Some of the underlying reasons why it might be hard to classify the Freebase domains in table 15 under a Dewey class are related to the general structure of Dewey, but also to frequent broadness and unclarity of Freebase domains.

- Some of the listed domains are hard to put into the DDC's structure, like a3) "*Conferences and conventions*". Items in traditional classification systems like the Dewey Decimal Classification are initially being classified on the basis of their discipline. Therefore a book about a certain conference or convention, should be classified under the discipline it is referring to (for instance economics, mathematics or law, all located in different divisions in Dewey).
- Many of the domains on the list are very generic. Domains like "*Awards*", "*Events*" or "*Symbols*" could exist in multiple sections of the Dewey Decimal Classification. The domains under the category "*Time & Space*" are very broad as well (e.g. Events, Location, Measurement Unit and Time).
- A third reason that a some of these domains are hard to classify is related to multiple possible meanings of the names of Freebase domains. The "*Broadcast*" domain seems to contain media-related items; however, this is not clear from the name of the domain, that could also be interpreted as broadcasting

(the communications channel). The "*Influence*" Freebase domain is not clear from its title, and even if you look at the domain description it is hard to define it precisely: "The goal of the Genealogy of Influence is to document the creative influences of historic figures" [96]. Classes like this are not easily captured in the DDC's structure.

To conclude, there are some Freebase domains that maybe cannot be mapped to Freebase, due to inherently different properties compared to DDC classes. However, a few domains might still be mapped in a manual way, like the "*Exhibitions*" and "*Broadcast*" domain, that could be placed in the 7: *Arts & Entertainment* class of the Dewey Decimal Classification, and the "*People*" domain, that probably could be placed under the 92: *Biography & Genealogy* division of the DDC. Therefore, a manual (or intellectual) mapping, in addition to the performed statistical mapping, might be useful to improve the mapping, but was not included in this thesis.

5.2 Analysis

After the descriptions and visualizations of the results of the mapping, this section continues with a basic analysis of the mapping of Freebase domains to the Dewey Decimal Classification, and it aims to answer this chapter's research question.

5.2.1 contains a basic summary of the quantitative results of the mapping, and subsequently 5.2.2 analyzes the results of the mapping, with the research question in mind. Finally, 5.2.3 describes future improvements to the mapping methodology, based on the experiences during the creation of the mapping.

5.2.1 Results summary

5.2.1.1 General results

As we have seen in 5.1, a high percentage of Freebase domains can be mapped to classes of the DDC. If we do not count the internal *System* domains of Freebase, a percentage between 73.5% and 91.7% of all Freebase domains could be mapped to one or more DDC classes. The location of these domains in the Dewey Classification differs, however. Some domains are mapped to broad Dewey classes (e.g. 2: *Religion*), and some others are mapped to more narrow classes (e.g. 796.962: *Ice Hockey*).

5.2.1.2 Results for different Freebase domains

The difficulty of mapping source Freebase elements to Dewey classes can vary for different Freebase domains. In general, Freebase domains from the following domain categories could be mapped relatively easily:

- the *Sports* domain category (e.g. American Football, Soccer, ..)
- the *Science & Technology* domain category (e.g. Biology, Chemistry, ..)

[96] http://schemas.freebaseapps.com/domain?id=/influence [retrieved: 23-06-2012]

- the *Arts & Entertainment* domain category (e.g. Film, Music, ..)

This can be related to the fact that these domains seem to have clear, distinct boundaries. In the case of *Science & Technology*, this might be related to the fact that they are exact disciplines, that have a well-defined set of concepts and terminology.

The Freebase domains that were hard, or even impossible to map are the following:

- the *System* domain category (Freebase, Data world, ..)
- the *Society* domain category (Awards, People, ..)
- the *Time & Space* domain category (Event, Location, ..)

These domains are used for internal knowledge representation in Freebase ("*Data World*"), or contain vague notions, like "*Awards*". In addition to that, there are facet-like domains in the *Time & Space* category, which are not covered by the DDC. So these concepts are hard to place under particular Dewey classes.

