On Combining Alignment Techniques

Anna Tordai
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

Promotiecommissie:
prof.dr. A.Th. Schreiber (promotor)
prof.dr. B.J. Wielinga (promotor)
dr. J. van Ossenbruggen (co-promotor)
dr. J. Euzenat
prof.dr. Y. Gil
prof.dr. F.A.H. van Harmelen
prof.dr. P.T.J.M. Vossen
prof.dr.ir. A. de Vries

ISBN 978-90-8659-629-4
Copyright © 2012 by Anna Tordai
Cover photo by Marjan Verkerk
Painting by Anna Tordai
On Combining Alignment Techniques

VRIJE UNIVERSITEIT

On Combining Alignment Techniques

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. L.M. Bouter,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Exacte Wetenschappen
op maandag 3 december 2012 om 15.45 uur
in de aula van de universiteit,
De Boelelaan 1105

doort
Anna Tordai

geboren te Szeged, Hongarije
promotoren: prof.dr. A.Th. Schreiber
          prof.dr. B.J. Wielinga
copromotor: dr. J.R. van Ossenbruggen
## Contents

1 Introduction ................................................................. 1
   1.1 Context: Integrating Cultural Heritage Collections .......... 1
   1.2 The Field of Ontology Alignment ................................ 2
   1.3 Research Questions ............................................... 3
   1.4 Approach .......................................................... 5
   1.5 Structure of the thesis .......................................... 6
   1.6 Publications ...................................................... 7

2 A Process Model for Vocabulary and Metadata Conversion and Alignment 9
   2.1 Introduction ....................................................... 9
   2.2 Related Work ...................................................... 11
   2.3 Approach ........................................................ 11
   2.4 Bibliopolis Data .................................................. 13
   2.5 Vocabulary Conversion ........................................... 14
   2.6 Metadata Schema Conversion .................................... 17
   2.7 Metadata Value Mapping ......................................... 20
   2.8 Vocabulary Alignment ............................................ 22
   2.9 Discussion ........................................................ 24

3 Combining Vocabulary Alignment Techniques 27
   3.1 Introduction ....................................................... 27
   3.2 Related Work ...................................................... 29
   3.3 Study Setup ....................................................... 30
   3.4 Alignment Generation ............................................ 33
   3.5 Evaluation Results ............................................... 38
   3.6 Alignment Disambiguation ....................................... 41
   3.7 Analysis of Alignment Combinations ............................. 46
   3.8 Discussion ........................................................ 48
   3.9 Acknowledgements ............................................... 50

4 Combining Vocabulary Alignment Techniques on Large Vocabularies 51
   4.1 Introduction ....................................................... 51
   4.2 Related Work ...................................................... 52
   4.3 Experimental Setup .............................................. 53
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.4</td>
<td>Predictions and Hypotheses</td>
<td>59</td>
</tr>
<tr>
<td>4.5</td>
<td>Alignment Results</td>
<td>61</td>
</tr>
<tr>
<td>4.6</td>
<td>Evaluation</td>
<td>67</td>
</tr>
<tr>
<td>4.7</td>
<td>Discussion and Conclusions</td>
<td>73</td>
</tr>
<tr>
<td>5</td>
<td>Composing Mappings from Existing Alignments</td>
<td>75</td>
</tr>
<tr>
<td>5.1</td>
<td>Introduction</td>
<td>75</td>
</tr>
<tr>
<td>5.2</td>
<td>Related Work</td>
<td>76</td>
</tr>
<tr>
<td>5.3</td>
<td>Materials and Methods</td>
<td>77</td>
</tr>
<tr>
<td>5.4</td>
<td>Results</td>
<td>83</td>
</tr>
<tr>
<td>5.5</td>
<td>Discussion</td>
<td>88</td>
</tr>
<tr>
<td>5.6</td>
<td>Conclusion</td>
<td>89</td>
</tr>
<tr>
<td>6</td>
<td>Manual Assessment of Vocabulary Alignments</td>
<td>91</td>
</tr>
<tr>
<td>6.1</td>
<td>Introduction</td>
<td>91</td>
</tr>
<tr>
<td>6.2</td>
<td>Related Work</td>
<td>92</td>
</tr>
<tr>
<td>6.3</td>
<td>Experimental Setup, Tools and Methods</td>
<td>93</td>
</tr>
<tr>
<td>6.4</td>
<td>Experimental Results</td>
<td>97</td>
</tr>
<tr>
<td>6.5</td>
<td>Discussion and Conclusion</td>
<td>104</td>
</tr>
<tr>
<td>7</td>
<td>On the Assessment of Vocabulary Alignments</td>
<td>107</td>
</tr>
<tr>
<td>7.1</td>
<td>Introduction</td>
<td>107</td>
</tr>
<tr>
<td>7.2</td>
<td>Related Work</td>
<td>108</td>
</tr>
<tr>
<td>7.3</td>
<td>Experimental Setup</td>
<td>110</td>
</tr>
<tr>
<td>7.4</td>
<td>OAEI Food track 2007 experiment</td>
<td>115</td>
</tr>
<tr>
<td>7.5</td>
<td>RKD - Cornetto experiment</td>
<td>120</td>
</tr>
<tr>
<td>7.6</td>
<td>Analysis of Mapping Types, Vocabularies and Mappings</td>
<td>124</td>
</tr>
<tr>
<td>7.7</td>
<td>Predicting Agreement in Evaluation</td>
<td>127</td>
</tr>
<tr>
<td>7.8</td>
<td>Testing the Prediction: AGROVOC-WordNet</td>
<td>128</td>
</tr>
<tr>
<td>7.9</td>
<td>Discussion</td>
<td>129</td>
</tr>
<tr>
<td>8</td>
<td>Conclusions</td>
<td>133</td>
</tr>
<tr>
<td>8.1</td>
<td>Revisiting Research Questions</td>
<td>134</td>
</tr>
<tr>
<td>8.2</td>
<td>Discussion</td>
<td>141</td>
</tr>
<tr>
<td>8.3</td>
<td>Future Work</td>
<td>146</td>
</tr>
<tr>
<td>A</td>
<td>Guidelines used for Manual Evaluation</td>
<td>149</td>
</tr>
<tr>
<td>A.1</td>
<td>RKD-Cornetto Experiment Guidelines</td>
<td>149</td>
</tr>
<tr>
<td>A.2</td>
<td>Guidelines for the Evaluation Experiments in Chapter 6</td>
<td>151</td>
</tr>
</tbody>
</table>
B Examples of Disagreements Between Raters 155
  B.1 Disagreements from Chapter 6 155
  B.2 Disagreements from Chapter 7 157

Bibliography 161

Summary 169

Samenvatting 171

SIKS Dissertation Series 175
This work would not have been possible if not for the help and support of a large number of people. First of all I would like to thank my supervisors, Guus Schreiber, Bob Wielinga and Jacco van Ossenbruggen. Guus you let me find my own way and provided guidance whenever I needed it. I’m also grateful to Bob for his persistence and precision, in particular at the end of the writing process. Thank you Jacco for finding this topic interesting despite your initial misgivings and for all your excellent advice and help.

My thanks also to Borys Omelayenko for his supervision in the early days of my Ph.D., and for entertaining me with his thoughts on life, research and everything.

During my stint as a Ph.D. student I had the wonderful opportunity to spend some time at Stanford. I would like to thank Mark Musen for giving me a temporary spot at the Stanford Center for Biomedical Informatics Research (BMIR). Thank you Natasha Noy for supervising me during my stay. It was very inspiring working with you and with Amir Ghazvinian. My stay in Palo Alto would have been a lonely one without Tania and Csongor who befriended me, let me cook for them, took me to all those lovely places in California and introduced me to geocaching. Thank you to all the colleagues at BMIR for making me feel welcome.

Most of my time however was of course spent sharing a room with several people who deserve special thanks for putting up with my incessant chatter. Rena Bakshi, Willem van Hage and Veronique Malaise, you guys are the best office mates anyone could ever wish for.

My wonderful colleagues past and present from the Web and Media and Knowledge Representation group and those from CWI who participated in the E-Culture project, it’s been fun! Thanks for letting me drag most of you to the coffee place all the way in the medical building. They do have the best coffee on campus.

I also want to thank my colleagues at Elsevier for helping me out with this final version of the thesis and giving me time to finish it. Thank you David Kuilman, Rob Schrauwen, Ton Bos, Ralph Lupton and Rob van Fucht for your patience and time.

My friends and paranimfen Laura Hollink, Mark van Assem, I now follow in your footsteps. Laura, I have known you during all my adult life my friend, and it has been a privilege to be both students and colleagues together. Mark, your positive attitude towards life despite any amount of pain is inspiring. I am happy to count you as my friend.

Lennart, Jan Jaap and Feike, thank you for all the games and fun times. I would like to thank ALL my friends for standing by me during these past years.

Meg szeretném kőszönni a családomnak hogy ilyen jól felneveltek, ilyen szuper géneket adtak és mindenben segítettek. Apu, Anyu, Gergő, Manyika és Ákos, igaz hogy pici a családunk de összetartunk még országhatárokon keresztül is, és ez a legfontosabb.

Zoli, köszi hogy rábeszéltél hogy nem baj ha nagy a távolság, és utánam jöttél annak ellenére hogy csúnya a nyelv és furák itt az emberek. De főkent annak örülök hogy egy ilyen mozgalmas év után is továbbra velem szeretnél lenni. Nélküled nagyon is magányos lenne az élet.
1 Introduction

1.1 Context: Integrating Cultural Heritage Collections

Cultural heritage institutions such as museums, archives and libraries have a long standing tradition of gathering objects of cultural and historic significance into collections. These institutions have spent considerable time and effort on carefully describing and indexing the content of these collections.

In order to support the indexing of collections, many institutions have built controlled vocabularies in various sub-domains of cultural heritage such as materials, techniques, artists and locations. Many of these vocabularies are often small and custom-built, with the purpose of annotating specific collections. There are also large domain specific vocabularies, such as the Getty Vocabularies (Peterson and Jackman-Schuller 1996)\(^1\).

In the past decades the World Wide Web has grown tremendously in size and importance, and collection owners have become increasingly interested in making their data available on the Web. There is particular interest in making this data available through portals spanning over multiple collections, such as the Europeana portal\(^2\).

There are several challenges in the creation of such a large virtual collection. The collections not only differ in the type of objects they contain (e.g., paintings, sculptures, books and photographs), but also in the annotation of these objects and the manner in which this data is structured. The disparate formats and content of the metadata describing the objects, and the vocabularies used to annotate the metadata require conversion and integration before they can be added to a virtual collection.

With the growth of the Semantic Web, mechanisms have become available for sharing and reusing data, which make the semantics of data explicit in a machine readable way. Data is represented in triple graphs which can be traversed. Programs can then derive facts from these graphs. Representation languages such as RDFS, OWL and SKOS further facilitate semantic and syntactic integration of data.

The project context of this thesis is provided by the MultimediaN E-Culture project (Schreiber et al. 2008). The aim of the project is to demonstrate how Semantic Web technology can be used to support indexing and searching in a large virtual collection of cultural heritage resources.

\(^1\)http://www.getty.edu/research/tools/vocabularies/
\(^2\)http://www.europæana.eu
In the project we use data from existing collections and their vocabularies from diverse sources such as the Rijksmuseum Amsterdam, the Royal Tropical Institute, the National Library of the Netherlands, the internet archive, Artchive, and others. The collections are either in English, Dutch or both, and cover objects ranging from paintings and books to ethnographic objects. In addition to converting collections, we also convert a number of independent vocabularies such as the Getty vocabularies, and lexical resources such as Princeton WordNet (Fellbaum 1998)\(^3\) and Cornetto (Vossen et al. 2008)\(^4\). By linking these vocabularies to collection vocabularies, we create more semantic links providing an added level of integration.

The conversion and integration of the collections and vocabularies presents multiple challenges rooted in the heterogenous nature and the scale of the data, as well as its multilingualism. The converted metadata and vocabularies allow semantic search using complex queries that would otherwise be impossible.

The Europeana Connect\(^5\) (Hennicke et al. 2011) project can be seen as a follow up project to the E-Culture project, as the aim also is to make cultural heritage collections available through a virtual portal, but on a significantly larger scale. The project provides technological solutions to Europeana which include multilingual searching, semantic enrichment of content and promoting interoperability standards. The project aims at creating a single point of entry to European cultural heritage, gathering the description of resources from museums, archives, libraries, and audio-visual archives across Europe. The challenges listed in the E-Culture project, such as multilinguuality, disparate data formats and the amount of data are in Europeana much more substantial. Europeana (europeana.eu) demonstrates the increasing interest in opening up cultural heritage data. The Europeana portal now includes over 20 million objects from more than 1,500 institutions and from 32 countries all converted to a single meta-model.

This thesis is concerned with studying mechanisms to achieve such an integration of large data sets.

### 1.2 The Field of Ontology Alignment

The field of ontology alignment, also called ontology matching, is concerned with methods and techniques for linking ontologies (Euzenat and Shvaiko 2007). An ontology is a structured vocabulary that models (parts of) the world using concepts and relationships in varying degrees of formality. Ontologies come in many forms, describing a wide variety of domains. Many different types of alignment techniques are available, including lexical comparison, graph based matching and instance based matching (Shvaiko and Euzenat 2005). With the rise of the Semantic Web, ontology alignment has become the key to making data interoperable. As a result of this interest, a plethora of alignment tools has been built using various techniques and their combinations.

\(^3\)http://wordnet.princeton.edu/
\(^4\)http://www2.let.vu.nl/oz/cltl/cornetto/
\(^5\)http://www.europeanaconnect.eu/
The goal of the Ontology Alignment Evaluation Initiative (OAEI)\(^6\) is to evaluate and compare automatic alignment systems. This is achieved by organizing a yearly evaluation platform (Caracciolo et al. 2008, Euzenat et al. 2009, 2010, 2011) where the performance of alignment systems is compared in tracks with varying tasks and vocabularies. The tracks include benchmark tests, the alignment of medical ontologies, the linking of large web directories and aligning large-scale thesauri. However, in the past years the number of participants in tracks using cultural heritage vocabularies or similar large-scale vocabularies has been low. These tasks tend to be time-consuming and cumbersome, and the vocabularies are too large for many of the systems to process. The systems that do participate tend to perform worse on these tracks than in others, such as the anatomy track.

In ontology matching the methods and techniques are closely linked to (formal) representation languages such as OWL. Because the vocabularies in the cultural heritage domain use simpler representations we use the term \emph{vocabulary alignment} instead of ontology matching.

In this thesis we borrow from terminology used in the ontology alignment field (Euzenat and Shvaiko 2007). In particular, we use the following terms:

\textbf{Mapping} \hspace{1em} A mapping is a single relation between two concepts from two different vocabularies. In ontology alignment literature a mapping is called a \emph{correspondence}.

\textbf{Ambiguous mappings} \hspace{1em} Ambiguous mappings are two or more mappings linking a single concept in one vocabulary to several concepts in another vocabulary.

\textbf{Mapping type} \hspace{1em} A mapping type is the kind of relation between two concepts. A mapping type is part of a pre-defined set of relations and can be part of a controlled vocabulary such as skos:broadMatch in SKOS matching relations. We also use the term \emph{mapping category} when we discuss the set of relations used by raters in a manual evaluation.

\textbf{Alignment} \hspace{1em} An alignment is a set of mappings between two vocabularies, the result of a matching process.

\section*{1.3 Research Questions}

In this thesis we are interested in the process of aligning vocabularies as part of the activities necessary for integrating cultural heritage collections into virtual collections. In projects such as MultimediaN E-Culture project and Europeana Connect project, multiple cultural heritage collections are linked to each other, allowing inter-collection search. Each collection has its own content and format. In order to perform integrated searches across multiple collections they need to be converted to a common format. Within the E-Culture project a number of data sets have been converted in an ad hoc fashion, but we need a framework for integrating these heterogeneous collections in an organized manner. Europeana, a project where the goal is to integrate thousands

\footnote{http://oaei.ontologymatching.org/}
of collections and make them interoperable, emphasizes the practical need for such a framework which can be applied systematically on a large scale. Therefore, the general problem for these types of projects can be stated as follows: *What is a good framework for integrating different cultural heritage collections into a virtual collection?*

Such collections generally contain metadata describing cultural heritage objects and vocabularies used to annotate them. A methodology for integrating a collection thus needs to address the conversion and linking of both the metadata and the vocabularies. In Chapter 2 we explore these steps in a case study and establish the context for vocabulary alignment, which is the main focus of this thesis. There are numerous techniques available for linking vocabularies. The comparison of such techniques, more specifically the comparison of tools that implement them is addressed by the OAEI. However, there is no methodological advice on how we need to link vocabularies from start (selecting tools/techniques) to finish (assessing the results). Therefore, our main research question is:

*How can we combine vocabulary alignment techniques, assess their performance, and evaluate alignments?*

We have split our main research question into four sub-questions:

1. **How can we combine vocabulary alignment techniques and assess their performance?** 
   
The rationale behind the first question is that vocabularies used in the cultural heritage domain tend to be too large to map manually. There is a plethora of alignment tools available using combinations of alignment techniques that are well established in literature. The manner in which a tool combines alignment techniques follows a recipe or method, where the aim for the tool is to perform well in as many scenarios as possible. As a result, tools cannot be (easily) adapted to use specific techniques that would work best on a particular data set, and it is difficult to pinpoint which technique, or combination thereof, generated a specific mapping. By implementing techniques in separate steps, we can control the process of alignment. As such, our goal is not to create a new alignment technique or tool, but to study the performance of techniques, and gain insight into how they can be combined and applied to vocabularies.

   Linking each vocabulary with every other vocabulary creates an enormous overhead, whereas linking a vocabulary to an already connected vocabulary offers a “cheaper” solution. Chaining mappings by reusing existing vocabulary alignments is a simple alternative to generating mappings from scratch. However, there has been little research on whether the quality of chains of mappings remains adequately high. Mapping composition is particularly relevant as more and more vocabularies and resources are becoming available as Linked Open Data (LOD)⁷(Bizer et al. 2009). As this data is becoming increasingly linked in the LOD more mapping compositions are formed. We need to examine the characteristics of these composed mappings.

   In both cases, alignment generation and mapping composition, the quality of alignments has to be assessed. When we combine several alignment techniques, the question is how to assess the quality of the resulting alignments.

---

⁷http://linkeddata.org/
2. How do vocabulary and mapping characteristics influence manual assessment? In our second question we consider a specific aspect of alignment evaluation: manual assessment. Manual evaluation is often used for determining the quality of alignments. It is a difficult task as raters may disagree in their assessment. The number of disagreements can be particularly high when concepts are ambiguous and vague. This vagueness can be caused by the contextual information in the vocabulary or by the high degree of ambiguity in certain mappings. Thus, these aspects and their influence on manual assessment need to be studied both quantitatively and qualitatively.

3. How can we evaluate large alignment sets based on manually assessed mappings? In practice, it is impossible to check alignment sets in their entirety, as automatic alignment techniques applied to large vocabularies invariably produce a high number of mappings. The application of statistical methods such as random sampling and stratified sampling allows us to evaluate smaller samples of mappings. Random sampling may be insufficient if the sample is not representative of the data. A large enough sample is generally sufficiently representative, however, by defining strata and performing stratified sampling we can identify sets of mappings (in these strata) with specific characteristics. By fully evaluating a smaller alignment set we can establish a way for defining strata.

4. What is the influence of manual assessment on quality measures for alignments? Finally, disagreements between raters may have an impact on the assessment of the overall quality of an alignment set. Some raters may be more strict than others, causing large variations in measures such as precision. We need to analyze the effect of such variations and determine the level of agreement necessary for acceptable stable quality measures.

Thus, our goal is to develop a methodological approach for the alignment of vocabularies, which includes manual assessment of the resulting alignment(s) by multiple raters.

1.4 Approach

First, we study the steps necessary for successfully aligning vocabularies using existing techniques, rather than comparing the performance of each individual technique in an exhaustive manner. We also focus on quality assessment and the methods for performing evaluations as successful integration of collections into a large whole hinges on the quality of the links between collections.

Our approach in Chapter 2 is to perform a detailed case study which defines the context for the application of vocabulary alignment: the integration of a collection into a large virtual collection. As such, our goal is to understand the processes involved in such an integration based on a small case. We use this case study to define the role of vocabulary alignment within the process integration of several cultural heritage collections into a single large collection. We select a small, relatively uncomplicated collection of around 1,500 objects annotated with a small vocabulary (1,000 concepts) and study the process of conversion and integration in detail. Our goal is to discuss the available techniques and methods for each step of this process and identify areas where
more work is needed.

In the rest of the thesis we perform case studies in vocabulary alignment which are driven by hypotheses. These studies involve experiments and we analyze the results not only quantitatively but also qualitatively. In each of our experiments we use real world data and are thus constrained in our choice of data sets by availability.

In Chapter 3 and Chapter 4 we perform experiments on a limited scale where we study the process of vocabulary alignment to address our first research question. We analyze our input vocabularies, generate mappings using existing tools and techniques in several iterations, and assess the quality of the mappings. In these experiments we focus on the steps needed to (successfully) align vocabularies rather than the choice of specific alignment techniques. Thus, we are not interested in the alignment of specific content, but rather in studying the process of alignment. In our experiments we select combinations of vocabularies we expect to be aligned in projects such as Europeana Connect: small collection-specific vocabularies aligned to a large vocabulary, and large domain-specific vocabularies to general vocabularies.

In Chapter 5 we study the reuse of existing vocabulary alignments through the composition of mappings. We perform experiments in different domains: medical, cultural heritage and library domain. In these experiments we perform quantitative and qualitative analyses by measuring the quality of the original alignments (input) and the composed alignments (output). We examine individual input and output mappings in detail in order to determine how the quality of the former influences the quality of the latter.

In order to assess the quality of mappings in Chapter 3 we perform a full evaluation of all generated mappings, whereas in Chapter 4 and Chapter 5 we use the method of stratified sampling to determine the quality of specific subsets of the alignments.

In Chapter 6 and Chapter 7 we perform experiments in manual assessment. In these experiments we study human behavior in a categorization task. Here, our goal is to gain insight into how humans perform evaluations of alignments and what causes disagreements between them, and therefore our experiments are exploratory rather than exhaustive. These experiments vary with regards to the type of aligned vocabularies (domain specific vs. generic), the type of mappings and raters. We compare the evaluations quantitatively, by measuring inter-rater agreement, and qualitatively by analyzing the assessments in detail and cataloguing the results.

The work done in this thesis is mainly focused on the cultural heritage domain with some notable exceptions: we use medical ontologies in our mapping composition experiment in Chapter 5 and data from the OAEI in the food domain in Chapter 7.

1.5 Structure of the thesis

This thesis is organized as follows. In Chapter 2 we present a case study where we explore the steps necessary to integrate a small cultural heritage collection and vocabulary into a larger whole. In the remainder of the thesis we zoom in on one of these steps: vocabulary alignment.
In Chapter 3 and Chapter 4 we study the process of creating alignments by combining simple techniques and the assessing of the results. In Chapter 3 we align a small collection specific vocabulary to a large lexical resource in a step-by-step manner, and perform a full evaluation and detailed analysis of the resulting mappings. Next, in Chapter 4 we perform similar experiments on a larger scale and explore the effect of vocabulary characteristics on the performance of alignment techniques. In Chapter 5 we study the quality of mapping compositions in three different domains. In Chapter 6 and Chapter 7 we examine the process of manual assessment and perform evaluation experiments with human raters where we study their level of agreement. More specifically, in Chapter 6 we study what causes disagreement in raters. In Chapter 7 we perform further experiments in evaluation, focusing on vocabulary and mapping characteristics that affect agreement between raters. We also look into how disagreements affect quality measures such as precision.

In Chapter 8 we present the overall conclusions of this thesis, propose a method for aligning vocabularies, and a method for assessing alignments. We discuss the implications of this work and suggest avenues for future research.

1.6 Publications

Publications on which the chapters in this thesis are based:


- Chapter 6 was published as: Anna Tordai, Jacco van Ossenbruggen, Guus Schreiber and Bob Wielinga. Let’s Agree to Disagree: On the Evaluation of Vocabulary Alignment. In *Proceedings of the Sixth International Conference on Knowledge Capture (K-CAP 2011)*, pages 65-72, Whistler, Canada, 2011.
The alignment evaluation tool used in Chapters 3 – 7 was built by Jacco van Ossenbruggen. In Chapter 5 the composition of Bioportal mappings and their analysis was performed by Amir Ghazvinian.
A Process Model for Vocabulary and Metadata Conversion and Alignment

In our projects we are dealing with various collections of metadata describing cultural heritage objects, accompanied by vocabularies that are varied in content and form. They require some modification to enable integration into a virtual collection and facilitate inter-collection search. In this chapter we focus on the problem statement of the E-Culture and Europeana Connect projects: What is a good framework for integrating different cultural heritage collections into a virtual collection? We describe a general framework for the conversion of a cultural heritage collection metadata and vocabulary into a semantically interoperable collection. The conversion process includes four steps: (I) vocabulary conversion, (II) conversion of the collection metadata schema, (III) metadata value mapping and (IV) vocabulary alignment. We illustrate the process with a case study and describe existing methods and approaches for each step. There is ongoing work on each of the four steps of the process but in the context of this thesis we are focusing on a methodology for vocabulary alignment.

This chapter is based on the paper coauthored with Borys Omelayenko and Guus Schreiber, “Semantic Excavation of the City of Books” (Tordai et al. 2007), which was presented at the SAAKM workshop co-held with the fourth Knowledge Capture Conference (K-CAP 2007) in Whistler, Canada.

2.1 Introduction

In this chapter we present a case study where we focus on the activities necessary for converting cultural heritage data into RDF/OWL. The context of this work is the MultimediaN E-Culture project (Schreiber et al. 2006)\(^1\), a leading Semantic Web project that won the Semantic Web Challenge in 2006. The objective of this project is to create a large virtual collection of cultural heritage objects that supports semantic search. Metadata and vocabularies are represented in RDF/OWL. The project demonstrator (see the demonstrator at the project website) includes multiple vocabularies which are partially semantically aligned.

\(^1\)http://e-culture.multimedian.nl
This chapter builds on earlier conversions of metadata and vocabularies and their commonalities. There are currently 5 collections and 6 vocabularies that are part of the E-Culture demonstrator. Among them are the collections from the Royal Tropical Institute (KIT)\(^2\) in Amsterdam and the National Museum of Ethnology (RMV)\(^3\) in Leiden. The vocabularies include three from Getty\(^4\): the Art and Architecture Thesaurus (AAT), the Thesaurus of Geographical Names (TGN) and the United List of Artist Names (ULAN), as well as the Dutch Ethographic Collection Foundation (SVCN)\(^5\) thesaurus. These form “standard” vocabularies in the cultural heritage field, that is, various institutions have agreed upon, and approved their usage. “Local” vocabularies on the other hand are often created or maintained by a single institution or person.

The objective of the present work is to describe the conversion of the Bibliopolis\(^6\) collection (Latin for city of books) and its alignment to existing vocabularies performed within the E-Culture project. The goal is to convert the vocabulary and metadata such that these become an interoperable part of the virtual collection. The Bibliopolis collection consists of images related to book-printing, and range from photographs of publishing houses to illustrations of the printing process. The collection also includes a local thesaurus of keywords. It is a good example of the range of data we come across when dealing with cultural heritage collections and vocabularies.

To represent the collections the E-Culture project uses a specialization of Dublin Core (DC)\(^7\) for visual resources (all objects in the virtual collection are required to have an image as their data representation) as the guiding metadata scheme. This Dublin Core specialization is named the Visual Resources Association Core (VRA)\(^8\) scheme which follows the Dublin Core dumb-down principle (i.e. it is a proper specialization and does not contain extensions). Likewise, we model collection-specific metadata schemes as specializations of VRA.

For the representation of vocabularies the project uses the SKOS Core Schema (Miles and Bechhofer 2009)\(^9\). It is a World-Wide Web Consortium (W3C) standard designed to support vocabulary interoperability. SKOS has been widely adopted by the Semantic Web community.

This chapter is organized as follows. We discuss related work in Section 2.2. We present our approach in Section 2.3 followed by a short presentation of the Bibliopolis data in Section 2.4. Next, we devote four sections to describe the case study based on the following four activities: vocabulary conversion, metadata schema conversion, metadata value mapping and vocabulary alignment. Finally, we conclude this chapter with a discussion in Section 2.9.

\(^2\)http://www.kit.nl/
\(^3\)http://www.rmv.nl/
\(^4\)http://www.getty.edu/research/conducting_research/vocabularies/
\(^5\)Acronym for Stichting Volkenkundige Collectie Nederland http://www.svcn.nl/thesaurus.asp
\(^6\)http://www.bibliopolis.nl/
\(^7\)http://dublincore.org
\(^8\)http://www.vraweb.org
\(^9\)http://www.w3.org/2009/08/skos-reference/skos.html
2.2 Related Work

In the area of vocabulary conversion Miles et al. (2004) propose guidelines for migrating vocabularies to the Semantic Web using the SKOS Core schema. They distinguish between standard and non-standard vocabularies, and propose to preserve all information in the vocabulary by using sub-class and sub-property statements where necessary. The work of van Assem et al. (2006a) is based on these guidelines, and they propose a three step method consisting of the analysis of the vocabulary, mapping to the SKOS schema and the creation of the conversion program. Their case studies do show however, that non-standard vocabularies are more difficult to convert completely, as some features cannot be mapped to the SKOS schema.

The problem of interoperability between two collections has been discussed by Butler et al. (2004). Within the SIMILE project they report on the conversion and enrichment of two data sets (a visual images data set and a data set used to support learning) using XSLT. The first data set was converted using the VRA schema and the second using Dublin Core, although non-standard properties were created as extensions. The steps in the conversion process included the creation of URIs, name normalization, re-creating term hierarchies, and data cleanup. The enrichment process involved identifying resources and enriching the data by looking up terms in external resources.

Hyvönen et al. (2005) describe the MuseumFinland project encompassing multiple collections and ontologies. The collections of various Finnish museums and additional ontologies were converted into RDF/OWL. The metadata of the collections was transformed using a common term ontology, while the additional ontologies form added semantic links between the collections and were further enhanced by manual editing and enrichment.

In the area of ontology alignment much work has been done on the development of alignment techniques and tools (Shvaiko and Euzenat 2005, Giunchiglia et al. 2005). In the past years, the Ontology Alignment Evaluation Initiative (OAEI) has become an integral part of the ontology alignment community. The OAEI is an evaluation platform where the performance of alignment systems can be compared on various types of data, from simple benchmark tests to medical ontologies (Euzenat et al. 2007b). However full integration of collections requires more than just alignment tools and techniques. In the subsequent sections, we will elaborate on a more general approach.

2.3 Approach

In order to integrate multiple collections within the context of the E-Culture project we need to achieve syntactic and semantic integration of data. Accordingly, we follow a practical bottom-up approach where we enrich real-world data with a layer of semantics to achieve interoperability. This approach may be seen as an alternative to the top-down approach that is common in the Semantic Web community. With the top-down approach we would first need to develop a conceptual model of the cultural heritage world in order to be able to perform semantic enrichment of the data. This ontology development effort has not been started yet and such efforts would take several years
to be finished. However, there are a number of vocabularies available at the moment which are widely used by the cultural communities. In our approach we perform syntactic integration and take the first step towards semantic integration by performing terminological integration. The task of integrating collections and vocabularies from both a structural and terminological perspective has evolved into four activities which are summarized in Fig. 2.1:

**Figure 2.1** The four activities for converting a collection.

1. Vocabulary conversion, including vocabulary schema mapping. This step is a relatively well-researched area, e.g. (van Assem et al. 2006a), with SKOS being the default option for the vocabulary schema.

2. Metadata schema conversion. Here we use generic schemas like Dublin Core and its specializations to the cultural domain, such as VRA.

3. Metadata value mapping. In this step we enrich the data value which is generally in the form of a string. The enrichment involves replacing these string values by resources where possible. We create URIs for each resource, which is then either part of the local vocabulary, or linked to external vocabularies using information extraction techniques.

4. Vocabulary alignment. Here we align the vocabulary to external (standard) vocabularies with ontology alignment techniques.

Structural integration is performed during vocabulary schema conversion for vocabularies, and metadata schema conversion for collections. The terminological integration performed during metadata value mapping and vocabulary alignment is dependent on the schema conversion activities, which we denote with vertical arrows. As vocabularies tend to be used in collection metadata, making this link explicit is part of the semantic enrichment process. Collection metadata, in turn, may contain implicit vocabularies hidden in data values that are candidates for vocabulary alignment.
2.4 Bibliopolis Data

The Bibliopolis data from the Koninklijke Bibliotheek (KB), the National Library of the Netherlands, consists of two XML files: collection and thesaurus. The collection file contains the metadata of 1,645 images related to the printing of books and book illustrations. The thesaurus contains 1,033 terms used as keywords for indexing images. These two files drive the Bibliopolis website. Both the thesaurus and the metadata are bilingual (English and Dutch). Note that in both files \textit{inm} is the namespace for Bibliopolis. We omit its use in the body text.

\textbf{Thesaurus} The thesaurus contains terms accompanied by their synonyms in both plural and singular, along with a descriptive note. Each record may also contain related, broader and narrower terms. Additionally, a record contains some administrative data: initials of the record creator, the date of entry, and the date of modification. A sample XML element for the term \textit{UNIVERSITY PRINTER} is shown in Fig. 2.2.

\begin{verbatim}
<inm:Record>
    <inm:NUM>2</inm:NUM>
    <inm:TWOND>academiedrukkers</inm:TWOND>
    <inm:TWVAR>academiedrukker</inm:TWVAR>
    <inm:TWVAR>universiteitsdrukker</inm:TWVAR>
    <inm:DEF>aan een universiteit verbonden...</inm:DEF>
    <inm:TWRT>academische geschriften</inm:TWRT>
    <inm:TWRT>overheidsdrukkers</inm:TWRT>
    <inm:ENG>university printer</inm:ENG>
    <inm:INVOERDER>emo</inm:INVOERDER>
    <inm:INVDAT>12/13/01</inm:INVDAT>
    <inm:TWSYN>universiteitsdrukkers</inm:TWSYN>
    <inm:TWBT>drukkers</inm:TWBT>
    <inm:TWNT/>
    <inm:TWOND_EN>university printers</inm:TWOND_EN>
    <inm:TWVAR_EN>university printer</inm:TWVAR_EN>
    <inm:TWVAR_EN>academy printer</inm:TWVAR_EN>
    <inm:TWVAR_EN>academic printer</inm:TWVAR_EN>
    <inm:DEF_EN>a printer appointed by...</inm:DEF_EN>
    <inm:TWSYN_EN>academy printers</inm:TWSYN_EN>
    <inm:TWSYN_EN>academic printers</inm:TWSYN_EN>
</inm:Record>
\end{verbatim}

\textbf{Figure 2.2} Thesaurus record for term UNIVERSITY PRINTER.

\textbf{Metadata} The metadata forms the description of images related to book printing. The data consists of titles and descriptions of the objects, names of their creator(s) with signatures of their roles, such as \textit{a} for \textit{author}. The works are also classified according to the technique used, their type, and a library classification of the subject matter. The metadata includes copyright information, measurements and other administrative information. An example collection object is shown in Fig. 2.3 and its corresponding metadata is shown in Fig. 2.4.
2.5 Vocabulary Conversion

Vocabulary schema mapping and conversion is a relatively well-researched area. In our work we use the method for vocabulary conversion proposed by van Assem (van Assem et al. 2006a). As for the vocabulary schema, we use SKOS within the E-Culture project. Van Assem et. al, list the following steps. First, the vocabulary needs to be analyzed with respect to the vocabulary features
Figure 2.4 A fragment of the XML record depicting a Delft Bible (Fig. 2.3) dated 10 January 1477, originated from Delft, classified with category ‘bibles’. (Certain fields may be empty).

and constraints; this includes analysis of the digital format as well as the documentation. Second, the vocabulary fields need to be mapped to the SKOS schema. When a vocabulary contains an element that is more specific than the SKOS schema, a specialization can be introduced. The last step is the implementation step where the conversion program is constructed.

Mapping the Bibliopolis thesaurus schema is relatively straightforward as it fits the SKOS template. Table 2.1 shows the details of the mapping of the thesaurus representation of the record shown in Fig. 2.2 to SKOS. Each RECORD from the original XML file contains a concept. The creation of the URI deserves special mention. There are several options available for generating URIs, such as the use of keys (numeric identifiers), labels or their combination. In the E-Culture project we chose to make URIs readable for humans.