5.2.1.3 Results in different Dewey sections

When we look at the sections of Dewey that are the target elements of the mapping, some areas of the Dewey classification are relatively easy to map to. These classes contain a variety of Freebase domains, that are covering a major degree of the class:

- the `7` class (*Arts & Recreation*)
- the `3` class (*Social Sciences*)[97]
- the `6` class (*Technology*)

The opposite is true for some other Dewey classes, that have no, or few domains assigned to them:

- the `1` Dewey class, *Philosophy and Psychology*, is missing from the mapping

A reason that this class could not be mapped, might be related to a lack of user-added content and structure for this domain, or to a (lack of) availability of datasets.

In addition to *Philosophy and Psychology* class, substantial parts of other classes are not covered, an example being the `0` class (*Computer science, information & general works*), and to a lesser degree the `6` (*Technology*), and `9` class (*History & geography*).

5.2.2 Results analysis

5.2.2.1 Research question

To answer the research question of this section, "To what extent can Freebase domains be mapped to classes of the Dewey Decimal Classification?", we look at the general

[97] The assigned domains to the Social Sciences DDC class are of a very varied nature; frequent criticism to the DDC is related to this issue, and to the uneven structure of the DDC in general (Chan, 2007).

results of the mapping, and at the coverage of Freebase domains and Dewey sections by the mapping.

If we look at the results from a quantitative perspective, a major degree of the Freebase domains (73.5-91.7%) could be mapped to Dewey classes, using the statistical methodology. Therefore we can say that it is generally possible to do a mapping between Freebase domains and DDC classes. This thesis showed the general feasibility of carrying out a mapping in this style, although the mapping can still be improved in multiple ways (see section 5.2.3), for example by having access to more data.

On the other hand, if we focus on the target elements of the mapping, we can see that certain ranges of the DDC are not, or only partially covered by the mapping. Some of the source elements of the mapping cannot be connected to Dewey classes, probably not even using an intellectual mapping, due to structural differences between these two knowledge organization systems. The next section will discuss some of these structural differences in relation to the performed mapping.

5.2.2.2 Further analysis

To look at the research question of this chapter from a broader perspective, we will look again at some of the criteria defined in the previous chapter: *domain, complexity, size, formality, usage* and *general modeling principles*, and discuss which possible influence they have on the mapping from Freebase to the Dewey Decimal Classification.

Both the DDC and Freebase try to capture a very broad domain: “the world's knowledge”. Therefore they have a very general knowledge structure, capable of representing many knowledge domains. In the mapping this is visible, since the Freebase domains cover major parts of the Dewey Decimal Classification, so both knowledge organization systems seem to offer a wide spectrum of knowledge. In a relative sense, the Freebase domains, however, seem to have more domains devoted to popular culture (for example sports), while the DDC seems to cover a broader spectrum of knowledge.

The complexity of the Dewey Classification System and Freebase differs. Important elements of an ontology, the relationships between elements, are included in Freebase, something which cannot be fully captured using the hierarchical Dewey Decimal Classification. Freebase's structure also allows one to represent knowledge from multiple perspectives, and to use multiple access points. A concept like “Amsterdam” can be classified by multiple dimensions, while arrays of classes on a DDC level are divided on the basis of one characteristic.

For the mapping, we chose to focus on the domains in Freebase, and as the mapping has showed, all domains only have one parent category (except for two domains). This implies that they can be mapped more easily to the structure of the DDC than the multifaceted elements of the knowledge structure of Freebase.

We can also look at the size of the Freebase structure compared to the DDC. As the focus in the mapping in this thesis was on the domains, only 80 elements were mapped to classes of the DDC (that contains more than 50,000 classes). This means that it will not be possible to cover the DDC's full knowledge structure with the limited set of domains. However, the aim would be to cover most of the higher-level elements of the Dewey classification. This worked up to a certain degree, but as the analysis showed, there are some gaps, and the Freebase domains can be located on different levels of the DDC hierarchy.

The purposes of the DDC and Freebase are different. Both are used for indexing materials and for retrieval. Freebase, is also used for reasoning, as it is possible to derive new knowledge from it's highly structured knowledge ("facts"). We have focused on the domains, that are higher-level groupings of information contained in Freebase, and are not directly used for reasoning. Therefore they could be represented to a large degree in the DDC's structure.