TWOND field is used in the metadata to refer to concepts from the thesaurus, and thus forms the link between the thesaurus and the metadata. The TWOND values are unique and we use them in the creation of the URI for each concept. The content of the TWOND field is also converted into the preferred label of the concept (skos:prefLabel). The TWSYN field contains synonymous terms and is converted into the alternative label (skos:altLabel). The TWVAR field contains the singular form of the terms in TWOND and TWSYN and is converted into an alternative label as well. Unfortunately, there is no direct link between the plural and singular field, although by using a stemmer or lemmatizer the link can be made explicit. While SKOS does not provide sufficient semantics to encode the plural-singular relation, SKOS-XL schema does. In this case however, we decided that the SKOS semantics are adequate. The DEF field is linked to the skos:definition property. All label and definition fields in Dutch have an English equivalent. We make the language explicit by using XML language tags. The TWBT, TWNT and TWRT

10http://www.w3.org/TR/skos-reference/skos-xl.html
<table>
<thead>
<tr>
<th>Data Item</th>
<th>Function</th>
<th>Activity</th>
<th>Source and Target Property/Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>NUM</td>
<td>Internal identifier</td>
<td>Create literal</td>
<td>source: 2 &lt;br&gt; target: vra:location.refId &quot;2&quot; .</td>
</tr>
<tr>
<td>TWOND</td>
<td>Preferred term in Dutch</td>
<td>Create URI, literal and language tag</td>
<td>source: academiedrukkers &lt;br&gt; target: bp:academiedrukkers rdf:type skos:Concept . skos:prefLabel &quot;academiedrukkers&quot;@nl .</td>
</tr>
<tr>
<td>TWSYN</td>
<td>Synonym in Dutch</td>
<td>Create literal and language tag</td>
<td>source: universiteitsdrukkers &lt;br&gt; target: skos:altLabel &quot;universiteitsdrukkers&quot;@nl .</td>
</tr>
<tr>
<td>TWVAR</td>
<td>Term in singular in Dutch</td>
<td>Create literal and language tag</td>
<td>source: academiedrukker &lt;br&gt; target: skos:altLabel &quot;academiedrukker&quot;@nl .</td>
</tr>
<tr>
<td>DEF</td>
<td>Definition in Dutch</td>
<td>Create literal and language tag</td>
<td>source: aan een universiteit verbonden... &lt;br&gt; target: skos:definition &quot;aan een universiteit verbonden...&quot;@nl .</td>
</tr>
<tr>
<td>TWBT</td>
<td>Broader term</td>
<td>Look up and add concept URI</td>
<td>source: drukkers &lt;br&gt; target: skos:broader bp:drukkers .</td>
</tr>
<tr>
<td>TWNT</td>
<td>Narrower term</td>
<td>Look up and add concept URI</td>
<td>source: narrower term &lt;br&gt; target: skos:narrower bp:narrower_term .</td>
</tr>
<tr>
<td>TWRT</td>
<td>Related term</td>
<td>Look up and add concept URI</td>
<td>source: overheidsdrukkers &lt;br&gt; target: skos:related bp:overheidsdrukkers .</td>
</tr>
<tr>
<td>TWOND_EN</td>
<td>Preferred term in English</td>
<td>Create literal and language tag</td>
<td>source: university printers &lt;br&gt; target: skos:prefLabel &quot;university printers&quot;@en .</td>
</tr>
<tr>
<td>TWSYN_EN</td>
<td>Synonym in English</td>
<td>Create literal and language tag</td>
<td>source: academy printers &lt;br&gt; target: skos:altLabel &quot;academy printers&quot;@en .</td>
</tr>
<tr>
<td>TWVAR_EN</td>
<td>Term in singular in English</td>
<td>Create literal and language tag</td>
<td>source: university printer &lt;br&gt; target: skos:altLabel &quot;university printer&quot;@en .</td>
</tr>
<tr>
<td>DEF_EN</td>
<td>Definition in English</td>
<td>Create literal and language tag</td>
<td>source: a printer appointed by... &lt;br&gt; target: skos:definition &quot;a printer appointed by...&quot;@en .</td>
</tr>
<tr>
<td>ENG</td>
<td>English translation of term</td>
<td>Not converted, duplicate information</td>
<td>source: university printer</td>
</tr>
<tr>
<td>INVOERDER</td>
<td>Entered by</td>
<td>Not converted</td>
<td>source: emo</td>
</tr>
<tr>
<td>INVDAT</td>
<td>Date of entry</td>
<td>Not converted</td>
<td>source: 12/13/01</td>
</tr>
</tbody>
</table>

Table 2.1 Mapping thesaurus data to SKOS. The table includes the data item (XML field name), the function of the field, the necessary activities for conversion and example data from Fig. 2.2 in its original and converted form
form the links between concepts and are converted into skos:broader, skos:narrower and skos:related respectively.

Two XML elements have not been converted, as they contain bookkeeping information and are not meant to be public. One XML element (see last column in Table 2.1) is a duplicate piece of information and is therefore omitted.

### 2.6 Metadata Schema Conversion

<table>
<thead>
<tr>
<th>Data Item</th>
<th>Function</th>
<th>Target Property</th>
<th>Target property is subproperty of</th>
</tr>
</thead>
<tbody>
<tr>
<td>NUMMER</td>
<td>Record Id</td>
<td>vra:Work</td>
<td>-</td>
</tr>
<tr>
<td>TITEL</td>
<td>Title in Dutch</td>
<td>vra:title with Dutch language tag</td>
<td>-</td>
</tr>
<tr>
<td>TITEL.EN</td>
<td>Title in English</td>
<td>vra:title with English language tag</td>
<td>-</td>
</tr>
<tr>
<td>MAKER</td>
<td>Creator and his marker for role</td>
<td>bp:role vra:creator</td>
<td></td>
</tr>
<tr>
<td>OBJECT</td>
<td>The type of the depicted object</td>
<td>vra:type</td>
<td>-</td>
</tr>
<tr>
<td>TECHNIEK</td>
<td>Technique used to create object</td>
<td>vra:technique</td>
<td>-</td>
</tr>
<tr>
<td>DATERING</td>
<td>Date</td>
<td>vra:date</td>
<td>-</td>
</tr>
<tr>
<td>ORIGINEEL</td>
<td>Title of the original (book)</td>
<td>bp:origineel vra:title</td>
<td></td>
</tr>
<tr>
<td></td>
<td>containing the image</td>
<td></td>
<td></td>
</tr>
<tr>
<td>REPRODUCTIE</td>
<td>Title of the reproduction (book)</td>
<td>bp:reproductie vra:title</td>
<td></td>
</tr>
<tr>
<td></td>
<td>containing the image</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLASSIFICATIE</td>
<td>Library classification of the work</td>
<td>bp:classification vra:subject</td>
<td></td>
</tr>
<tr>
<td>TWNAAM</td>
<td>Person used as subject for work</td>
<td>vra:subject. personalName</td>
<td>-</td>
</tr>
<tr>
<td>TWOND</td>
<td>Thesaurus term used as subject</td>
<td>vra:subject</td>
<td>-</td>
</tr>
<tr>
<td>TWGEO</td>
<td>Place used as subject for work</td>
<td>vra:subject. geographicPlace</td>
<td>-</td>
</tr>
<tr>
<td>OMSCHRIJVING</td>
<td>Dutch description</td>
<td>vra:description with Dutch language tag</td>
<td>-</td>
</tr>
<tr>
<td>OMSCHRIJVING.EN</td>
<td>English description</td>
<td>vra:description with English language tag</td>
<td>-</td>
</tr>
<tr>
<td>AFMETINGEN</td>
<td>Size of the work</td>
<td>vra:measurements. dimensions</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 2.2 Overview of the conversion of the metadata XML elements (Data Item) to target VRA properties and non-VRA properties that are specializations of VRA
In this step we map the original record fields (see Fig. 2.4) to a metadata schema. In the E-Culture project we use the VRA Core scheme which is a specialization of Dublin Core\textsuperscript{11} for visual resources (our target type of resources).

Before mapping to the schema we analyze the metadata; this includes examination of any additional documentation, websites, and interviews with experts. The meaning of the fields needs to be understood to find a correct correspondence within the target schema. The first impression of the meaning of a field might be misleading. For example, the TWGEO field was initially mapped to \texttt{vra:location}, i.e., the DC/VRA element indicating where the work was created. However, the documentation showed that the field actually gives information about the location depicted in the work, and not the creation place. We finally used the VRA Core v4 element \texttt{vra:subject.geographicPlace}, which gives the correct interpretation. This element is a subproperty of DC/VRA \texttt{subject}.

An important additional consideration is that certain records or fields may contain confidential or administrative information such as acquisition or bookkeeping information. For example, the amount for which an object is insured should not be publicly visible. This situation did not occur with the Bibliopolis data.

Table 2.2 shows an overview of the mapping from the XML record fields to a VRA metadata schema. Here we encounter two situations. First, in the simplest case, there is a exact semantic match between an original field and a VRA field. Second, if this is not the case, the field should be specified as a specialization of an existing VRA element. In the Bibliopolis case this occurs with the ORIGINAL\textsuperscript{12}, REPRODUCTION and CLASSIFICATION fields. The first two are specific (book) “titles”, the third one is a specific “subject” description. In Table 2.2 we see that the RDF/OWL specification contains properties in the Bibliopolis namespace (bp:). In the table we have paired the Bibliopolis properties with the VRA element of which they are the subproperty. Note that we use bp: instead of the original \texttt{inm:} namespace in the conversion.

One field requires some deeper study. The MAKER field not only contains the creator of the work, but also a character indicating the role that the person played in creating the work. As shown in the example record in Fig. 2.4 the MAKER field has the value YEMANTSZOON, MAURICIO : D, where “d” stands for “drukker”, Dutch for “printer”. To preserve the roles of the creators we specialize the VRA property \texttt{vra:creator} with the properties that correspond to the roles found in the Bibliopolis data.

Dublin Core has excellent general coverage. In all collections we tackled so far, we were able to find for each field a Dublin Core / VRA property which was either an equivalent, or could act as a superproperty of a local specialization. This characteristic makes Dublin Core a powerful tool for metadata interoperability.

\textsuperscript{11}\url{http://dublincore.org/}
\textsuperscript{12}For readability we use the English in the text, in cases where it is close to the Dutch equivalent (“original” vs. “origineel”)
<table>
<thead>
<tr>
<th>Data Item</th>
<th>Activity</th>
<th>Source and Target Property/Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>NUMMER</td>
<td>Create URI and type resource as vra:Work</td>
<td>source: 6 rdf:type vra:Work .</td>
</tr>
<tr>
<td>TITEL</td>
<td>Create literal and language tag</td>
<td>target: bp:6 vra:title &quot;Delftse Bijbel...&quot;@nl .</td>
</tr>
<tr>
<td>OMSCHRIJVING</td>
<td>Create literal and language tag</td>
<td>source: Eerste bijbel die... target: bp:6 vra:description &quot;Eerste bijbel die...&quot;@nl .</td>
</tr>
<tr>
<td>AFMETINGEN</td>
<td>Create literal</td>
<td>source: 27 x 20 cm target: bp:6 vra:measurements.dimensions &quot;27 x 20 cm&quot; .</td>
</tr>
</tbody>
</table>

**Table 2.3** Metadata value mapping activities for Biblopolis XML fields (data field) using values of the Delft Bible record (Fig. 2.4). The meaning of the fields can be found in Table 2.2. The resulting RDF/OWL is in Turtle format. Note that the URI of the record is not valid Turtle for the sake of compactness.
2.7 Metadata Value Mapping

After creating the schema we need to map the field values. We have two kinds of fields: those that contain free-text literal values, such as a description field, and those that contain values that can be turned into resources, such as the fields for keywords or geographic places. The values of the latter may come from vocabularies (local or external) although the origin of the values may not always be specified. (From discussions with the KB we found that they had used external vocabularies for entering values in the Bibliopolis metadata, however, no external identifiers were used and thus this link is not explicit.) We distinguish four kinds of the field values:

1. A term from a local vocabulary.
2. A term from a vocabulary that is implicitly present in the field values.
3. A term that may be linked to an external vocabulary.
4. A free text literal.

In this step the essential choice to be made is whether the metadata value is turned into a resource or a literal. By converting values into resources we perform value enrichment. We now describe in detail the four types of metadata value mappings. Table 2.3 illustrates the value mapping step for a single Bibliopolis record shown in Fig. 2.4.

2.7.1 Converting to a Local Vocabulary Concept

This first case is exemplified by the values of the field TWOND which represent thesaurus concepts. This relationship is explicitly present in the source data and is preserved during the metadata value conversion. We look up values of this field in the converted thesaurus and replace them by the appropriate URI.

2.7.2 Converting to an Implicit Vocabulary Concept

In this case we map field values to resources which form new vocabulary concepts implicitly present in the data. In the Bibliopolis data there were two fields whose values formed an implicit vocabulary.

In Table 2.3 we see the value D in the field CLASSIFICATIE. Further analysis revealed that these single-letter values actually represent a small vocabulary for library-type classifications of the subject. This information is not part of the XML data, and is only shown on the website of Bibliopolis. This classification vocabulary also has some broader/narrower relations. We represented this vocabulary using SKOS and mapped the field values to concepts from this vocabulary.

The RDF example in Fig. 2.5 shows the SKOS specification of a subset of such classification subjects, including the D concept. The M concept ("secondary subjects") has a hierarchical substructure.
The other implicit vocabulary present within the data is that of roles. The field `MAKER` contains the name of the creator along with its role (e.g., `YEMANTSZOO`, `MAURICIUS` : `d` where `d` stands for printer) which is one of the 14 roles. We create RDF representations of these roles as sub-properties of `vra:creator`.

### 2.7.3 Converting to a Typed Resource

We create new RDF resources from field values that are potentially part of some vocabulary. Once again, these URIs are composed of text as the records refer to the (unique) label of the concept. We also include the field name in the URI to distinguish between concepts (cf. Section 2.5). For example, for values of the field `TECHNIQUE` this results in `bp:techniek_boekdruk`, which is part of the `bp:` namespace. The reason for this is that the values of `TECHNIQUE` and `OBJECT` sometimes coincide, for example, `FOTO` is a technique as well as an object type.

We convert the values of the fields `TECHNIQUE`, `OBJECT` and `TWGEO` into resources and
Table 2.4  Mappings between the Bibliopolis vocabularies and external vocabularies. The table displays the total number of mappings generated, the number of mapped Bibliopolis concepts (source), the number of target concepts and the number of correct mappings with the percentage of correctly mapped Bibliopolis concepts in parenthesis

<table>
<thead>
<tr>
<th>Source Vocabulary</th>
<th>Target Vocabulary</th>
<th>Total Generated Mappings</th>
<th>Mapped Source Concept</th>
<th>Mapped Target Concept</th>
<th>Correct Mappings</th>
<th>Correctly Mapped Source Concept</th>
</tr>
</thead>
<tbody>
<tr>
<td>BP Thesaurus</td>
<td>AAT</td>
<td>567</td>
<td>397</td>
<td>503</td>
<td>325 (57%)</td>
<td>322 (31%)</td>
</tr>
<tr>
<td>Metadata technique</td>
<td>AAT</td>
<td>16</td>
<td>14</td>
<td>16</td>
<td>3 (18%)</td>
<td>3 (12%)</td>
</tr>
<tr>
<td>Metadata object type</td>
<td>AAT</td>
<td>17</td>
<td>15</td>
<td>17</td>
<td>15 (88%)</td>
<td>18 (83%)</td>
</tr>
<tr>
<td>Metadata subject place</td>
<td>TGN</td>
<td>20</td>
<td>19</td>
<td>20</td>
<td>19 (95%)</td>
<td>19 (68%)</td>
</tr>
</tbody>
</table>

add the field value as label. These resources can then be linked to external vocabularies such as AAT and TGN, which is discussed in Section 2.8.

Additionally, the values of MAKER and TWNAAM contain person names. We create resources out of these names with URIs in the bp: namespace, removing invalid characters and spaces. We use the ULAN schema and create local collection specific resources of ulan:person type. These names can possibly be linked to the ULAN vocabulary by collection owners.

2.7.4 Converting into a Literal

Finally, pieces of text such as titles and descriptions are converted to literals. In Bibliopolis the values of TITLE and DESCRIPTION fields were converted into literals with language tags. Both title and description are in English as well as Dutch. The values of AFMETINGEN and DATERING are also converted into literals. Additional information could be extracted from these fields, for example, the date in DATERING could be converted into a standardized form to allow temporal reasoning within the collection.

2.8 Vocabulary Alignment

By aligning the collection specific vocabularies with standard vocabularies we can increase interoperability with collections that link to these vocabularies. In addition, standard vocabularies may contain more labels and relations than concepts occurring in collection specific vocabularies, which would improve search within the collection. We aligned the Bibliopolis vocabularies with
AAT, AATNED\textsuperscript{13}, the Dutch version of the AAT, and TGN using the simple technique of label comparison.

We aligned the Bibliopolis thesaurus to AAT by matching the English skos:prefLabel to the AAT preferred terms. The result was 567 mappings between 397 Bibliopolis concepts (38\%) and 503 AAT concepts, as presented in Table 2.4. Thus, some of the Bibliopolis concepts are matched to multiple AAT concepts and vice versa.

We use the SKOS Mapping Vocabulary specification\textsuperscript{14} which was created for the purpose of linking thesauri to each other. It specifies relationships, such as skos:exactMatch, skos:broadMatch, skos:narrowMatch among others, for aligning vocabularies. For this alignment the mappings are based on the exact string matching of term labels, and so we assigned skos:exactMatch relations.

We also assessed the quality of the mappings manually and found that of the 567 mappings, 325 were correct skos:exactMatch, which is a precision of 0.57. Many Bibliopolis concepts concerning book printing techniques were linked to both the technique, and the resulting object from the technique (eg: STEEL ENGRAVING and ETCHING). In these cases only the technique is correct, although the resulting object is semantically related to the technique. There were also mappings where the concepts were not related at all, for example, LEAVES (a single thickness of paper) in a book and LEAVES as plant material.

We also aligned the concepts in the field TECHNIQUE and OBJECT, the labels of which are in Dutch, to AATNed, the Dutch version of the AAT. The TECHNIQUE contains terms describing the type of technique used for creating a Bibliopolis object. Unfortunately, out of the 16 mappings only 3 were correct because the Dutch term for technique is the same as for the object. For example, the term for AQUARELLEN (watercolor) can mean both the technique of creating watercolors and the object itself. However, in AATNed the term AQUARELLEREN (watercoloring) is used to describe the technique. Label matching in this case is insufficient for finding the correct concept in AATNed. The OBJECT field contains terms describing the type of Bibliopolis object. We generated 17 mappings, 15 of which were correct. In this case most of the Bibliopolis object type labels matched the AATNed object type labels correctly.

The field TWGEO contains geographic names which we mapped to a part of TGN which models Europe. As the labels of this field are in Dutch, some concepts were not found in TGN, such as the concept PARIJS (Paris). However, of the 28 geographical names 20 were mapped, and 19 of these were correct. Most of these terms are towns and cities in the Netherlands, the names of which are mostly unique in Europe. However some geographic names are more ambiguous, for example, had we used TGN in its entirety, PHILADELPHIA would have been mapped to 55 TGN concepts.

Label comparison is a simple technique that is easy to apply. However, it is not always effective. In the alignment of the Bibliopolis thesaurus to AAT we achieved a precision of 0.57 and mapped less than half of the thesaurus concepts. The alignment of the Bibliopolis technique

\begin{footnotes}
\item[14]http://www.w3.org/2004/02/skos/mapping/spec/
\end{footnotes}
concepts was even less successful (precision of 0.18). The issue here was that although AAT is the right type of vocabulary (cultural heritage specific), the labels were used in a different manner. Several techniques could be used to improve the quality and/or quantity of mappings, such as stemming, or targeting specific facets/subtrees in the target vocabularies.

In this case study, mappings for the technique, object type and geographical names could be created by hand as there are not many concepts to map. However, the Bibliopolis collection is quite small, and in a specific sub-area of the cultural heritage domain. We need methods and guidelines for aligning larger vocabularies where mapping manually is no longer an option.

2.9 Discussion

Interoperability is becoming one of the key issues in the open Web world. Many research programs, such as the IST program of the EU, have interoperability high on the agenda. However, real interoperability between collections is still scarce. Until now, many approaches have focused on interoperability as a problem between two collections.

In this chapter we take a different approach. We assume a multitude of collections will become part of the interoperable space; the activities we present can to a large extent be carried out by studying an individual collection. Mapping to existing other vocabularies can be performed using simple techniques, such as label matching, although the effectiveness of such techniques is somewhat limited. For vocabulary alignment, the adage “a little semantics goes a long way”\textsuperscript{15} holds. Also, one should not view this as a one-shot thing. Metadata and vocabularies change, so extensions will take place at regular intervals in time. This also means that tool support should be in place to support this process, allowing updates to be generated semi-automatically, similar to the AnnoCultor\textsuperscript{16} tool that is being currently developed within the E-Culture project.

For the E-Culture virtual collection we have now carried out this process a number of times. This chapter should be viewed as a post-hoc rationalization of this work. Our goal is to provide a set of methods and tools that allow collection owners (museums, archives) to carry out this process. Cultural heritage institutions are now often bound to closed content management systems; the “three-O” paradigm (open access, open data, open standards) is gaining support, but we have to provide the owners of collections with the necessary support facilities.

We see two potential weaknesses of this work. Firstly, our process still requires much more tool support. In particular for vocabulary alignment we need to explore how existing tools, such as the ones participating in the OAEI contest, perform on this data set. Our current work is still too much based on manual work and only uses simple syntactic tools.

Secondly, the use of Dublin Core as a “top-level ontology” for the metadata structure can also be perceived as a risk. What if the collection has metadata fields that fit with none of the DC elements? However, this was not a problem in either of the collections we converted. For the

\textsuperscript{15}quote from J. Hendler

\textsuperscript{16}http://annocultor.sourceforge.net/
moment it seems Dublin Core is indeed a key resource in information interoperability. However, it is a challenge to construct reasoners that make use of the collection-specific specializations.

The advantage of a converted and integrated collection is that it becomes searchable through common search terms. The Bibliopolis collection on the KB website forms an isolated data set, and is not cross-searchable even with other KB data sets. Because of the conversion it can now become part of virtual collections such as the E-Culture demonstrator.

In this chapter we considered the steps necessary for the conversion and integration of a collection and vocabulary into a large virtual collection. In the last step we discussed the alignment of collection vocabularies to external “standard” vocabularies. The alignment task involved linking several small domain-specific vocabularies to large domain-specific vocabularies in diverse domains such as geographic names, materials, or more generic subject description. The vocabularies thus contain various types of data with varying amounts of contextual information, such as a full hierarchy and term definition in the Bibliopolis thesaurus, and no contextual information for techniques and locations. The alignment of such vocabularies thus requires different strategies, implementing different alignment techniques. An increasing amount of work is being done in the development of alignment techniques and tools that utilise them. However, there is a lack of methodological work on the process of aligning vocabularies. In other words: what is a good method for aligning vocabularies, in particular in the cultural heritage domain? Which vocabulary and alignment technique features do we need to take into account in the alignment process, and how can we assess the results? In the remainder of this thesis our aim is to answer these questions.

**Acknowledgments**

We are grateful to our colleagues from the Multimedian E-Culture team: Alia Amin, Lora Aroyo, Victor de Boer, Lynda Hardman, Michiel Hildebrand, Antoine Isaac, Marco de Niet, Annelies van Nispen, Marie France van Orsouw, Jacco van Ossenbruggen, Annemiek Teesing, Jan Wielemaaker and Bob Wielinga. We would also like to thank Mark van Assem for his input. The project is a collaboration between the VU University Amsterdam, the Centre of Mathematics and Computer Science (CWI), the University of Amsterdam, Digital Heritage Netherlands (DEN) and the Netherlands Institute for Cultural Heritage (ICN). The MultimediaN project is funded through the BSIK programme of the Dutch government.

We are especially thankful to Marieke van Delft of the Koninklijke Bibliotheek (National Library of the Netherlands) for her cooperation in the Bibliopolis case.
Combining Vocabulary Alignment Techniques

In the field of ontology alignment many tools have been developed. These tools often use combinations of mapping techniques, and thus the “behavior” of the tool becomes difficult to predict and track. In this chapter we want to gain insight into the performance and characteristics of mapping techniques by applying them separately. Given these insights we can then investigate how techniques can be combined and address our first research subquestion: How can we combine vocabulary alignment techniques and assess their performance? We therefore perform steps comparable to those used by an off-the-shelf tool, and analyze the effects of each step. This work should be viewed within the context of conversion and integration of several cultural heritage collections and their vocabularies. As part of this integration work we align a small collection specific vocabulary to a large lexical resource which forms a typical step in the process of integration.

We first apply a lexical matching technique, as we expect the return of a high number of potential mappings between synonym-rich vocabularies used in cultural heritage. Since many of the mappings are likely to be ambiguous, linking one concept to many other concepts, we also need to perform a disambiguation step using the hierarchical structure of the vocabularies. We compare the performance of these techniques to off-the-shelf tools. We also perform an elaborate manual evaluation of all the generated mappings. In our analysis we look at the way alignments are generated by examining how different techniques overlap. This allows us to compare the performance of each technique.

This chapter is based on an improved version of the paper coauthored with Jacco van Ossenbruggen and Guus Schreiber, “Combining Vocabulary Alignment Techniques” (Tordai et al. 2009), which was presented at the fifth Knowledge Capture Conference (K-CAP 2009) in Redondo Beach, California.

3.1 Introduction

In past years there has been tremendous activity in the ontology alignment field. A large number of techniques and algorithms have been developed (Euzenat and Shvaiko 2007, Euzenat et al. 2007a).
Within the Ontology Alignment Evaluation Initiative (OAEI) (Euzenat et al. 2007b) alignment techniques are applied to benchmark data. However, despite these efforts, there is still a clear lack of methodological support for selecting an appropriate (subset of) alignment technique(s) for a given dataset. This chapter presents a case study in ontology alignment. The application context is the MultimediaN E-Culture project (Schreiber et al. 2008). This project deploys a large number of vocabularies of different cultural heritage collections. These include large vocabularies such as the Getty thesauri (Peterson and Jackman-Schuller 1996) and large lexical resources such as Princeton WordNet (Fellbaum 1998), but also smaller collection-specific vocabularies. Vocabulary alignments are a crucial element of the semantic interoperability realized by these systems.

In this chapter we investigate the application of three alignment tools to two vocabularies from the E-Culture repository. The general objective of this study is to gain insight into methodological issues related to alignment-technique selection. In particular, we are interested in the following research questions:

1. How can we measure the performance of alignment techniques within the context of vocabulary characteristics?

2. Can we show added value of the combined use of different alignment techniques? The OAEI study has shown that the performance of techniques is dependent on the application context and that no single group of techniques can be identified as being superior. Therefore, combining techniques appears to be the obvious way to go, in particular to increase recall.

3. Can we improve recall and maintain high precision by deploying techniques for disambiguating mappings? Higher recall is likely to lead to lower precision. One way of improving precision again after increasing recall is to use disambiguation. In this chapter we examine techniques for pruning the set of ambiguous candidate mappings and identify likely correct mappings.

This chapter is structured as follows: In Section 3.2 we discuss related work relevant for methodological issues in ontology alignment. Section 3.3 describes the setup of our alignment study. Section 3.4 describes the results of the combined application of alignment techniques. In Section 3.6 we look at techniques for improving precision using disambiguation techniques. In Section 3.7 we analyze the effects of combining alignment techniques. Section 3.8 discusses what we have learnt with respect to a future alignment methodology. We postulate a number of potential avenues of future research.

---

1 http://oaei.ontologymatching.org/
2 http://www.getty.edu/research/tools/vocabularies/
3 http://wordnet.princeton.edu/
3.2 Related Work

There is comparatively little work on procedures and guidelines in ontology alignment. Euzenat et al. (2007a) identify application requirements and propose a case-based method for recommending alignment techniques, whereby application dimensions are correlated with properties of alignment tools to determine the best fit. Euzenat discusses the applicability of existing alignment tools based on results of the OAEI. At the time, RiMOM (Zhang et al. 2008), closely followed by Falcon (Hu and Qu 2008) had the best fit for each application scenario. The description of these application scenarios however is very general, such as “schema integration” and “multi agent communication”.

Aleksovski et al. (2007) performed a survey of techniques for alignment problems and based the alignment cases on existing ontology alignment applications. They list several applications, among these are the STITCH Cultural Heritage browser (van Gendt et al. 2006)\(^4\), The Unified Medical Language System (UMLS) (Lindberg et al. 1993) and Internet Directories (Ichise et al. 2003). The applications and their more abstract types are categorized according to priorities regarding the quality of alignment and the complexity of representation of the ontologies. The authors suggest alignment techniques for each application type based on the techniques used in these specific applications. For example, they type the STITCH browser as a “unified view over collections” and suggest that lexical alignment techniques (used for STITCH) can be used in similar applications.

The OAEI workshops’ aim is the comparison of ontology matching tools on predefined test sets. The tools use various techniques and their combinations for performing ontology alignments. An overview of alignment techniques can be found in Euzenat and Shvaiko (2007). Tools implementing alignment techniques or combinations of techniques take part in a number of tracks. These include a “benchmark track”, an “expressive ontologies track”, and a “directories and thesauri track”. The challenges for the tools vary depending on the track. For example, one task in the expressive ontologies track is to align anatomical ontologies which are complex and use specialized vocabularies. The vocabularies of the library task in the directories and thesauri track contain less structural information and are in the Dutch language. Of the systems that participate in several OAEI tracks some perform consistently better than others. In the 2007 OAEI workshop (Euzenat et al. 2007b), Falcon stood out with a consistently good performance across most tracks. Falcon uses a combination of lexical comparison and statistic analysis with structural similarity techniques and graph partitioning. In the 2008 OAEI workshop (Caracciolo et al. 2008), where Falcon no longer took part, the top performing systems, such as RiMOM and DSSim (Nagy et al. 2008), all use combinations of techniques for generating alignments.

Thus while alignment tools are being compared within the OAEI we are given little insight into why certain tools perform better than others on certain data sets.

\(^4\)http://www.cs.vu.nl/STITCH/
3.3 Study Setup

In this section we describe our study setup for the alignment of two vocabularies. We first present the outline of the study and continue by explaining the study components in more detail.

**Figure 3.1** Process model of the experimental setup.

**Study design and data collection**  Fig. 3.1 displays the steps of our experimental setup. In **Step 1** we preprocess the data sets by converting them to the formats required by the alignment tools. In **Step 2** we apply the three alignment techniques to the vocabularies. The tools are used in parallel and independently of each other. We record the time each tool takes to perform the alignment process. In **Step 3** we perform manual evaluation of the data by classifying each mapping into one of six categories: exact, broader, narrower, related, unrelated and unsure. We explain these categories in Section 3.4. Since this is a time consuming task we record the amount of time the entire process takes. In **Step 4** we have independent raters evaluate a random sample of mappings in order to get inter-rater agreement statistics (Cohen’s Kappa). We consider the results of the manual evaluation to be a Gold standard. We can now assess the performance of each tool. We can also assess the added value of combining their results by looking at the amount of overlap between the tools. Here, our focus is on correct exact match mappings. Finally, in **Step 5** we apply two disambiguation techniques. We measure the number of correctly kept mappings (true positives) and mappings that were correctly filtered out by each of the techniques (true negatives). We also measure the number of incorrectly kept (false positives) and incorrectly removed mappings (false negatives).

**Data sets**  In the E-Culture project we have a large number of small collection-specific vocabularies and a few large general purpose vocabularies. Aligning each vocabulary to all of the others would be time consuming and inefficient. As a rule, we want to map small vocabularies to the
large ones. Small vocabularies are generally used in a specific way by collection specialists, while the large vocabularies have more widespread use and contain more synonyms and relations. For this study we use The Netherlands Institute for Art History (RKD)\(^5\) subject thesaurus as the small source vocabulary. We chose a subject thesaurus because users tend to search on the subject of artworks rather than say, on materials, therefore, linking it to a vocabulary with more synonyms creates more access points to the collection. For the target thesaurus we chose Cornetto\(^6\) (Vossen et al. 2008), which can be best understood as the Dutch version of Princeton WordNet with additional relations. Both vocabularies are in Dutch, and an extra added value of using Cornetto, in addition to its large coverage, is that it has links to WordNet.

The original thesauri were in XML format. For project purposes, the RKD thesaurus had already been converted to SKOS\(^7\) and Cornetto to the Princeton W3C schema (van Assem et al. 2006b). The source thesaurus contains 3,342 concepts with 3,342 preferred labels and 242 alternative labels and has broader, narrower and related relations. Cornetto contains 70,434 synsets and a large number of relations such as hypernym, hyponym and meronym, as well as \texttt{skos:exactMatch} links to the WordNet. Since the source thesaurus is much smaller than the target thesaurus we are likely to find one-to-many mappings. One benefit of aligning a small thesaurus to a large one, as opposed to aligning large vocabularies to each other, is that, due to the smaller number of possible mappings, the results can be evaluated manually.

**Selection of alignment techniques** We selected three alignment techniques and their implementations for generating exact match relations.

First, we use a simple exact string match technique as a baseline. This technique is easy to implement. We used a home-grown tool for performing case independent exact matching on concept labels. To improve precision, the tool ignores all concepts where multiple (ambiguous) mappings are possible.

Second, we chose more complex linguistic techniques. We use the “in-house” STITCH tool (Isaac et al. 2008) that employs lexical matching techniques such as compound splitting and lemmatization. The vocabularies we want to align are semantically rich where concepts have several labels. By applying lexical techniques such as lemmatization we expect to generate more mappings than with simple string matching.

Third, we selected an off-the-shelf alignment tool which is freely available and can be deployed on any data set. We chose Falcon-AO\(^8\) (Hu and Qu 2008) which uses the structure of vocabularies besides other techniques for finding mappings. It is also “state of the art” giving one of the best performances at the 2007 OAEI workshop.

Both the STITCH tool and Falcon require some preprocessing of the vocabularies (Step 1 in Fig. 3.1).

\(^5\)http://english.rkd.nl/
\(^6\)http://www2.let.vu.nl/oz/cornetto/index.html
\(^7\)http://www.w3.org/2004/02/skos/
\(^8\)http://iws.seu.edu.cn/projects/matching/
Quality measures for alignment  We use precision, recall and F-measure for measuring the quality of the alignment between vocabularies. These are standard measures in information retrieval (van Rijsbergen 1979) and in ontology matching (Euzenat et al. 2005). We use the following formula to define the precision $P_T$ of the technique where $E_T$ is the number of correct mappings generated by the technique, and $M_T$ is the number of mappings generated by the technique.

$$P_T = \frac{E_T}{M_T}$$ (3.1)

In order to measure recall, we pool all exact mappings generated by all techniques. In the pool we count all unique mappings, thus even if several techniques have found the same mapping, it is only counted once (distinct total). We use the following formula to measure recall $R_T$, where $E_T$ is the number of correct mappings generated by a technique and $E_{DT}$ is the total number of distinct correct mappings found by all techniques.

$$R_T = \frac{E_T}{E_{DT}}$$ (3.2)

Our recall measure is by definition incomplete as we can not take into account additional correct mappings that were not found by any of the techniques. This is a known problem in information retrieval (Voorhees 2002, Büttcher et al. 2007) and in ontology matching.

In order to form a more complete picture of the quality of the techniques, we introduce two additional measures of our own. We report on the coverage of a technique $C_T$, where $M_T$ is the number of mappings found by a technique, and $M_{DT}$ is the distinct total number of mappings found by all techniques. We define coverage in the following manner:

$$C_T = \frac{M_T}{M_{DT}}$$ (3.3)

In our experiment we align a small vocabulary (RKD subject thesaurus) to a large vocabulary (Cornetto). As we are interested in linking as many source concepts (RKD concepts) as possible, we measure the source coverage of a technique $SC_T$, where $S_T$ is the number of source concepts mapped by the technique, and $S_V$ is the total number of source concepts in the vocabulary.

$$SC_T = \frac{S_T}{S_V}$$ (3.4)

Lastly, we use the F-measure in order to be able to compare the performance of the techniques both with respect to precision and recall. We use the following formula for calculating the F-measure:

$$F = \frac{2 \times P_T \times R_T}{(P_T + R_T)}$$ (3.5)

Manual evaluation  We performed an exhaustive manual evaluation of the alignments in order to acquire a deep insight in the nature of the mappings (Step 3 in Fig. 3.1). All proposed mappings were classified using the SKOS Matching properties: skos:exactMatch, skos:broadMatch,
skos:narrowMatch, skos:relatedMatch, with unrelated or unsure as additional properties. In order to check the reliability of the evaluation, which was performed by a single person, external raters were asked to rate at least some part of the mappings (Step 4 in Fig. 3.1). By measuring inter-rater agreement between raters we can determine the level of consensus in the evaluation. We measure agreement between two raters at a time (main rater and outside rater), and use Cohen’s Kappa (Cohen 1960). The level of agreement that is necessary for an acceptable gold standard is not clear cut. In the field of content analysis Krippendorff (2007) recommends agreement levels of 0.8 or higher with 0.7 being the minimum. According to Landis and Koch (1977) a value higher than 0.61 amounts to substantial agreement. The evaluated and validated mappings form a gold standard set. We describe the evaluation and validation in more detail in Section 3.4.3.

Techniques for improving precision We want to develop techniques for improving precision by disambiguating mappings automatically, and evaluate the performance of these techniques on our gold standard (Step 5 in Fig. 3.1). We aim at reducing the number of one-to-many mappings by removing incorrect mappings using the structure of the vocabularies. An example of an ambiguous mapping is the concept QUEEN (royalty) mapped to QUEEN (royalty) and QUEEN (chess piece). We evaluate two home-brewed techniques for disambiguation described in detail in Section 3.6.

3.4 Alignment Generation

3.4.1 Preprocessing

Most tools have various preprocessing needs, including the STITCH tool and Falcon-AO. Before using the STITCH tool, Cornetto needed to be converted to SKOS, the RKD thesaurus already being in SKOS format. The wn20s:senselabels were converted to skos:altLabel and hypernym/hyponym relations to skos:broader/narrower relations. All other relations between synsets were ignored by the STITCH tool.

For Falcon-AO, both vocabularies needed to be converted into an RDF/OWL representation. SKOS labels and senselabels were converted to rdfs:label. As a result, the distinction between preferred and alternative labels was lost in the source thesaurus (RKD). Each concept became an owl:Class and broader/hyperonym relations were converted to rdfs:subClassOf property statements.

3.4.2 Alignments

We generated alignments using the three tools discussed in Section 3.3. Running the baseline tool took approximately 10 minutes, including loading time of the vocabularies. Alignments were generated using both preferred and alternative labels, with no distinctions being made between the two. In order to create mappings the label of the RKD concept and Cornetto concept must match exactly. We only create mappings based on labels that are unique within both RKD thesaurus
and Cornetto. Thus for example, if concept $C_1$ from the RKD thesaurus, and concept $C_2$ from Cornetto have the same label $L$, and no other concept from RKD or Cornetto has $L$ as label, $C_1$ and $C_2$ are mapped. If $C_1$ has an additional label $L_{alt}$ which is shared by concept $C_3$ in RKD thesaurus the mapping between $C_1$ and $C_2$ is still created, because although $L_{alt}$ is ambiguous, label $L$ which formed the basis of the mapping is not. As a result, although mappings between RKD thesaurus and Cornetto are created based on unique labels within these vocabularies, it is nevertheless possible to have several RKD concepts mapped to the same Cornetto concept, or vice versa. In practice however no ambiguous mappings were created by the baseline tool in this experiment. As both vocabularies have labels in singular form we only match non-ambiguous labels, we expect this method to generate relatively few but correct mappings.

Generating an alignment with the STITCH tool took approximately 2 minutes. The tool generates one-to-one and one-to-many mappings and aligns nouns and adjectives, but not verbs. The tool distinguishes between preferred and alternative labels, mappings based on the latter get a lower confidence rating. Cornetto contains only alternative labels, while the RKD subject thesaurus contains both preferred and alternative labels. The mappings were also separated according to the technique used, exact match or exact match combined with compound splitting and lemmatization based on the Dutch CELEX lexicon (Burnage 1990). The results are four sets of alignments: exact match of preferred label to alternative label(s) (PrefAlt), match of compound split preferred label lemma to compound split alternative label lemma(s) (PrefAltLemma), match of alternative label to alternative label(s) (AltAlt) and match of compound split alternative label lemma to compound split alternative label lemma(s) (AltAltLemma).

Obtaining results from the Falcon-AO tool took some time. The first runs with varying parameters generated no mappings. Falcon is optimized for English, and Dutch XML language tags in our vocabularies were the reason for finding no mappings. After removing the language tags we ran the Falcon tool with default parameters on the two vocabularies, and generated an alignment after approximately 20 hours of runtime.

Table 3.1 displays the mappings generated by each tool. The baseline string matching algorithm found the lowest number of mappings, 1,306 mappings for 1,306 source concepts and has a coverage of 0.30. This was expected due to the restrictive nature of the technique. The STITCH tool generated 3,766 mappings for 2,043 source concepts and has the highest coverage of the three tools (0.86). The STITCH tool generated an average of 2 mappings for each source concept, meaning a large portion of the mappings is ambiguous and possibly incorrect. More than four-fifth of the mappings were found using compound splitting and exact matching, with few mappings found using lemmatization. There were also more mappings found for preferred labels than alternative labels of concepts in the source thesaurus, due to the higher number of the first type. Falcon found 2,580 mappings for 2,473 concepts and has few one-to-many mappings. Falcon has the highest source coverage out of the three tools mapping 74% of RKD concepts, and its coverage (0.65) is lower than that of the STITCH tool. Falcon aims for higher precision by only returning mappings above a certain threshold. In this respect the STITCH tool is more indiscriminate generating all possible mappings for homonyms. In total, the three tools generated 4,375 mappings for 2,492 distinct
### Table 3.1

<table>
<thead>
<tr>
<th>Method</th>
<th>Total Mappings</th>
<th>Source concepts mapped</th>
<th>Coverage</th>
<th>Source coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>1,306</td>
<td>1,306</td>
<td>0.30</td>
<td>0.39</td>
</tr>
<tr>
<td>PrefAlt</td>
<td>3,184</td>
<td>1,901</td>
<td>0.73</td>
<td>0.57</td>
</tr>
<tr>
<td>PrefAltLemma</td>
<td>380</td>
<td>176</td>
<td>0.09</td>
<td>0.05</td>
</tr>
<tr>
<td>AltAlt</td>
<td>174</td>
<td>113</td>
<td>0.04</td>
<td>0.05</td>
</tr>
<tr>
<td>AltAltLemma</td>
<td>28</td>
<td>13</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>STITCH total</td>
<td>3,766</td>
<td>2,043</td>
<td>0.86</td>
<td>0.61</td>
</tr>
<tr>
<td>Falcon-AO</td>
<td>2,580</td>
<td>2,473</td>
<td>0.65</td>
<td>0.74</td>
</tr>
<tr>
<td>Distinct total</td>
<td>4,375</td>
<td>2,493</td>
<td>1.00</td>
<td>0.75</td>
</tr>
</tbody>
</table>

Number of mappings and number of RKD subject thesaurus concepts mapped by the alignment tools along with their coverage and source coverage. The table also includes a breakdown of the number of mappings generated by the four individual lexical techniques applied by the STITCH tool. Distinct total is the number of mappings generated by all three tools together, where each mapping is counted once even if it has been generated by several methods.

source concepts (75% of all RKD thesaurus concepts). This is a smaller number of mappings than the sum generated by the three tools (7,652), which indicates that the mappings overlap.

To investigate the added value of using multiple alignment techniques we look at their degrees of overlap. Fig. 3.2 displays all three techniques in a Venn diagram with the number of mappings in each segment. The figure shows that 1,145, approximately a quarter of the total of mappings, is found by each of the techniques. These mappings are the easiest to find, the “low hanging fruit”. The segments with no overlap show that the baseline technique adds 10 mappings, less than 1% of the total, on top of what Falcon and the STITCH tool generate. Falcon on the other hand generates 507 extra mappings, 11% of the total, while the STITCH tool 1,726 extra mappings, 40% of the total. These numbers seem to confirm the added value of combining techniques for generating mappings.

849 RKD subject thesaurus concepts were not mapped to the target thesaurus. A portion of these concepts is formed by multiple words or short sentences and tend to be at the top level of the thesaurus. Examples of these are **levensfasen van de mens** (life-phases of man) and **fysieke en/of psychische toestand** (guideterm) (physical and/or mental state (guideterm)). There were also terms that cannot be found in Cornetto such as **zangvogel** (songbird) and **scheepspортret** (ship portait). The latter is an example of a domain-specific term found in the source thesaurus, targeting the description of the content of artworks.
Figure 3.2 Venn diagram representing mappings per alignment technique and their overlaps.

Figure 3.3 A screenshot of the tool used for evaluating mappings by hand.

3.4.3 Alignment Evaluation

Manual evaluation of the alignments

In order to assess the quality of the mappings we performed a manual evaluation of each of them. Concepts with the same or synonymous labels are categorized as skos:exactMatch. We make allowances for differences in hierarchies of the vocabularies, and so it is not necessary for concepts to have the parent concepts with the same labels. When a concept is mapped to a more specific concept, the relation is categorized as skos:narrowMatch, a mapping to a more general concept is marked as skos:broadMatch. When the concepts are clearly related, such as for example the concept of CARITAS (the allegory of charity) and CHARITY, the relation is labeled as
In some cases the relationship is not clear, often due to ambiguity in the vocabularies. In such cases the mapping is categorized as unsure. Finally, mappings which are none of the above are marked as unrelated.

We created a tool for performing evaluation shown in Fig. 3.3. For each source concept it displays all the available mappings. The parent concepts are also displayed for each concept, and more information can be viewed about the concept by clicking on “detail panel”. If a concept has been used for annotating artworks, up to 5 thumbnails of artworks are displayed to help in the decision process. As the target thesaurus, Cornetto, has not been used for annotation no thumbnails can be displayed for the target concept. For each mapping, the evaluator has to select one of the 6 matching categories.

Performing the evaluation of 4,375 mappings for 2,493 RKD subject concepts took slightly longer than 26 man-hours. On average, evaluating a single mapping costs 20 seconds. Correct mappings and obvious rejects took the shortest amount of time, while mappings with other relations generally took a bit longer. In some cases the usage of a source concept was investigated by looking at artworks more closely.

Validation

In order to validate the manual evaluation of the mappings, we asked 5 raters to each evaluate mappings for 50 source concepts. The number of mappings varied between 82 and 93 as a single concept can have multiple mappings. Fixing the number of source concepts as opposed to mappings provided a more natural cutoff point, as the number mappings per source concept varies. We selected source concepts randomly from the pool of aligned concepts. The raters were provided with guidelines (Appendix A.1). It took the raters on average 19 minutes to evaluate the mappings. We then compared their ratings with our evaluation of the same mappings.

We measured Cohen’s Kappa (Cohen 1960) between the main rater and each external rater for 6 categories. The average kappa was $\kappa = 0.58$, which is interpreted as moderate agreement. Although the mappings were evaluated over 6 categories, for the purposes of this study we are mostly interested in skos:exactMatch relations. Therefore, we also measured Cohen’s Kappa for two categories: skos:exactMatch, and an aggregation of the 5 other categories. We measured an average $\kappa = 0.70$ which we found acceptable.

When looking at the disagreements between raters, especially when one rater marks a mapping as skos:exactMatch and the other as unrelated, we found two main causes. The first is human error, a mapping was categorized falsely by one or both raters. The second cause is disagreement in the interpretation of the vocabularies. In Cornetto, sometimes different meanings of the same concept have not been disambiguated, or concepts are in the wrong hierarchy. While some raters classify mappings as correct even if the meaning is ambiguous, or the concept is in the wrong hierarchy, others reject such mappings. For example, the concept SPIERING is a type of fish in the source thesaurus. In the target thesaurus the fish SPIERING is under the hierarchy for MUSCLE but has a gloss stating it is a sea-fish. In the future we plan to provide guidelines for interpreting
Table 3.2 Result of the manual evaluation. The number of mappings are displayed for each alignment method and evaluation category.

such errors in the thesaurus during evaluation.

3.5 Evaluation Results

We present the results of the evaluation per technique in Table 3.2. The rows “Baseline”, “STITCH total” and “Falcon-AO” display the result for each technique that was used. We find that overall, 56% of mappings were evaluated as exact matches and 40% as unrelated. For approximately 1% of the mappings the rater was unsure of the relationship and in 3% of the cases the mapped concepts had some semantic relationship other than exact match. The largest portion of these semantic mappings were generated by Falcon.

Table 3.3 presents the statistics on quality measures per technique. The baseline technique has the highest precision of the three techniques with 94% exact match mappings and the lowest recall at 30%. The Falcon tool scored a precision of 67% and recall of 72%, whereas the STITCH tool scores lowest precision of 53%, and the highest recall of 86%. The results of the STITCH tool are also displayed according to the lexical technique used. From these details we find that the techniques using lemmatization (PrefAltLemma and AltAltLemma) have the lowest performance both with respect to precision and recall. The average precision PrefAltLemma and AltAltLemma is 5%, which is much lower than the precisions of PrefAlt (59%) and AltAlt (44%) (AltAlt). The lower precision for alternative labels (AltAlt) supports the general view that alternative labels are less informative than preferred labels. With respect to F-measure, we find that the Falcon has the highest score and the baseline technique the lowest. The last row of Table 3.3 shows statistics for the distinct total number of mappings. At first glance the level of precision appears very low. In Fig. 3.2 we found that the mappings generated by the three techniques overlap to a large degree.
Table 3.3 The quality measures: precision, recall and F-measure for each alignment method. The last two columns provide statistics on the level of ambiguity: the number of ambiguous mappings generated by each method, and the number of mapped source concepts that thus have at least two mappings to target concepts. The distinct total is the total number of unique mappings.

<table>
<thead>
<tr>
<th>Method</th>
<th>Total Mappings</th>
<th>Exact Match</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
<th># Ambiguous Mappings</th>
<th># Source Concepts with Ambiguous Mappings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>1,306</td>
<td>1,222</td>
<td>0.94</td>
<td>0.54</td>
<td>0.64</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PrefAlt</td>
<td>3,184</td>
<td>1,893</td>
<td>0.59</td>
<td>0.85</td>
<td>0.70</td>
<td>1,958</td>
<td>675</td>
</tr>
<tr>
<td>PrefAlt-Lemma</td>
<td>380</td>
<td>19</td>
<td>0.05</td>
<td>0.01</td>
<td>0.01</td>
<td>288</td>
<td>84</td>
</tr>
<tr>
<td>AltAlt</td>
<td>174</td>
<td>76</td>
<td>0.44</td>
<td>0.03</td>
<td>0.06</td>
<td>95</td>
<td>34</td>
</tr>
<tr>
<td>AltAlt-Lemma</td>
<td>28</td>
<td>1</td>
<td>0.04</td>
<td>0.00</td>
<td>0.00</td>
<td>22</td>
<td>7</td>
</tr>
<tr>
<td>STITCH total</td>
<td>3,766</td>
<td>1,989</td>
<td>0.53</td>
<td>0.89</td>
<td>0.66</td>
<td>2,527</td>
<td>804</td>
</tr>
<tr>
<td>Falcon</td>
<td>2,580</td>
<td>1,741</td>
<td>0.67</td>
<td>0.78</td>
<td>0.72</td>
<td>213</td>
<td>106</td>
</tr>
<tr>
<td>Distinct total</td>
<td>4,375</td>
<td>2,232</td>
<td>0.51</td>
<td>1.00</td>
<td>0.68</td>
<td>2,712</td>
<td>860</td>
</tr>
</tbody>
</table>

Most of the mappings generated by the baseline technique are also generated by the STITCH tool and Falcon. In the evaluation we found that the baseline mappings have high precision. The additional mappings generated in particular by the STITCH tool have lower precision, as many of these mappings are ambiguous. Thus by aggregating all mappings, we find that the higher quality mappings are "diluted" by the addition of low quality mappings resulting on a low overall precision.

In order to determine the differences between the techniques with respect to mapping quality we examine the overlap between the techniques. By determining the quality of the overlap we can select mapping sets that fit our requirements with respect to precision, recall or F-measure. Fig. 3.4 displays a legend identifying segments in the overlap of the three alignment techniques. We present the statistics of these segments in Table 3.4.

From this table we find that the segment with the highest recall and precision, and thus the highest F-measure, is the overlap of all three techniques (G). Additional segments with mappings generated by the baseline technique (segments A, D and E) also have high precision, but the number of mappings is much smaller and therefore the recall is low.

Many of the incorrect mappings in the segments with mappings generated by the baseline
technique (A,D,E and G) are between homonymous concepts. The baseline technique does not return mappings between ambiguous concepts, and these errors occur only if the correct concept is not part of the target thesaurus. An example is CITROENTJE which is a kind of butterfly in the RKD subject thesaurus, and an alcoholic drink in Cornetto. Neither vocabulary contains both concepts.

The overlap between Falcon and the STITCH tool (F) adds a significant amount of mappings, but has a lower precision of 0.51, than the segments with mappings from the baseline technique (0.90 or more). Some of the mappings in this segment are ambiguous which contributes to the low precision.

Both segments B and C have low precision. In segment C we have mappings generated only by Falcon. An analysis of these mappings showed that the mapped concept do not share labels. Falcon uses edit distance algorithms and can also deal with concepts whose labels are composed of several terms and compound terms. This resulted in some correct mappings, such as matching DRIEKONINGENFEEST (Epiphany) to DRIEKONINGEN. However in most cases this approach generated errors. For example, Vogelkooi (birdcage) is matched to Kooivogel (caged bird) and Streekkleding (regional clothing/wear) to Strobedekking (thatch). We also find that the level of ambiguity in segment C is relatively low, as Falcon generates mostly unique mappings by internally selecting the best candidate.

Segment B, which is formed by mappings generated only by the STITCH tool, is the largest and has the lowest precision of 0.25. This is caused by the high number of ambiguous mappings. In this segment 391 RKD concepts are linked to 1,320 Cornetto concepts. Many of these links are incorrect as the linked concepts are homonyms. For example BALCONY, which is part of a building is linked to BALCONY, a type of seating in theatre. In both segment B and segment F precision could be improved by disambiguating mappings.
Table 3.4 The number of mappings and exact matches found broken down per number of techniques that generated them. The Venn diagram in Fig. 3.4 provides a legend for the segments. We display the precision, recall and F-measure for each segment. The last two columns provide statistics on the level of ambiguity: the number of ambiguous mappings generated by each method, and the number of mapped source concepts that thus have at least two mappings to target concepts.

In summary, we find that the baseline technique implementing non-ambiguous string matching generates the least number of mappings with the highest precision. The STITCH tool generates the most mappings with the lowest overall quality due to the large number of ambiguous mappings. The performance of Falcon is in the middle both with respect to the number of generated mappings and their quality. Falcon generated few ambiguous mappings because it uses a strategy that prefers non-ambiguous mappings. Many of the incorrect mappings by Falcon were returned due to erroneous substring matching. An analysis of the overlaps showed that the many of the mappings found by the STITCH tool (segment B) and mappings in the overlap between Falcon and the STITCH tool (segment F) are ambiguous. In the next section we will perform a disambiguation step in order to improve the precision of the mappings and analyze its effect.

### 3.6 Alignment Disambiguation

In the previous section we found that a large number of mappings, in particular those generated by the STITCH tool, are ambiguous. An example of an ambiguous mapping is the concept KING as royalty in the RKD thesaurus which has three mappings. The first mapping is to a playing card KING, the second is to the chess piece KING and the third is to royalty KING. The first two mappings are false positives and we need some disambiguation technique to detect them.

If we consider the total set of 4,375 mappings generated by all three tools we find that 2,712
of these are ambiguous, mapping 860 RKD concepts to 2,500 Cornetto concepts. Thus, on average one RKD concept is mapped to three Cornetto concepts, and therefore ambiguity is a serious problem. Fig. 3.6 shows the distribution of the ambiguous mappings per technique. These numbers differ from those in Table 3.4 where the number of ambiguous mappings was calculated per segment. In Fig. 3.6 we have pooled all mappings and present the overall level of ambiguity. As a result, for example mappings that were not ambiguous in the baseline segment become ambiguous because additional mappings were generated for the same concept by Falcon or the STITCH tool through other means (e.g., lemmatization and edit-distance). This is illustrated by Fig. 3.5. Fig. 3.6 shows that most of the ambiguous mappings are found in the overlap between STITCH and Falcon and in the mappings found only by STITCH.

As we have evaluated all existing mappings, we can use our gold standard evaluation to assess the performance of our disambiguation techniques. We have implemented two disambiguation techniques exploiting the structure of the vocabularies: disambiguation by counting “child” mappings (Child Match), and by counting “parent” mappings (Parent Match).

### 3.6.1 Disambiguation Techniques

In the Child Match technique (see Fig. 3.7), for each concept with multiple mappings we follow the hierarchy “down” using narrower relations, and count the number of mappings in the lower reaches between the two vocabularies. We assume that concepts which are equal in meaning will have similar hierarchies below them. This means there are more mappings between their children, than for concepts which may be lexically similar but differ in meaning. We then count the number of mappings that have at least one or more child mappings and consider them to be correct exact match. If multiple mappings for a single concept have more than one child mapping we choose the mappings with the highest number of child mappings. However, in some cases both mappings have the same (highest) number of child mappings and then both are chosen. If the mapped concepts have no mapped child mappings, then no disambiguation can be performed.
Parent Match is a mirroring of the Child Match counting technique. We want to find correct mappings by exploiting the levels of the hierarchy that are above the leaf nodes. For each ambiguous mapping we count the number of mappings that could be reached from each concept through broader relations. Alignments with at least one “parent” mapping are considered to be correct exact match. Note that neither the RKD thesaurus and nor Cornetto have a single top concept, as a result we did not have to exclude the top concepts when using this technique. Nor were the top concepts mapped, meaning we did not have to exclude any mappings when applying the Parent Match technique.

### 3.6.2 Results

Table 3.5 displays the results of the implementation of both techniques. Using the Child Match technique we were able to disambiguate 449 mappings. Of these, we kept 120 mappings which had the highest number of (at least one) child mappings linking 112 RKD concepts. Our assumption is that these 120 mappings are correct, and therefore we removed the remaining 329 mappings. We evaluated the effect of this technique using our gold standard. We found that 91 of the 120 mappings were correctly selected (true positives). We also counted the number of false negatives, or mappings that were removed but that were in fact correct. We found that less than 10%, that is 32 out of 329 removed mappings were false negatives. Examining the false negatives, we found that the main reason for excluding them was because they had no child concepts or the child concepts were organized differently. For example, the concept FACTORY in the RKD thesaurus has multiple child concepts such as STEEL FACTORY and BRICKYARD, while in Cornetto the child concept is FACTORY HALL.
Figure 3.7 Diagram of the Child Match technique. Concept A is mapped to both concept B and C. There is a mapping between D and E, the child concepts of A and C respectively. The child concepts of B are not mapped to child concepts of A. Therefore the mapping between A and B is removed and the mapping between A and C is kept. The Parent Match technique is the mirroring of the Child Match technique.
<table>
<thead>
<tr>
<th></th>
<th>Disambiguated Concepts</th>
<th>Mappings Total</th>
<th>Mappings Kept</th>
<th>Mappings Removed</th>
<th>True Positives</th>
<th>False Positives</th>
<th>False Negatives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Child Match</td>
<td>112</td>
<td>449</td>
<td>120</td>
<td>329</td>
<td>91</td>
<td>29</td>
<td>32</td>
</tr>
<tr>
<td>Parent Match</td>
<td>185</td>
<td>561</td>
<td>234</td>
<td>327</td>
<td>182</td>
<td>52</td>
<td>41</td>
</tr>
<tr>
<td>Distinct total</td>
<td>279</td>
<td>951</td>
<td>336</td>
<td>615</td>
<td>256</td>
<td>80</td>
<td>72</td>
</tr>
</tbody>
</table>

Table 3.5 Results of disambiguation techniques. Out of 2,712 ambiguous mappings 951 mappings were successfully disambiguated.

Figure 3.8 The distribution of the kept mappings in the overlapping segments of the techniques. The number in parentheses displays the number of mappings that were removed from that segment as a result of the disambiguation.

Using the Parent Match technique we were able to disambiguate 561 mappings. Of these, we kept 234 mappings which we assume to be correct, and removed 327 mappings. Again, we used the gold standard to evaluate the results of this technique and found that 182 of the 234 mappings were true positives. We also examined the mappings that were removed and found that 41 out of 327 mappings were false negatives. The reason these mappings were not returned is usually because of differences in hierarchies. For example, the concept ALMANAC has as parents BOOK and PRINTED WORK in the RKD thesaurus and EXPRESSION, DESCRIPTION and CHRONICLE in Cornetto. There was a small overlap in the mappings found by the Child and Parent Match techniques. Fig. 3.8 shows the distribution of the disambiguated mappings across the overlap of mappings. We find that the segments with the highest ambiguity, B and F are the segments
with the highest number of disambiguated mappings and where the most ambiguous mappings were removed. We also find that the proportion of kept mappings in segment F is higher than the proportion of removed mappings, whereas as in segment B the number of removed mappings (452) is more than three times as many as the proportion of kept mappings. This indicates that more mappings in segment F have a hierarchical basis than in segment B.

![Diagram of overlapping segments and mappings]

**Figure 3.9** The distribution of the remaining ambiguous mappings in the overlapping segments of the techniques. The number in parentheses displays the original number of ambiguous mappings in that segment.

Fig. 3.9 shows the effect of the disambiguation step by displaying the remaining ambiguous mappings in the overlap between techniques. The largest decrease (of 35%) in ambiguous mappings is in segment B (cf. Fig. 3.4), followed by segment F (31%). Overall, the two disambiguation techniques together reduced the number of ambiguous mappings by a third.

### 3.7 Analysis of Alignment Combinations

Previously we have found that the quality of subsets of mappings vary significantly with respect to precision and recall. We have also applied a disambiguation step whereby the number of ambiguous mappings was reduced by a third, thus boosting precision of the disambiguated segments. Based on these steps we find that each mapping segment (A to G) is composed of four types of mappings: non-ambiguous mappings, disambiguated mappings that were returned as correct using disambiguation techniques (kept mappings), disambiguated mappings that were returned as incorrect (removed mappings), and mappings where the disambiguation technique failed. We now examine how these sets and subsets of mappings can be combined.

The use of mappings depends on the use case and users can decide whether they prefer to have high precision mappings, high recall, or a high F-measure. The highest recall can be achieved
by selecting all generated mappings. A high precision or high F-measure can be achieved by selecting subsets of mappings that have the desired quality. In this section we provide an example of gradual selection and combination of subsets of mappings with the goal of improving recall without a significant drop in precision. This gradual selection process follows the following steps:

1. Select segment with the highest precision
2. Disambiguate mappings
3. Select subsegments for examination:
   - Non-ambiguous subsegment
   - Disambiguated subsegment (kept mappings)
   - Remaining non-disambiguated subsegment
   Add subsegments above a certain precision threshold
4. Repeat steps 1 to 3 until a sufficiently high level of recall or F-measure is reached

Table 3.6 displays the number of mappings in each subsegment along with the precision of the subsegment in parentheses. We can select and combine mappings based on the step by step process described above. For our example of mapping combinations we have chosen a required minimum precision of 0.6 for the selection of subsegments. Table 3.7 shows quality measures of the combined subsets of mappings in each step.

<table>
<thead>
<tr>
<th>Segment</th>
<th>Non-Ambiguous</th>
<th>Ambiguous</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Disambiguated</td>
<td>Non-disambiguated</td>
</tr>
<tr>
<td></td>
<td>Kept</td>
<td>Removed</td>
</tr>
<tr>
<td>A</td>
<td>0 (n. a.)</td>
<td>0 (n. a.)</td>
</tr>
<tr>
<td>B</td>
<td>31 (0.26)</td>
<td>136 (0.64)</td>
</tr>
<tr>
<td>C</td>
<td>379 (0.30)</td>
<td>9 (0.44)</td>
</tr>
<tr>
<td>D</td>
<td>71 (0.94)</td>
<td>1 (1.00)</td>
</tr>
<tr>
<td>E</td>
<td>15 (0.93)</td>
<td>7 (1.00)</td>
</tr>
<tr>
<td>F</td>
<td>65 (0.61)</td>
<td>169 (0.85)</td>
</tr>
<tr>
<td>G</td>
<td>1,102 (0.94)</td>
<td>14 (0.93)</td>
</tr>
</tbody>
</table>

Table 3.6 Breakdown of each segment according to mapping type based on ambiguity and on whether the ambiguous mappings were disambiguated. The table includes the size of each subsegment and precision of the mappings are in parentheses.

In the first step we select the segment D, which has the the highest precision (0.95). No mappings were removed from this segment through disambiguation and the precision of each subsegment is higher than 0.6. Next we select segment G, with the second highest level of precision. In
segment G all subsegments, including the removed mappings, have a precision above the threshold. However, following our selection process we do not include these removed mappings. By adding mappings from segment G to D we have an average precision of 0.94, a recall of 0.52 and an F-measure of 0.67. Segments A and E have the third highest precision of 0.90. We again add mappings from all subsegments except those that were removed in the disambiguation process (cf. rows 3 and 4 in Table 3.7). The combined segments D, G, A and E form the mappings generated by the baseline technique. At this stage, the precision of the mappings is 0.93, the recall is 0.55, and the F-measure 0.69.

The segment with the next highest precision is segment F (0.51) (Table 3.4). In this segment only the non-ambiguous and disambiguated mappings have a precision that is higher than the threshold. We add these mappings to our baseline mappings and reach a precision of 0.91, recall of 0.63 and F-measure of 0.74. All subsegments in segment C have a precision that is lower than the threshold and thus no mappings are added from this segment. In segment B only the disambiguated mappings that we have kept have a precision higher than 0.60. By adding these mappings to our previous set we achieve a precision of 0.89, recall of 0.67 and an F-measure of 0.76 (row 6 in Table 3.7).

The procedure described above increases the precision from the original overall 51% (Table 3.3) to 0.89, while keeping recall at a reasonable level of 0.67 and increasing the F-measure from 0.68 to 0.76. Thus a significant improvement in the quality of the mappings has been achieved by the combination procedure.

As an additional step, not part of the procedure described above, some subsegments may be fully evaluated manually in order to raise recall without affecting precision. For example, the non-disambiguated subsegment in segment F (Table 3.6) contains 460 mappings, of which 226 are correct. This subsegment was not included in our combining procedure as its precision (0.49) is too low. However, if we performed a manual evaluation of this subsegment, which based on our experiences would take 2.5 man-hours, the addition of 226 correct mappings to our previous combined set would raise precision to 0.90, recall to 0.77 and F-measure to 0.83 (row 7 in Table 3.7).

In this study we had already performed a full evaluation, but in cases where this is not possible, the decision to perform additional manual evaluation can be made based on evaluated samples.

### 3.8 Discussion

In this study we have applied and evaluated a number of typical state-of-the-art techniques for ontology alignment. We found that our baseline technique, a simple non-ambiguous exact string matching algorithm, yielded relatively few but high precision mappings. The STITCH tool generated a significant amount of additional correct mappings improving recall, but also a large set of incorrect mappings which means its level of precision was relatively low. The STITCH tool relies on the richness of lexical information, such as the number of synonyms to generate mappings. In a lexically poor environment its performance would decrease. The performance of Falcon is the
Table 3.7 Precision, recall and F-measure values for mapping combinations following the step by step procedure of including subsegments with precision higher than 0.60.

most balanced with comparable precision and recall figures, however it comes with high computational costs. Using Falcon to create an alignment between a relatively small vocabulary and a large vocabulary took 20 hours of runtime. Thus, creating alignments between larger vocabularies would likely be a problem.

The two disambiguation techniques, Child Match and Parent Match, successfully disambiguated only a third of the ambiguous mappings. The level of success in disambiguation greatly depends on the similarity of the hierarchy of the aligned vocabularies. In domains, for example medicine, where the organization of hierarchies follows a more established pattern such structure based disambiguation techniques would likely perform better.

All three tools (baseline, STITCH and Falcon) generate mappings that other tools do not, as well as mappings that the other tools generate. The more tools that generate mappings, the stronger the support for the creation of the mappings. Thus, mappings generated by all three techniques are of the highest quality, whereas mappings generated by a single tool are generally of low quality. Counting the number of techniques that generate a mapping provides the possibility of rating a mapping and requires further study.

Separating the mappings into overlapping and non-overlapping sets provides a clear view of the performance of each technique. Furthermore, it allows us to select subsets of mappings and combine them in order to increase the quality of the mappings.

In addition to using standard information retrieval measures (precision, recall and F-measure), we measured the coverage of tools with respect to each other. This measure provides a fast way of assessing the power of the alignment tools prior to evaluation. Given that we were able to perform a full evaluation of all mappings the usefulness of this measure is limited in comparison to precision and recall, as it does not take into account the number of correct mappings.

With respect to the source coverage of the three alignment-generation techniques, in total 75%
of RKD concepts were mapped. An inspection of the remaining 25% showed that 655 out of the remaining 949 (77%) concepts had a direct hierarchical (broader and narrower) or related link with a directly mapped concept in the 75% set. For the application context of this study (the alignments are used as part of a semantic network for information retrieval, the E-Culture semantic search engine) such a coverage is fine, while finding more mappings would be too time consuming.

In conclusion, we can assert that each technique performs poorly with regard to either precision, recall or both. Much improvement can be gained by a procedure as follows:

1. Apply the baseline method and accept all results.
2. Apply additional mapping techniques (lexical and structural) to generate additional mappings.
3. Select the overlapping mappings generated by the additional techniques.
4. Apply the disambiguation techniques and split the segment into accepted and not accepted mappings.
5. Apply additional full manual evaluation to non-accepted mapping segments of reasonable size.

In the next chapter we will investigate whether these results also hold for vocabularies with other characteristics.

### 3.9 Acknowledgements

The data sets have been kindly provided by RKD and the Cornetto project. We thank Antoine Isaac and the STITCH project, Borys Omelayenko and Wei Hu for helping us using their tools, and Mark van Assem, Willem van Hage, Laura Hollink and Jan Wielemaker for their contributions on the alignment evaluation. We also thank Bob Wielinga for his comments on earlier versions of this paper. This research was supported by the MultimediaN project funded through the BSIK programme of the Dutch Government.
Combining Vocabulary Alignment Techniques on Large Vocabularies

In the previous chapter we have shown that combining alignment techniques in a step-by-step manner and examining their overlap gives us a clear insight into how effective these techniques are. Whereas previously we have aligned a small vocabulary with a large one, in this chapter we look at alignments between large vocabularies and investigate whether our findings are scalable. We address the same research subquestion from the previous chapter (How can we combine vocabulary alignment techniques and assess their performance?). As the vocabularies are much larger than in the previous chapter, the number of mappings generated is too high to perform a full manual evaluation. This raises the following additional question: How can we evaluate large sets of mappings effectively? Instead of randomly sampling from a large set of mappings, we sample the overlap between mappings generated by the different alignment techniques. This gives us an insight into portions of alignments of similar quality allowing us to perform a strategic sampling and evaluation.

This chapter is based on a paper coauthored with Jacco van Ossenbruggen, Guus Schreiber and Bob Wielinga, “Aligning Large SKOS-like vocabularies: Two Case Studies” (Tordai et al. 2010a), which was presented at the seventh Extended Semantic Web Conference (ESWC 2010) in Heraklion, Greece.

4.1 Introduction

As Semantic Web technology gains prevalence the field of ontology alignment is becoming more and more important. Within the MultimediaN E-Culture project (Schreiber et al. 2008) we use a large number of vocabularies for the annotation of artwork metadata. Despite the large amount of work done on developing ontology alignment techniques, in a practical setting it is still hard to predict for two given vocabularies, which combination of techniques can best be used to create an alignment between them.

In the previous chapter we took a first step towards a methodology for selecting such a combination. We applied three alignment tools to two vocabularies, the RKD subject thesaurus and
Cornetto, from the E-Culture repository, and looked at which combination of techniques gave the best results.

In this chapter we take a second step by questioning to what extent we can use an analysis of the characteristics of the vocabularies to predict the performance of the different techniques, and to predict which combination will generate the best results. In addition, we have selected large vocabularies for our alignment experiments where a full manual evaluation is impractical. Thus, we also need a method for evaluating a large alignment set by manually evaluating samples of mappings. Consequently, we want to answer the following research questions:

1. How can we select alignment techniques based on vocabulary characteristics?
2. How can we combine vocabulary alignments?
3. How can we evaluate large sets of mappings effectively?

To answer these questions, we perform two experimental studies. In each experiment we align a large domain-specific vocabulary to a lexical resource. We analyze the vocabularies being aligned and select alignment techniques based on the vocabulary characteristics. We focus on the analysis of vocabulary characteristics that would influence the alignment techniques. We make predictions regarding the performance of these alignment techniques. We then apply alignment techniques to generate mappings, followed by a manual evaluation of representative samples to assess the performance of each technique. Finally, we discuss our findings and compare them to our initial predictions.

4.2 Related Work

Work on procedures and guidelines for ontology and vocabulary alignment is still limited. The Ontology Alignment Evaluation Initiative (OAEI\(^1\)) campaigns provide a standardized way of comparing alignment tools with tools such as Falcon (Hu and Qu 2008) and RiMOM (Zhang et al. 2008) which had the best performance in the 2007 (Euzenat et al. 2007b) and 2008 (Caracciolo et al. 2008) campaigns. Unfortunately, there are no clear selection criteria for these tools, and many of the off-the-shelf tools are either unavailable or do not work on data other than that of the OAEI campaigns.

Euzenat et al. (2007a) identified application requirements and proposed a case-based method for recommending alignment techniques but the work remains at a high level of abstraction. In a survey of alignment techniques, Aleksovski et al. (2007) listed techniques for alignment problems based on real world applications. They recommend using those techniques for similar applications. In both cases there is a lack of a systematic method of comparison and evaluation of techniques with respect to domains and vocabulary characteristics.

\(^1\)http://oaei.ontologymatching.org/
Eckert et al. (2009) used machine learning techniques for alignment generation but found that combining the results of multiple alignment tools by a system of voting works just as well as machine learning. This result suggests that machine learning techniques, while useful in other areas such as natural language processing, have little added value in the field of ontology alignment.

Ghazvinian et al. (2009a) compared the performance of AROMA (David 2008), an alignment tool that had participated in the OAEI, and a simple lexical algorithm for creating alignments between medical ontologies using the OAEI gold standard. Ghazvinian et. al., concluded that their simple lexical algorithm outperformed AROMA with respect to precision.

Our conclusion is that we need to develop our own methods for aligning vocabularies with clear criteria for selecting alignment techniques based on the data set characteristics.

4.3 Experimental Setup

In this section we start by describing the rationale behind our choice of alignment techniques and how we expect vocabulary characteristics may influence the alignment process. We continue with our experimental setup and describe each step of the experiment in detail.

4.3.1 Study Rationale

In Chapter 3 we learned that for vocabularies containing many synonyms or alternative labels, simple string matching techniques can already yield relatively good results at low computational costs. For the current study, we tested Falcon on our data set, but it ran out of memory. A test run on a relatively small subset of a source vocabulary required 15 Gb of memory and 46 hours of runtime to generate an alignment. Ghazvinian et al. (2009a) reported similar problems with Falcon, and other off-the-shelf tools. In this chapter we focus on relatively simple alignment techniques based on string matching. We extend exact string matching by normalizing labels. We also distinguish between techniques that result in non-ambiguous mappings comparable to the baseline technique in Chapter 3 by either selecting labels that are unique to a single concept, and allowing ambiguous mappings which is comparable to the STITCH tool. We expect that these variations will have a significant influence on the quality of the mappings generated. In this study we attempt to predict which alignment techniques will perform best, based on an analysis of the characteristics of the vocabularies being aligned.

There are a number of vocabulary characteristics that may be the source of potential alignment problems. In this chapter we use vocabularies that are represented in SKOS (Miles and Bechhofer 2009) or can be easily mapped to the SKOS model. The use of SKOS may introduce some bias. For example, the ISO standard for thesauri (International Organization for Standardization 1986) prescribes the use of the plural form for the preferred term of nouns, whereas in SKOS the preferred term may be plural or singular. In lexical resources, such as Princeton WordNet (Fellbaum 1998) the labels are generally singular.

Another possible problem for alignment is the use of preferred and alternative labels versus
the use of synonyms of equivalent status. The first is common practice in many domain-specific thesauri, the latter is commonly found in dictionaries and other lexical resources.

Other potential sources of matching problems are the spelling conventions of words with upper case characters, diacritics and hyphens. For example, “Fin-de-siècle” may be spelled in this way in one vocabulary but as “fin de siècle” in another.

Finally, vocabularies tend to differ in the treatment of homonyms, that is, terms with the same label that have different meanings. Some vocabularies prevent homonyms by explicitly adding qualifiers to labels so that each label is unique. Others allow multiple concepts to have the same label, and rely on the concept’s place in the hierarchy or its scope note to clarify its meaning.

For all the examples given above, it is a priori not clear how the different characteristics will influence alignment techniques.

### 4.3.2 Study Setup

To answer our research questions we perform two experiments following the steps shown in Fig. 4.1. In the first step we preprocess the data sets by removing qualifiers from the preferred labels. Next, we apply alignment techniques to the vocabularies to generate candidate mappings. In the third step we apply disambiguation techniques that use the structure of the vocabularies. In the fourth step we take samples of the resulting alignments. In step 5 we perform manual evaluation of samples of data classifying each alignment into one of seven categories: `skos:exactMatch`, `skos:closeMatch`, `skos:broadMatch`, `skos:narrowMatch`, `skos:relatedMatch`, `unrelated` and `unsure`. We also record the amount of time the evaluation takes. In step 6, independent raters evaluate random samples of evaluated mappings in order to assess reliability by
calculating inter-rater agreement statistics using Cohen’s Kappa. In step 7 we estimate based on the results of the evaluated mappings the performance of the alignment and disambiguation techniques. The focus here is on mappings evaluated as skos:exactMatch and skos:closeMatch. Lastly, we examine the quality of varying combinations of alignments.

There are some differences in our alignment process with respect to that in Chapter 3. First, we cannot perform a full manual evaluation, as we expect the number of mappings to be too large. Therefore, we need an additional sampling step. Second, for practical consideration we perform the evaluation step after mapping alignment generation (Step 2) and disambiguation step (Step 3). In this manner we draw samples from all subsegments we expect to be of interest in a single step. We describe the steps of the experiments in more detail in the following section.

4.3.3 Vocabularies and their Characteristics

For the two case studies we use Getty’s Art and Architecture Thesaurus (AAT)\(^2\) (Peterson and Jackman-Schuller 1996) and its Dutch version, AATNed\(^3\). The two vocabularies are closely linked, in fact the AATNed was based on the AAT and extended with additional terms. We chose to align the AAT with Princeton WordNet version 2.0, and AATNed with Cornetto\(^4\) (Vossen et al. 2008), a WordNet-like lexical resource for the Dutch language. For WordNet we used the RDF version published by W3C. The other vocabularies were originally in XML but were converted to SKOS\(^5\) by the MultimediaN E-Culture project. We describe the vocabulary characteristics in more detail below.