The usage of the structure of both the DDC and Freebase have a direct influence on these KOS: in the case of the DDC this is based on feedback to the DDC editorial committees by libraries and institutions that use the system. In the case of Freebase, this is done in a more direct manner, in the form of forum discussions, but also by the creation of structure by users themselves, in the user-generated "bases". We can also argue that the structure of the datasets that are used in Freebase have a direct influence on the structure of Freebase, since Freebase has to be able to accommodate these structures. The domains, that were mapped in this chapter, have been added gradually, most likely based on the needs of the users, and the properties of the imported datasets.

The DDC is rather static: if you make use of the printed edition, the knowledge structure remains the same until the next edition, even while society can change quickly. The approach of Freebase in this regard is quite different: it uses a dynamic structure, that can be changed at any moment. The fluid nature of the knowledge structure of Freebase is related to the fact that it is editable by all of its users. By mapping the domains approved by Freebase's staff (called "Commons"), we tried to avoid the problem of rapidly changing structure. These domains in Freebase, as opposed to some other elements of the knowledge structure, generally have been quite stable, and are approved by Freebase's staff. This can be seen in the dates of addition of domains, that have mainly been added between 2006 and 2009, with the latest additions being in 2010.

To conclude, as the criteria to compare KOS showed, there are some fundamental differences in the structure and principles of Freebase and the DDC. These differences mean that it is not possible to do a one-to-one mapping of Freebase's structure in general. However, by focussing on the domains, it was possible to create a mapping of a

substantial number of Freebase domains, and to circumvent some of the issues that are caused by the differences in structure of approaches of Freebase and the DDC.

5.2.3 Methodology improvements

Some results of the mapping created in this thesis might be influenced by the mapping methodology. To improve the mapping based on statistical methods, a few approaches can be taken. Among other things, it would be possible to:

- Lower the threshold of items that are included in the mapping, to improve the precision of the mapping. A Freebase domain can be represented by multiple classes in the DDC, and if the threshold is lowered, possibly more classes are assigned to one Freebase domain. To achieve this lower threshold in a feasible way, this could be done using the same basic methods, but in an automated way, instead of entering the values manually in an Excel sheet.
- Improve the precision of method 1 and 2, by including also narrower classes than the section ("hundreds") level. This is something which could not be done in this thesis, because of (possible copyright-related) restrictions in the Dewey-Browser interface, that only shows results up to the section level.
- Apply another mapping step, in which an intellectual mapping is carried out, in order to refine the results of the mapping, and to add unmapped domains to the mapping. This could work up to a certain degree; however, some Freebase domains might still be hard to map, for example because a lack of clear domain descriptions.
- Take the Freebase domain descriptions and domain categories into account for improving the reliability and coverage of the mapping. The statistical methods could possibly retrieve more matching DDC classes if more elaborate information about a Freebase domain would be used. However, an issue with Freebase domain descriptions and categories is that they are not always very accurately specified, and therefore might not help in retrieving more matching DDC classes.
- An element of the DDC that was not taken into account in the mapping in this study are the "tables", which allow one to do notational synthesis. It would also be possible to see if these tables can be used to improve mappings from Freebase to the DDC[98].

Some of the methods to improve the mapping methodology could be carried out if more (and unrestricted) data is available. Finally, it would be possible to do a qualitative evaluation of the mapping that was created in this thesis, something which could not be done in the scope of this thesis.

[98] A problem with the notations that are derived from the tables is that they are not listed in the schedules, but are created on the basis of the particular needs of libraries (using the instructions provided in the DDC's class notes and tables).

5.3 Summary

This chapter has shown the results of applying a basic statistical methodology for mapping Freebase domains, in order to show the feasibility of doing a mapping in this style, and to show to what extent Freebase domains can be mapped to DDC numbers. The chapter shows that a general mapping can be done using these methods (covering a major part of all Freebase domains), but that there would also be room for improvement of these methods, for example by gaining access to the full WorldCat database and hierarchical structure of Dewey.

6 Conclusions

Recent developments have added new dimensions to the traditional approaches of knowledge organization. The systematic organization of information is not just the domain of higher-level experts anymore, but also the users of knowledge organization systems themselves can influence this organization. New and hybrid knowledge structures have emerged, like Freebase, an online semantic encyclopedia, in which users play a central role in creating both structure and content, that can be contrasted with traditional classification systems like the Dewey Decimal Classification, which involve editorial decisions by the creators of these systems.