**AAT** is a structured vocabulary in English containing terms related to fine art, architecture and archival materials. It is organized in 7 facets with 36 hierarchies and contains 2,949 guide terms and 27,992 concepts. There are broader/narrower and related relations between the concepts, and each concept has exactly one preferred label and possibly multiple alternative labels with a total of 92,089 alternative labels for concepts. Ambiguous preferred labels are distinguished from each other with the use of qualifiers. An important feature in terms of alignment is that the preferred labels are in plural form if a plural form is linguistically possible, whereas the singular form is captured as an alternative label.

**AATNed** is a structured vocabulary in Dutch, closely related to the English AAT. It is organized in 34 hierarchies with 2,873 guideterms and 30,817 concepts. There are broader/narrower and related relations between concepts. Each concept has one preferred label. Similarly to the AAT, qualifiers are used to distinguish homonymous preferred labels. Concepts may also have alternative labels. AATNed contains 24,817 alternative labels in total, a significantly lower number than the AAT. Similarly to the AAT, preferred labels are in plural form where possible. We found 20,457 singular labels for the same number of concepts which are made explicit within AATNed.

---

\(^2\)http://www.getty.edu/research/conducting_research/vocabularies/aat/
\(^3\)http://www.aat-ned.nl/
\(^4\)http://www2.let.vu.nl/oz/cornetto/index.html
\(^5\)http://www.w3.org/2004/02/skos/
in contrast to AAT where the relation is implicit. The remaining 10,360 concepts describe processes, states, or certain materials, such as “marble” or “cement”, and thus have no plural form.

**WordNet** is a large lexical database for the English language. It contains 115,424 synsets with 203,147 labels. A synset may contain over thirty labels and one label may appear in multiple synsets (homonyms). There are 17 relations between synsets such as the hyponym and meronym relations. Important differences with the AAT are that all labels are equivalent in the sense that there is no distinction between a preferred label and alternative label; labels tend to be in singular form and contain no diacritics.

**Cornetto** is a large lexical database in Dutch containing 70,370 synsets and 103,762 labels. There are 57 relations between the synsets, significantly more than in WordNet. Examples of relations that do not occur in WordNet are “involved instrument” and “patient role”. An important distinction between Cornetto and WordNet is that Cornetto has fewer synsets and significantly fewer labels than WordNet.

Finally, an important difference between the source vocabularies (AAT and AATNed) and the target vocabularies (WordNet and Cornetto) is that the first describe the cultural heritage domain while the latter describe more general perceptions of the world, which is often visible in the different way the hierarchies are organized. The difference in ontological commitments means that even lexical matches do not necessarily have the same meaning. One example is the concept **artist** in AAT referring exclusively to artists in the fine arts, while in WordNet the meaning also includes musical and other types of artist. As a result, we expect that many mappings will not be true exact matches.

### 4.3.4 Techniques for Generating Alignments

We use three techniques which generate non-ambiguous (one-to-one) mappings and one technique that generates ambiguous (one-to-many) mappings (Step 2 in Fig. 4.1). All four techniques are based on simple string matching. The non-ambiguous techniques apply either exact string matching, string matching after normalization of hyphens and diacritics, or string matching after conversion of plurals to singular form. The non-ambiguous techniques only match AAT or AATNed **preferred labels** to WordNet or Cornetto, whereas the fourth technique, which may generate ambiguous mappings, also uses alternative labels in AAT or AATNed. Thus, the latter technique may generate a large number of ambiguous mappings due to the many homonyms and polysemes, and these mappings will need to be disambiguated in a separate step.

Our Baseline technique is based on simple exact string-matching. We used the same baseline technique as in Chapter 3 in order to compare performances. It generates mappings between unique preferred labels of the source vocabularies (AAT and AATNed) and unique labels of the target vocabularies (WordNet and Cornetto). The labels are unique in the sense that only one concept within the vocabulary has that label. If several concepts share the same label, that label is not used in the matching process. We show example mappings for each technique in Table 4.1.
The second technique matches unique singular labels. If a concept has no singular label we use the preferred label instead (provided it is also unique). Concepts with homonymous labels are simply ignored. For AATNed, we used the 20,457 singular labels present in the original vocabulary. Since in AAT the relation between preferred labels and the singular alternative label is not explicit, we generated them in a preprocessing step (Step 1 in Fig. 4.1). We applied the built-in Porter stemmer of SWI Prolog\(^6\) to the preferred label of each concept. We then matched the resulting stem to the alternative labels of the same concept. If the stem matched an alternative label, we added the label as a singular preferred label to the concept. This yielded 9,129 singular labels, which is just a third of the total number of concepts, significantly less than in AATNed. The main reason for this is that the Porter stemmer does not work perfectly, stemming for example the word “houses” to “hous”, which would subsequently not match the alternative label “house”. We refer to this technique as the Singular Non-ambiguous technique or SN for short.

The third technique matches unique normalized singular preferred labels, or normalized preferred labels if no singular label is available. These are matched to unique normalized labels from WordNet or Cornetto. For the AAT normalization is performed after singular labels were made explicit using the Porter Stemmer in the process described for the SN technique. Normalization includes replacing diacritics with a non-diacritic character ( “ö” to “o”), replacing hyphens and underscores by spaces, and turning each label into lower case. Note that normalization may infrequently introduce ambiguity. For example, after normalization the Indian style “Amber” and the material “amber” have the same preferred label. Concepts with homonymous labels are again ignored. We call this technique the Normalized Non-ambiguous technique or NN for short.

With the fourth technique, called Lexical\(^7\), we match all normalized labels of the source vocabularies to normalized labels of the target vocabularies, regardless of whether they are unique. Our goal is to generate as many candidate mappings as possible.

We applied all four techniques to generate mappings from AAT to WordNet, and from AATNed to Cornetto. Before applying the techniques, we removed qualifiers from the preferred labels of the AAT and AATNed as neither WordNet nor Cornetto have qualifiers. This means that we introduce ambiguity in the AAT labels, and we need to rely on the disambiguation techniques to repair this in a later phase. We then apply the Child Match and Parent Match disambiguation techniques described in Section 4.3 on the set of ambiguous mappings (Step 3 in Fig. 4.1).

### 4.3.5 Disambiguation Techniques

In the previous chapter we described two types of disambiguation techniques for ambiguous candidate mappings. Both techniques use the broader/narrower relationships of the source and target vocabularies (hyper/hyponym in lexical sources).

In the Child Match technique for each source concept with multiple mappings in the target vocabulary we follow the hierarchy “downwards”, and count the number of mappings between

---

\(^6\)http://www.swi-prolog.org/

\(^7\)Strictly speaking all four techniques are lexical techniques. Here we use the term lexical for alignment techniques that are allowed to generate ambiguous mappings
Alignment Technique | AAT Term | Meaning of AAT Term | WordNet Term | Meaning of WordNet Term
--- | --- | --- | --- | ---
Baseline | irons | heatable devices used to press cloth | irons | metal shackles
Baseline | dampness | state of an artifact or atmosphere when it contains moisture | dampness | a slight wetness
SN | masons | construction workers who lay stone amongst others | mason | craftsman who works with stone or brick
NN | papier mâché | material made of repulped paper and adhesive binder | papier-mache | substance made from paper pulp
Lexical | quarters | regions in a town or city | quarter | a fourth part of a year

| quarters | regions in a town or city | mercy shown to an opponent |

Table 4.1 Examples of typical (some correct, some incorrect) mappings between AAT and WordNet generated by each alignment technique along with the meaning of the mapped concepts. The alignment techniques, Baseline, Singular Non-ambiguous (SN), Normalized Non-ambiguous (NN) and Lexical, are described in detail in Section 4.3.4.

“child” concepts of aligned concepts. We assume that concepts with similar meaning will have similar hierarchies below them. This means there should be more mappings between their children than for homonymous concepts which may be lexically similar but differ in meaning. We then count the number of mappings that have at least one or more child mappings and consider them to be correct exact matches. If multiple concepts have more than one child mapping we choose the mapping with the highest number of child mappings. In some cases where the mappings to be disambiguated have the same (highest) number of child mappings no choice can be made and all mappings are kept. The Parent Match technique is a mirroring of the Child Match technique. We find correct mappings by following the hierarchies “upward”, and count the number of aligned “parent” concepts to distinguish the correct target concept from its homonyms. As neither the AAT and WordNet, nor the AATNed and Cornetto have a single top concept or similar top concepts, there is no need to exclude any mappings when using the Parent Matching technique.

4.3.6 Evaluation and Quality Measures for Alignment

Unlike in previous work where the entire set of generated mappings (4,375) was evaluated manually, we expect in the current case studies to generate significantly more mappings. Evaluating a large number of mappings manually is not feasible. We sample mappings from various subsets we expect to have different properties (Step 4 in Fig. 4.1). These samples are then evaluated manually (Step 5 in Fig. 4.1). We also perform inter-rater agreement evaluations to check the quality of the manual evaluation (Step 6 in Fig. 4.1). Subsequently, we extrapolate from the results of the
evaluated samples to estimate the precision of the subsets using the method described by van Hage et al. (2007) (Step 7 in Fig. 4.1).

Instead of using random sampling, where we estimate the precision of a set of $N$ mappings by establishing the precision of a random subset of $n$ mappings, we use stratified random sampling. We estimate the precision of the entire set of $N$ mappings by establishing subsets of mappings, or strata. Each stratum $h$ has a size $N_h$. From each stratum we take a random sample $n_h$, and after evaluation of the sample, establish its precision $p_h$. The precision $p_h$ of the sample is used as an estimate for the precision $P_h$ of the entire stratum $h$. Thus $P_h = p_h$. The precision of the entire set of $N$ mappings is measured as follows:

$$P_N = \frac{1}{N} \sum_{h=1}^{L} N_h P_h$$  \hspace{1cm} (4.1)

where $L$ is the number of strata. In this study we do not measure recall as we will likely not generate all possible correct mappings, because we are limiting the type of alignment techniques to variations on string matching. In addition, due to the large size of the vocabularies we would be unable to evaluate all mappings. Instead, we measure the power of the alignment techniques with respect to each other by calculating their coverage $C_T$, where $M_T$ is the number of mappings found by a technique, and $M_{DT}$ is the distinct total number of mappings found by all techniques. We define coverage in the following manner:

$$C_T = \frac{M_T}{M_{DT}}$$  \hspace{1cm} (4.2)

In addition, we are interested in linking as many concepts as possible from the AAT and AATNed to WordNet and Cornetto. We measure the source coverage of a technique $SC_T$, where $S_T$ is the number of source concepts mapped by the technique, and $S_V$ is the number of source concepts in the vocabulary.

$$SC_T = \frac{S_T}{S_V}$$  \hspace{1cm} (4.3)

### 4.4 Predictions and Hypotheses

Based on the analysis of the characteristics of the vocabularies in our data set, we make the following predictions.

First, AAT, AATNed and Cornetto contain diacritics in their labels, while WordNet does not. The vocabularies also differ in their use of capital letters and finally they also differ in the use of hyphens. We predict that these differences will have a significant negative effect on all alignment techniques that use simple string matching without normalization of the labels (baseline and SN techniques). We also predict that any negative effect of normalization caused by creating ambiguity in labels where previous there was none (SN technique vs. NN technique) will be outweighed by the advantage of normalizing.
Table 4.2  Number of mappings generated between AAT and WordNet by each alignment technique and their coverage. We include the number of mapped AAT concepts and the fraction they represent of all AAT concepts (source coverage). The Lexical additional represents mappings that were generated by the Lexical technique but not the non-ambiguous techniques.

Second, both the AAT and AATNed contain lexical variations of their preferred label as alternative labels of the same concept. We predict that an alignment tool not restricted to preferred labels would therefore generate significantly more mappings using the lexical variations, but at the cost of lower precision.

Third, the large number of synonym labels in the target vocabularies increases the likelihood of finding mappings and therefore increasing coverage. However, we also expect the precision to be low as both WordNet and Cornetto contain a large number of homonyms.

Finally, WordNet and Cornetto labels are mostly in singular form, while in AAT and AATNed many of the preferred labels are in plural. We predict that if we use string matching algorithm without converting plural terms into singular form we will find fewer mappings, and many of these will be incorrect matches, especially between Cornetto and AATNed. This is because in Dutch, the plural form of many nouns is the same term as the derived verb. For example, the plural of the noun *WERK* (a work) is the same as the verb form *WERKEN* (to work). This is contrary to our results from Chapter 3, where the aligned vocabularies only had singular labels, and therefore a simple string matching algorithm generated mappings with high precision.
4.5 Alignment Results

4.5.1 AAT-WordNet

Alignment Generation

We generated four alignment sets using the four alignment techniques. Table 4.2 displays the number of mappings per technique with the number of aligned AAT concepts, the coverage of the techniques and the source coverage. The Baseline technique has the lowest coverage (0.05), generating the fewest mappings, which was expected. This is caused by the large number of preferred labels in plural form. The SN and NN techniques generate more mappings and have a coverage twice as high (0.10) as the Baseline technique. The NN technique did not generate many more mappings than the SN technique, which indicates that normalizing labels does not appear to increase the number of mappings. Combined, the three non-ambiguous tools generate 4,592 distinct mappings which map 16.4% of the AAT concepts. The Lexical tool has a coverage of 1.00 which means all mappings generated by the Baseline, SS and NN tools are also generated by the Lexical tool. The tool generates almost ten times more mappings (42,039) and for three times the amount of concepts the non-ambiguous tools generate. This is because Lexical tool also generates ambiguous mappings in addition to non-ambiguous mappings, and as the number of non-unique labels in WordNet are numerous and we include alternative labels from AAT, the potential for matches is increased.

Figure 4.2 Venn diagram showing the overlapping and non-overlapping mappings between AAT and WordNet generated by the four tools. The Lexical tool generated all mappings that the three non-ambiguous tools generated.
Alignment Overlap

Examining the overlap between the alignment sets generated by the four tools revealed that all mappings generated by the non-ambiguous tools were also generated by the Lexical tool. Fig. 4.2 shows the overlaps between the three non-ambiguous tools. There is a large overlap among these three tools and a large overlap between the SN and NN tools. The figure also shows that 191 NN mappings were not found by the Baseline and SN tools. Most of these mappings are upper case labels matched to lower case labels or normalized diacritics e.g.: Venetian Blind matching Venetian blind. An example of a mapping only found by SN is Maltese cross. When this label is normalized it matches two concepts in WordNet (a type of flower and a cross), and label is no longer unique to a single concept which means the NN technique will discard the mapping. Normalizing labels may therefore introduce ambiguity. The Lexical tool generated an additional 37,447 mappings for 8,990 concepts (see Table 4.2).

Disambiguation

In the disambiguation process we consider the mappings generated by the Lexical tool. The three mapping sets generated by the three other techniques are not ambiguous by themselves, however when combined with additional mappings found by the Lexical tool 870 mappings become ambiguous.

Fig. 4.3 displays the ambiguous and non-ambiguous mappings found by the Lexical tool along with their distribution in the aggregated mappings found by the three non-ambiguous techniques. Of the total 42,039 mappings found by the lexical tool, 36,201 mappings are ambiguous (the total grey area in Fig. 4.3) linking 6,887 AAT concepts to WordNet. Out of these 870 are also generated by one or more non-ambiguous techniques. Of the non-ambiguous mappings, 3,722 were generated by the non-ambiguous techniques, and an additional 2,116 non-ambiguous mappings were generated by the Lexical tool, leading to a total 5,838 non-ambiguous mappings.

When we apply the Child Match and Parent Match techniques to disambiguate mappings, mappings found by the three non-ambiguous tools are also affected as shown in Fig. 4.4. The figure shows that 74 mappings found by the three non-ambiguous tools are removed in the disambiguation process, although the bulk of the removed mappings is from the additional mappings found by the Lexical tool. We also find that 20,449 mappings found by the Lexical tool remain ambiguous.

For the analysis of the performance of the two disambiguation techniques we only consider mappings that were not found by the three non-ambiguous tools. Our focus is thus the result of the disambiguation that improves the mappings additionally returned by the Lexical tool. Table 4.3 displays the number of mappings that were kept, the number of disambiguated concepts and the mappings that were rejected. The Parent Match technique disambiguated three and a half times more concepts than the Child Match technique. There is a small overlap between the two. An example of a correctly disambiguated mapping is the concept Vehicle (motorized vehicle), which is disambiguated from Vehicle (expression or medium), because of mappings between its children
such as AIRCRAFT and TRICYCLE. Of the 6,887 ambiguous concepts, 2,665 concepts (39%) were disambiguated.

4.5.2 AATNed-Cornetto

Alignment Generation

The result of the alignment process is shown in Table 4.4. Similarly to the AAT-WordNet case, Baseline generated the fewest number of mappings and has the lowest coverage (0.06). The SN
### Table 4.3
The number of disambiguated mappings between AAT and WordNet. The table includes the number of mappings kept, and the number of mappings that were removed using the Child Match and Parent Match techniques. The table also shows the number of overlapping mappings between the two techniques.

<table>
<thead>
<tr>
<th>Segment</th>
<th>Kept Mappings</th>
<th>Disambiguated Concepts</th>
<th>Removed Mappings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Child Match only</td>
<td>590</td>
<td>554</td>
<td>3,205</td>
</tr>
<tr>
<td>Parent Match only</td>
<td>2,485</td>
<td>2,011</td>
<td>7,035</td>
</tr>
<tr>
<td>Overlap</td>
<td>236</td>
<td>234</td>
<td>1,331</td>
</tr>
<tr>
<td>Distinct Total</td>
<td>3,311</td>
<td>2,665</td>
<td>11,571</td>
</tr>
</tbody>
</table>

### Table 4.4
Number of mappings generated by each alignment technique between AATNed and Cornetto and their coverage. The table includes the number of AATNed concepts that were mapped, and their fraction of the total AAT concepts (source coverage).

<table>
<thead>
<tr>
<th>Alignment Technique</th>
<th># of Mappings</th>
<th># of Mapped AATNed Concepts</th>
<th>Coverage</th>
<th>Source Coverage of AATNed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>1,980</td>
<td>1,980</td>
<td>0.10</td>
<td>0.06</td>
</tr>
<tr>
<td>SN</td>
<td>6,563</td>
<td>6,563</td>
<td>0.32</td>
<td>0.21</td>
</tr>
<tr>
<td>NN</td>
<td>6,644</td>
<td>6,644</td>
<td>0.33</td>
<td>0.21</td>
</tr>
<tr>
<td>Distinct total non-ambiguous</td>
<td>6,914</td>
<td>6,856</td>
<td>0.34</td>
<td>0.22</td>
</tr>
<tr>
<td>Lexical additional</td>
<td>13,417</td>
<td>4,414</td>
<td>0.66</td>
<td>0.14</td>
</tr>
<tr>
<td>Lexical</td>
<td>20,331</td>
<td>10,773</td>
<td>1.00</td>
<td>0.34</td>
</tr>
<tr>
<td>Total</td>
<td>20,331</td>
<td>10,773</td>
<td>1.00</td>
<td>0.35</td>
</tr>
</tbody>
</table>

tool generated over three times as many mappings as the Baseline technique, as the singular labels were explicit in AATNed and thus matching to Cornetto which contains only singular labels was more successful. In total, the three non-ambiguous techniques generated 6,914 mappings for the same number of concepts, aligning a little over 22% of AATNed concepts.

The Lexical tool has a coverage of 1.00, which means it generates all mappings generated by the three non-ambiguous techniques. The Lexical technique generated 20,331 mappings for over a third of the total AATNed concepts. Of these mappings 13,417 additional mappings were generated on top of the mappings generated by the non-ambiguous techniques. This is a much lower number than the mappings generated than for AAT-WordNet. This difference is caused by the lower number of alternative labels in AATNed than in AAT, and fewer sense labels in Cornetto than in WordNet.
Figure 4.5 Venn diagram showing the number of overlapping and non-overlapping mappings between in AATNed and Cornetto generated by the alignment tools. The Lexical tool includes all mappings generated by the non-ambiguous tools.

Alignment Overlap

Again, all mappings generated by the three non-ambiguous tools were also generated by the Lexical tool. Fig. 4.5 shows the overlap between the alignment tools. There is a large overlap between the three non-ambiguous tools. However, the number of mappings found only by Baseline tool is higher than in the AAT-WordNet case. There is an even larger overlap between SN and NN tools and the NN tool only generates 100 additional mappings.

The Lexical tool generated an additional 13,417 mappings for 4,414 concepts. An analysis showed that a small subset of these additional mappings (569) is not ambiguous (Fig. 4.6). This is a smaller number than in AAT-WordNet, caused by the fewer alternative labels in AATNed.

Disambiguation

Similarly to the AAT-WordNet case we focus on the mappings generated by the Lexical tool. Fig. 4.6 displays the number of ambiguous and non-ambiguous mappings and their overlap with the aggregated mappings found by the three non-ambiguous techniques. A small set of 609 mappings found by the non-ambiguous mappings becomes ambiguous when aggregated with the Lexical mappings. A further 12,848 mappings found by the Lexical tool are ambiguous resulting in a total of 13,457 ambiguous mappings linking 3,899 AATNed concepts to WordNet.

After applying the Child Match and Parent Match techniques we found that 234 mappings found by the non-ambiguous tools played a role in the disambiguation. Fig. 4.7 shows that 75 mappings were removed by the disambiguation process from this set and that 159 mappings were kept. As we are interested in improving the quality of the mappings generated additionally by
Figure 4.6 Diagram showing the types of mappings generated by the Lexical tool divided into ambiguous and non-ambiguous AATNed-Cornetto mappings. The diagram also shows the number of aggregated mappings found by the three non-ambiguous techniques.

Figure 4.7 Diagram showing the result of the disambiguation in the ambiguous set of AATNed-Cornetto mappings. The figure shows the overlap of the disambiguated mappings with the mappings generated by the three non-ambiguous tools along with the number of kept and removed mappings, and the remaining ambiguous mappings.

the Lexical tool, we do not include these 234 mappings in our analysis. The results of the disambiguation are shown in Table 4.5. Again, there is a small overlap between the two disambiguation techniques. Overall 1,297 concepts were disambiguated, which is a third of the total number of ambiguous concepts. In comparison to the AAT-WordNet case we see that a smaller percentage of the aligned concepts are ambiguous. This is caused by fewer number of alternative labels in AATNed and also fewer labels per concept in Cornetto.
<table>
<thead>
<tr>
<th>Segment</th>
<th>Kept Mappings</th>
<th>Disambiguated Concepts</th>
<th>Removed Mappings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Child Match only</td>
<td>342</td>
<td>327</td>
<td>1,140</td>
</tr>
<tr>
<td>Parent Match only</td>
<td>1,281</td>
<td>920</td>
<td>1,667</td>
</tr>
<tr>
<td>Overlap</td>
<td>106</td>
<td>104</td>
<td>200</td>
</tr>
<tr>
<td>Distinct Total</td>
<td>1,729</td>
<td>1,297</td>
<td>3,007</td>
</tr>
</tbody>
</table>

Table 4.5 The number of disambiguated mappings between AATNed and Cornetto. The table includes the number of mappings kept, and the number of mappings that were removed using the Child Match and Parent Match techniques. The table also shows the number of overlapping mappings between the two techniques.

### 4.6 Evaluation

![Venn diagram](image)

Figure 4.8 Venn diagram of segments representing the overlaps of the four tools. This diagram serves as a legend for Table 4.6.

We manually evaluated samples of generated mappings. The samples were selected from all segments of the Venn diagram shown in Fig. 4.8. Segment H represents the mappings that were additionally generated by the Lexical technique. We sample several subsets from segment H (Fig. 4.9). The samples are from mappings kept by the two disambiguation techniques and their overlap, as well as mappings removed by the two techniques, in order to assess the number of correct mappings that were removed (false negatives). We also took a sample of remaining ambiguous mappings that were not disambiguated to assess the overall performance of the Lexical tool. Finally, we also sampled non-ambiguous mappings generated using unique alternative labels. In total we selected mappings for 1,000 source concepts from AAT and 1,000 concepts from AATNed. These mappings were scattered along all segments: 400 source concepts and their associated mappings were sampled from segments A to G, and 600 source concepts and their associated mappings from segment H.
Segment H

Non-ambiguous mappings

Remaining ambiguous mappings

Child Match

Kept mappings

Parent Match

Removed disambiguated mappings

Figure 4.9 Diagram of the composition of segment H representing the additional mappings found by the Lexical technique. This diagram serves as a legend for Table 4.7.

The mappings were manually evaluated by the first author using the evaluation tool used in Chapter 3. Each mapping was rated as one of the seven following categories: skos:exactMatch, skos:closeMatch, skos:broadMatch, skos:narrowMatch, skos:relatedMatch, unrelated, or unsure. In these studies we consider skos:exactMatch and skos:closeMatch to be correct equivalent mappings, and the remaining five categories to be incorrect. The evaluation of 1,000 AAT concepts and their associated 1,870 mappings to WordNet, and of 1,000 AATNed concepts and their associated 1,481 mappings to Cornetto took approximately 13 person-hours.

Three extra raters evaluated small samples of mappings and we subsequently measured agreement between each rater and the first author. These three raters each evaluated two different samples of 50 source concepts with mappings from AAT-WordNet and AATNed-Cornetto. We measured Cohen’s Kappa (Cohen 1960) between each extra rater and the first author and found that agreement varied between 0.59 and 0.86. The average agreement between the raters and the first author was a $\kappa$ of 0.67 for AAT-WordNet and 0.71 for AATNed-Cornetto. These values are considered to be moderate agreement according to the interpretation of Landis and Koch (1977). We found this level of agreement acceptable for our experiments, and use the evaluation of the first author as a gold standard in the remainder of this chapter.

An analysis of the disagreements showed that some disagreements are caused by human error. Other disagreements are caused by different interpretations of concepts where one rater is more strict than the other. This result shows that the evaluation task is difficult even for humans.
Table 4.6 Results of the sampled evaluation for the segments from the non-ambiguous tools for both AAT-WordNet and AATNed-Cornetto. The table includes the size of the segment, the size of the evaluated sample, the number of mapping judged as correct and the precision of the sample. See Fig. 4.8 for the meaning of the letters denoting the segments.

### 4.6.1 Results for AAT-WordNet

Table 4.6 displays the result of the manual evaluation. In the columns for the AAT-WordNet mappings we see that the overlap between the SN and NN tools marked by segment F has the highest precision (0.90), followed by the segment G at 0.81. As predicted, the precision of the mappings generated only by the Baseline tool (segment A), is lower than most other segments at 0.38. This is caused by plural nouns in AAT incorrectly matching verbs in WordNet.

Table 4.7 shows the results of the sampling of mappings from the set found only by the Lexical tool. The disambiguation with the Child Match technique performed better with a precision of 0.74, while the mappings Parent Match technique generated have a precision of 0.46. The overlap between the two disambiguation techniques has the highest precision at 0.84. The number of false negatives in the discarded segments is low, between 3% and 8%. The remaining ambiguous mappings that were not disambiguated have a precision of 0.12. This means that of the 20,449 ambiguous mappings an estimated 2,650 mappings should be close- or exact-matches. The non-ambiguous mappings generated using alternative labels have a precision of 0.65. This is lower than for mappings with preferred labels, supporting the view that alternative labels yield worse mappings than preferred labels. We estimate the precision of all the mappings between AAT and WordNet without disambiguation at 0.17. Thus, only applying lexical alignment without disambiguation yields an unacceptably low precision. The estimated precision of all mappings generated by the Lexical technique is 0.24. As the Lexical technique generated all mappings generated by the three non-ambiguous techniques, 0.24 is the precision of all mappings between AAT and WordNet.
Table 4.7 The results of the evaluation of the subsegments of segment H (see Fig. 4.9), which are the additional mappings generated by the Lexical technique. The table includes the number of mappings in each segment, the size of the evaluated sample, the number of mappings judged to be correct and the precision of the sample. The overall precision of segment H is based on the weight and precision of each subsegment, thus we only include the estimated precision of the sample.
4.6.2 Results for AATNed-Cornetto

The results of the evaluation displayed in Table 4.6 and Table 4.7 show that the overall precision of the Dutch mappings is higher than the English language mappings. This is caused by the lower number of labels per concept in both AATNed and Cornetto resulting in fewer mappings per concept.

Similarly to the results of AAT-WordNet, half of the sample mappings generated only by the Baseline tool (segment A) are incorrect. There are more mappings in this segment than in the AAT-WordNet, and the evaluation revealed that 60% of these mappings are to Cornetto verbs, which are incorrect. For example, the concept handwerken (needle-work) was mapped incorrectly to the verb handwerken (needle-working). Some of these mappings are correct however, as the labels of processes in AATNED are often verbs. Again, just as in AAT-WordNet, the mappings generated by the NN only (segment E) have relatively low precision at 0.48, although this is higher than the 0.36 precision of the same segment in AAT-WordNet. Most of the erroneous mappings are due to concepts that describe styles and periods mapped to the nationality, or language the style gets its name from. These concepts are related but are not the same (e.g. the pueblo (style) and pueblo (house)). The overlap between SN and NN (segment F) has the highest precision, following the trend we have seen in AAT-WordNet.

Table 4.7 shows that the disambiguation techniques performed slightly better in Dutch than in English. Although the exact cause is difficult to pinpoint, this is possibly due to more similar hierarchies in the Dutch vocabularies than the English vocabularies. The overlap in mappings generated by the two disambiguation techniques has the highest precision of 0.91 followed by the mappings found only by the Child Match technique at 0.89, and the Parent Match at 0.68. The number of false negatives ranges from 6% to 12%. The precision of the sample of the subset of 569 non-ambiguous mappings generated by the Lexical tool is 0.55. Finally, we estimate the average precision of all mappings between AATNed and Cornetto at 0.47. This is significantly higher than in the AAT-WordNet case, caused by fewer alternative labels in AATNed and fewer labels per concept in Cornetto.

4.6.3 Results of Alignment Technique Combination

We now look at the precision of some combinations of techniques for AAT-WordNet and AATNed-Cornetto. Table 4.8 displays the number of mappings, their precision, the number of source concepts mapped, and the percentage they represent of all source concepts. The mappings generated by the non-ambiguous techniques (segment 1), which is the combination of mappings by the Baseline, SS and SN techniques, have relatively high precision (0.82 for AAT-WordNet and 0.88 for AATNed-Cornetto). The precision of the non-ambiguous Lexical mappings (segment 2) and disambiguated mappings (segment 3) is much lower for AAT-WordNet than the precision of segment 1. If we wish to increase coverage of the AAT by adding segments 2 and 3 to segment 1, we can raise it from 16.4% of the AAT concepts to 32.9%, although precision would drop to 0.69. The addition of subsegments from the Lexical set seems to be an acceptable trade-off for the boost in
<table>
<thead>
<tr>
<th>Segment</th>
<th>AAT-WordNet</th>
<th></th>
<th>% of Mapped</th>
<th>AATNed-Cornetto</th>
<th></th>
<th>% of Mapped</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mappings</td>
<td># of</td>
<td>Precision</td>
<td>Mappings</td>
<td># of</td>
<td>Precision</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AAT</td>
<td></td>
<td>AATNed</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Concepts</td>
<td></td>
<td>Concepts</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Non-ambiguous techniques</td>
<td>4,592</td>
<td>4,592</td>
<td>0.82</td>
<td>16.4%</td>
<td>6,914</td>
<td>6,914</td>
</tr>
<tr>
<td>2. Non-ambiguous Lexical</td>
<td>2,116</td>
<td>2,116</td>
<td>0.65</td>
<td>7.6%</td>
<td>569</td>
<td>569</td>
</tr>
<tr>
<td>3. Disambiguated Lexical</td>
<td>3,311</td>
<td>2,665</td>
<td>0.53</td>
<td>9.5%</td>
<td>1,729</td>
<td>1,297</td>
</tr>
<tr>
<td>1 + 2</td>
<td>6,708</td>
<td>6,695</td>
<td>0.70</td>
<td>23.9%</td>
<td>7,483</td>
<td>7,425</td>
</tr>
<tr>
<td>1 + 2 + 3</td>
<td>10,019</td>
<td>9,208</td>
<td>0.69</td>
<td>32.9%</td>
<td>9,212</td>
<td>8,621</td>
</tr>
<tr>
<td>Lexical (random sampling)</td>
<td>42,039</td>
<td>12,725</td>
<td>0.27</td>
<td>45%</td>
<td>20,331</td>
<td>10,773</td>
</tr>
</tbody>
</table>

Table 4.8 Precision of combinations of alignment techniques along with the number of mapped source concepts and the percentage they form of the source vocabulary. The non-ambiguous techniques stand for the combination of Baseline, SN and NN techniques (Segment 1). Segment 2 represents the non-ambiguous mappings generated additionally by the Lexical technique. Segment 3 represents the mappings that were disambiguated using the Child and Parent Match techniques. We also include the estimated precision of the entire set of Lexical mappings measured by taking a random sample of 100 mappings for comparison.

coverage, in particular when compared to the overall precision of the Lexical set, which is 0.24 for AAT-WordNet. The increase in coverage by adding segments 2 and 3 for AATNed-Cornetto is less dramatic: 22.4% to 27.9%, but the precision remains relatively high at 0.84.

Overall, we find that the same techniques yield different results in the two experiments. Whereas combining different sets of mappings between AAT and WordNet doubles the total number of mappings, it reduces precision significantly. This is caused by the relatively large size of segments 2 and 3 and their lower precision. The precision of the combined segments 1, 2 and 3 for AATNed-Cornetto remains high because segments 2 and 3 are much smaller than segment 1.

We can conclude that alignment tactics that perform well on one data set do not necessarily work as well on another data set, even when these data sets are very similar. The combination of different sets of mappings needs to be done on a case-by-case basis. We could have applied the same detailed selection process as in Chapter 3, but as the dominant segments are large (segments
F and G) and due to their size determine overall precision, we did not.

The last row in Table 4.8 displays the precision of the Lexical set, based on a random sample. Based on such precision values, 0.27 for AAT-WordNet and 0.44 for AATNed-Cornetto the Lexical set would have been considered too inaccurate to be of any practical use for alignment. By identifying subsets of mappings sharing similar characteristics we are able to select sets with higher than average precision.

### 4.7 Discussion and Conclusions

We make the following conclusions about the performance of the techniques.

1. The simple non-ambiguous string matching techniques work well with a high precision but low source coverage.

2. The lexical matching technique allowing for ambiguous mappings improves source coverage but reduces precision significantly.

3. Disambiguation of the lexical matches improves the precision of these mappings.

4. By combining disambiguated mappings with mappings generated by simple non-ambiguous techniques we can improve coverage and keep precision at an acceptable level.

Our key findings with respect to the characteristics of the vocabularies are:

First, the selection of alignment techniques is mainly influenced by the characteristics of the vocabularies. For example, in this case, the use of plural preferred labels in AAT and AATNed made it necessary to find the singular version of the label to improve the matching process. Some differences in the performance of matching techniques was caused by the explicit availability of singular labels in AATNed as opposed to implicit labels in AAT. This was not the case in the previous case study in Chapter 3 where all labels where in singular form.

Second the number of aligned source concepts is influenced by the difference in domain of the vocabularies, the source vocabularies being specialist cultural heritage vocabularies and the target vocabularies covering a “common sense” domain.

We found that a combination of non-ambiguous and ambiguous string matching techniques with disambiguation works relatively well given the differences in vocabularies.

We also find that by examining the manner in which mappings generated by the alignment techniques overlap, we are able to assess the value of each technique in clear terms. These subsets also provide us with natural strata which we can sample for evaluation purposes, rather than randomly sampling alignment sets generated by each technique. As a result, we have clear perception of the quality of alignments, which we can compare to the overall precision of all mappings, which was in this case the Lexical mappings.

In this study, measures such as coverage and source coverage were more useful than in Chapter 3, because we were unable to measure precision precisely, and establishing recall was impractical. The measures provide some insight into the performance of alignment techniques, in particular.
in establishing whether modifications in alignment techniques based on vocabulary characteristics increase the amount of mappings.

For future work, we would like to make the observation that the mappings between AAT and WordNet, and AATNed and Cornetto can be used to create mapping chains. AATNed is closely related to AAT and a large portion of AATNed concepts (27,077) is mapped to AAT. These are good quality mappings with very high precision (see Chapter 5). By combining mappings between AAT and WordNet, and AATNed and Cornetto we can create mappings between WordNet and Cornetto, and thus we would not need to generate mappings from scratch. Nevertheless, the quality of such chained mappings needs to be examined. This issue will be explored in the next chapter.

Acknowledgement

The data sets have been kindly provided by RKD and the Cornetto project. We thank Marieke van Erp for her contributions to the alignment evaluation. This research was supported by the MultimediaN project funded through the BSIK programme of the Dutch Government.
Having identified key aspects of vocabulary alignment techniques we now focus on the resulting alignments. In Section 4.7 we have seen that when multiple vocabularies are mapped, chains of mappings can be formed. In this chapter we investigate the quality of these chains. We study mapping compositions in cultural heritage as well as medical vocabularies and in various languages by manually evaluating samples. We examine the quality of the composite mappings relative to the quality of the input mappings, and analyze how characteristics of the input mappings and the ontologies influence the composition.

This chapter is based on a paper coauthored with Amir Ghazvinian, Jacco van Ossenbruggen, Natalya F. Noy and Mark A. Musen “Lost in Translation? Empirical Analysis of Mapping Compositions for Large Ontologies” (Tordai et al. 2010b) presented at the Fifth International Ontology Matching workshop co-held with the International Semantic Web Conference (ISWC 2010) in Shanghai, China.

5.1 Introduction

Researchers typically study ontology alignments in the context of a single source and target ontology. However, as more and more of such alignments are being created and published, longer chains of equivalent or otherwise related concepts start to emerge in our data sets. In this chapter, we analyze the quality of a subset of such chains, focusing on short chains of equivalence and near equivalence links. Most of us have clear intuitions about the properties of such chains. For example, equivalence relations such as owl:sameAs and skos:exactMatch (Miles and Bechhofer 2009), are defined as being transitive, so it should be safe to assume that if term A is equivalent to B, and B is equivalent to C, then A should also be equivalent to C. We will test this hypothesis empirically by determining to what extent such transitivity actually holds in our data sets, and if not, what is going wrong. Furthermore, for relations such as skos:closeMatch, which are not defined as being transitive, we might ask how often chains of these relations turn out to be transitive after all.

Given a mapping from A to B and from B to C, where concepts A, B and C are part of three different ontologies, we call the mapping from A to C a composite mapping. Although mapping composition is related to the use as background knowledge where concept B would be part of
the background ontology (Aleksovski et al. 2006), we do not predefine ontologies as a source of background knowledge. We analyze the properties of such composite mappings on real life data sets, addressing the following two research questions:

- What is the quality of composite mappings relative to the quality of input mappings?
- Does the quality of composite mappings depend on other characteristics of the input mappings or ontologies?

In order to answer these research questions, we study composite mappings for ontologies in different domains, using input mappings generated in different ways (Section 5.3.1). We analyzed the precision of composite mappings by sampling them and having human experts verify the samples (Section 5.3.3). In some cases, we already had pre-existing alignments for the sets of ontologies for which we analyze composite mappings. In these cases, we compared the precision of the composed mappings with the precision of existing mappings. We then analyzed our results (Section 5.5) and made observations regarding the quality and quantity of composed mappings, trying to identify reasons for correct and incorrect mapping compositions based on characteristics of the data and the input mappings.

The main contribution of this chapter is a large-scale empirical analysis of the nature of composite mappings given varied sets of input ontologies and mappings.

5.2 Related Work

In ontology matching, Euzenat (Shvaiko and Euzenat 2008) discusses mapping composition in a theoretical paper on algebras of relations as a means for validating existing mappings and creating new mappings. This work considers composition through equivalence mappings to be a trivial case because the result is an equivalence relation, and because we can assume that equivalence is transitive. In practice, however, automatically generated mappings are usually similarity mappings at best, and therefore the composition of such mappings is not trivial. We look at such automatically generated mappings and analyze results of composition to find out whether they are interesting or truly lost in translation.

Researchers have already developed a plethora of tools for generating mappings and compared their performance at the OAEI. These off-the-shelf tools, such as ASMOV (Jean-Mary et al. 2009), RiMOM (Zhang et al. 2009), Falcon-AO (Hu and Qu 2008), and DSSim (Nagy et al. 2009) perform well on OAEI benchmarks and on certain specialized tracks. However, the results of the 2009 library track showed that current tools largely fail on extremely large vocabularies and vocabularies that use multiple languages (Euzenat et al. 2009).

Mapping composition has some parallels to the use of background knowledge by mapping tools. Tools such as SAMBO (Lambrix et al. 2008) and ASMOV use background knowledge (UMLS Metathesaurus, WordNet) to improve the quality of mappings. When mapping two domain ontologies, these tools either use existing mappings from these domain ontologies to some
background source, such as UMLS or WordNet, or create these mappings “on the fly” through lexical comparison or other means. The tools then use these mappings to a single source of background knowledge for creating mappings for the domain ontologies. This method is related to mapping composition because we use a mapping to a third ontology or vocabulary. In this sense, in mapping composition any ontology becomes a source of background knowledge.

The COMA (Do and Rahm 2002) and COMA++ (Aumueller et al. 2005) tools combine several matching techniques including composition of mappings. The evaluation of the tools demonstrated the effectiveness of mapping composition without going into a more detailed analysis of the results.

5.3 Materials and Methods

In this section, we describe the ontologies and existing mappings that we used for mapping composition (Section 5.3.1), the method for creating compositions and its complexity (Section 5.3.2), and our methodology for assessing the precision of the composed mappings (Section 5.3.3).

5.3.1 Data: Ontologies and Input Mappings

In order to get a comprehensive analysis of mapping composition under different conditions, we consider three sets of ontologies and mappings. We have ontologies in three different domains: biomedicine, cultural heritage and library subject headings (Table 5.1). The terms in these ontologies have labels in four languages: English, Dutch, German and French, and the input mappings we use for composition were generated using two types of methods: lexical string matching, and instance-based matching.

BioPortal Data

Our first set of ontologies came from BioPortal (Noy et al. 2009), a Web-based repository of biomedical ontologies. At the time we collected the data, BioPortal contained 151 ontologies with more than 2.5 million concepts among them. We generated mappings between these ontologies using simple lexical comparisons of preferred names and synonyms after normalization (Ghazvinian et al. 2009a,b).

Because the Bioportal data set contains a large number of ontologies, we cannot describe the input mappings with respect to their source ontology in our quantitative analysis. Instead, we have divided the input mappings according to the type of label used to create the mapping. The BioPortal mappings can be divided into three groups: Preferred–Preferred, Preferred–Synonym, and Synonym–Synonym mappings. The Preferred–Preferred mappings are based on exact string matching between the preferred labels of concepts. The Preferred–Synonym mappings are based on exact string matching between the preferred label and synonyms, and the Synonym–Synonym mappings are based on the matching of synonyms. We manually evaluated samples from each of these sets adding up to a total of 1,000 mappings. Table 5.2 shows the number and precision
<table>
<thead>
<tr>
<th>Set</th>
<th>Domain</th>
<th>Ontologies</th>
<th>Language</th>
<th>Ontology size</th>
</tr>
</thead>
<tbody>
<tr>
<td>BioPortal</td>
<td>Biomedicine</td>
<td>151 ontologies</td>
<td>English</td>
<td>Ranging from under 100 concepts to 380K concepts</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Mean size=17,805 (SD= 61,614)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Total concepts: 2,688,609</td>
</tr>
<tr>
<td>CH</td>
<td>Cultural-Heritage</td>
<td>AAT</td>
<td>English and Dutch</td>
<td>27,077 concepts with English and Dutch labels</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WordNet</td>
<td>English</td>
<td>115,424 synsets with 203,147 English labels</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cornetto</td>
<td>Dutch</td>
<td>70,370 synsets and 103,762 Dutch labels</td>
</tr>
<tr>
<td>Library</td>
<td>General</td>
<td>LCSH</td>
<td>English</td>
<td>339,612 concepts</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rameau</td>
<td>French</td>
<td>157,287 concepts</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SWD</td>
<td>German</td>
<td>163,175 concepts</td>
</tr>
</tbody>
</table>

Table 5.1 Sets of ontologies used for mapping composition and their characteristics.

based on the evaluated sample of each set. We use Preferred–Preferred and Preferred–Synonym mappings as input for composition. We do not to include Synonym–Synonym mappings in our input mappings because they have low precision (0.36).

Cultural-Heritage Data

The second set of mappings links four large vocabularies in the cultural heritage domain: Getty’s Art and Architecture Thesaurus (AAT\(^1\) (Peterson and Jackman-Schuller 1996), Princeton WordNet\(^2\) version 2.0 (Fellbaum 1998), AATNed\(^3\), the dutch version of the AAT and Cornetto,\(^4\) (Vossen et al. 2008) a WordNet-like lexical resource for Dutch. Most of the concepts in AATNed (22,077) are linked to AAT concepts, as AATNed was created based on AAT. Our assumption is that these mappings are 100% correct, and therefore in the remainder of this chapter we treat the AATNed concepts as an extension of the AAT in Dutch. We only use the 27,077 concepts in AAT that are linked to AATNed, and we refer to the combination of these vocabularies as AAT. We generated mappings between AAT and WordNet, and between AATNed and Cornetto using lexical comparison in Chapter 4. The Cornetto project (Vossen et al. 2008) created mappings between Cornetto and different versions of WordNet using a combination of manual and automatic methods.

\(^1\)http://www.getty.edu/research/conducting_research/vocabularies/aat/
\(^2\)http://wordnet.princeton.edu/
\(^3\)http://www.aat-ned.nl
\(^4\)http://www2.let.vu.nl/oz/cornetto/index.html
Table 5.2 The number and precision of Bioportal input mappings along with the evaluated sample size. The mapping sets are divided according to the type of label that was used for creating the mapping (preferred label or synonym). The Synonym–Synonym mappings were not used as input mappings.

The input mappings between AAT-Cornetto and AAT-Wordnet are non-ambiguous mappings with relatively high precision values. (These mappings correspond to the union of morphological tools in Table 4.8) These mappings form a subset of all mappings created in Chapter 4. The evaluation of these mappings is described in detail in Chapter 4.

We also selected a subset of existing mappings between Cornetto and WordNet. The original Cornetto project mappings contain two types of equivalence mappings: 3,144 eqSynonym (equal synonym) mappings and 79,313 eqNearSynonym (near equal synonym). We manually evaluated a sample of 107 eqSynonym mappings and 173 eqNearSynonym mappings and measured a precision of 0.95 for the first and 0.7 for the second set. We only use eqSynonym as input mappings as the precision of the eqNearSynonym is much lower than the precision of the AAT-Cornetto and AAT-Wordnet mappings. Table 5.3 shows the number of input mappings per vocabulary pair along with the estimated precision of the mappings.

Table 5.3 The number of input mappings per vocabulary pair, and their precision for the cultural heritage data.
Library Data

We used a set of ontologies and mappings from the Library track in the OAEI 2009. This set contains three lists of subject headings for describing content of books: the Library of Congress Subject Headings (LCSH); Rameau, a list used by the French National Library; and the Subject Heading Authority File (SWD), which is used by the German National Library. Each list contains from 150,000 to 300,000 concepts.

For the library data we used mappings that Wang et al. (2009) created using instance-based matching based on books that were classified using terms from more than one vocabulary. This method for generating mappings ranks the resulting mappings according to confidence level. Although there are a total of almost 2 million mappings, over 90% of them have confidence measure lower than 0.1. For the purpose of composing mappings, we selected only those mappings that had a confidence measure greater than 0.7. These mappings involve fewer than 1.5% of the concepts in the vocabularies. We estimated the precision of these input mappings by manually evaluating samples of 100 to 113 mappings per set. Table 5.4 shows the number of input mappings per vocabulary pair along with the estimated precision of the mappings.

<table>
<thead>
<tr>
<th>Set</th>
<th>Mapping set</th>
<th>Number of mappings</th>
<th>Precision</th>
<th>Evaluated Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Library</td>
<td>LCSH–Rameau</td>
<td>2,242</td>
<td>0.95</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>SWD–LCSH</td>
<td>2,334</td>
<td>0.54</td>
<td>113</td>
</tr>
<tr>
<td></td>
<td>Rameau–SWD</td>
<td>685</td>
<td>0.72</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 5.4 The number of input mappings, their estimated precision and the size of the evaluated sample for the library-track data.

In the cultural heritage and OAEI library track the number of input mappings is significantly lower than in the BioPortal case, as our aim was to select high-quality mappings. We chose a representative subset in order to analyze the properties of mapping composition.

5.3.2 Computing Mapping Composition

In this chapter, we consider only the composition of two mappings. The BioPortal compositions were computed using a relational database, and the cultural heritage and OAEI library track composition algorithms were written in SWI-Prolog⁵.

BioPortal Mapping Composition

From the Preferred–Preferred and Preferred–Synonym mappings we are able to compose 6 groups of composed mappings which are displayed in Figure 5.1.

⁵http://www.swi-prolog.org/
Figure 5.1  Methods for composing mappings between concepts in three different ontologies ($C_1 \in O_1$, $C_2 \in O_2$, $C_3 \in O_3$) using mappings between preferred labels ($P$) and synonyms ($S$). Figure A illustrates the PPP mappings: a composition of a mapping from a preferred name of $C_1$ to preferred name of $C_2$ with the mapping between preferred names of $C_2$ and $C_3$. Figure B illustrates PSP mappings: a match of $C_1$ preferred name to $C_2$ synonym with a match of $C_2$ synonym to $C_3$ preferred name. Figures C—F illustrate the remaining possible cases.

For instance, Fig. 5.1A illustrates the case where we compose a mapping from a preferred name for the concept $C_1$ to a preferred name for $C_2$ with a mapping from the preferred name for $C_2$ to the preferred name of $C_3$. We refer to this case as PPP. Note that this composition produces a subset of the Preferred–Preferred input mappings between $O_1$ and $O_3$. PSP mappings (Fig. 5.1B) also produce a subset of the Preferred–Preferred input mappings. Similarly, PPS mappings (Fig. 5.1C) and SPS mappings (Fig. 5.1D) produce subsets of the Preferred–Synonym and Synonym–Synonym mappings between $O_1$ and $O_3$, respectively. We analyze these subsets and compare their precisions to those of the original Preferred–Synonym and Synonym–Synonym mappings that were generated directly by comparing $O_1$ and $O_3$. Fig. 5.1E and F illustrate the other two cases, PSPS and PSSP, which produce mappings that we cannot obtain by comparing preferred names and synonyms directly. That is, in Fig. 5.1E and F concepts $C_1$ and $C_3$ have no label in common.

Cultural Heritage and Library Mapping Composition

For the cultural heritage and library-track data the composition is easier to track, as in each cases we have three mapped vocabularies. Fig. 5.2 shows the number of input mappings and their precision per aligned vocabulary. In both cases we can generate three sets of compositions. Thus for example by combining mappings between AAT and WordNet and AAT and Cornetto we can compose mappings between WordNet and Cornetto.

5.3.3 Sampling

In order to evaluate the precision of the composed mappings as well as the precision of input mappings (see Section 5.3.1), we sampled the mappings and evaluated the samples manually. Because of the scale of our data—with hundreds of thousands of mappings to verify—evaluating all the mappings manually was not feasible. Furthermore, because of the size of the ontologies themselves, creating a complete set of mappings so that we can evaluate recall was not feasible.
either. In addition, the recall of mapping composition is necessarily limited by the recall of the input mappings used for composition. Thus, we focus in this evaluation on estimating only the precision of the composed mappings.

For BioPortal mappings, we used stratified sampling (van Hage et al. 2007) to select mappings for manual evaluation. We divided the mappings into 6 subsets depending on the methods used for composing mappings shown in Fig. 5.1 (PPP, PSP, PPS, SPS, PPS and PSSP). Among the BioPortal ontologies, there is a large number of ontology pairs that have only one or two composed mappings between them. Therefore, we constructed a number of strata to ensure that our samples include mappings between ontology pairs with only a few mappings between them, as well as mappings between ontology pairs with thousands of mappings, and clusters in between. We divided all mapping subsets based on composition method into 7 strata according to the number of mappings between ontology pairs: ontology pairs with over 10,000 mappings, ontology pairs with 3,001 to 10,000 mappings, ontology pairs with 1,001 to 3,000 mappings, ontology pairs with 501 to 1,000 mappings, ontology pairs with 101 to 500 mappings, ontology pairs with 3 to 100 mappings and ontology pairs with up to 2 mappings. In total we sampled 2,350 mappings from the different BioPortal mappings sets. For example, the sample for the evaluation of PPP mappings is made up of 400 mappings sampled from the 7 strata.

In the case studies involving cultural heritage and library subject headings we manually evaluated all mapping sets containing fewer than 500 mappings and took samples of 100 mappings from larger sets. We sampled the total of approximately 1,000 mappings from these sets.

Figure 5.2 Precision and number of input mappings (in parentheses) for the cultural heritage data (A) and library-track data (B).
5.3.4 Evaluation

Human experts evaluated the samples using the evaluation tool used in Chapter 3 for the cultural heritage and Library track data, and a similar tool for the BioPortal data. The raters categorized each mapping into one of six categories: exact match, close match, broader match, narrower match, related match, or incorrect. For measuring precision, we considered only exact and close matches as correct. A detailed analysis of the broader, narrower and related matches is out of scope of this chapter. We measured agreement using Cohen’s Kappa on subsets of samples between raters, finding substantial agreement for BioPortal (0.72) and cultural heritage evaluation (0.70) and almost perfect agreement with the manually evaluated mappings used in the OAEI library track (0.85).

5.4 Results

In this section, we present the precision of mapping composition for the three sets of ontologies in our study. We discuss these results in Section 5.5.

5.4.1 Results: Biomedical Ontologies

Fig. 5.3A shows the results for the overall precision of composed mappings. Using 575,642 input mappings with precision 0.94, we generated 599,625 composed mappings with an average precision of 0.92. Figures 5.3B, 5.3C, and 5.3D show the precision of composition for different cases from Fig. 5.1. We group these cases by the sets of input mappings that they used. Composing Preferred–Synonym mappings, which have a precision of 0.76, yielded 147,438 composed mappings with precision 0.84. Other combinations (Figures 5.3C and 5.3D) resulted in sets of composed mappings with precisions similar to the precisions of the input mappings.

Fig. 5.4 provides additional information on the precision of the individual cases. The two cases that resulted in the subset of what we could have obtained directly by comparing preferred names lexically (PPP and PSP), provided mappings with the highest precision, 0.99. The SPS mappings constitute a subset of the Synonym–Synonym mappings for $O_1$ and $O_3$. We did not use these types of mappings as input mappings because they have very low precision, 0.36. However, the SPS mappings have a higher precision (0.6) than the overall Synonym–Synonym mappings.

Additionally, using composition, we identified mappings without lexical similarity in their preferred names or synonyms (PSPS and PSSP mappings). Such mappings can be identified by composition through a concept with lexical similarity to both mapped concepts. These two cases produced 50,353 new mappings with the precision of 0.68. For example, we found a PSSP mapping between the concept CRANIAL_SKELETON from the Amphibian gross anatomy ontology and SKULL from the Foundational Model of Anatomy. These two concepts each map to the concept CRANIUM from the Teleost anatomy and development ontology, which has the synonyms CRANIAL SKELETON and SKULL.
5.4.2 Results: Cultural Heritage

Fig. 5.5 shows the results of mapping composition for the cultural heritage domain. The precision of composed mappings is at least 0.8 in all three cases, with the number of mappings identified through composition ranging from 263 to 1,774. The precision of the composed mappings between Cornetto and WordNet is the highest at 0.9.

Because we have lexical mappings available for this set, we can compare the composed mappings to the lexical ones, and analyze how many non-lexical mappings we generate by composing lexical mappings.

Upon closer examination of the mappings, we found that 134 (30%) of the composed mappings between AAT and WordNet have little or no lexical similarity. For example, through composition we mapped TOBACCONISTS’ SHOP to TOBACCO SHOP and WATCHMEN to GUARD. This subset has a precision of 0.56. Similarly, we found 110 non-lexical mappings between AAT and Cornetto
Figure 5.4  Mapping composition results for BioPortal ontologies. The bar graph shows precision for composed mappings. The (blue) left bar shows precision of exact and close matches, the (red) right bar shows the precision if we include broader, narrower, and related matches. Numbers in parentheses indicate the total number of mappings.

Figure 5.5  Mapping composition results for cultural heritage data. The numbers in bold outside the triangle show the precision of the composed mappings. The number of composed mappings is in parentheses. The numbers inside the triangle show the precision and number of input mappings.

with a precision of 0.57. Examples of such a mapping is the concept BADKLEDING mapped to BADKOSTUUM, both of which mean “bathing suit” in Dutch. Both of these subsets have a low precision compared to the overall precision of the composed mappings (0.82 and 0.80).

Between Cornetto and WordNet, 1,208 of the 1,774 composed mappings overlap with eqNearSynonym mappings in the original Cornetto-WordNet mappings of the Cornetto project. However, the precision of this subset of these 1,208 mappings is higher (0.98) than the overall precision of the original set of eqNearSynonym mappings (0.70). An additional 448 composed mappings do not overlap with existing Cornetto-WordNet mappings and have an average precision of 0.7.
Figure 5.6  Mapping composition results for the cultural heritage data. The bar graph shows precision for composed mappings. The (blue) left bar shows precision of exact and close matches, the (red) right bar shows the precision if we include broader, narrower, and related matches. Numbers in parentheses indicate the total number of mappings.

5.4.3 Results: The OAEI Library Track

Fig. 5.7 shows the results of mapping composition using the library subject headings mappings. Precision of the composed mappings is higher than 0.74 and the number of generated mappings ranges from 132 between the Subject Heading Authority File (SWD) and the Library of Congress Subject Headings (LCSH) and 266 between SWD and Rameau (a list used by the French National Library).

Figure 5.7  Mapping composition results for library-track data. The numbers in bold outside the triangle show the precision and the number of composed mappings in parentheses. The numbers inside the triangle show the precision and number of input mappings.
In two cases—mappings between SWD and LCSH and mappings between Rameau and SWD—the composed mappings had higher precision than the input mappings.

We also compared the composed mappings to the input mappings. We found that, of the 132 mappings between SWD and LCSH, 13 (10%) mappings do not overlap with any of the original instance-based mappings, including those that had a confidence measure lower than 0.7. In other words, for these 13 mappings, there were no instances (books) available. For LCSH and Rameau, we found 8 (5%) such “new” mappings, and for Rameau and SWD, 65 (24%) mappings. The high number of new composed mappings between Rameau and SWD is due to the low number of instances available for creating the original mappings. However, the precision of these subsets is lower: 0.37 between LCSH and Rameau, 0.54 between Rameau and SWD, and 0.92 between SWD and LCSH.

![Figure 5.8](image_url)  

*Figure 5.8* Mapping composition results for the library-track data. The bar graph shows precision for composed mappings. The (blue) left bar shows precision of exact and close matches, the (red) right bar shows the precision if we include broader, narrower, and related matches. Numbers in parentheses indicate the total number of mappings.

### 5.4.4 Broader, Narrower, and Related Mappings

During the evaluation of composed mappings, we also recorded whether each mapping represented a narrower, broader, or related mapping, rather than a close or exact match. Figures 5.4, 5.6 and 5.8 show the increase in precision of composed mappings if we also count broader, narrower, and related mappings as correct. The increase in precision in both cultural heritage and library track mappings is less dramatic than for the biomedical ontologies. For these mappings, the average increase in precision was 11%, whereas for BioPortal ontologies the average increase was 14%, with the most significant increase (30%) in the PSSP case.
5.5 Discussion

In this chapter, we have presented the results of our analysis of mapping composition in three different domains.

Our experimental results indicate that the quality of composed mappings is influenced by ontology characteristics, and the content and quality of the input mappings.

In the BioPortal experiment we found that the characteristics of the ontologies, such as when concepts are an aggregate of narrower terms and a broader term, lead to lower quality composed mappings. For example, in Medical Subject Headings (MeSH) concepts often have narrower terms as synonyms. The concept *TREMORS* in MeSH has a synonym *Nerve tremor*, which in reality is a narrower term, not a synonym. As a result, many of the composed mappings that involved MeSH terms are not close matches, but rather broader or narrower mappings.

The content of the ontologies also influences the quality of the mapping compositions. When the content overlaps, meaning the domains of the ontologies are the same or very similar, the meaning of the concepts is also closer, and the composed mappings are likely to be equivalence mappings rather than broader, narrower or related mappings. In the cultural heritage case study Cornetto and WordNet are unlikely to cover art and architectural concepts, reducing the chance of creating equivalence compositions between AAT and Cornetto, and AAT between AAT and WordNet.

We found large variations in the ratio of number of input mappings and to the number of composed mappings. We can measure this ratio by dividing the number of composed mappings by the smallest set of the two sets of input mappings. Our assumption is that the smaller set of the input mappings provides the upper boundary to the number of composed mappings. This ratio is easy to measure in the cultural heritage and Library track case studies. For example, in the cultural heritage case, the composed mappings between Cornetto-WordNet (1,774) represent 39% of the smallest AAT-WordNet input mappings (4,592) which is smaller than the AAT-Cornetto set (6,914). The composed mappings between Cornetto-WordNet (263) represent only 8% of Cornetto-Wordnet input mappings (3,144). There is a similar variation in ratio of input and composed mappings in the Library track case study. This is partly due to the size of the ontologies and partly due to the content overlap of the input mappings. In the Library track case study the input mappings map less than 1.5% of all concepts in the vocabularies, therefore, the likelihood that the two sets of input mappings overlap and thus form compositions is low. Thus, the vocabulary coverage of input mappings appears to influence the quantity of composed mappings.

Finally, we found the quality of composed mappings to be unexpectedly high. Intuitively we expect the precision of composed mappings to be the product of the precision of input mapping, and thus lower than those of the input mapping. However, we found that in almost all cases the precision of composed mappings is higher than expected, and in some cases significantly higher. Examples of this are the *SPS* mappings in the Bioportal case study (precision of 0.75 vs. the expected 0.58), and the SWD-LCSH mappings in the Library track case study (precision of 0.89 vs. the expected 0.51). We are unable to account for these phenomena as it would require an in
depth qualitative analysis of large portions of input and composed mappings that is out of scope of this chapter. However, it appears that composed mappings are formed more often from higher than average quality input mappings. There is possibly a minimal quality level for input mappings below which the quality of composed mappings starts to decline.

We also found that many of the composed mappings though not exact or close matches, nevertheless represent a semantic relationship such as broader, narrower or related (Figure 5.4 and 5.8). For example, the concept **BLURRED VISION** from the “Suggested Ontology for Pharmacogenomics” maps to the concept **VISION ABNORMAL** in “MedDRA”, forming a narrower relationship between the two concepts. This composition occurs because both concepts were originally mapped to a single concept in in the “WHO Adverse Reaction Terminology” that has both **BLURRED VISION** and **VISION ABNORMAL** as labels. This kind of semantic drift between concepts seems to arise often through mapping composition caused by ontology characteristics, or concepts deviating in meaning in different languages.

### 5.6 Conclusion

In this chapter, we presented an empirical analysis of the quality of mapping composition in three case studies. Although the domain, ontologies and type of mappings was different in each study we can draw a number of conclusions. First, we find that ontology characteristics, such as the way concepts are modeled influences the quality of composed mappings. If an ontology contains concepts where broader and narrower terms are aggregated into a single concept, then the composed mapping created with mappings to this ontology will not be an equivalent mapping. Such mappings lead to “semantic drift”. Second, an overlap in content of the ontologies, both with respect to domain and granularity, is likely to yield higher quality composed mappings. Third, the quantity of composed mappings partly depends on the coverage of the input mappings with respect to the mapped ontologies and on the domain overlap of the ontologies. Our results confirmed some of our intuitions on mapping composition with one notable exception: The precision of composed mappings is often higher than the expected product of input mapping precision. Although we are unable to account for this phenomenon in this chapter, it provides interesting avenues for further research in this area.

### Acknowledgments

Many thanks to Guus Schreiber and Bob Wielinga for their help and comments on this paper. We thank Antoine Isaac, Shenghui Wang and the Telplus project for providing the library track data and their mappings, and for the many explanations they provided. We also thank the Cornetto project for providing data and mappings and Babette Heyer for helping with the evaluation of the mappings. This work was supported in part by the National Center for Biomedical Ontology, under roadmap-initiative grant U54 HG004028 from the National Institutes of Health, and partly by the EC eContentplus programme via de EuropeanaConnect project.
Manual Assessment of Vocabulary Alignments

The evaluation of alignments is an integral part of the vocabulary alignment process. Manual evaluation is an important method for assessing the quality of mappings. In order for the evaluation to be consistent, multiple raters need to assess mappings in the same way; they need to agree, at least to some point, on the quality. In various fields such as psychology and computational linguistics, consensus between raters is measured using statistics such as Cohen’s Kappa. In these fields a Kappa of 0.8 or higher is considered good agreement, and a Kappa between 0.7 and 0.8 is considered an acceptable level of agreement. Yet, in Chapter 3 and 4 we found that agreement between raters was frequently lower than this minimum level of 0.7. In the vocabulary alignment field the general assumption is that humans are good at evaluating alignments. We need to analyze the differences between raters and study the causes for disagreement.

This chapter is based on a paper coauthored with Jacco van Ossenbruggen, Guus Schreiber and Bob Wielinga, “Let’s Agree To Disagree: On the Evaluation of Vocabulary Alignments” (Tordai et al. 2011), which was presented at the sixth Knowledge Capture Conference (K-CAP 2011) in Banff, Alberta, Canada.

6.1 Introduction

In this chapter we study the process of manual evaluation of vocabulary alignments. Manual evaluation is a fundamental method for establishing quality in ontology and vocabulary alignment and many other fields such as information retrieval and linguistic research. In vocabulary matching evaluators rate the quality of mappings by assigning them into categories, thus creating a gold standard, also called a reference alignment, that is then used to assess the overall quality of an alignment or alignment algorithms. An established method of validating the gold standard is to have multiple raters evaluate the same set of mappings into categories. Agreement between raters is then measured by correcting for chance agreement using measures such as Cohen’s Kappa (Cohen 1960). Given a high enough inter-rater agreement measure the results of the manual evaluation can be used as a gold standard. However, what the threshold of agreement should be is not clear cut and also depends on the research field in question (Landis and Koch 1977, Carletta 1996, Artstein and Poesio 2008).
While evaluation by multiple raters is a preferred validation method, it is not always documented in practice. The focus of evaluation reports frequently lies on the performance of evaluated tools. In cases where inter-rater agreement measures have been used in the manual evaluation, the reported levels of agreement diverge greatly. For example, in the Very Large Cross-Vocabulary track of the Ontology Alignment Evaluation Initiative (OAEI)\(^1\) organizers reported perfect agreement between raters (Euzenat et al. 2009). Halpin et al. (2010) however reported very poor agreement levels in their experiments evaluating **owl:sameAs** mappings sampled from Linked Data. In Chapter 3 and Chapter 4 we also measured interrater agreement and found only moderate levels of agreement between raters which we found unexpectedly low. As manual evaluation is such an integral part of the evaluation process we have asked ourselves why raters find it so difficult to agree on relationships between concepts. In this paper we will focus on the following research questions:

1. What is the level of agreement between raters when evaluating alignments?
2. If agreement is low, what are the reasons behind it?

To this end we perform three evaluation experiments on mappings between two sets of vocabularies and analyze the results quantitatively as well as qualitatively. Because our experiments were explorative in nature, we only evaluate small sets of mappings and focus on qualitative analysis in particular. As part of our experimental setup we create specific guidelines detailing evaluation categories and provide examples and further explanations to raters. We perform a quantitative analysis by using established measures such as Cohen’s Kappa and Krippendorff’s Alpha and analyze data from “think aloud” sessions during the experiments.

### 6.2 Related Work

There are relatively few research papers on vocabulary alignment that detail an evaluation by multiple raters and include inter-rater agreement measurements. In Chapters 3 and 4 we described case studies of alignments between various vocabularies, and manually evaluated resulting mappings. In these case studies we validated our evaluation by asking three raters to evaluate samples of the alignments. We used an evaluation tool to display mapped concepts along with their immediate hierarchies, scope-notes and labels, and had raters select a SKOS matching relation (Miles and Bechhofer 2009) to categorize the mapping. To support raters in this task we provided a set of guidelines which included a short description of each matching relation based on the W3C recommendation and examples of mappings. We measured Cohen’s Kappa and found moderate agreement (0.56) between raters in our first report (Chapter 3), and just slight agreement (0.36) in Chapter 4. As our goal in both case studies was to assess precision with regards to equivalence, we reduced the number of categories into equivalent or not equivalent. With just two categories the inter-rater agreement rose to substantial agreement (0.70 and 0.67). From these values we

\(^1\)http://oaei.ontologymatching.org/
concluded that the evaluation task is difficult even for humans in particular when more than two categories of agreement are used. Further study revealed that raters’ understanding of SKOS matching relations varied from person to person. We also found that the lexical richness of vocabularies such as WordNet may contribute to the difficulty level of the evaluation task, as closely related senses are separated into different concepts. Also, a clearer delineation between mapping relations would likely raise agreement.

Halpin et al. (2010) also reported low levels of agreement in their paper. They analyzed the use of owl:sameAs mappings in Linked Data and defined a similarity ontology to differentiate between various degrees of similarity. In their evaluation experiment they defined 5 levels of similarity relations between entities and used them to evaluate mappings. The agreement level between raters was very low with Kappa of 0.16, which the authors attributed to different styles of judgments. After a recombination of the rating categories into three the agreement increased to 0.32, which is still lower than what we experienced. They found that raters had the most difficulty in defining whether two entities were the same, and that background knowledge has an impact on decisions. They concluded that this inability to rate entities as the same stemmed from not knowing how the entities would be used. In the mapping categorization instruction Halpin et al. used variations on the same type of entity to illustrate each mapping category (descriptions of performances of Bohemian Rhapsody by Queen or some other band). In richly varied data such as Linked Data mapping categories need to be defined in more general terms with examples varying in domain and type. Raters then have less need to interpret examples themselves.

While guidelines with clearer descriptions of categories could improve inter-rater agreement, it is clear from these reports that the task of manual evaluation is difficult. Manual evaluation of mappings is a type of categorization task. Studies in cognitive science in general (Lakoff 1987) and linguistic categorization in particular (Taylor 2003) have shown that humans do not categorize according to the classic Aristotelian view, where each category is clearly defined and categories are mutually exclusive. Instead, Lakoff argues in his book (Lakoff 1987) that prototype theory is at the core of cognitive categorization whereby some members of a category are more central (prototypical) than others. For example, chair is a more prototypical member of the category furniture than side-table. Categories thus form a graded cloud with fuzzy boundaries where member concepts do not necessarily share common properties. They are defined by culture and experience and therefore vary from person to person. This fuzzy nature of categories provides an insight into why categorization tasks can be difficult.

6.3 Experimental Setup, Tools and Methods

6.3.1 Experimental Setup

Our first experiment, AATWordNet is a replication of our mapping evaluation described in Chapter 4. Because in our earlier evaluation the inter-rater agreement was low, one of the objectives for this experiment is to increase agreement by improving the description of matching categories.
As summarized in Table 6.1, in this experiment we ask 5 raters to evaluate a sample of 74 lexical mappings between the Getty’s Art and Architecture Thesaurus (AAT)\(^2\) (Peterson and Jackman-Schuller 1996) and Princeton WordNet version 2.0 (Fellbaum 1998). The lexical mapping is based on string matching between preferred and/or alternative labels of concepts (see Lexical tool in Chapter 4).

In the second experiment, \textit{GTTinstance}, we aim to rule out lexical mappings to WordNet as the cause of low inter-rater agreement due to WordNet’s ambiguous word senses. We choose a different set of mappings created by a different alignment technique for our second experiment. Raters have to evaluate 70 mappings, which were created using instance-based matching between the Dutch Royal Library’s Gemeenschappelijke Trefwoordenthesaurus (GTT) and Brinkman Thesaurus, two subject heading thesauri. Instance-based matching is based on instances, in this case books commonly annotated by thesauri. In this case books commonly annotated by concepts from both vocabularies.

In our last experiment, \textit{GTTlexical}, we want to study lexical mappings between less ambiguous vocabularies than WordNet, and determine whether evaluation is easier when the two vocabularies are from the same domain. In this experiment we have 5 raters evaluate 75 lexical mappings between GTT and Brinkman.

In each experiment we provide raters with written guidelines on how to categorize mappings which include descriptions of the mapping categories and example mappings. We ask raters to evaluate mappings using our evaluation tool into 7 different categories. Additionally, we ask raters to “think aloud” by explaining their choice of categories and their application of the guidelines, which we transcribe. We then calculate the inter-rater agreement measurements for 7 categories, and for 2 categories by aggregating the original categories into equivalent and non-equivalent mappings. We then perform detailed analysis of the evaluations and of the raters’ comments. We describe the matching categories, guidelines and vocabularies in more detail in the next section.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Vocabularies</th>
<th>Alignment technique</th>
<th>Number of raters</th>
<th>Number of mappings</th>
</tr>
</thead>
<tbody>
<tr>
<td>AATWordNet</td>
<td>AAT and WordNet</td>
<td>lexical matching</td>
<td>5</td>
<td>74</td>
</tr>
<tr>
<td>GTTInstance</td>
<td>GTT and Brinkman</td>
<td>instance based matching</td>
<td>5</td>
<td>70</td>
</tr>
<tr>
<td>GTTlexical</td>
<td>GTT and Brinkman</td>
<td>lexical matching</td>
<td>5</td>
<td>75</td>
</tr>
</tbody>
</table>

\textit{Table 6.1} Overview of the three evaluation experiments.

\(^2\)http://www.getty.edu/research/conducting_research/vocabularies/aat/
6.3.2 Tools and Methods

SKOS relations and guidelines

We use the SKOS mapping properties to categorize the type of mappings. The `skos:exactMatch` and `skos:closeMatch` properties make statements about the degree of equality between two concepts. Hierarchical relations are expressed using `skos:broadMatch` and `skos:narrowMatch`, and `skos:relatedMatch` expresses an associative relation between mapped concepts. In addition to these relations, we define a property to indicate that there is no relation between the mapped concepts: `unrelated`. We also give raters the option to choose `unsure` when they are unable to choose between the relations above. (In the remainder of this chapter we refer to these properties in short form, i.e., exact instead of `skos:exactMatch`.)

As remarked earlier, we found in previous experiments (Chapters 3, 4) that raters diverged greatly in the way they selected mapping properties. We attributed this divergence to an unclear differentiation between mapping relations. For example, we found in earlier experiments that raters varied greatly in their application of related. Some raters were more strict in their use of the related category than others. For this reason we wrote guidelines on the use of each mapping property for these experiments. Our rationale was to differentiate between each property as much as possible by describing them both in general terms, and by giving specific examples. For example, here is an excerpt from the guidelines for related:

“**Related:** The two concepts have an associative relationship and are of two different (ontological) types. For example: a material and an object made from that material, such as milk and cheese, or an activity and object involved in the activity, such as the game volley ball, and a volley ball. Generic examples of such relationships are: process and agent, action and property (e.g., environmental cleanup and pollution), action and product (e.g., weaving and cloth), cause and effect, object and origin, material and object, and object and practitioner.”

We also define the difference between exact and close, and instructed raters to use the close relation when the two concepts share the same label, but their parent concepts are different, as the vocabularies have different organizational schemes. An example of a close relation is the concept BLOWGUN, where in one vocabulary it is a conduit and in the other it is a weapon. The two vocabularies present blowgun in different views: a structural view versus a functional view. The full guidelines can be found in Appendix A.2.

In typical alignment evaluation settings researchers are interested in equality relations between concepts. In such cases the evaluation categories are equivalent and non-equivalent. Although we did not perform separate experiments with these categories, we reduce the number of categories to two by summing up ratings of exact and close into the equivalent category and the remainder into the non-equivalent category.
Vocabularies and mappings

In our AATWordNet experiment we used a sample of mappings between AAT and WordNet. We generated the mappings using an exact string matching technique to match preferred and alternative labels in AAT to labels in WordNet (Lexical tool in Chapter 4). AAT is an ISO standard compliant vocabulary (International Organization for Standardization 1986) that we converted to SKOS, where each concept has preferred and alternative labels and is often accompanied by scope notes. WordNet contains synonymous labels grouped into synsets with no distinction between preferred and alternative labels. The meaning of each synset is clarified by glosses containing example sentences or a definition. Because in WordNet multiple synsets may share the same label, many of the lexical mappings between AAT and WordNet are ambiguous.

For our GTTInstance and GTTlexical experiments we used samples of mappings between the GTT and Brinkman thesaurus. Both thesauri are subject-heading vocabularies used to annotate the Dutch Royal Library’s book collection, and both include not only general descriptors but also geographic terms. Thus, the two vocabularies have the same purpose, although they differ in size and granularity: GTT is five times larger than Brinkman. Brinkman contains 13,025 concepts while GTT contains 65,297 concepts. The mappings we used in the second experiment were created within the STITCH project, using an instance-based matching technique described by Isaac et al. (2007). The sample of mappings we used in the GTTInstance experiment have no linguistic similarity, because we filtered out all concepts with matching labels. We excluded mappings based on lexical similarity from the second experiment, because we want to study agreement between raters evaluating concepts that do not share labels. For the third experiment, GTTlexical, we generated mappings between the two thesauri through lexical comparison of concept labels.