Combining these novel and traditional systems could yield new possibilities, for example by using the rich knowledge base of Freebase to provide context to the DDC, or to augment the contents of Freebase with items classified using the Dewey Decimal Classification.

The aims of this thesis were to establish a theoretical grounding in terms of classification systems and community-driven ontologies, to carry out a mapping, and to analyze and visualize it, all in order to create a better understanding of these different types of knowledge systems. In order to do so, the study answered various research questions, and the literature review provided a theoretical grounding of the issues related to mapping knowledge organization systems.

The theoretical analysis compared the DDC and Freebase in terms of structure. On this theoretical level, both systems have the same general goal: to capture the world's knowledge. However, they do this in very different ways, as there are some fundamental differences in the knowledge structure of Freebase and the DDC. The ontology-based nature of Freebase allows one to define the relations between elements in the knowledge structure very precisely. The DDC, in its turn, allows one to define the hierarchical structure explicitly, while other relations are included more implicitly (located in for example class notes and see-also references). This means that some depth of Freebase's structure will be lost in translation.

The practical analysis of this study focused on the possibility of mapping Freebase domains to classes of the Dewey Decimal Classification. On this practical level, there were some issues with the definition of domains in Freebase, both on the level of their descriptions and on the level of the concepts they represent. They sometimes overlap each other (for example some of the sports-related domains), and have been divided based on different dimensions. The fundamental ideas of mutually exclusive and totally exhaustive classes that are (theoretically) applied in classifications, have not been used in Freebase. Furthermore, the collection of Freebase domains seemed to contain more narrow entries for “popular” categories (for instance sports and entertainment), as compared to less popular domains. This may be related to the fact that user-created

"bases", that organize information in the same way as domains, have regularly been upgraded to become an official Freebase domain, and these bases generally contain subjects that are more popular to users of the system.

In the end, the literature review, the theoretical and practical analysis have given insights that can help to answer the main research question, that looks at the mapping the structure of semantic collaborative knowledge bases in general to traditional classification systems. There are many issues that make it difficult to map classification schemes in general. As Koch et al. (2003) put it, the structures, details, vocabularies, languages and cultural contexts of classifications can vary, making it hard to map classes between systems. Mapping between a semantic collaborative knowledge base and a classification system is even harder, because of the difference in their structure. In ontologies, that form the basis of semantic knowledge bases, it is possible to define relationships between data elements very precisely, while classification systems are less flexible in this regard (they have a lower level of complexity, as indicated by Weller (2010)). Still, it is possible to look at the structure of a semantic collaborative knowledge base, and use the stable elements of this structure to make connections (mappings) between this knowledge base and a classification system, especially if the aim of the knowledge base is to capture the same knowledge domain as the classification system.

As this thesis has shown, instead of focusing on the differences, it is possible to focus on the similarities of different types of knowledge structures. It might not be possible to incorporate all features of an ontology-based structure in a mapping to a classification system, but there are elements that *can* be mapped. By focusing on the similarities, new insights into the structures of semantic community knowledge bases can be obtained, that can help in their development. Furthermore, also classification schemes could gain from this exchange: performing a mapping creates pointers to the elements that might need improvement, and can help classification systems to stay relevant. The long tradition of classification systems, paired with the new insights that user-created knowledge organization systems provide, could combine the best of both worlds.

6.1 Discussion

6.1.1 Implications

This study has shown that the division between a classification system like the DDC and a knowledge base like Freebase is not as clear-cut as it might seem at first glance. Major differences between the DDC and Freebase are related to the role of the user in the process of the creation of the KOS, to modeling guidelines, to complexity of the structure and to legacy aspects of the systems. The role of the user in Freebase reflects the current Web 2.0 approach, in which the power of influencing knowledge systems

has shifted from traditional experts to ordinary users. The users have a major influence on Freebase, as they define its structure, something which in the case of the DDC is mainly done in editorial committees (that the users of the DDC only have an indirect influence on). However, even though Freebase has been devised as a knowledge structure that is fully created by its users, its development over the years has resulted in semi-formal modeling guidelines available to users, moderation done by expert users, and a set of "Commons" domains that cannot be changed by users, but only by moderators. So Freebase is not as freely defined as for instance folksonomies. This implies that there it is not possible at the moment to create ontologies that are totally free and unmoderated.

The results of this thesis point to changes that can be made in the knowledge structures of the featured knowledge systems, that would result in better possibilities for mappings. First of all, it would be valuable to make implicit relationships in the DDC explicit, in order to find corresponding relations in other KOS, to be able to connect them, and to make the meaning of classes in the Dewey Decimal Classification more clear. Also, in terms of openness, some improvements can be done. The Linked Data of the DDC that is released via Dewey.info includes the captions and DDC numbers, but not the class notes, that are essential in understanding the meaning of classes. In order to properly use the DDC for new purposes, and to include it in the Linked Data "cloud", it is necessary to release more data under an open license.

The data of Freebase is already of an open nature, and freely available for download in many different formats. The results from this thesis indicate that improvements could be done on the level of the structure of Freebase, related to the definition of domains, and the associated documentation. Furthermore, it would be valuable if the full platform and knowledge structure of Freebase would be based on open, standardized formats, though recent developments also have showed that it is possible to convert Freebase's contents into standardized semantic Web formats like RDF.

To sum up, this thesis implies that there are opportunities to improve the structure and availability of data of both Freebase and the DDC. The main recommendation for knowledge organization systems in general, as derived from this thesis, would be to provide more extensive documentation about their structure and to release data in fully open formats, that are available for download (Linked Open Data). This will open up new possibilities to connect and reuse the data, and to develop new applications that use the data of both formal and informal knowledge organization systems in new ways.

6.1.2 Suggestions for further Research

This study has shown that a basic mapping can be created between Freebase and the Dewey Decimal Classification using the chosen approach to a statistical mapping. Fu-

ture research could use this technique in more advanced ways, also described in the previous chapter. For example, by lowering the threshold of items to be included in *1:n* mappings, the precision of the mapping could be improved. Having access to more data (for instance by using the full, unrestricted WorldCat database), could also result in more precise mappings. Furthermore, to improve on the mapping that has been done on a statistical level, an intellectual mapping could be done. At the level of the structure of Freebase, one could look at mapping other elements of its knowledge organization structure to the DDC, for example Freebase's types, topics and properties. Furthermore, it could be valuable to create a mapping of the Freebase "bases", user-created domains, to the Dewey Decimal Classification, since these bases can be freely defined by their users. It would be possible to compare the feasibility of performing such a mapping with the mapping of Freebase domains to DDC classes that was created in this thesis.

In short, the findings of this study indicate that there might be ways to connect socially and formally created knowledge organization systems. The used statistical method to create a mapping between Freebase and the Dewey Decimal Classification, by using WorldCat data in different ways, might be used on a broader scale in the future, to perform similar mappings, making it possible to map Freebase, but also other formally and informally created knowledge organization systems to the Dewey Decimal Classification.

7 Bibliography

- Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. *The Semantic Web*, 722–735.
- Bellahsene, Z., Bonifati, A., Duchateau, F., & Velegrakis, Y. (2011). On evaluating schema matching and mapping. *Schema Matching and Mapping*, 253–291.
- Bollacker, K., Cook, R., & Tufts, P. (2007). A platform for scalable, collaborative, structured information integration. *Intl. Workshop on Information Integration on the Web (IIWeb'07)*.
- Bollacker, K., Evans, C., Paritosh, P., Sturge, T., & Taylor, J. (2008). Freebase: a collaboratively created graph database for structuring human knowledge. *Proceedings of the 2008 ACM SIGMOD international conference on Management of data* (pp. 1247–1250).
- Boulos, M. N. K. (2009). Semantic Wikis: A comprehensible introduction with examples from the health sciences. *Journal of Emerging Technologies in Web Intelligence*, *1*(1), 94–96.
- Breslin, J., Passant, A., & Decker, S. (2009). *The Social Semantic Web*. Heidelberg, Germany: Springer.
- Buckland, M. K. (1991). Information as thing. *Journal of the American Society for information science*, *42*(5), 351–360.
- Buffa, M., Gandon, F., Ereteo, G., Sander, P., & Faron, C. (2008). SweetWiki: A semantic wiki. *Web Semantics: Science, Services and Agents on the World Wide Web*, *6*(1), 84–97.
- Chan, L. M. (2007). *Cataloging and Classification - An Introduction* (Third ed.). Plymouth, UK: Scarecrow Press, Inc.
- Chowdhury, G. G., & Chowdhury, S. (2007). Subject heading lists and thesauri in information organization. *Organizing Information - from the shelf to the Web*. London, England: Facet Publishing.
- Colillas, M. G. (2012). UDC on the Internet: Theory and project in evolution for use of indexing and retrieval systems. *IFLA Journal*, *37*(4), 305–313. doi:10.1177/0340035211430195
- Deutsche Nationalbibliotek. (2011). *The Linked Data Service of the German National Library*.
- Dewey, M. (2011). *Dewey Decimal Classification and Relative Index*. (J. S. Mitchell, J. Beall, R. Green, G. Martin, & M. Panzer, Eds.) (23rd ed., Vols. 1-4, Vol. 1). Dublin, OH: OCLC.