Interrater Agreement Measures

The simplest method for measuring agreement between raters is the percentage of agreement: observed agreement. Unfortunately, this measure is difficult to interpret and compare across multiple experiments (Carletta 1996, Artstein and Poesio 2008), because it does not take into account agreement that occurs by chance. A number of measures exist that do correct for chance agreement. Cohen’s Kappa (Cohen 1960) is used to measure agreement between two raters on nominal data. Fleiss’ Kappa (Fleiss 1971) is a generalization of Scott’s Pi Scott (1955) and measures agreement between multiple raters on large sample sizes. Krippendorff’s Alpha (Krippendorff 2007), a more versatile measure, can be used with nominal, ordinal and interval type categories and even with missing data (for example when raters select unsure). Weighted Cohen’s Kappa allows us to count disagreements differently by using a weight matrix. The latter can be used, for instance, to count disagreement between exact and unrelated heavier than a disagreement between exact and close. All agreement measures use the observed agreement (or disagreement in the case of Krippendorff’s Alpha) that is the number of times raters agree, and an estimate of what the agreement would be if raters had assigned categories randomly.

3http://www.cs.vu.nl/STITCH/
These measures have two known problems: prevalence bias and annotator bias. Prevalence bias occurs when data falls into mostly one category: even if observed agreement between raters is high, the agreement measure may turn out low. In order to have a high measure the raters must agree on rare categories. Annotator bias occurs when the distribution of disagreement is highly skewed, leading to lower measures than when disagreements are more uniformly distributed.

While these measures are widely used, their interpretation is not clear cut. In the social science field Landis and Koch (1977) suggest the set of intervals displayed in Fig. 6.1 based on their personal opinion. For content analysis Krippendorff (2004) recommends Alpha values of 0.8 or higher, although values higher than 0.7 can be acceptable as the absolute minimum. In the field of computational linguistics Artstein and Poesio (2008) also set the minimum agreement threshold at 0.8 because they found that at that level annotations in corpora were of reasonable quality.

![Figure 6.1 Interpretation of Kappa values according to Landis and Koch. The scale is from -1 to 1](image)

Evaluation Tool

We used an evaluation tool shown in Fig. 6.2 to support raters in their judgment. Its user interface presents mappings with context information, such as the concept hierarchy, labels and scope notes. Raters can select mapping relations between concepts. The resulting choices are stored in RDF using the OAEI alignment format (Euzenat 2004) along with provenance information. This tool is a newer version of the one we used in Chapter 4. It is open source and available for download.

6.4 Experimental Results

6.4.1 Quantitative Results

The inter-rater agreement measures for our three experiments are displayed in Table 6.2. The AATWordNet experiment is a replication of the experiment described in Chapter 4 where we measured Cohen’s Kappa between pairs of raters and reported average measures of 0.36 for 7 categories and 0.67 for 2 categories with three raters. In the AATWordNet experiment we have an average Cohen’s Kappa of 0.564 for 7 categories which is a considerable increase over 0.36. We attribute this increase to a better description of the mapping categories in the guidelines, as we used the same evaluation tool and had five raters instead of three. There was however no improvement.

---

4[http://semanticweb.cs.vu.nl/amalgame/]
Figure 6.2 A partial screenshot of the evaluation tool used by raters. The screenshot shows a mapping between the concept RESTORERS from AAT and PRESERVER from WordNet. The labels and scope-notes are found in the upper boxes. This mapping caused high disagreement between raters.

in the average Cohen’s Kappa for 2 categories, which suggests that the guidelines did not help raters in making a distinction between equality and inequality.

Overall, we found higher agreement measures for 2 categories than for 7 categories which suggests that it is easier for raters to reach agreement over fewer categories.

The inter-rater agreement in the GTTlexical experiment is the lowest of all our experiments, despite the highest observed agreement (0.85). This is caused by prevalence bias, as 85% of the ratings fall into the exact match category (see Table 6.3). The prevalence of one category causes the disagreement on the rare categories to weigh more heavily when measuring agreement. In the other experiments the distribution in the use of relations is less extreme than in GTTlexical.

We also found that the value of Fleiss’ Kappa is close to the average Cohen’s Kappa in all three experiments. Krippendorff’s Alpha is a bit higher for both AATWordNet and GTTInstance experiments because it does take into account missing values, in this case the use of the unsure category (see Table 6.3 for the distribution of mapping relations per experiment). In the GTTlexical experiment, where the unsure category was not used by raters, the value of Krippendorff’s Alpha is equal to the value of Fleiss’ Kappa.

We measured Cohen’s Kappa for each pair of raters. Table 6.4 shows the values for the GTTInstance experiment with the highest value between Rater 4 and Rater 5 (0.783) and the lowest between Rater 3 and Rater 4 (0.483). The large difference is due to Rater 3’s tendency to select related match for mappings that Rater 4 and 5 considered to be unrelated. We found a similarly high variation in the Cohen’s Kappa in AATWordNet and GTTlexical experiments which leads us to conclude that two raters are not enough to provide a consistent evaluation result.

Table 6.3 shows that in each experiment raters selected categories in very different distributions. In the GTTInstance experiment raters rarely selected the exact or close match categories. This was caused by the mapped concepts having no labels in common, as such mappings had been filtered out, therefore equivalent mappings were rare.
<table>
<thead>
<tr>
<th>Experiment</th>
<th>Observed agreement</th>
<th>Avg. Cohen’s $\kappa$</th>
<th>Fleiss’ $\kappa$</th>
<th>Krippendorff’s $\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>7 categories</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AATWordNet</td>
<td>0.69</td>
<td>0.564</td>
<td>0.565</td>
<td>0.575</td>
</tr>
<tr>
<td>GTTInstance</td>
<td>0.72</td>
<td>0.606</td>
<td>0.604</td>
<td>0.617</td>
</tr>
<tr>
<td>GTTlexical</td>
<td>0.85</td>
<td>0.473</td>
<td>0.475</td>
<td>0.475</td>
</tr>
<tr>
<td>2 categories</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AATWordNet</td>
<td>0.84</td>
<td>0.666</td>
<td>0.669</td>
<td>0.679</td>
</tr>
<tr>
<td>GTTInstance</td>
<td>0.94</td>
<td>0.706</td>
<td>0.698</td>
<td>0.699</td>
</tr>
<tr>
<td>GTTlexical</td>
<td>0.95</td>
<td>0.514</td>
<td>0.538</td>
<td>0.538</td>
</tr>
</tbody>
</table>

**Table 6.2** Inter-rater agreement table for 7 mapping categories and for 2 categories. Measures include observed agreement between raters, the average of Cohen’s Kappa measured between each pair of raters, Fleiss’ Kappa over all raters and Krippendorff’s Alpha over all raters.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Exact</th>
<th>Close</th>
<th>Broad</th>
<th>Narrow</th>
<th>Related</th>
<th>Unrelated</th>
<th>Unsure</th>
</tr>
</thead>
<tbody>
<tr>
<td>AATWordNet</td>
<td>31.9</td>
<td>6.5</td>
<td>4.1</td>
<td>5.4</td>
<td>10.8</td>
<td>40.8</td>
<td>0.5</td>
</tr>
<tr>
<td>GTTInstance</td>
<td>5.7</td>
<td>3.7</td>
<td>7.7</td>
<td>4.6</td>
<td>39.1</td>
<td>38.3</td>
<td>0.8</td>
</tr>
<tr>
<td>GTTlexical</td>
<td>85.0</td>
<td>9.1</td>
<td>1.0</td>
<td>1.8</td>
<td>1.0</td>
<td>2.1</td>
<td>0.0</td>
</tr>
</tbody>
</table>

**Table 6.3** Distribution of SKOS matching relations used by raters in percentages. The ratings in each category are summed over all 5 raters.

<table>
<thead>
<tr>
<th>Cohen’s $\kappa$</th>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Rater 3</th>
<th>Rater 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater 2</td>
<td>0.736</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rater 3</td>
<td>0.534</td>
<td>0.665</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rater 4</td>
<td>0.634</td>
<td>0.566</td>
<td>0.483</td>
<td></td>
</tr>
<tr>
<td>Rater 5</td>
<td>0.577</td>
<td>0.592</td>
<td>0.491</td>
<td>0.783</td>
</tr>
</tbody>
</table>

**Table 6.4** Cohen’s Kappa between each pair of raters for 7 categories from the GTTInstance experiment. The highest and lowest agreement is displayed in bold.
Table 6.5  The matrix of relations is the sum of the coincidence matrices from each experiment. The matrix shows the number of pairs of rating used by two raters for the same mapping. We consider it worse when raters mark opposing categories such as exact and unrelated than the categories exact and close. The numbers are percentages of the total amount of observations in the three experiments (4,420) and the numbers in bold represent agreements. Note, the table is symmetric across the diagonal.

We examined the judgment of raters focusing on disagreements. The interrater-agreement measures when used on nominal data assume independence between categories. However, intuitively it is worse if raters disagree whether a mapping is an exact match or unrelated, than when they disagree on whether it is related or unrelated. Table 6.5 shows a matrix of coincidence of relations summed over all experiments. The agreements are along the diagonal, while the disagreements occupy the other cells. In Table 6.6 we isolated the pairs of relations raters disagreed upon, and ordered them according to the number of times they occurred in the experiments. The table shows that disagreements that can be considered the least “harmful” (i.e., the smallest semantic distance between the mapping relations involved) are the most frequent, such as the disagreements on related-unrelated and on exact-close. An analysis of the cases where raters selected “opposed” categories for the same mapping showed that it was mostly caused by one rater making a mistake.

Our main observation is that the inter-rater agreement measures are stable across our experiments. Although the inter-rater measures are relatively low, our analysis showed that most disagreements between raters are of the less “harmful” type.

### 6.4.2 Qualitative Results

We analyzed the use of SKOS matching relations by raters along with the reasons raters gave for their choices transcribed in the “think-aloud” sessions.

We found that overall raters selected different types of relations for lexical mappings than for non-lexical mappings. In the AATWordNet experiment mapped concepts share at least one label. Table 6.3 shows that most of the mappings were rated as either exact match or unrelated while the hierarchical relations (broader and narrower) were least frequently used. When mapped

<table>
<thead>
<tr>
<th>Category</th>
<th>Exact</th>
<th>Close</th>
<th>Broad</th>
<th>Narrow</th>
<th>Related</th>
<th>Unrelated</th>
<th>Unsure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exact</td>
<td>37.30</td>
<td>4.88</td>
<td>0.59</td>
<td>1.72</td>
<td>1.54</td>
<td>0.63</td>
<td>0.13</td>
</tr>
<tr>
<td>Close</td>
<td>4.88</td>
<td>2.17</td>
<td>0.86</td>
<td>1.17</td>
<td>1.36</td>
<td>0.72</td>
<td>0.04</td>
</tr>
<tr>
<td>Broad</td>
<td>0.59</td>
<td>0.86</td>
<td>2.22</td>
<td>0.04</td>
<td>0.68</td>
<td>1.49</td>
<td>0.04</td>
</tr>
<tr>
<td>Narrow</td>
<td>1.72</td>
<td>1.17</td>
<td>0.04</td>
<td>1.63</td>
<td>0.68</td>
<td>0.81</td>
<td>0.09</td>
</tr>
<tr>
<td>Related</td>
<td>1.54</td>
<td>1.36</td>
<td>0.68</td>
<td>0.68</td>
<td>11.10</td>
<td>5.84</td>
<td>0.32</td>
</tr>
<tr>
<td>Unrelated</td>
<td>0.63</td>
<td>0.72</td>
<td>1.49</td>
<td>0.81</td>
<td>5.84</td>
<td>21.82</td>
<td>0.27</td>
</tr>
<tr>
<td>Unsure</td>
<td>0.13</td>
<td>0.04</td>
<td>0.04</td>
<td>0.09</td>
<td>0.32</td>
<td>0.27</td>
<td>0.00</td>
</tr>
<tr>
<td>Categories disagreed upon</td>
<td>Occurrences (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------------</td>
<td>-----------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. related-unrelated</td>
<td>258 (24.39)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. exact-close</td>
<td>192 (20.42)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. exact-narrow</td>
<td>76 (7.18)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. exact-related</td>
<td>68 (6.43)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. broad-unrelated</td>
<td>66 (6.24)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. close-unrelated</td>
<td>32 (3.02)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12. exact-unrelated</td>
<td>28 (2.65)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18. broad-narrow</td>
<td>2 (0.19)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6.6  Partial list of disagreements in mapping categories ordered by total number of occurrences in the three experiments. The total number of disagreements is 1,068 and the number in parentheses is the percentage of total disagreements.

Concepts are not equivalent; they are either polysemes or homonyms. Homonyms have labels with the same spelling but the concepts are unrelated (e.g., BENDS; the act of bending and the decompression sickness). Polysemes are terms with different but related meanings, such as MILK, being the product and the act of milking. In our experiments, raters found polysemes difficult to rate because the boundary between relatedness and unrelatedness is not clear. In particular in WordNet, concepts are specifically divided into word senses thus distinguishing between various polysemic and homonymic forms. As a result, when raters evaluate mappings to WordNet they are confronted with multiple related and unrelated word senses. In our AATWordNet experiment raters found the concept FLOW from AAT one of the most difficult to evaluate, because it was mapped to 14 word senses in WordNet. This problem of distinguishing between meanings is not restricted to polysemes. In the GTTinstance experiment raters had the most difficulty in deciding whether mapped concepts were related or unrelated. For example, some raters found the concept ARID, DRY TERRITORY related to EROSION, while others thought the link too remote to be useful. The fuzziness of concept boundaries and category boundaries makes agreement in evaluation more difficult to achieve. They are a manifestation of prototype theory where concepts far from the prototype become more difficult to categorize.

We also found that the contextual information such as hierarchy, multiple labels and scope note can increase the difficulty in judgments in particular when they are contradictory. For example, the categorization of the mapping of concept MANTEL between AAT and WordNet, both referring to the thing around a fireplace, was complicated by the AAT scope note “Decorative frames around fireplace openings” and the WordNet gloss “shelf that projects from wall above fireplace”. Three out of 5 raters judged the mapping exact match. The fourth rater judged it related because the AAT parent concept is FURNITURE COMPONENT while the WordNet parent concept is SHELF. The
fifth rater judged the AAT term as broad as she considered the frame in the AAT scope note to be a broader than a shelf.

In comparison to AATWordNet, the GTTlexical experiment was judged “easier” by raters, as they were confronted with very few ambiguous mappings and made quicker judgments. In addition, both GTT and Brinkman contain few alternative labels and scope notes limiting the amount of contextual information. In the GTTlexical experiment, where mapped concepts had the same label, raters tended to select exact match due to lack of context. In the GTTinstance experiment, however, the lack of context meant that concepts that were related were sometimes rated as unrelated by some raters. Both GTT and Brinkman vocabularies cover a wide range of subjects from “general culture” to economy, physics, history and even medicine. The evaluation of concepts from more specific domains was more difficult due to lack of context. For example, the mapping between the drug DAPSONE and the disease LEPROSY was rejected by some raters because the parent concept of Dapsone is ANTI-EPILEPTIC DRUG. Other raters looked up Dapsone on Wikipedia, found that it is also a drug used for leprosy and selected related match because the therapy and disease have an associative relationship. In this case, if we had prohibited the use of outside sources such as Wikipedia, all raters would have most likely selected unrelated based on the available information (unless one of them was a medical expert), which would have led to higher inter-rater agreement. However, a related match between the two concepts can be useful in some applications. Our experiments have shown that raters behave differently and some are more inclined to look up information than others.

We found that for some mappings raters thought the SKOS matching relations inadequate. Although raters could use the unrelated category whenever the mapping was not a SKOS relation, they were reluctant to reject mappings with some semantic link. In particular in the AATWordNet and GTTlexical experiments some aligned concepts (partially) overlapped each other in meaning, therefore warranting some sort of equivalence relation that could not be defined as exact or close match. An example from AATWordNet is the concept RESTORER as shown in Fig. 6.3. The WordNet concept for restorer also included the labels “refinisher”, “renovator” and “preserver”, whereas RESTORERS and PRESERVATIONISTS were separate concepts in AAT. Raters were reluctant to reject the mapping from AAT’s RESTORER to WordNet’s RESTORER, but felt neither exact nor close match was appropriate.

A complementing explanation of possible differences between raters could be based on the variability of subjective guidelines that raters appear to construct during the evaluation task. This view is supported by the notion of situated cognition (Clancey 1997) that stipulates that people construct their knowledge “on the fly” in a specific context. When raters were confronted with a non-prototypical mapping they formed their own interpretation of the guidelines and applied that particular rule to similar mappings. For example, one rater created the following rule during the GTTinstance experiment: “if two concepts are not on the same level of specialization they cannot be related”. The rater continued to apply this rule throughout the evaluation, even though our guidelines did not contain such a specific rule, and none of the other raters formulated it so clearly.

The background of raters also had an impact on their process of categorization. Two of the
Figure 6.3 Fuzzy boundaries between AAT and WordNet concepts. The WordNet concept of *preserver* overlaps with multiple concepts from AAT through its labels: *Preservationists* and *Restorers*, but not with *Conservators*.

Raters had a strong thesaurus background that influenced some of their choices. For example, one of these raters would not categorize mappings as broad or narrow match if they shared the same label, commenting that it is not proper ISO standard practice (International Organization for Standardization 1986). Raters without this background had no reservations in using hierarchical categories on mappings that shared the same label. We did not specify a purpose or task for the alignments but it seems that raters with a thesaurus background thought of the mappings in terms of a thesaurus merging task, while other raters thought of mappings in terms of an annotation task. Our findings are similar to those reported by Bailey et al. (2008) in the field of information retrieval, where they found that judges with different levels of specialization in the task had low agreement.

In the experiments we found that certain disagreements were caused by different interpretations of differences in thesauri, in particular in their hierarchy. GTT is organized according to is-a type relations, while in Brinkman concepts are organized according to a mix of is-a and part-of relations. This sometimes caused problems when the relation a rater wanted to choose contradicted the relation in the thesaurus. For example, there were disagreements on the mapping between *Waste products* from Brinkman to *Environmental pollution* in GTT. Some raters
chose related match in accordance with our guidelines about cause and effect, while other raters chose broader because in Brinkman ENVIRONMENTAL POLLUTION is the parent concept of WASTE PRODUCTS. Such disagreements can be avoided by adding additional guidelines, but in practice it is often impossible to foresee the effect specific differences between thesauri have on evaluation. Further examples of disagreements can be found in Appendix B.1.

6.5 Discussion and Conclusion

Manual evaluation is a method for establishing the quality of vocabulary alignments. High agreement between raters is a requirement for being able to make conclusive statements about the quality of alignment methods. However, there are a number of factors that influence the judgment of raters. In this chapter we studied the process of manual evaluation, and found that there are aspects of the evaluation setup that can be controlled, and aspects that make the task inherently difficult.

One aspect that can be controlled is the provision of clear guidelines to the raters. Guidelines should include clear examples, precise descriptions of the categories, and instructions how to deal with thesaurus errors. The granularity of the categories is another factor where choices can be made. In our experiments we chose to use the SKOS mapping categories, but our results show that a two category system (match/no match) leads to higher reliability measures. On the other hand, some of our raters indicated that they found the SKOS categories too limited.

Another aspect that can be controlled is the nature of the sample. One can simply choose to select a random sample of alignments, or one can construct a sample that contains certain types of alignments as we did in the GTTinstance evaluation. Although our results indicate that different samples lead to similar values for inter-rater reliability, the choice of a specific sample can circumvent certain problems such as prevalence bias.

Aspects largely beyond control are lexical ambiguity and rater characteristics. It is well known from studies in lexical semantics that the boundary between polysemy and homonomy is vague, and that the classification of different types of polysemy is still a matter of debate among linguists. Humans rarely have problems with disambiguating the meaning of words in a discourse context. However, in an ontology alignment task this context is usually much more limited than in discourse.

The evaluation process can also be influenced by the background of the raters of alignments. Domain specialists (e.g. in a medical or cultural heritage domain) may use different evaluation criteria than raters with a linguistic or a computer science background. Of course one can choose to select raters with a similar background.

A related factor is the purpose of the ontology alignment. For example, if the aligned concepts are used to retrieve documents annotated by different ontologies in the domain of medicine, the difference between organs of a human and a mouse may not be of great importance. In other applications such differences may be essential. Of course the guidelines can be adapted to the nature of the application, but this make comparison of the quality of alignment methods much
more complex.

In summary, our results indicate that the manual evaluation of ontology alignments is by no means an easy task and that the ontology alignment community should be careful in the construction and use of reference alignments. We recommend that the OAEI community starts establishing best practices and guidelines for constructing reference alignments. Based on this paper we suggest to include at least the following elements in such an evaluation methodology:

- Select one of the three interrater-agreement measures used in Table 6.2 as the prescribed standard. Although from the results reported in this paper there is no clear winner, we suggest using Krippendorff’s Alpha for its versatility as it can be used with any number of raters, with incomplete data and on different sample sizes. The use of an inter-rater agreement measure will make comparison between experiments of different authors easier.

- Prescribe a minimum set of raters for manual evaluation. This minimum should not be lower than 3. A range of 3-5 raters appears reasonable.

- Agree on a set of alignment relations. The SKOS relations are attractive candidates mainly because they are part of a heavily used standard for publishing thesauri on the Web. However, the set of equivalence relations used by Halpin et al. (2010) has a more formal underpinning.

- Agree on a set of guidelines for helping to decide which mapping relation to use. The guidelines provided by us (cf. Appendix A.2) might serve as a place to start.

Having said this, we agree with Lakoff’s view on categorization and its consequence: we should not expect full agreement on reference alignments.

**Acknowledgements**

We would like to thank our raters: Mark van Assem, Victor de Boer, Marieke van Erp, Michiel Hildebrand, Veronique Malaisé, Lourens van der Meij, Carmen Reverté and Roxane Segers for their participation in our experiments. This research was supported by the MultimediaN project funded through the BSIK programme of the Dutch Government.
On the Assessment of Vocabulary Alignments

In Chapter 6 we studied the causes of disagreement between raters evaluating alignments. We found that there are aspects in an evaluation that can be controlled, leading to higher agreement between raters, but the inherent fuzziness of concepts is a characteristic that is always present in some form. In this chapter we analyze the process of evaluation in more detail by performing two evaluation experiments using data from 2007’s OAEI Food track and an alignment between RKD’s subject thesaurus and Cornetto. We also attempt to predict the level of agreement between raters based on vocabulary and mapping characteristics. We perform a small additional experiment to test our prediction. In this chapter we also study how rater disagreements affect quality measures such as precision and the ranking of alignment systems, which corresponds to our fourth research subquestion.

7.1 Introduction

A recommended practice is to have multiple raters evaluate mappings manually, and then measure agreement between raters using inter-rater measures such as Cohen’s Kappa (Cohen 1960) or Krippendorff’s Alpha (Krippendorff 2007). High agreement between raters indicates that the assessment is reliable. The impact of disagreements between raters has been studied in fields such as Information Retrieval (IR) (Bailey et al. 2008, Carterette and Sobooff 2010, Osman et al. 2010) but has yet to be explored in ontology alignment.

In Chapter 6, we have shown that the evaluation of alignments is a difficult task for humans influenced by multiple factors such as the nature of vocabularies, the goal of the alignment task and the inherent fuzziness of concepts. In this chapter our goal is to establish which vocabulary characteristics influence the evaluation of alignments, and attempt to predict the level of agreement raters will achieve based on them. We also study how disagreements affect the assessment of the quality of alignments. This is of particular importance for the Ontology Alignment Evaluation Initiative (OAEI)\(^1\), where the performance of tools is compared to reference alignments that were created and/or evaluated by humans. We will focus on the following research questions:

\(^1\)http://oaei.ontologymatching.org/
1. Which vocabulary and mapping characteristics influence the evaluation process and agreement level between raters?

2. How do disagreements between raters affect quality measures such as precision values of alignment systems?

3. Can we accurately estimate the level of agreement between raters based on small samples of mappings between vocabularies?

To answer these questions we perform two large evaluation experiments. In each experiment three raters evaluate large sets of mappings between two vocabularies. We then measure the agreement between raters and analyze the data quantitatively and qualitatively. We focus chiefly on the types of disagreements between raters with respect to vocabulary characteristics, and also look at the variations in precision measures depending on the rater. Based on our analysis we can formulate an expected agreement level and predict the difficulties that will arise for a new evaluation. To test our prediction we perform a third evaluation experiment with three raters and assess agreement levels. We close this chapter with a discussion.

7.2 Related Work

Human evaluation tasks, such as discourse analysis, relevance assessment and alignment evaluation, are all categorization tasks. Taylor (2003) gives an account of the developments in the field of categorization and linguistic categorization in particular. Early theories on categorization followed the Aristotelian view that each category is clearly defined, and mutually exclusive, following the rules of set theory. In the 1970’s Rosch (1973), Rosch et al. (1976) performed a number of experiments and proved the existence of prototypical categories where certain members of the category are more central than others. An example is the category furniture, of which CHAIR is more prototypical than SIDE-TABLE. Lakoff (1987) expanded prototype theory with additional theories of categorization and argued that categories form a graded cloud with fuzzy boundaries, where member concepts do not necessarily share common properties. According to Lakoff, categories are defined by culture and experience and therefore vary from person to person. These theories imply that when humans perform an evaluation task, their perspective on whether something fits into a category will vary depending on their experience and background, and when a category is fuzzy two people can disagree on its members.

Taylor also discusses the differences between monosemous, polysemous and homonymous categories. The distinction between a monosemous and polysemous category is generally clear: a monoseme has one meaning and a polyseme multiple meanings, although some cases are not clear cut. In general, the distinction is that a monosemous category can be vague, whereas a polysemous category is ambiguous. Taylor gives the example of the monoseme BIRD, i.e. referring to a category of flying creatures, and the concept SCHOOL which is polysemic in meaning and could refer to a school of thought, or a place where children are educated. The distinction between polysemy
and homonymy is also non-trivial in some cases. Polysemic concepts share a non-trivial relation, whereas homonyms are unrelated concepts that share the same form. Some homonyms are easy to identify and others are not. For example, the concept TO TIRE (being fatigued) and TIRE (of a car) are clearly homonyms and unrelated, whereas EYE OF THE NEEDLE is a metaphor of the HUMAN EYE. The latter relation however is more subjective, as some people do not perceive the metaphorical relation and consider the two concepts to be homonyms. Taylor argues that homonymy is an accidental phenomenon that is unlikely to appear in another language, whereas polysemy tends to cross over into other languages. Nonetheless, distinguishing between monosemes and polysemes, and polysemes and homonyms is not easy in certain cases.

The relatedness between polysemes may also be a metonymical relation. Metonymy is a type of figure of speech where part of a concept stands for an entire concept. For example, when we say PUT THE KETTLE ON, we refer to the water in the kettle that needs to be heated and not an empty kettle. Metaphorical relations also represent a type of relatedness, although they are not necessarily classified as a polysemic relation. As a result of the imprecise distinction in the type of relations between concepts it is difficult for humans to determine the specific relationship in some cases.

An alternative theory for distinguishing categories that removes the problem of identifying monosemes, polysemes and homonyms is the two level approach (Bierwisch and Schreuder 1992), that separates the linguistic level of meaning from the conceptual level. It states that the meaning of the concept is only given once it has been provided a context, for example when it appears in a sentence. Thus polysemes do not exist in our “mental lexicon”, but are rather linguistic representations of various interpretations of the concept. Therefore, the amount and type of context is important in a categorization task.

A related theory is situated cognition (Clancey 1997) that stipulates that people construct their knowledge “on the fly” in a specific context, rather than possess complete models in their mind. Thus, when humans are confronted with similar situations in evaluation tasks they develop rules that apply in that type of situation.

Categorization is a complex task where humans do not necessarily agree with each other. This disagreement can be measured in a number of ways. The simplest method is to calculate the percentage of agreement (observed agreement), which has a number of drawbacks. The percentage is hard to interpret and compare across experiments of varying size and type (Carletta 1996, Artstein and Poesio 2008), because it does not take chance agreement into account. There are various measures that do correct for chance agreement, such as Cohen’s Kappa (Cohen 1960), Fleiss’ Kappa (Fleiss 1971) and Krippendorff’s Alpha (Krippendorff 2007). In Chapter 6 we compared these measures and found that the values are very close to each other across experiments. However, Cohen’s Kappa can only be used with two raters, whereas Fleiss’ Kappa works with any number of raters, and neither of them works with missing data. Krippendorff’s Alpha on the other hand is more versatile, because it can be used with any number of raters, missing data, and nominal, ordinal and other type of categories.

Prevalence bias needs to be taken into account when using chance-corrected inter-rater agree-
ment measures. Prevalence bias occurs when data falls into mostly one category, making the
distribution between categories highly skewed. Even if observed agreement (the ratio of agree-
ments and total ratings) between raters is high, the agreement measure may turn out low. In order
to have a high measure the raters must agree on rare categories.

The interpretation of agreement measures is not entirely clear. In computational linguistics
only Krippendorff’s Alpha values higher than 0.8 are considered acceptable for content analysis
tasks (Artstein and Poesio 2008), with 0.7 being the absolute minimum. The nature of the catego-
ration task could influence the interpretation of the measure, and the question remains whether
lower agreement values are acceptable in alignment evaluation tasks.

There has been substantial work done in IR on the subject of relevance assessment and the
effect of disagreements in retrieval evaluations. Trotman and Jenkinson (2007) showed that judg-
ments by experts are interchangeable with respect to IR system scores. Bailey et al. (2008) per-
formed experiments using the IR Enterprise track data and compared the performance of raters of
varying degrees of expertise using Cohen’s Kappa. They found that there is substantial disagree-
ment between expert and non-expert raters. In addition, this difference in evaluation also affected
the score of IR systems, where non-expert judgments lead to higher scores than expert judgments.
In IR, disagreements between human raters only affect system scores when their level of expertise
varies. We still need to determine whether the same situation applies in vocabulary alignment
evaluation.

Carterette and Soboroff (2010) examined the effect of rater errors in the TREC Million Query
track by creating models of typical rater behavior, and running simulations. Their suggestions
for adjusting errors in ratings include the re-evaluation of certain documents, the use of voting
to select the correct judgment and the assessment of the extent to which errors may have been
introduced in the judgments.

Osman et al. (2010) investigated the validity of using a single assessor to evaluate the relevance
of documents on a single topic in the TREC blog opinion track. They found a great deal of
variation in agreement between each pair of raters, and call for the use of majority voting and the
use of prior testing to remove dissenting raters. They discuss the impact of the variance in opinion
on the data, in particular when the evaluated data are used for training in machine learning settings.

In summary, there are many theories on non-classical categorization in the field of cognitive
linguistics. The experimental results in tasks requiring consensus of human assessors are not clear
cut, in particular what constitutes an acceptable level of agreement. We will now perform experi-
mental research to determine why raters disagree and how it affects the application of assessment
results on vocabulary alignment.

### 7.3 Experimental Setup

In this chapter we describe two detailed evaluation experiments where we study the process of
manual evaluation focusing on two aspects: first, on data characteristics that influence rater deci-
sions; second, the manner in which rater disagreements affect quality measures such as precision.
Our approach in these experiments is driven by some practical considerations, such as the availability of data and human raters. In these experiments our goal is not to perform a full evaluation of mappings, but rather to focus, in particular in our second experiment, on a subset of mappings we expect to cause some difficulty for raters. We now describe our experimental setup in more detail.

### 7.3.1 Overall Approach

For our first experiment we want to use data from one of the OAEI tracks. The data needs to be relatively easy to evaluate, and not require expertise in a specific field. This rules out a number of tracks, such as the anatomy and directory tracks. Another requirement is that all the data, including system runs, original vocabularies and the reference alignment or gold standard, needs to be available. We decided to use the vocabularies and mappings from the 2007 OAEI Food track (Euzenat et al. 2007b) which meet all our requirements. The task in the Food track was to align two vocabularies from the food domain: the United Nations Food and Agriculture Organization AGROVOC Thesaurus\(^2\) and the United States National Agricultural Library Agricultural thesaurus (NALT)\(^3\). Five systems participated in this track in 2007 and a gold standard had been created by domain experts from a sample of submitted mappings. Our goal in this experiment is twofold: first, to have three raters re-evaluate the gold standard mappings and analyze the process and results; second, to determine whether different evaluations have a significant impact on the assessment of the quality of mappings of participating systems.

In our second experiment we want to evaluate “difficult” mappings, that is mappings that would likely lead to more disagreement between raters. Our intuition is that the evaluation of ambiguous mappings, where a concept is linked to multiple closely related concepts, is the most difficult task for an rater. We want to establish how such a difficult task compares to a more general evaluation task, such as our first experiment, and what its effects are on determining the quality of a set of mappings. We selected ambiguous mappings between Netherlands Institute for Art History (RKD)\(^4\) subject thesaurus and Cornetto\(^5\) (Vossen et al. 2008), a lexical resource similar to Princeton WordNet\(^6\) (Fellbaum 1998). Cornetto is rich in concepts with related meanings, whereas the RKD subject thesaurus is small, and has relatively broad coverage in subject matter. We expect this combination to result in low agreement between our raters.

In each experiment we ask raters to evaluate mappings into four categories and comment on their choices. We then calculate inter-rater agreement measures for 4 categories and 2 categories (by aggregation). We analyze the results of the experiments, focusing in particular on disagreements. We also analyze mapped concepts and their labels in terms of their characteristics, and determine how these characteristics influences agreement between raters. Based on these charac-

---

\(^2\)http://www.fao.org/agrovoc/
\(^3\)http://agclass.nal.usda.gov/
\(^4\)http://english.rkd.nl/
\(^5\)http://www2.let.vu.nl/oz/cornetto/index.html
\(^6\)http://wordnet.princeton.edu/
teristics, we analyze the characteristics of a small sample of mappings between two vocabularies, and predict the level of agreement. In order to validate our prediction, we have three raters evaluate a larger sample of mappings between these vocabularies.

We now describe the vocabularies, matching categories and guidelines in more detail.

7.3.2 Vocabularies and mappings

OAEI Food track 2007

The task in the OAEI Food track was to align AGROVOC and NALT. AGROVOC contains 28,439 concepts with a preferred label each, 10,104 alternative labels, and 1,180 scope notes. NALT contains 42,326 concepts also with a preferred label each, 25,985 alternative labels and 1,025 scope notes, and is thus one and a half times larger than AGROVOC. In both thesauri both preferred and alternative labels are unique, and homonymous labels are distinguished through the use of qualifiers. The vocabularies are organized into hierarchies through the use of broader/narrower relations and make use of the related relation as well. The vocabularies contain content specific to the agriculture and food domain, and both contain a significant number of concepts that describe organisms, chemical compounds and agricultural methods and technologies.

In 2007 there were five participating systems: RiMom, Falcon, Dssim, Xsom and Scarlet. Most participants submitted skos:exactMatch mappings although Falcon and Scarlet also submitted skos:narrowMatch and skos:broadMatch mappings. Table 7.1 shows the number of mappings and their type per system. Each system applies different mapping strategies or their combinations to create mappings. Falcon (Hu and Qu 2008) for example uses a combination of lexical similarity and structural techniques in order to find mappings. As a result, whereas a large number of mapped concepts share a label or at least part of a label, some concepts have no lexical link and were found using structural techniques or lexical/domain resources.

In the original evaluation (van Hage et al. 2010) a gold standard sample was constructed by applying random stratified sampling. We recount the method as it will be relevant in later sections. The total submitted mappings were separated according to mapping type, and according to the number of systems that found them. There were four types of mappings: taxonomical (23,023), chemical (3,965), geographical (1,354) and miscellaneous (13,625). There were also four sets of mappings based on the number of systems that found it: found by four systems (1,462), three systems (3,944), two systems (7,142) and a single system (29,419). The rationale behind these differentiations is the following. First, certain types of mappings require different strategies and mapping techniques than others, for example, chemical concepts are often described using synonyms and lexical variants whereas taxonomical mappings can be aligned by knowing the rules in the naming system. Second, when multiple systems return the same mapping it is more likely to be correct than when a single system returns it and therefore the overall quality of these subsets may vary significantly. The samples were drawn from the overlap of the mapping type sets and the shared system sets with the exception of taxonomical mappings. The latter were excluded, as taxonomic concepts use the latin name as preferred or alternative label and all correct mappings
could be easily identified automatically. As a result, this set could be fully evaluated. The set of mappings found by a single system was the largest. During sampling care was taken to sample mappings from each system to be able to differentiate between them. As a result of this sampling method the gold standard evaluation contained 756 mappings, with 196 chemical, 86 geographical and 474 miscellaneous mappings. For the distribution of the sample among the systems see Table 7.5.

We decided to exclude Scarlet mappings from our experiment and analysis. Scarlet mappings are underrepresented in the evaluated sample because they had been submitted after the original OAEI deadline.

<table>
<thead>
<tr>
<th>System</th>
<th>Total</th>
<th>Exact</th>
<th>Broad</th>
<th>Narrow</th>
<th>Unrelated</th>
</tr>
</thead>
<tbody>
<tr>
<td>RiMom</td>
<td>18,420</td>
<td>18,420</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Falcon</td>
<td>15,300</td>
<td>14,615</td>
<td>127</td>
<td>558</td>
<td>0</td>
</tr>
<tr>
<td>Dssim</td>
<td>14,962</td>
<td>14,962</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Xsom</td>
<td>6,583</td>
<td>6,583</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Scarlet</td>
<td>6,764</td>
<td>81</td>
<td>2,328</td>
<td>3,708</td>
<td>647</td>
</tr>
</tbody>
</table>

Table 7.1 OAEI Food track mappings generated by each system participating in the track (total and per mapping category). This is original input data we were given for our experiment

RKD-Cornetto

The RKD subject thesaurus is a relatively small vocabulary containing 3,342 concepts with 3,342 preferred labels, 242 alternative labels and 344 definitions or scope notes. The concepts are linked through broader/narrower relations, as well as the related relation. Cornetto is a much larger lexical resource containing 70,434 synsets and 103,762 labels. The labels in Cornetto are not separated into preferred and alternative labels and multiple concepts may share the same label. As a result, a large number of mappings from RKD to Cornetto are one concept to many or many to one.

For our second experiment we use mappings described in Chapter 3. The mappings were created using a variety of methods including lexical string matching. Of the original 4,375 mappings a large portion is ambiguous. There are 875 RKD thesaurus concepts that are mapped to at least two Cornetto concepts, with a total of 2,755 mappings, and there are 297 Cornetto concepts mapped to at least two RKD thesaurus concepts with a total of 642 mappings. We randomly sampled 303 mappings where an RKD concept is linked to multiple Cornetto concepts. We call this set RKDCornetto. We randomly sampled another 100 mappings where a Cornetto concept is linked to multiple RKD concepts and call this set CornettoRKD. There is a small overlap of 6 mappings between the two samples.
7.3.3 Evaluation categories

We used SKOS mapping relations (Miles and Bechhofer 2009) to categorize mappings in our evaluation. We restricted the choice of evaluation categories on the OAEI 2007 Food track data to the same four categories used for the creation of the original gold standard: `skos:exactMatch`, `skos:broadMatch`, `skos:narrowMatch` and `unrelated`. In addition, raters could select `unsure` when they were unable to decide on a category.

In the RKD-Cornetto experiment raters could additionally use the `skos:closeMatch` and `skos:relatedMatch` categories to indicate other types of relations between concepts. We took these extra matching relations into account in our qualitative analysis.

7.3.4 Interrater Agreement Measure

We use Krippendorff’s Alpha for measuring agreement because we have missing data (unknown category) in our evaluations. In our experiments our matching categories were on the nominal scale. We measure inter-rater agreement for four categories: `skos:exactMatch`, `skos:narrowMatch`, `skos:broadMatch` and `unrelated`. In the RKD-Cornetto experiment where raters also used `skos:closeMatch` and `skos:relatedMatch`, we merge `skos:closeMatch` with `skos:exactMatch`, and `skos:relatedMatch` with `unrelated` to reduce the number of categories to four.

As alignment tools often only generate equivalence mappings we also measure agreement for two categories: `equivalent` and `inequivalent` mappings. The `equivalent` mappings are `skos:exactMatch` mappings, whereas the `inequivalent` mappings are a merging of `skos:narrowMatch`, `skos:broadMatch` and `unrelated` mappings.

7.3.5 Evaluation Tool and Guidelines

We used Amalgame (van Ossenbruggen et al. 2011) to support raters in the evaluation task. Amalgame’s user interface presents mappings with hierarchical information for each concept, labels and scope notes. Raters can select the appropriate relation category for each mapping. The choices are then stored in RDF using the OAEI alignment format (Euzenat 2004) along with provenance information. This tool is a newer version of the tool used in Chapter 6. It is open source and available for download\(^7\).

In our evaluation experiments we used the same guidelines as in the previous chapter (Appendix A.2). These guidelines form a frame of reference for raters. During analysis of the evaluation we found that raters created and applied their own meta-guidelines during evaluation. We asked them to describe these rules and discuss them in more detail in our analysis section.

\(^7\)http://semanticweb.cs.vu.nl/amalgame/
7.4 OAEI Food track 2007 experiment

7.4.1 Agreement measurements

We re-evaluated the gold standard with three different raters and compared the results to the gold standard evaluation from 2007. As shown in Table 7.2, the overall agreement level is higher than 0.7, and for 2 categories even higher than 0.8. We also found that agreement among the three raters is higher than agreement with the gold standard. We observe large differences between pairs of raters: the highest agreement for 4 categories is 0.813 and the lowest 0.767. This variation suggests that at least three raters are needed for consistent results. We also applied majority voting on the evaluation results of the three new raters for 2 categories by selecting the category chosen by at least two raters. We measured agreement between this majority vote and the gold standard, and found the highest agreement overall. Majority voting could not be fully applied on four categories, because for some mappings all three raters selected a different category.

<table>
<thead>
<tr>
<th></th>
<th>Observed agreement for gold standard and 3 new raters</th>
<th>Observed agreement for 3 new raters</th>
<th>K’s ( \alpha ) for gold standard and 3 new raters</th>
<th>K’s ( \alpha ) for 3 new raters</th>
<th>K’s ( \alpha ) for gold and majority vote (3 new raters)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 categories</td>
<td>85.7%</td>
<td>88.2%</td>
<td>0.745</td>
<td>0.790</td>
<td>not applicable</td>
</tr>
<tr>
<td>2 categories</td>
<td>92.1%</td>
<td>92.7%</td>
<td>0.836</td>
<td>0.851</td>
<td>0.871</td>
</tr>
</tbody>
</table>

Table 7.2 Observed agreement and inter-rater agreement measures (Krippendorff’s \( \alpha \)) for 4 evaluation categories and 2 categories in the Food experiment. Agreement between the 3 new raters is higher than agreement with the gold standard. Agreement for 2 categories is higher than for 4 categories. Agreement between the gold standard and a majority vote applied on the 3 new raters for 2 categories is the highest.

7.4.2 Qualitative Analysis

For the qualitative analysis we disregard the original gold standard evaluation and focus on the evaluation of the three new raters, as we are familiar with the background of the raters, and we have information on their choice of categories. Consequently, we can perform a detailed analysis of the evaluation, in particular of the disagreements between raters. In the analysis one rater performs the task of meta-rater by examining mapped concepts in detail.

Of the 756 evaluated mappings, the raters agreed fully (all three raters chose the same category) on 632 (84%). Out of the remaining 124, for 118 mappings one rater chose a different category from two other raters. In the remaining 6 mappings each rater choose a different category. Thus, for most of the disagreements in four categories majority voting may be applied to decide on the mapping category. Table 7.3 shows the distribution of the disagreements per sample
type. Most of these fall into the miscellaneous type both numerically and as a percentage of all miscellaneous mappings. The least disagreement is in the geographical type mappings. We also looked at the number of disagreements for 2 categories (equivalent vs. inequivalent). There were 71 mappings with disagreements and in all cases a majority voting can be applied. For 2 categories the distribution of disagreements per sample type changes, and we find that the number of disagreements in the miscellaneous and chemical type mappings are halved. This is because many disagreements for 4 categories concerned broader/narrower relations.

<table>
<thead>
<tr>
<th></th>
<th>4 categories</th>
<th>2 cat.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total # mappings</td>
<td>Total # of mappings with disagreement</td>
</tr>
<tr>
<td>Miscellaneous</td>
<td>474</td>
<td>92 (19%)</td>
</tr>
<tr>
<td>Chemical</td>
<td>196</td>
<td>24 (12%)</td>
</tr>
<tr>
<td>Geographical</td>
<td>86</td>
<td>9 (10%)</td>
</tr>
</tbody>
</table>

Table 7.3 Disagreements between the three new raters per sample type for 4 categories and 2 categories. Disagreements for 4 categories occurs the most in miscellaneous type mappings. Majority vote can be applied to most of the disagreements. Note that for 2 categories there can be no full disagreement between three raters.

We continue our analysis of the disagreements with respect to the four matching categories. We examined each individual mapping and categorized the causes of disagreement. We found that in 46 mappings (almost 40% of the disagreements) the disagreement is caused by one or more raters making a mistake in their assessment. Such mistakes are either due to inattentiveness (human error), or caused by domain-specific concepts requiring more research on the part of the rater. Raters were free to look at sources, such as Wikipedia, to find out more about the meaning of a concept, but did not always do so. This introduced disagreement in cases where one rater looked up information and the other did not.

Using majority vote is a good way of eliminating errors but there are exceptions. We found five mappings where the evaluation required more research, and two raters made a mistake while the third chose the correct category. For example, the concepts NUCLEOSIDASES and NUCLEOTIDASES seem to be equivalent at first glance, but are in fact two different chemical compounds. The evaluation of this mapping is made more difficult because it is chemistry-domain specific, and there is but one letter difference between the labels. On average, each rater made 17 errors (14% of disagreements) and although the precise causes are difficult to pinpoint, some errors were due to unfamiliarity with the domain and others due to inattentiveness.

For further 37 mappings (31% of disagreements) the disagreement can be attributed to the “fuzziness” of a concept, in particular AGROVOC concepts. In AGROVOC we found that mapped
concepts at the bottom of the hierarchy may aggregate several distinct concepts. Such concepts have non-synonymous labels that indicate that two concepts were united into a single concept. An example of such a concept is MORPHINE, THEBAINE which unifies the chemical compounds morphine and thebaine, the latter is similar to morphine but not the same. Strictly speaking these concepts are siblings. Raters selected skos:exactMatch, skos:narrowMatch as well as unrelated in the evaluation of the mapping between MORPHINE, THEBAINE and the concept THEBAINE. Although each choice is understandable, from the raters perspective such compound concepts frequently cause disagreement.

Finally, for the remaining disagreements the precise meaning of concepts is unclear, making the category selection difficult. These are the mappings between concepts that are not equivalent, but can nevertheless have some relation. In some cases more domain knowledge is required to make a choice. One such example is the mapping between PLACENTA and CHORION. The concepts are related, as both are membranes around a developing fetus. However, raters were unable to agree on which concept is broader or whether they are sibling concepts. In other cases the contextual information of concepts can be interpreted multiple ways. For example, for the mapping between DIGESTIBLE CELLULOSE and CELLULOSIC MATERIALS two raters chose the broader category because in their opinion the digestible cellulose is a type of cellulosic material. The third rater however took into account that the parent term of digestible cellulose is NUTRITIVE VALUE, and decided that the two concepts are in two very different contexts, and therefore unrelated.

The mappings described above along with additional examples can be found in the appendix in Table B.4.

7.4.3 Precision Measures

The results of a manual evaluation is often used to determine the quality of an alignment. At the OAEI a gold standard evaluation is used to compare the performance of alignment tools. Consequently, disagreement between raters can have an effect on individual precision measures of alignment tools, and their ranking with respect to each other. We now examine the results of our three individual evaluations of the OAEI food track, along with the original gold standard.

<table>
<thead>
<tr>
<th>Rater</th>
<th>Equivalent</th>
<th>Inequivalent</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Exact</td>
<td>Broad</td>
</tr>
<tr>
<td>Gold</td>
<td>453</td>
<td>54</td>
</tr>
<tr>
<td>Rater A</td>
<td>441</td>
<td>20</td>
</tr>
<tr>
<td>Rater B</td>
<td>444</td>
<td>34</td>
</tr>
<tr>
<td>Rater C</td>
<td>442</td>
<td>34</td>
</tr>
</tbody>
</table>

Table 7.4 Distribution of mapping categories per rater. The difference between Raters A, B an C for equivalent mappings is very small.
Table 7.4 shows the distribution of aggregated mapping categories per rater. Based on this table it seems that the difference between the raters, in particular raters A, B and C, is minimal with respect to equivalent (exact) matches. However, the new raters disagreed on roughly 9% of the mappings for 2 categories. Although this is not reflected in Table 7.4, it could affect the final precision measures and ranking of systems depending on the distribution of these disagreements per system and per sample type.

![Image of the distribution of categories](image)

**Figure 7.1** The distribution of the categories in the 71 Food track mappings where raters disagree. For each individual mapping the decisions by each of the three raters is shown. We show the interval that represents a single mapping. Raters categorize each mapping either as equivalent (white) or inequivalent (black). The distribution of disagreements in the geographical mappings is the most skewed. Rater A rated most mappings as equivalent whereas rater B rated most mappings as inequivalent.

The evaluated sample of mappings contains three sample types (chemical, geographical and miscellaneous). Fig. 7.1 shows distribution of categories where raters disagreed per sample type, that is mappings where the three raters did not select the same category. From this figure it becomes apparent that the proportion of equivalent (white area in Fig. 7.1) vs. inequivalent matches (black area in Fig. 7.1) in the chemical and miscellaneous mappings is similar for each rater. For geographical mappings the proportions are more skewed; Rater B has fewer equivalent mappings than Rater A or C. However, the geographical sample type is relatively small, and when the number of equivalent mappings is added up per rater in Table 7.4, there is almost no evidence of disagreement.

The evaluated mappings represent small portions of mappings returned by four systems in three strata according to sample type (chemical, geographical and miscellaneous).

Table 7.5 shows the distribution of the evaluated samples per type and per system along with the number of mappings each system found per type. We excluded the taxonomical mappings from our analysis, because they were not evaluated manually, and we wish to measure the effect of disagreements on precision. For each system, miscellaneous type mappings form the largest stratum, although its proportion varies from 70.3% for Rimom to 44.5% for Xsom. The smallest stratum is formed by the geographical mappings for all four systems. Therefore, in theory, the result of the evaluation and the effect of disagreements is much more substantial for a large stratum (miscellaneous) than a small stratum (geographical). The size of the evaluated sample per stratum is also of
importance. The proportion of the evaluated geographical mappings compared to the geographical stratum is relatively high for Rimom (13.4%), Falcon (14.7%) and Xsom (26.2%) compared to the other types of evaluated mappings. As a result, the effect of disagreements is less enhanced for this stratum than for other strata where the evaluated samples represent a comparatively smaller part of the stratum.

We now look at the precision measure for each system per rater. To measure the precision of each system we use a similar method to that described by Van Hage, the organizer of the Food track, in (van Hage et al. 2010, Euzenat et al. 2007b). For each system we determined the number of evaluated mappings per sample type and the precision of this sample. We also determined the total number mappings the system returned per type, and calculated the weight of the evaluated sample. For example, Falcon found 1,795 chemical mappings and 145 of these have been assessed by raters. The precision of these 145 mappings is 0.903. The total number of mappings Falcon returned (non taxonomic) was 7,077 and therefore the weight of this stratum is $\frac{1,795}{7,077} = 0.253$ and the precision for this stratum is 0.903. The total precision for the system is the sum of the precision of each stratum times its weight.

The precision measures per system and per rater are shown in Table 7.6. We also applied majority voting to the disagreements among the three new raters (Raters A, B and C), and the precision falls between the values of the raters. Although there is some variation in the precision measures per rater, we found that these differences are not significant, nor are the differences in ranking. We used the Bernoulli distribution as a significance test on the precision scores of the system using the method from van Hage et al. (2007). In this case, the disagreements among raters

<table>
<thead>
<tr>
<th>System</th>
<th>Evaluated (Found)</th>
<th>Evaluated (%) of sample type found</th>
<th>Evaluated (%) of total found</th>
<th>Found (%) of total found</th>
<th>Evaluated (%) of total found</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rimom</td>
<td>97 (5.5%)</td>
<td>1,745 (22.3%)</td>
<td>76 (13.4%)</td>
<td>566 (7.2%)</td>
<td>242 (4.4%) (70.3%)</td>
</tr>
<tr>
<td>Falcon</td>
<td>145 (8.0%)</td>
<td>1,795 (25.4%)</td>
<td>85 (14.7%)</td>
<td>577 (8.2%)</td>
<td>237 (5.0%) (66.5%)</td>
</tr>
<tr>
<td>Dssim</td>
<td>44 (2.9%)</td>
<td>1,498 (33.9%)</td>
<td>54 (8.8%)</td>
<td>612 (13.9%)</td>
<td>105 (4.5%) (52.2%)</td>
</tr>
<tr>
<td>Xsom</td>
<td>66 (16.9%)</td>
<td>390 (40.9%)</td>
<td>34 (26.2%)</td>
<td>130 (13.6%)</td>
<td>125 (28.9%) (44.5%)</td>
</tr>
</tbody>
</table>

Table 7.5: Number of evaluated mappings found by each system per sample type, and total number of mappings found by each system per sample type. The percentage the evaluated set represents of the sample type is in parentheses for each evaluated sample type. The percentage the sample type represents of the total number of mappings found by a system is also in parentheses.
Table 7.6 Precision of each system per rater.

was reasonably well distributed across the systems, and have little effect on the precision measure of each system and their ranking.

7.5 RKD - Cornetto experiment

7.5.1 Agreement measurements

We evaluated mappings between the RKD subject thesaurus and Cornetto with three raters. In addition to the exact, broad, narrow and unrelated categories, raters also used related and close match in the evaluation. In order to compare the results of this evaluation with the OAEI Food track experiment, related mappings were merged with unrelated mappings, and close matches with exact matches. Table 7.7 shows the distribution of mappings for 4 and 2 categories used to measure inter-rater agreement.

Table 7.7 Distribution of merged mapping categories per rater for 4 categories and for 2 categories. The differences between raters is more marked than on the Food data in all categories.

Table 7.8 shows the Krippendorff’s Alpha measures per evaluation session for all raters and for each rater pair. As expected, the Alpha for this type of data is significantly lower than in the Food track experiment, 0.66 for four categories. The Alpha varies a great deal for individual pairs of raters, from 0.59 (Rater A-Rater B) to 0.73 (Rater A-Rater D). The Alpha for two categories is only slightly higher than four categories, with the exception of CornettoRKD, which suggests that there are disagreements on whether two concepts are equivalent or not. Our agreement measure of
0.667 for 2 categories is similar to the agreement measure on mappings between Getty’s Art and Architecture Thesaurus and Princeton WordNet in Chapter 6, where we measured 0.679 for two categories. The Alpha however, is much lower than than in the Food track experiment.

<table>
<thead>
<tr>
<th>Evaluation session</th>
<th>All raters</th>
<th>Rater A - Rater D</th>
<th>Rater A - Rater B</th>
<th>Rater B - Rater D</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 cats RKDCornetto</td>
<td>0.687</td>
<td>0.756</td>
<td>0.616</td>
<td>0.697</td>
</tr>
<tr>
<td>2 cats RKDCornetto</td>
<td>0.711</td>
<td>0.791</td>
<td>0.640</td>
<td>0.707</td>
</tr>
<tr>
<td>4 cats CornettoRKD</td>
<td>0.558</td>
<td>0.617</td>
<td>0.473</td>
<td>0.587</td>
</tr>
<tr>
<td>2 cats CornettoRKD</td>
<td>0.457</td>
<td>0.530</td>
<td>0.357</td>
<td>0.477</td>
</tr>
<tr>
<td>4 cats Both sessions</td>
<td>0.660</td>
<td>0.730</td>
<td>0.589</td>
<td>0.670</td>
</tr>
<tr>
<td>2 cats Both sessions</td>
<td>0.667</td>
<td>0.746</td>
<td>0.595</td>
<td>0.665</td>
</tr>
</tbody>
</table>

Table 7.8 Krippendorff’s Alpha per evaluation session for 4 categories and 2 categories for all raters and per rater pair.

### 7.5.2 Qualitative Analysis

We will now look at how raters evaluated these mappings and analyze their disagreements. For this analysis we also look at mappings where raters selected close and related relations.

The most significant difference between raters in this evaluation was their approach to the mapping task. Raters A and B viewed the mappings from a retrieval perspective, and Rater D from a lexical perspective. As a result each rater had different meta-guidelines. For example, Rater D evaluated mappings representing metaphorical or metonymic relations as related match, whereas raters A and B chose unrelated, as such mappings were not deemed useful in a retrieval scenario. This difference was not reflected in the inter-rater agreement measures, as related matches were merged with unrelated matches.

<table>
<thead>
<tr>
<th>Rater</th>
<th>Exact</th>
<th>Close</th>
<th>Broad</th>
<th>Narrow</th>
<th>Related</th>
<th>Unrelated</th>
<th>Unsure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater A</td>
<td>154</td>
<td>0</td>
<td>2</td>
<td>13</td>
<td>15</td>
<td>217</td>
<td>3</td>
</tr>
<tr>
<td>Rater B</td>
<td>117</td>
<td>0</td>
<td>19</td>
<td>41</td>
<td>33</td>
<td>190</td>
<td>4</td>
</tr>
<tr>
<td>Rater D</td>
<td>121</td>
<td>18</td>
<td>11</td>
<td>32</td>
<td>78</td>
<td>142</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 7.9 Distribution of the mapping categories per rater, including close and related relations.

The distribution of mapping categories in Table 7.9 shows that there is a large difference in the way raters evaluate mappings. Rater A selected the highest number of exact matches for two reasons. First, as the RKD concepts are frequently used in polysemic form (related meanings), she considered related concepts such as strawberry the plant and the fruit to be equivalent for annotation, whereas other raters selected related match. Second, concepts where one is strictly
speaking more specific than the other but share a label were also considered equivalent rather than narrow match. One example is the mapping between the generic PRIEST (female and male) and the mapping between a female priest. Rater A selected exact match, and raters B and D narrow match. As a result, Rater D and in particular Rater B, had selected narrow and broad much more frequently than Rater A. Raters created a meta-guideline for such situations, and when they were confronted with a similar type of mapping, they applied the same rule throughout the evaluation. The creation of such meta-guidelines is consistent with the theory of situated cognition, in that rules are created on the fly when a new situation arises.

We also analyzed disagreements between the raters. Out of the 404 evaluated mappings, all three raters agreed fully (on the 7 categories from Table 7.9) on 234 mappings (58%) by selecting the same matching category. For an additional 60 mappings the disagreement was very minor (related vs unrelated and close vs. exact) (15%). We focus our analysis on the remaining 109 mappings with disagreements. Of these 109 mappings, there were 77 where two raters selected the same mapping and the third rater another (majority vote can be applied), and there were 32 mappings where each rater selected a different category. Compared to the Food experiment the number of disagreements is greater (27% vs. 16%), and the number of disagreements where all raters say something different is also greater (7.9% vs 0.7%).

We found three main causes for disagreement:

**Human error**  We found that in 14 mappings disagreement was caused by human error by one or more raters, which is 12% of all disagreements. On average, raters made 5 errors per evaluation. For example, the mapping between a knight’s spur and larkspur (flower), both of which have the label RIDDERSPOOR, was rated as exact match by one rater, and unrelated by the two other raters. Certain errors were caused by errors or inconsistencies in the vocabulary. A mapping between ZWAAN the bird (RKD) and the constellation (Cornetto) was evaluated by Rater B as an exact match because the gloss for the Cornetto term states it is a large bird. However, the broader term in Cornetto is constellation, giving the concept its proper context. Although raters were familiar with this vocabulary characteristic from our guidelines, Rater B must have forgotten the guideline or overlooked this instance. Compared to the Food experiments however, raters made fewer mistakes, because most concepts required no domain-specific knowledge, with the exception of some cultural heritage concepts in the RKD thesaurus such as GRISAILLE, which is a monochrome style painting.

**Vagueness of concepts**  We also found that many disagreements were caused by a “vagueness” of concepts. We call concepts vague when the scope of one or both mapped concepts is not entirely clear from the context or their meaning overlaps greatly. One such example is the mapping between HOEDEN (herding) and OPPASSEN (watching out/to beware). The two concepts overlap in meaning, in particular as the concept OPPASSEN also had the term HOEDEN as an additional label. However, as Cornetto has an additional concept for herding, these concepts are not precisely equivalent but polysemes. During the evaluation two raters selected exact match and one rater selected broad match.
Difference in rater meta-guidelines  Raters created and applied different meta-guidelines during the evaluation. The application of such guidelines caused disagreement between raters. For example, a mapping between the concepts SCHRIJVER meaning “professional writer” and “someone who writes” was categorized as exact match by Rater A, and broad match by Rater B and D. Rater A chose exact match because she considered the RKD term to be applicable in both cases.

The mappings described above, along with additional examples, can be found in the appendix in Table B.5.

There were 6 mappings that occurred in both RKDCornetto and CornettoRKD set of mappings. We found that after raters evaluated the RKDCornetto mappings, they were not consistent in their choice of matching category for the same mappings in CornettoRKD. Raters stated that they had forgotten the meta-guidelines they had created in previous session(s). Intra-rater consistency can therefore also be a problem.

7.5.3 Precision Measure

In this experiment we focused on ambiguous (one to many) mappings where the precision of the mappings is low. As a result, the evaluation results are not generalizable to the original set of mappings generated between RKD and Cornetto. We calculate precision based on each rater’s evaluation in order to establish its variation, and attach no importance to its absolute value.

Again, we use the number of equivalent mappings divided by the total number of mappings to determine precision. We measured 0.381 based on Rater A’s evaluation, 0.290 for Rater B and 0.344 for Rater D. We also calculated precision by applying majority voting to determine the category of a mapping, and measured 0.339. We performed a significance test and found that the precision values between Rater A and B and between Rater B and D are significantly different, whereas between Rater A and D they are not.

![Figure 7.2](image)

Figure 7.2 The distribution of categories in the 89 mappings with disagreements in RKD-Cornetto for 2 categories per rater.

In contrast to the Food track experiment, the distribution of mapping categories (Table 7.7) and precision measures varies more per rater. Fig. 7.2 shows that the distribution of categories is not homogenous in the mappings where raters disagree; Rater A selects exact match more often than Rater C, which results in a higher precision score for the mappings when Rater A’s evaluation is used than Rater C’s. In addition the percentage of disagreements is also higher than in the Food
experiment, which was reflected by the lower inter-rater agreement. On this type of data consensus would need to be reached for instance by applying a majority vote, or using meta-raters to evaluate difficult mappings.

7.6 Analysis of Mapping Types, Vocabularies and Mappings

The results of the Food experiment and RKD-Cornetto experiment are different in terms of inter-rater agreement and the types of disagreements. In this section we look at the types of mappings we come across in alignment scenarios and how they can be interpreted during an evaluation. We apply this interpretation to the vocabularies and mappings used in our experiments and draw conclusions.

7.6.1 General Mapping Types

We can distinguish between mapping types based on the linguistic category they fall into. Fig. 7.3 shows these categories depending on whether concepts share the same label or not. A mapping between two concepts is usually either based on a lexical (string) match, or a non-lexical match. The latter may be created through the use of structural or instance-based matching techniques.

When concepts are mapped based on a partial or full lexical match, they can be monosemes; that is the mapped concepts have a single meaning. Since these concepts have the same label, we
often find that they are exact matches. However, the context of these concepts (scope notes and/or hierarchy) can introduce a fuzziness, or vagueness which can cause disagreement between raters.

One of the main sources of disagreement are polysemes. Strictly speaking polysemes are concepts that are related but not equivalent. However, when the context of a source concept is vague, it can cover multiple target concepts that are polysemes. Because polysemes are related concepts, the definition and degree of the relatedness is a source of disagreement. We found the SKOS matching relations to be inadequate in these cases.

Homonyms are concepts with different meanings. In most cases homonyms are easy to recognize as unrelated concepts. For example, *a row* (dispute) and *to row* (the activity) would be categorized as unrelated. There are homonyms however that border on the polysemous and where as a consequence, disagreement can arise on the relation category. For example, a *person’s leg* and a *chair’s leg* have a metaphorical connection.

Even if there is no lexical match between two concepts they can be equivalent when they are synonyms. In practice however, whereas synonyms mean almost the same thing, they are often not identical and thus are a source of disagreement in evaluation scenarios.

The fifth type of mapping is between non-lexically related terms where the relation between concepts is other than equivalence. Examples of such a relation are the sibling, broader and narrower relations. Disagreements in such cases are either caused by domain-specific terms where raters are unfamiliar with one or both concepts, or because SKOS matching relations are inadequate for describing the relationship, such as metaphor, metonym or sibling.

### 7.6.2 Vocabularies

Both AGROVOC and NALT are large vocabularies covering the domain of agriculture and food. NALT is one and a half times larger than AGROVOC and there is some difference in granularity between the two. In AGROVOC some concepts are merged at the lowest levels in the hierarchy, whereas these concepts are separate in NALT. Both vocabularies use unique labels, distinguished by the use of qualifiers and scope notes. Content-wise, the vocabularies cover the same or similar concepts. For example, AGROVOC contains around 22,000 concepts describing organisms and in NALT there are 26,615 concepts in the Taxonomic Classification of Organisms concept scheme. Both vocabularies also contain a large number of concepts describing chemicals and compounds, geographical locations and various farming tools and methods. These latter type of concepts were the focus of our evaluation experiment, as mappings between taxonomic concepts are trivial in terms of evaluation.

The RKD subject thesaurus is the smallest of the four vocabularies describing objects, activities, locations and characters that are often represented in cultural heritage media. The thesaurus contains a varied assortment of concepts such as body positions, plants, animals, activities, mental states, furniture, clothing, methods of transportation and many others. Thus, most terms are generic with respect to their domain, although there are a number of cultural heritage related concepts, such as the genre or a style of the painting, and abstract concepts used to describe allegories.
We looked at the manner in which concepts were used for annotation and found that in many cases the application of a concept is very broad as exemplified by the use of stone for a small stone as well as the building material, which are polysemes.

Cornetto is the largest vocabulary, and distinguishes itself from the other vocabularies in that it has no preferred labels, and all labels have the same status. Another distinguishing feature is that the labels are not unique: over 30,000 concepts share one or more labels with other concepts. Concepts that share a label are either homonyms, such as bank (financial institution) and bank (river), or polysemes, such as milk (the activity) and milk (the drink). The meaning of concepts, in particular homonymous and polysemous concepts, is sometimes clarified in the gloss. In Cornetto the meaning of concepts is separated almost to the extreme. For example, the label steen (stone) occurs in 10 different synsets, and some senses are difficult to distinguish between (see Table B.5 in the appendix).

<table>
<thead>
<tr>
<th>Aligned vocabularies</th>
<th>Source Domain</th>
<th>Target Domain</th>
<th>Vocabulary Sizes</th>
<th>Monosemy</th>
<th>Polysemy</th>
<th>Homonymy</th>
<th>Non lexical matches</th>
<th>Sample size</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGROVOC - NALT</td>
<td>food-, agricul-tural processes, chemical substances, taxa</td>
<td>food-, agricul-tural processes, chemical substances, taxa</td>
<td>28,439 - 42,326 concepts</td>
<td>68%</td>
<td>12%</td>
<td>4%</td>
<td>16%</td>
<td>23 mappings</td>
</tr>
<tr>
<td>RKD - Cornetto</td>
<td>physically conceptual/ cultural heritage</td>
<td>lexical, domain neutral</td>
<td>3,342 - 42,326 concepts</td>
<td>8%</td>
<td>60%</td>
<td>32%</td>
<td>0%</td>
<td>24 mappings</td>
</tr>
<tr>
<td>AGROVOC - WordNet</td>
<td>food-, agricul-tural processes, chemical substances, taxa</td>
<td>lexical, domain neutral</td>
<td>28,439 - 115,424 concepts</td>
<td>32%</td>
<td>60%</td>
<td>8%</td>
<td>0%</td>
<td>25 mappings</td>
</tr>
</tbody>
</table>

Table 7.10 Comparison of mapping characteristics per vocabulary pair based on analysis of samples.
7.6.3 Mapping Analysis

We examined 25 mappings between AGROVOC and NALT and RKD and Cornetto in terms of their mapping types (Fig. 7.3). Although distinguishing between the different types is not always easy, we estimate the proportions of mapping types using the samples in Table 7.10.

We found that most of the mappings between AGROVOC and NALT are monosemes or homonyms or not linguistically related at all. The greatest source of disagreement is with monosemes caused in particular by compound concepts in AGROVOC. For example, the concept of *EQUINE ENCEPHALOMYELITIS* in AGROVOC also has the label *VENEZUELAN EQUINE ENCEPHALOMYELITIS*, whereas in NALT the latter is a narrower term of the concept. Although the labels are not ambiguous, the granularity of the vocabularies is a source of disagreement in the evaluation.

The analysis of ambiguous mappings between RKD and Cornetto showed that over half are polysemes and the remainder are monosemes and homonyms. The latter are the easiest to identify, whereas vagueness of context is a problem with monosemes during evaluation. Again, some problems can be caused by RKD concepts that unify multiple polysemic concepts into one. For example, the term strawberry is used to annotate paintings with the fruit, but also paintings with the fruit attached to the plant, and can thus be mapped to the polysemic concepts of strawberry (fruit) and strawberry (plant) in Cornetto. However, there is no suitable SKOS relation to describe the relationship between these two concepts.

7.7 Predicting Agreement in Evaluation

Based on the results of our experiments and our analysis, we have found that mapping types and vocabulary characteristics influence the manner in which raters evaluate mappings. The purpose of the following small experiment is to investigate whether we can predict agreement based on vocabulary characteristics.

In the Food task raters had to evaluate two very similar vocabularies with similar amounts of information in the same domain. Given these factors the high level of agreement between raters is not surprising. A significant number of disagreements were caused by raters making errors in their judgment and the compound concepts in AGROVOC. In the RKD-Cornetto evaluation the disagreements were mostly caused by polysemic concepts in Cornetto, as well as the fuzziness of RKD concepts.

In order to test our conclusions, we analyzed a small set of 25 ambiguous mappings between AGROVOC and Princeton WordNet\(^8\). The mappings were all created using string matching, and therefore the mapping types are monosemes, polysemes and homonyms. Among the ambiguous mappings we found that over half the mappings are between polysemes, over a quarter are monosemes and a few homonyms. We summarize the results of this analysis in Table 7.10, along with the analysis of the other samples. Again, certain mapped concepts in AGROVOC combine multiple concepts, for example *BRAIN* with *CEREBELLUM* along with other regions of the brain.

\(^8\)http://wordnet.princeton.edu/
Compared to the RKD-Cornetto mappings we found the same number of polysemes but fewer homonyms and more monosemes. We predict that the agreement will be higher than in the RKD-Cornetto evaluation, because AGROVOC is a domain-specific vocabulary and its concepts have a clearer context. We also expect most of the disagreement to be on whether the mapping is of narrow/broad or unrelated type. Thus, when it comes to Krippendorff’s Alpha, the value for 2 categories will be significantly higher than for four categories.

7.8 Testing the Prediction: AGROVOC-WordNet

7.8.1 Agreement measures

We performed an evaluation of 100 mappings between 28 AGROVOC concepts and 100 WordNet concepts. The mappings are ambiguous (one to many) mappings comparable to the RKD-Cornetto mappings, where we mapped a smaller vocabulary to a lexical resource. We measured the inter-rater agreement for four categories and two categories and show results in Table 7.11.

<table>
<thead>
<tr>
<th>AGROVOC - Wordnet</th>
<th>Alpha for all raters</th>
<th>Rater A - Rater D</th>
<th>Rater A - Rater B</th>
<th>Rater B - Rater D</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 categories</td>
<td>0.713</td>
<td>0.731</td>
<td>0.673</td>
<td>0.728</td>
</tr>
<tr>
<td>2 categories</td>
<td>0.722</td>
<td>0.752</td>
<td>0.646</td>
<td>0.758</td>
</tr>
</tbody>
</table>

Table 7.11 Inter-rater agreement for 4 and 2 categories between all raters and pair of raters in the AGROVOC-WordNet evaluation. Agreement for 2 categories is not much higher than for 4 categories.

We found that the overall agreement is somewhat higher than in the RKD-Cornetto experiment; however, there is no greater gain in agreement for two categories. We do find that the differences between pairs of raters is smaller (0.058 vs. 0.14) although this could be accounted for by learning effect and raters becoming more adept at evaluating mappings.

Overall, agreement for 2 categories is slightly higher than for four categories, except for the agreement between Rater A and B. This is caused by prevalence bias; although the number of mappings the raters agreed on was slightly higher for 2 categories (84 vs. 83), in terms of Alpha measurement, agreement in a smaller category is much more important than in a large category, the latter being more likely probabilistically. The additional mappings the raters agreed on for two categories fell into the larger non-equivalent category, and thus agreement is comparatively lower.

In this experiment we were unable to predict the level of agreement between raters. We now perform a qualitative analysis in order to understand why this was the case.

7.8.2 Qualitative Analysis

We now look at the mappings in more detail. Of the 100 mappings, raters agreed fully on 27 mappings. For an additional 53 mappings the disagreement was very minor (related vs. unrelated
and close vs. exact). Of the remaining 20 mappings, two raters selected the same category for 15 mappings, and all raters selected a different relation for 5 mappings. When we reduce the four categories to two, we find the raters still disagree upon 17 mappings.

We found that very few disagreements were caused by human error. The main reason for disagreement were compound concepts in AGROVOC. For example, the concept of \textit{SHIPBUILDING} and \textit{NAVAL ENGINEERING} are merged in AGROVOC, and separate concepts in Wordnet. As raters had insufficient information on the perspective from which the concept is used, they chose different relations based on their internal meta-guidelines. Given that raters selected various relations such as exact, narrow/broad and related for these mappings, the SKOS matching vocabulary can be considered inadequate. Similar problems occur with vaguely defined concepts that are linked to more clearly defined concepts, such as \textit{MAURITIUS}, which in one vocabulary had no context, and in the other vocabulary was defined as two distinct concepts, the country (administrative region) and the island (geographic region). Instructions on how to deal with such concepts could be given in guidelines, although it is likely to be difficult to cover all cases. We would like to refer the reader to Table B.6 in the appendix for more examples of disagreements.

### 7.8.3 Precision measures

Again, we also measure precision for each rater where we focus on the differences in precision values rather than the level of the precision. Because we evaluated ambiguous mappings we expect the precision to be low. We calculate precision per rater by adding up exact and close match mappings from Table 7.12 and dividing it by the total number of mappings. We measured a precision of 0.25 for Rater A, 0.36 for Rater B and 0.32 for Rater D. We also calculated precision by applying majority voting on disagreements and found a value of 0.30. When testing for significance, we found that the measures were not significantly different. However, the distribution of categories for 17 disagreements shows some variation between raters as illustrated in Fig. 7.4.

<table>
<thead>
<tr>
<th></th>
<th>equivalent</th>
<th>non-equivalent</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Exact</td>
<td>Close</td>
</tr>
<tr>
<td>Rater A</td>
<td>17</td>
<td>8</td>
</tr>
<tr>
<td>Rater B</td>
<td>10</td>
<td>26</td>
</tr>
<tr>
<td>Rater D</td>
<td>20</td>
<td>12</td>
</tr>
</tbody>
</table>

\textbf{Table 7.12} Distribution of the mapping categories on the AGROVOC-Wordnet mappings per rater.

### 7.9 Discussion

In this chapter we described large scale experiments in alignment evaluation. Our goals were twofold:
Figure 7.4  Graphical representation of the disagreements between raters. Raters disagreed on 17 of the 100 AGROVOC-WordNet mappings on 2 categories. The equivalent matches are in white and inequivalent matches in black. The distribution of equivalent vs. inequivalent matches varies greatly per rater.

First, to gain insight into which mapping and vocabulary characteristics influence manual evaluation of alignments, and whether we can predict rater agreement based on these characteristics.

Second, to study how disagreements between raters affect quality measures such as the precision of an alignment, and the ranking of alignment systems.

In our experiments we focused on data that made the evaluation task subjective. For example, in our RKD-Cornetto experiment we used mappings between vocabularies that model (parts of) the world in different ways, thus making the categorization task more difficult. We have not been able to determine a priori whether three raters will agree to such a degree that one rater would be sufficient for an evaluation. We were also unable to predict the level of agreement between raters. Our analysis based on the vocabulary characteristics, mapping types, and the available mapping relation scheme did not lead to a correct prediction because we do not know how these different characteristics interact with one another. There are a number of other factors that can not be fully analyzed or predicted, such as errors (human and vocabulary errors) and the effect of evaluator bias. Manual evaluation is a difficult task and more research is needed into the effects and interactions of these factors.

Although some disagreements, such as those caused by human error cannot be prevented, detailed guidelines help improve agreement by reducing disagreements caused by vocabulary error, differences in vocabulary models, and the application of matching relations in particular mapping scenarios. Following the evaluation, majority voting can be applied to eliminate human error, and to reach consensus on the matching category of mappings.

Based on our experiments we propose the following principles for manual evaluation:

1. Analyze the alignment vocabularies and mappings with respect to their type based on a small sample.
2. Analyze mappings qualitatively with respect to the type of relations that occur.
3. Define matching relations based on step 2. Relations may include SKOS matching relations or tailor made ones.
4. Select at least three raters for the evaluation.
5. Measure and report agreement using an inter-rater measure such as Krippendorff’s Alpha.