- Dextre Clarke, S. G. (2010). In Pursuit of Interoperability: Can We Standardize Mapping Types? In F. Boteram, W. Gödert, & J. Hubrich (Eds.), *Concepts in Context* (pp. 91–109). Presented at the Cologne Conference on Interoperability and Semantics in Knowledge Organization, Cologne, Germany: Ergon Verlag.
- Dextre Clarke, S. G. (2011). Knowledge Organization System Standards. *Encyclopedia of Library and Information Sciences*. New York: Taylor and Francis.
- Efrati, A. (2012, March 15). Google Gives Search a Refresh. *Wall Street Journal*. Retrieved from http://online.wsj.com/article_email/SB10001424052702304459804577281842851136290-lMyQjAxMTAyMDEwNDExNDQyWj.html
- Falconer, S., & Storey, M. A. (2007). A cognitive support framework for ontology mapping. *The Semantic Web*, 114–127.
- Fluit, C., Sabou, M., & Van Harmelen, F. (2003). Ontology-based information visualization: towards Semantic Web applications. *Visualizing the semantic web*, 45.
- Green, R., & Panzer, M. (2009). The Ontological Character of Classes in the Dewey Decimal Classification. *Paradigms and conceptual systems in knowledge organization*, 171–179.
- Hickey, T. B., & Vizine-Goetz, D. (2001). The Role of Classification in CORC. *Journal of Library Administration*, *34*(3-4), 423–432. doi:10.1300/J111v34n03_22
- Hirsch, C., Grundy, J., & Hosking, J. (2008). Thinkbase: A visual semantic wiki. *Demo Session of the 7th International Semantic Web Conference*.
- Hodge, G. (2000). *Systems of knowledge organization for digital libraries*. The digital library federation, council on library and information resources.
- Hubrich, J. (2010). Intersystem Relations: Characteristics and Functionalities. In F. Boteram, W. Gödert, & J. Hubrich (Eds.), *Concepts in Context* (pp. 91–109). Presented at the Cologne Conference on Interoperability and Semantics in Knowledge Organization, Cologne, Germany: Ergon Verlag.
- Huurdeman, H. C. (2011). Semantic Classification Search: Enhancing the Dewey Decimal Classification using Freebase. Retrieved from http://www.timelessfuture.com/download/semantic_classification_search.pdf
- Huurdeman, H. C. (2012). Visualizing the Semantic Web 2 - InfoSpace2D: a FreeBase visualization. In W.-F. Riekert & I. Simon (Eds.), *"Information in e-motion". Proceedings BOBCATSSS 2012 – 20th International Conference on Information Science*. Bad Honnef, Germany: Bock+Herchen Verlag.
- Jacobs, J.-H., Mengel, T., & Müller, K. (2010). Insights and Outlooks: A Retrospective View on the CrissCross Project. In F. Boteram, W. Gödert, & J.

Hubrich (Eds.), (pp. 38–49). Presented at the Cologne Conference on Interoperability and Semantics in Knowledge Organization, Cologne, Germany: Ergon Verlag.

- Koch, T., Neuroth, H., & Day, M. (2001). *DDC mapping report* (p. 15).
- Koch, T., Neuroth, H., & Day, M. (2003). Renardus: Cross-browsing European subject gateways via a common classification system (DDC).
- Langridge, D. W. (1989). *Subject Analysis - Principles and Procedures*. London, England: Bowker-Saur.
- Lau, A. J. (2008). Burning Down the Shelf: Standardized Classification, Folksonomies, and Ontological Politics. *InterActions: UCLA Journal of Education and Information Studies*, *2*(1).
- Lesk, M. (2005). *Understanding Digital Libraries.pdf* (Second ed.). San Francisco, CA: Morgan Kaufmann Publishers.
- Maltby, A. (1975). *Sayers' Manual of Classification for Librarians* (5th ed.). London, England: Andre Deutsch.
- Markoff, J. (2007, September 3). Start-Up Aims for Database to Automate Web Searching. *New York Times*.
- Mattison, D. (2008, February). The Freebase Experience. *Searcher*, *16*(2).
- McCulloch, E., & Macgregor, G. (2008). Analysis of mapping types for terminology services.
- Menzel, J. (2010). Deeper understanding with Metaweb | Official Google Blog. Retrieved March 14, 2012, from http://googleblog.blogspot.com/2010/07/deeper-understanding-with-metaweb.html
- Mika, P., & Greaves, M. (2008). Editorial: Semantic Web & Web 2.0. *Journal of Web Semantics*, (6), 1–3.
- Mitchell, J. S. (2001). Relationships in the DDC. In C. A. Bean & R. Green (Eds.), *Relationships In The Organization Of Knowledge* (pp. 211–226). Elsevier.
- Mitchell, J. S., Beall, J., Matthews, W., & New, G. (1996). Dewey Decimal Classification. *Encyclopedia of Library and Information Science*.
- Mitchell, J. S., & Vizine-Goetz, D. (2009). The DDC and OCLC. *Journal of Library Administration*, *49*(6), 657–667. doi:10.1080/01930820903238867
- Morville, P., & Rosenfeld, L. (2006). *Information Architecture for the World Wide Web* (Third ed.). Sebastopol, CA: O'Reilly.
- O'Reilly, T. (2007). Freebase Will Prove Addictive - O'Reilly Radar. Retrieved March 6, 2012, from http://radar.oreilly.com/archives/2007/03/freebase-will-p-1.html
- OCLC. (2003). Dewey Summaries. OCLC.

- OpenBusiness. (2007, May 23). Wikipedia for Data – Freebase. Retrieved March 14, 2012, from http://www.openbusiness.cc/2007/05/23/wikipedia-for-data-freebase/
- Peters, I. (2009). *Folksonomies: indexing and retrieval in Web 2.0*. Berlin, Germany: De Gruyter.
- Pickard, A. J. (2007). *Research Methods in Information*. London, England: Facet Publishing.
- Ranganathan, S. R., & Gopinath, M. A. (1987). *Colon Classification* (7th Ed.). Bangalore, India: Sarada Ranganathan Endownment.
- Saeed, H., & Chaudhry, A. S. (2002). Using Dewey decimal classification scheme (DDC) for building taxonomies for knowledge organisation. *Journal of Documentation*, *58*(5), 575–583.
- Salah, A. A., Cheng, G., Suchecki, K., & Scharnhorst, A. (2011). Generating Ambiguities - Mapping Categoring Names of Wikipedia to UDC Class Numbers. *Critical Point of View - A Wikipedia Reader*. Amsterdam, the Netherlands: Institute of Network Cultures.
- Salah, A. A., Gao, C., Suchecki, K., & Scharnhorst, A. (2011). Need to categorize: A comparative look at the categories of the Universal Decimal Classification system (UDC) and Wikipedia. *Arxiv preprint arXiv:1105.5912*.
- Satija, M. P. (2007). *The Theory and Practice of the Dewey Decimal Classification System*. Oxford, England: Chandos Publishing.
- Satija, M. P., & Singh, J. (2009). Colon Classification (CC). *Encyclopedia of Library and Information Sciences*. New York: Taylor and Francis.
- Shirky, C. (2005, May 17). Shirky_Ontology is Overrated. *Ontology is Overrated: Categories, Links, and Tags*. Retrieved from http://www.shirky.com/writings/ontology_overrated.html
- Taylor, A., & Joudrey, D. (2009). *The organization of information*. Westport, Conn: Libraries Unlimited.
- Tiropanis, T., Davis, H., Millard, D., & Weal, M. (2009). Semantic Technologies for Learning and Teaching in the Web 2.0 era-A survey.
- Vizine-Goetz, D. (2001). Dewey in CORC: *Journal of Internet Cataloging*, *4*(1/2), 67–80. doi:10.1300/J141v04n01_07
- Vizine-Goetz, D. (2006). Dewey Browser. *Cataloging & Classification Quarterly*, *42*(3-4), 213–220. doi:10.1300/J104v42n03_10
- Von Ahn, L. (2006). Games with a purpose. *Computer*, *39*(6), 92–94.
- Weller, K. (2007). Folksonomies and ontologies: two new players in indexing and knowledge representation. *Applying web*, *2*, 108–115.
- Weller, K. (2010). *Knowledge Representation in the Social Semantic Web*. Berlin, Germany: De Gruyter.

- Zeng, M. L., Panzer, M., & Salaba, A. (2009). Expressing classification schemes with OWL 2 Web Ontology Language. *Proceedings of the ISKO 2010 International Conference: "Paradigms and conceptual systems in KO"* (pp. 356–362). Presented at the ISKO 2010, Rome, Italy: Ergon Verlag.
- Zhang, Y., Peng, J., Huang, D., & Li, F. (2011). Design of automatic mapping system between DDC and CLC. *Digital Libraries: For Cultural Heritage, Knowledge Dissemination, and Future Creation*, 357–366.

8 Appendix

This section shows the workflow that has been used to record and analyze the collected data in this study, using Microsoft Excel.

Excel workflow:

domain name	#	domain id	freebase category	freebase description	# types	# instances	creation date
American football	1	/american_football	Sports	American Football types pertain to the sport of Football in the style played	24	87881	22-12-2006
Amusement Parks	2	/amusement_parks	Special Interests	n/a	15	3466	15-9-2008
Architecture	3	/architecture	Special Interests	This is the domain for Architecture and the Built Environment. The key type is	35	189361	22-12-2006
Astronomy	4	/astronomy	Science & Technology	The Astronomy domain holds information on all celestial objects and phenomena,	64	92068	5-3-2007
Automotive	5	/automotive	Transportation, Products & Se	Welcome to the Automotive domain. Here you'll find anything and everything about	30	83544	11-5-2007
Aviation	6	/aviation	Transportation	The aviation domain currently has a lot of schemas, but very few instances. Very	27	37854	22-12-2006

Figure 18: Excel: basic information from Freebase (excerpt)

Basic information has been extracted from Freebase, that can be used to determine its basic categories, contents and size (figure 18).

The data collection has been carried out using the steps described in 3.2.3. This involves the searching for Freebase “domain” names in WorldCat titles and subject headings. The image on the next page is an excerpt from the data collection sheet, showing WorldCat results on a class, division and subdivision level for both titles and subject headings, and refining the results using text searches in the DDC's Relative Index, class notes and captions. Some automatic data validation is done using Excel (see figure 19).

data analysis part 1: WorldCat [ti:({classname})]										data analysis part 2: WorldCat [su:({classname})]										match?	data analysis part 3: caption search [dewey.info]		
Dewey #	class results	total	%	div results	total	%	subdiv results	total	%	Dewey #	class results	total	%	div results	total	%	subdiv results	total	%		prob	Dewey #	comments	Dewey caption
796	51	57	**0,895**	51	51	1	50	51	0,98	796	160	198	**0,808**	160	160	1	150	160	0,938	1	**0,8514**	796.332	match	American football
791	19	26	**0,731**	18	19	0,947	13	18	0,722	791	104	215	0,484	93	104	0,8942	80	93	0,86	1	**0,6072**	791.068	match	Amusement parks
72	6949	9020	**0,77**	5875	6949	0,845				72	14677	18787	**0,781**	12203	14677	0,8314				1	**0,7758**	72	match	Architecture
52	1350	1429	**0,945**	1286	1350	0,953				52	3435	3754	**0,915**	3139	3435	0,9138				1	**0,9299**	52	match	Astronomy
629	815	912	**0,894**	796	815	0,977	778	796	0,977	388	208	278	**0,748**	128	208	0,6154	106	128	0,828	0	**0,8209**	629.2	also 388.3	Motor land vehicles
629	456	1261	0,362	419	456	0,919	402	419	0,959	629	96	176	0,55	45	96	0,47	42	45	0,93	1	0,5554	629.13		Aeronautics

Figure 19: Excel: data collection sheet (for the Freebase domains above)