6. Assess the effect of disagreements between raters and select a method for handling them (e.g., majority vote, consensus or meta-rater).

7. Publish the details of each step.

By publishing the details of the manual evaluation, it will be easier for the alignment community to identify problems with alignments and the manual evaluation.

Our conclusion with respect to our second research question is that the effect of disagreements on the evaluation results cannot be predicted based on inter-rater agreement measures alone. We found that the distribution of disagreements needs to be analyzed based on the following orthogonal aspects:

1. Distribution of disagreements over the evaluation categories
2. Distribution of disagreements over the evaluated systems
3. Distribution of disagreements over the sampled strata

For each of these aspects there are a number of possible cases:

With respect to the disagreements in the evaluation categories, there can be a similar number of approved (equivalent or other) mappings for each rater, but the disagreement may be high. In this situation the disagreements cancel each other out with respect to precision, resulting in similar precision measures for each rater. Another possibility is that raters vary in strictness, but the overall ranking of systems remains the same.

Disagreements can be distributed evenly across various systems resulting in a similar relative ranking, but potentially different absolute precision measures. In this case the disagreements have no direct impact on precision measures. When disagreements are not distributed evenly, they may lead to entirely different ranking per rater.

Finally, when using stratified sampling, the ratio of the sample size and the stratum size can affect precision measures by enlarging or reducing the effect of disagreements. In our first experiment, disagreements had relatively small impact on precision and ranking of systems, despite disagreement affecting a significant portion of the evaluation. This situation is an exception rather than the rule.

More research is needed on the impact of different evaluations and disagreements between raters. However, in order to be able to perform this research, evaluations need to be properly described and data has to become available for re-evaluation. In the OAEI for example, the method used to create reference alignments should be described in detail. Although for ongoing tracks the reference alignments are not meant to be public, they should be made available for this type of research.
Conclusions

In this thesis we focus on processes and methods in vocabulary alignment, in the context of the broader problem of integrating cultural heritage collections into a single large collection. The model of the steps (Fig. 8.1) necessary for this integration is simple, but has proven to be useful in delimiting the necessary tasks.

In the Introduction we presented the problem statement of the MultimediaN E-Culture and Europeana Connect projects, which form the context of this thesis: How to integrate multiple cultural heritage collections into a single virtual collection. In Chapter 2 we propose a model that defines the activities necessary for integrating a collection into a larger virtual collection and illustrate the steps using the Bibliopolis collection and vocabulary. This model distinguishes between the following tasks:

1. Conversion of the vocabulary
2. Alignment of the metadata schema
3. Metadata value mapping (semantic enrichment of the metadata)
4. Vocabulary alignment

This four step model, shown in Fig. 8.1, has been used to guide the integration of collections in the E-Culture project and in Europeana. With an increasing number of data sets being made public as Linked Open Data (LOD)\(^1\) (Bizer et al. 2009), the LOD cloud has grown and is growing in size. At the beginning of our conversion and integration efforts we were linking collection vocabularies and metadata values to other vocabularies within the virtual collection. A number of these vocabularies and collections have become part of the LOD cloud (e.g., WordNet). Thus, linking to central data sets (hubs) in the LOD has become part of the method for integrating collections and vocabularies into a virtual collection, and indeed a larger whole.

The work in Chapter 2 defines the context for the remainder of the thesis where we focus on the fourth step: vocabulary alignment. This choice leads to our main research question:

\[ \textit{How can we combine vocabulary alignment techniques, assess their performance and evaluate alignments?} \]

\(^1\)http://linkeddata.org/
The first step is conversion of the vocabulary schema and data. The second step is the conversion of the metadata schema. The third step is the metadata value mapping. This includes replacing metadata values with concepts from the vocabulary. The last step is the alignment of the vocabulary to other (standard) vocabularies.

We split our main research question into four smaller research questions. In the following section we revisit each of these in detail. We then explore the implications of this work and conclude with a discussion of future work.

### 8.1 Revisiting Research Questions

#### 8.1.1 How can we combine vocabulary alignment techniques and assess their performance?

In Chapter 3 and Chapter 4 we aligned cultural heritage vocabularies to lexical resources using mostly simple techniques based on lexical similarity. These vocabularies are synonym rich, and lexical techniques can produce good results. Similarly, in the medical domain (Chapter 5), vocabularies contain concepts with numerous labels and large number of mappings can be found using simple lexical matching techniques. More complex alignment tools, such as those participating in the OAEI, generally combine multiple techniques in order to generate mappings. While it is often possible to change settings, and turn components on and off, determining the precise configuration of settings is cumbersome, and pinpointing the technique that generated a particular mapping is difficult.

Because we wanted to compare the performance of two off-the-shelf tools to a simple technique, we used a complex alignment tool (Falcon) along with a lexical matching tool (STITCH), and a simple string matching algorithm in Chapter 3. In Chapter 4 we were unable to use off-the-shelf tools due to the size of the vocabularies, and experimented exclusively with combinations of simple lexical techniques tailored to the characteristics of the vocabularies.
As the techniques we applied were simple, we needed several iterations to achieve a certain quality of the alignment and aligned the vocabularies in a step-by-step manner (Chapter 3, Chapter 4). Such an approach makes the analysis and assessment of the resulting mappings more straightforward. The individual techniques may perform poorly by themselves, but by combining them we can improve the quality and/or quantity of mappings. For example, we can improve the precision of an ambiguous set of mappings by applying a disambiguation step. We also found that due to the variation in vocabulary characteristics, the alignment techniques may require tailoring, as the characteristics can influence their performance. For example, in Chapter 4 the characteristic that determined the choice of technique was that the source vocabulary had preferred labels in plural form, whereas in the target vocabulary all labels were in singular form. In this case, the simple string matching technique resulted in low precision.

In Chapter 3 we separate distinct techniques to a limited degree. Falcon combines several techniques in a single matching tool, and the STITCH mappings are also the product of several lexical matching techniques. At the same time we generate mappings separately from the disambiguation step, thus gaining more insight into the quality of mappings in each step. In Chapter 4 we chose simple techniques and applied them in a step-by-step process, and thus are better able to determine the quality of a mapping in relation to the technique that generated it and the characteristic of the vocabulary.

Separating mappings into segments, for example by isolating overlapping and non-overlapping mappings, and distinguishing between ambiguous and non-ambiguous mappings has proven useful for assessing the value of mapping techniques. In Table 8.1 we present a summary of our quantitative results from Chapter 3 and Chapter 4. We use precision, coverage and source coverage for comparison as these measures were used in each experiment. For RKD-Cornetto and AATNed-Cornetto the level of precision and coverage of the non-ambiguous techniques, which include the Baseline from RKD-Cornetto, is comparable. The precision and coverage of the AAT-WordNet non-ambiguous techniques is somewhat lower than both RKD-Cornetto and AATNed-Cornetto. Overall, these techniques have the highest level of precision and lowest coverage and map the least number of source vocabulary concepts (source coverage). The ambiguous techniques (STITCH and Lexical) have the highest coverage and the lowest precision. The precision of the ambiguous techniques (0.24 to 0.53) varies more than the precision of non-ambiguous techniques (0.82 to 0.94). For the RKD-Cornetto alignment a semi manual selection and combination of mapping set increased coverage and source coverage, while maintaining high precision in comparison to the Baseline result. The semi-manual combination set also has a higher precision than Falcon, the off-the-shelf fully automated alignment tool. In Chapter 4 the data sets were too large for Falcon and we had to use simple alignment techniques. In this case the semi-manual selection and combination of mappings provided reasonable results, improving upon the coverage and source coverage of the non-ambiguous techniques, and had much higher precision than the Lexical mappings. Thus, where complex tools may fail, a strategic combination of mapping sets based on evaluated samples can lead to acceptable results.

From the results of our experiments (Chapter 3, Chapter 4) we found that a good workflow
Table 8.1 Overview of the results of the the RKD-Cornetto (Chapter 3), the AAT-WordNet and AATNed-Cornetto alignments (Chapter 4) in terms of precision, coverage and source coverage. For RKD-Cornetto we list the results of the three mapping generation techniques, the results of all mappings, and the results of the combination of mappings from Section 3.7. For AAT-WordNet and AATNed-Cornetto we show the results of the combined non-ambiguous techniques, the results of the Lexical technique and the results of the combinations of alignment techniques from Section 4.6.3 (Segments 1+2+3 from Table 4.8).
consists of the following steps: tailoring lexical matching techniques based on vocabulary characteristics, applying techniques, determining overlap between the different techniques, applying a disambiguation step and finally combining segments of mappings based on an analysis of the improvement in precision. Additional matching steps can be applied if necessary, and the resulting mappings can be selected based on the level of precision and coverage required. Combining alignments is thus an interactive process.

8.1.2 How do vocabulary and mapping characteristics influence manual assessment?

Manual assessment is necessary for determining the quality of alignments; however, we found in several experiments (Chapter 3, Chapter 4, Chapter 5) that agreement between raters is low. In Chapter 6 and Chapter 7 we studied the process of manual assessment focusing on the causes of disagreements between raters in particular. We found that three elements influence manual assessment: vocabulary characteristics, mapping characteristics and rater characteristics. Each element has predictable and unpredictable aspects. The following list of each element describes their effects on manual assessment, and on the level of agreement between raters in particular:

1. **Vocabulary characteristics**

   **Vocabulary error**  Inconsistencies and errors in vocabularies which cause discussion between raters lead to disagreement. Raters wonder whether the concept is meant to be interpreted in a particular way, or whether there is a mistake in the vocabulary. When these types of errors are known, we can create guidelines with instructions on how to deal with them. For example, we were aware that Cornetto (Chapter 7) has a significant number of concepts where the gloss was incorrect and contradicted the position in the hierarchy. We included the instruction in our guidelines that when raters detect such concepts they are to ignore the gloss and interpret the meaning of the concept based on the hierarchy. In other cases defining guidelines can be more difficult. For example, in several vocabularies, non-synonymous labels were aggregated into a single concept; for instance, the labels GEBOUW (building) and PERCEEL (plot) in Cornetto form one synset.

   **Vocabulary representation**  Varying vocabulary representations (eg: ISO standard, SKOS, OWL-DL) make “translation” from one representation to another difficult for raters. For example, the same concept is represented in a different manner in a thesaurus than in a lexical resource. In the first, the concept has a unique preferred label, perhaps with additional alternative labels, in the latter several synonyms are grouped together forming a synset, and these synonyms would not necessarily be unique. These different characteristics influence the way concepts are perceived and the relations between them.
**Vocabulary model**  The conceptualization of concepts, the manner in which they need to be interpreted, varies per vocabulary but also from concept to concept. For example, the RKD subject thesaurus (Section 7.5) was built for annotating cultural heritage objects, and thus its concepts are interpreted broadly. For example, the concept **stone** is used to annotate works where stone is the building material, works containing small pieces of stone, and works where stone is part of the landscape. In Cornnetto these are all separate concepts. Thus, the RKD thesaurus point of view of concepts is different from Cornnetto. Raters in this case selected various relations depending on what they thought was best. To prevent disagreement, guidelines should include examples of situations where the differences in vocabulary model require a certain interpretation and instruct which mapping relation needs to be selected.

**Vocabulary domain**  The domain of vocabularies, and the manner in which these domains overlap, influences how raters perceive relations between concepts. In some cases assessing relations between concepts may require domain specific knowledge, or clear understanding of the intended use of concepts. There is also a difference in the type of relations between two domain specific vocabularies and between a domain specific and a generic vocabulary. For example, two domain specific vocabularies from the same domain likely use the same terms that mean the same thing. Linking vocabularies from different domains may require expert knowledge.

The type of relations between two domain specific vocabularies from different domains are unlike relations between a domain specific vocabulary and a generic vocabulary, or relations between two domain specific vocabularies from the same domain. In the last case, concepts with the same label are more likely to be evaluated as the same than in the two first cases. For example, in the evaluation of mappings between Agrovoc and NALT (Section 7.4) raters had the highest agreement of all our experiments. Agrovoc and NALT both describe the same agricultural and food domains, whereas AAT is a cultural heritage vocabulary and WordNet covers a more generic, albeit detailed domain. Although we have no examples of evaluation analysis between two vocabularies from different domains, in Chapter 5 we used mappings between various medical ontologies. These ontologies are all in the medical domain but describe different areas such as anatomy (Foundational Model of Anatomy), genes (Gene Ontology) and other medical subjects. The relations between concepts from these ontologies can also be a source of disagreement, even amongst expert raters.

A method for dealing with disagreements caused by disparate domains or scope includes an analysis of the coverage of vocabularies, and guidelines for assessing relations between domain specific and generic concepts. For example, in Fig. 6.3 we gave the example of multiple concepts in AAT (**preservationists**, **restorers** and **conservators**) linked to a single WordNet concept (**renovator**). Guidelines could include the rule that in such cases the AAT concepts are narrower than the WordNet concept.
2. Mapping characteristics

**Large scale ambiguity** Mappings to resources with many ambiguous labels tends to overwhelm raters. When a concept is linked to a multitude of concepts raters are unable to distinguish between them, and the choice of categories becomes almost random. In Chapter 6 we had the AAT concept of flow mapped to 14 WordNet concepts, each of which had flow as sense-label. Although there was agreement on some of these mappings, for many mappings raters had difficulty in telling the difference between WordNet synsets.

**Lack of a relation metric** In our evaluation experiment (Chapter 6 and Chapter 7) we found that categories do not have strict boundaries. Rather than being one relation or the other, possible relations between concepts form a continuum. For example, in our experiments one rater would consider the same pair of concepts a close match, while another rater considered them an exact match. Thus the boundary of what constitutes an exact or close match varied per rater. Raters were also inconsistent in their own definition of boundaries between categories, and judged the same or similar mappings in a different manner. The arguments for such choices were usually expressed as a “feeling” that the distance between the concepts was too great for a relation to be true. Guidelines could define examples of “distances” between concepts; however the “grey area” in between two mapping relations would remain.

**Lack of suitable matching relation** We have found multiple cases where a matching category was found inadequate or too broad to be useful. The “related” category in our experiments with alignments to WordNet and to Cornetto (Chapter 6 and Chapter 7) was used for varying relationships, such as metaphor and metonym relations, for lack of a better category. 

In several experiments we had mappings between one concept combining several terms from one vocabulary, to two distinct concepts in the other vocabulary (e.g.: Agrovoc in Chapter 7, MESH in Chapter 5). In order to consistently evaluate such mappings we need to define a new matching relation such as a sibling relation.

3. Rater characteristics

**Human error** In each of our experiments (Chapter 6, Chapter 7) we found examples of human error where raters selected an “incorrect” category due to inattentiveness or lack of domain knowledge. Typical errors include raters not noticing labels or scope notes, clicking on the wrong button, or general inattention caused by fatigue. The number of human errors could be decreased by reducing the number of mappings that need to be evaluated in one session, or by adding more supportive features to the evaluation tool. However, human error will always occur in evaluations to some degree.

**Evaluator bias** The behavior of raters is influenced by their professional and cultural background and their experiences, which in turn influence the meta-guidelines they
create on the fly. In evaluations we can select raters based on their background, but the other aspects of their knowledge is beyond our control. The differences between raters manifest themselves in their meta-guidelines, the rules they create when confronted with certain mappings. In our analysis of the experiments we also found that raters sometimes disagreed with themselves and had difficulty understanding their previous choices. Intra-rater reliability is thus also an issue to contend with.

Thus, the vocabulary, mapping and rater characteristics each influence manual assessment in multiple ways. Some problems (e.g., caused by vocabulary error, vocabulary model) can be prevented by creating guidelines that provide instructions for problematic cases. Others, such as the lack of a relation metric and the rater bias are more difficult to deal with as fuzziness and inconsistency is inherent in humans in their perception of concepts and matching categories.

8.1.3 How can we evaluate large alignment sets based on manually assessed mappings?

In Chapters 4 to 7 we needed to assess the quality of large sets of mappings. As a full manual assessment was unfeasible because manual evaluations are time-consuming, we needed to select smaller sets of mappings to evaluate manually. (The evaluation of 400 mappings by one rater took on average 5 hours in Chapter 7.) The evaluation problem thus becomes one of selecting the right statistical method. The simplest method is drawing a random sample from a heterogeneous set of mappings. If the random sample is large enough, the likelihood that it is a representative sample increases.

However, a random sample only gives us a global picture with no indication of interesting subsets or ways to improve mappings. Instead, we need to identify subsets of mappings with comparable characteristics and sample each of these sets using stratified random sampling. In Chapter 3 and Chapter 4 we identified mappings with similar characteristics by isolating the overlapping and non-overlapping mappings from each alignment technique used and selecting subsets of ambiguous and non-ambiguous mappings. By identifying these subsets we are better able to understand the differences caused by each alignment technique. We can retain certain combinations of mappings and leave out others; therefore the resulting quantity and quality of mappings can be adapted on a case by case basis. In comparison, off-the-shelf tools produce confidence thresholds which are difficult to interpret and vary per tool. Identifying the ideal threshold can take a lot of effort as mappings from each confidence interval need to be analyzed and assessed.

8.1.4 What is the influence of manual assessment on quality measures for alignments?

In Chapter 7 we studied how differences in the opinion of raters influences quality measures such as precision. We found that while low inter-rater agreement measure is a good indicator that the disagreements will result in significant variance, a high agreement (Krippendorff’s alpha of 0.8)
is not a guarantee for stable quality measures. Disagreements between raters do not necessarily affect quality measures in an obvious manner. Their distribution across the alignment needs to be analyzed in detail. There are three situations to consider:

1. The distribution of the disagreements across mapping categories directly influences the precision measure per rater. Here are two examples with similar levels of disagreement manifested in a very different manner. In the first case, if raters have evaluate mappings in similar proportions, that is, the ratio of correct/incorrect mappings is the same for each rater, the precision measures are comparable for each rater. In this case, even though the raters may disagree on the entire set of mappings, this disagreement is not apparent as it is cancelled out with respect to precision. In the second case, when one rater is more strict than the other rater, the precision measures can be very different although the level of disagreement is still the same as in the previous case. Thus, we need to not only measure the inter-rater agreement level but also determine how the disagreements are distributed per matching category and per rater.

2. The distribution of disagreements across samples from each stratum (representing an overlap section) needs to be examined. The sample size and the stratum size can affect precision measures by enlarging or reducing the effect of disagreements. A large number of disagreements from a small stratum may have a smaller impact on overall precision than a relatively small number of disagreements from a large stratum.

3. If the purpose of the manual assessment is to compare the quality of various alignment systems we need to analyze the distribution of disagreements across systems. For example, if disagreements are evenly distributed across systems, the ranking will remain stable for each rater. However, if disagreements are concentrated in mappings found by only one system, the ranking of this system can vary depending on which raters’ assessment is used as a gold standard.

8.2 Discussion

In this section we discuss the implications of the work in this thesis. We consider the nature of mappings, practical implications for the OAEI and give methodological advice based on our experiences.

8.2.1 What’s in a Mapping?

Before discussing the nature of mappings we must consider theories in categorization. A mapping between two concepts is defined by the relation between them. When algorithms generate mappings or humans assess the quality of mappings, they perform categorization tasks.

Aristotelian (classical) categorization defines categories as distinct entities characterized by properties shared by all members. Categories are clearly defined, mutually exclusive and collec-
tively exhaustive. Although alignments and mappings are not explicitly defined in these terms we derive some assumptions from classical categorization. For example, our intuition is that a mapping is either equivalent or non-equivalent. There should not be a grey area. In carefully constructed ontologies, if a mapping results in an inconsistency then either the mapping is incorrect or some parts of the ontology need to be adjusted. In other words, intuitively we feel that mappings can be categorized in an unambiguous way as long as the mapping category which describes the relation between concepts is defined in a clear and unambiguous manner.

There are cases where we can agree on the category, for example where categories have been painstakingly defined by a group of individuals. In chemistry there is agreement on the naming of chemical elements, compounds and techniques. The same is true in many disciplines from engineering to mathematics. There are also categories that are internalized by everyone, such as physical objects. For example, we generally agree on what an apple is. In modern theories on categorization a distinction is made between basic level categories (the level at which people mostly conceptualize cognitively and linguistically), the superordinate level (a more abstract level where similarities between concepts are harder to define) and the subordinate level (a more specific level than the basic level). Thus, an apple is at the basic level, fruit at the superordinate level and Granny Smith (green apple) at the subordinate level. According to basic level theory, humans find it easier to categorize at the basic level than at the superordinate and subordinate levels. In our evaluation experiments we found that raters tended to agree more on tangible concepts than abstract concepts, which are often higher level and more arbitrary. At the same time concepts on the subordinate level tend to be domain specific, requiring domain knowledge for correct categorization.

In our experiments we found that there are cases with reasonable consensus. For example in the Agrovoc-NALT evaluation raters agreed mostly on mappings on geographical locations. There were however cases where the concepts and their relation was more difficult to define. For example, while most people would agree that there is a relation between AAT concepts restorer, preservationist and the WordNet concept of restorer as illustrated by Fig. 6.3, pinpointing the precise relation is found to be more difficult. The labeling of this relation is likely to vary from person to person. This is because these concepts have fuzzy boundaries forming “clouds”. In modern theories of categorization such as prototype theory described in detail by George Lakoff in his book (Lakoff 1987) concepts are viewed as irregular clouds with no clear boundaries. Prototype theory suggests that categorization is based on prototypes with degrees of membership and fuzzy boundaries. The background, knowledge and experience of a person each affect how a concept is perceived and where it is placed within a category. Categories thus form irregular clouds without fuzzy edges, the “shape” of which is unique for each person. Thus, concepts and mappings have various degrees of prototypicality that also depend on the person dealing with the concepts. For example, to use the example of the restorer, a person specialized in cultural heritage data such as the curator of a museum would consider the AAT concept of restorer to be different, or at the very least more specific than the WordNet concept whereas, for a layman, these two concepts may overlap.

This cloud aspect of concepts means that when we consider mappings we must think in terms
Figure 8.2 This figure represents an abstract example of relationships some of which are difficult to pinpoint in mapping evaluations. The two concepts (cloud 1 and 2) overlap slightly. A, B, C, and D are representations of concepts in vocabularies.

of distance between concepts. We illustrate this case with an abstract example shown in Fig. 8.2. Each cloud represents a concept or notion as it could be viewed by a person, and the points stand for specific representations in a vocabulary. Some relations are easier to categorize than others. Concept A is more central to Cloud 1 than the concepts B and C, and as both B and C have the same distance to A, we would say that both A and B, and A and C are close matches.

A more difficult question is then what the relation is between concept B and C. Technically, they are part of the same concept cloud however both are at the edge at similar semantic distance from the center. At the same time they are distant from each other and here the question is how this distance translates into a relation. These types of mappings tend to be difficult to evaluate as raters are unable to express the distance between two peripheric concepts.

Concept D is in a different cloud than concepts A or C and therefore we can assert that they are unrelated. Concept B on the other hand is on the edge of both Cloud 1 and Cloud 2. When raters have already assessed the relation between A and B as a close match, the relation between B and D becomes unclear. Concepts A and D are clearly different, and so how can B be close to A as well as D?

These are some of the types of cases raters struggled with in our assessments. In our use cases we find that it is not always possible to define a category or relation. Whereas there is consensus on some concepts, such as materials, there is less agreement in other areas (art styles).
This is an inherent feature of the cultural heritage domain; there is more room for varying opinions depending on the point of view.

8.2.2 Implications for the OAEI

Our work on the evaluation of alignments has practical implications on the OAEI and similar endeavors.

We recommend that gold standards and reference alignments be discussed and reviewed regularly. These discussions should cover how the gold standard can be created by whom and to what purpose. Our work on manual evaluation of alignments (Chapter 6 and Chapter 7) has shown that manual evaluation is a difficult task, where the level of difficulty is linked to the type of vocabularies and mappings. Thus, prior to the evaluation these characteristics need to assessed. The selection of matching relations needs to be given careful thought and raters have to be selected based on task requirements. Based on our analysis we find that at least three raters need to evaluate mappings. The inter-rater agreement has to be measured and published. Furthermore, an analysis of disagreements and their distribution across the sampled data can point towards issues with quality measures such as precision and the ranking of techniques or systems.

1. Analyse aligned vocabularies and available mappings
2. Analyse sample of mappings in detail
3. Define matching relations
4. Set up evaluation guidelines
5. Select raters (minimum 3)
6. Perform evaluation
7. Measure inter-rater agreement
8. Assess effect of disagreements
9. Publish detail of the evaluation

**Figure 8.3** Method for manual evaluation of alignments.

In Fig. 8.3 we present a method for the manual evaluation of alignment. We describe each step in more detail:
1. Perform a quantitative study of the total set of mappings. This study may include coverage, ambiguity (one to many, many to one, many to many) and type of the mappings (lexical or otherwise).

2. Perform a qualitative study of a sample of mappings. Perform an analysis of the type of relations (see Fig. 7.3 in Chapter 7).

3. Define matching relations. The choice of relations depends on the goal and the vocabularies. Possible choices are SKOS matching vocabulary, OWL or tailor made relations. Note that with a high number of matching relations the agreement levels will likely be lower. Define possible aggregations of relations. (e.g.: skos:closeMatch merged with skos:exactMatch).

4. Set up guidelines for the evaluation. The guidelines should cover prototypical mappings per matching relation, detail aspects of the vocabulary model that impact the assessment and instructions on dealing with vocabulary errors.

5. Select raters based on the level of expertise necessary for the evaluation task.

6. Evaluate (sample of) mappings by at least three raters.

7. Measure observed agreement and inter-rater agreement using Krippendorff’s alpha and publish agreement levels.

8. Publish disagreements between raters and report and discuss what has been done with them and why. For example, whether the mappings with disagreements were excluded, consensus was reached or majority voting was applied.

9. Assess the effect of disagreements across the matching categories, evaluated systems and/or the strata.

We hope that these proposals will be used in the alignment community and will help to identify issues in alignment and manual assessment.

8.2.3 Proposed Method for Aligning Vocabularies

Our work on alignment of vocabularies shows the added value of creating alignments in explicit steps. In a step-by-step process it is easier to keep track of the mappings and fine-tune them further if necessary. By making each step in the alignment process explicit the resulting alignment set can be tailored for each data set.

We propose the following method for aligning vocabularies (Fig. 8.4):

1. Select vocabularies to be aligned

2. Analyze vocabularies with respect to size, domain, vocabulary model and overlap of content.
3. Select alignment technique(s) and the order in which they are applied based on vocabulary characteristics and availability of other data (e.g.: instances annotated by vocabularies and external resources).

4. Apply alignment technique(s) to the vocabularies.

5. Determine how mappings created by different techniques overlap.

6. Draw samples from each set in the overlap.

7. Assess the quality of the sampled mappings.

8. Analyze the resulting mappings and determine whether another alignment technique needs to be applied on the vocabularies or (subset of) mappings.

9. Repeat steps 2 to 8 if needed.

### 8.3 Future Work

Our goal in this thesis was to use existing tools and techniques and develop working methodologies for aligning and evaluating vocabularies that are not grounded in specific techniques, because each alignment problem is unique.

There are several questions that remain unanswered in this thesis, in particular with respect to manual evaluation, and can provide the basis for future work.
First and foremost is the problem of dealing with disagreements. In each evaluation experiment there was some disagreement between raters (on at least 10% of the mappings). This is likely to be the case in all evaluation scenarios involving multiple raters. Regardless of the purpose of the evaluation something needs to be done with disagreements. We have no single solution to this problem but can offer some suggestions. The first option is to apply majority vote to select the “correct” mapping category. This is a useful method for filtering out errors by raters. Another (complementing) solution is to seek consensus between raters, which can also be done on mappings where majority voting cannot be applied. Third, a meta-rater with seniority can have final say on the mapping relation. Finally, the mappings with disagreements can be excluded from the alignment set. We do not recommend this last option as these contested mappings can have added value.

In the evaluation experiments we found that the SKOS matching relations were not always applicable to the relation between concepts. One solution is to extend the existing relations with new relations or define a new set of matching relations that are a better representation of actual relations between concepts. Defining relations and applying them is a difficult task, as demonstrated by the low agreement in the work of Halpin et al. (2010). Furthermore, more matching categories lead to lower agreement when measuring Krippendorff’s alpha (or any other agreement measure) as the chance that raters will diverge in opinion grows. More work is needed in particular in studying the suitability of SKOS matching relations.

We have discussed basic level categories in Section 8.2.1. In theory basic level categories are more easily recognized by humans than subordinate or superordinate categories. While we found some evidence in our experiments to support this, we have not studied the matter in detail. Based on categorization theory our hypothesis is that basic levels are easier to categorize for raters. An exception would be for domain specialists who find subordinate level categories (in their field) easy to categorize. Experiments would have to involve resources where basic level concepts have been identified.

In Chapter 7 we presented Fig. 7.3 which depicts the types of mappings based on the alignment technique used. In our analysis we found that the possible relations between mapped concepts are constrained by whether the mapped concepts share a label or not. In terms of SKOS relations for example, synonyms are likely to be exact or close matches and unlikely to be broader/narrower matches. Homonyms are likely to be unrelated whereas polysemes are likely to be close or related matches. A study of the connection between the type of mapping the relations would improve the process of selecting matching relations in a manual evaluation. For example, if all mappings are based on string matches there may be a core set of relations that would cover most possibilities.

In the composition of mappings, most of our results confirm our expectations. Compositions among mappings with overlapping content are of higher quality and more likely to occur. The way concepts are modeled in vocabularies influences the quality of composed mappings. However, the relationship between the quality of input and composed mappings remains unclear. In our research we limited ourselves to compositions of length two. However, with more and more vocabularies being added to LOD longer mapping chains are possible. How would such longer chains of map-
pings behave? Would the relation between concepts at either end degrade or transform into other kinds of relations. Some work has been done in this area by Halpin et al. (2010). They analyzed `owl:sameAs` mappings in LOD and found that these mappings cover a wide range of similarity relations. What would be the outcome of chaining mappings with different similarity relations?

In September 2011 the Linked Open Data cloud contained 295 data sets with over 31 billion RDF triples connected by 504 million links. By the time this thesis is printed these numbers will likely have increased substantially. However, the number of links between data sets is quite small compared to their relative size. The methods presented in this thesis provide guidance to those interested in creating links between their vocabularies to the vocabularies in LOD, and in assessing the quality of those links.

Creating and evaluating ontology alignments is a complex endeavor that requires technical, methodological and theoretical insights to be solved. Heterogeneity is at the core of this complexity: heterogeneity in the ontological sources, the variety in techniques and diversity of viewpoints that evaluators can entertain. This thesis has explored some of the issues in alignment and can serve as a basis for improving the science of ontology alignment.
A.1 RKD-Cornetto Experiment Guidelines

Evaluating mappings by hand

Dear participant,

Thank you for agreeing to evaluate mappings by hand.

The goal is to evaluate candidate “exact match” mappings between RKD subject thesaurus concepts and Cornetto, a Dutch WordNet-like vocabulary. You will be asked to evaluate approximately 50 RKD concepts. The number of mappings may vary as some concepts only have a single mapping, but others may have as many as six. The RKD subject thesaurus is used by RKD for annotating the subject matter, in other words whatever is depicted in cultural heritage objects. The objects are mostly paintings but there are photographs, etchings, as well as photographs of various artworks such as frescos. Each mapping has to be sorted into one of the following categories:

**Approved:** The source and target concepts both mean the same thing, i.e., it is a proper skos:exactMatch relation.

**Broader:** You think the target concept should be a broader term than the source concept. For example, “Cornetto milk product” is broader than “RKD cheese”.

**Narrower:** You think the target concept should be a more specific term than the source concept. For example: a “Cornetto waist-jacket” is a type of “RKD jacket”.

**Related:** The two concepts are not an exact match but you think are closely related. For example a ”silver can” is related to the concept ”silver”. Another example is the fruit ”fig” and a ”fig tree”. Related is one of the more difficult relationships to define.

**Not sure:** You feel there is a relationship between the two concepts but none of the above relations are appropriate. Another option is that the term is used in a confusing or contradictory fashion.

**Rejected:** The two concepts are definitely not the same, nor do they have any other direct relationship with each other as listed above.
Some additional comments:
The RKD thesaurus is meant to annotate visual imagery. Although the thesaurus contains nonvisual terms such as spring, februari and happiness if you look in the higher level of the hierarchy you can see that these are allegories or personifications. Therefore the Cornetto concept of “happiness” would not be an exact match to “happiness” in RKD.

The RKD thesaurus itself is occasionally used in an ambiguous way. Some terms could have multiple meanings. For example: ”pink” can be a ”koe” (cow) or ”schip” (ship); and artworks could be annotated erroneously in both ways. Please ignore such erroneously used terms.

How to use the tool

You may access the evaluation tool at eculture.cs.vu.nl/exp/session/mappings Now you should have the view in Fig. A.1.

Figure A.1 Annotated screenshot of the evaluation tool. The description of the elements is found in the “How to use the tool” section.

Before you start with the evaluation, please login by clicking on the Login link on the top (marked by 1 on the screenshot). Your username is your first name in lower case letters, and the
password is “valide”. After logging in you see the same screen as in Fig. A.1 except you are logged in.

First, select the file with your name in the top list (2). The mappings to be evaluated are now loaded on the web page. The left hand area (3) displays the RKD source concept in bold. The higher levels of the hierarchy are also shown to provide context. We also display thumbnails of at most 5 artworks annotated with this concept. In addition, you may click on detail panel (6) to see all the information connected to this concept and navigate the RDF graph.

The target Cornetto concepts (4) are displayed on the right. You can click on one of the 6 buttons (5) to sort each mapping. We display the parent concepts for each target concept to provide context. As Cornetto has not been used for annotating artworks we cannot display any artworks.

Clicking on one of the matching category buttons (5) makes the target concept vanish. If there are other mappings the next target concept jumps up. You can click on detail panel (6) here to see all information about the new concept. For Cornetto the additional information can be useful as the tree only shows one label and some concepts have several labels.

Some additional comments:

• The concepts were selected randomly but are presented alphabetically.
• There is no undo button.
• You need to finish your evaluation session in one stretch as after one hour the session is finished and you would have to start over again.

You can now proceed to evaluate the mappings assigned to you.

Hope you have fun!

A.2 Guidelines for the Evaluation Experiments in Chapter 6

A.2.1 AAT-Wordnet Alignment Evaluation

Dear participant,

The goal is to evaluate candidate exact/close match mappings between the Gettys Art and Architecture Thesaurus (AAT) and Princeton WordNet version 2.0. You will be asked to evaluate 73 mappings. Some AAT concepts can have candidate mappings to multiple WordNet synsets. Each mapping has to be sorted into one of the following categories:

**Exact Match:** The source and target concepts both mean the same thing. The concepts can be used interchangeably in most applications. E.g.: “Orange (fruit)” and “orange (type of fruit)”
Close Match: The source and target concepts almost mean the same thing; they are synonyms and can be used interchangeably in some applications. There are multiple situations where this occurs. First, the two concepts can be in a somewhat differently ordered hierarchy representing different viewpoints. Second, specific to the cultural heritage domain, styles, techniques and adjectives for that style or technique can be used interchangeably and are thus close matches. An example of the first case is clergy where one hierarchy is from the perspective of a group and the other from the perspective of person/people. An example of the second case is “pointillism”, which is an art style and a painting technique. While the two are different ontological types, they are nevertheless used interchangeably and should be close matches.

Broader: You think the target concept should be a broader term than the source concept. E.g.: “milk product” (target) is broader term than ”cheese” (source).

Narrower: You think the target concept should be a more specific/narrower term than the source concept. Artist in one vocabulary refers only to visual arts (painter, sculptor) while in the other vocabulary it also refers to musicians and performing artists. E.g.: a ”waist-jacket”(target) is a narrower term than “jacket” (source).

Related: The two concepts are have a relationship but are of two different types. For example: a material and object made from it such as “milk” and “cheese”, an activity and an object such as “volley ball” and “volley ball game”.

Rejected: The two concepts are definitely not the same, nor do they have any other semantic relationship of the sort listed above.

Not sure: One or both concepts are used in the thesauri in a confusing or contradictory fashion making the exact relationship unclear.

Other semantic relations you might encounter are:

1. Sibling relation where the two concepts share a common parent. Although this is a strong semantic link, these matches should be rejected. Examples are “fig tree” and “orange tree” sharing a common parent “tree”.

2. Another relationship is where one concept is a colloquial term derived from the other concept. Examples of this are “bakkie (colloquial term for cup) and bak (bucket or vat). The “mouth of a river” is also related to “mouth (organ)” in a similar fashion.

Thesaurus specific guidelines:
The gloss in Cornetto is not always paired correctly with the concept. If in doubt, use the position in the hierarchy to determine the meaning of the concept. An example is “dorpshuis” (cultural center or parish hall) which has a parent concept “gemeenschapshuis” (community house) and the gloss “huis in een dorp” (house in a village). The gloss suggests that the concept refers to
any house in a village but the parent concept indicates that it is a cultural center. Concepts with ambiguous meanings due to their position in the hierarchy should be considered carefully. E.g.: “visionair, ziener” (visionary, prophet) which is a narrower concept of “people by ideology and philosophy” would suggest the term is not about psychics but this is not entirely clear. Another example is “vakje” (compartment) where the subconcepts were “cell” and “facet”, which suggest a more Excel type compartment.

Specific examples and how to deal with them:

1. Two concepts have exactly the same label within the same context. For example the concepts have the label “dragger” which is a type of boat. The source concept has a parent concept “trawler” while the target concept both the label “trawler” and “dragger”. Because the source thesaurus has chosen to model boats in a much more specific way by separating the concept of trawler from the concept of a dragger, it would be inappropriate to choose exact or close match. Rather, the source concept “dragger” is a narrower term than the target concept which representing a set containing both trawlers and draggers making it a broader relationship.

2. AAT colors are sometimes a set of colors of similar hue with many alternative labels. As such, multiple WordNet/Cornetto colors can be linked to a single AAT color. This is a case of a small set (WN/CN) linked to a larger set (AAT) and a kind of part-of relation. For example, “very light green” has several other labels in AAT such as “blue green”, “emerald green”, “light mint”, whereas the concept “very light green” in WN has the labels “bluish green”, “blue green”, and “teal”. The two color sets do not fully overlap as teal (a dark blue green) is a different green than emerald green (a light blue green).

3. If the hierarchy and gloss (definition) of the term do not match you should ignore the gloss and use the position in the hierarchy to determine correctness. In other cases please select the “not sure” category.

In some cases more than one mapping may be correct. For example, the materials in AATNed can be mapped to adjectives and nouns.

You can now proceed to evaluate the mappings assigned to you.

A.2.2 GTT-Brinkman Alignment Evaluation

Dear participant,

The goal is to evaluate candidate exact/close match mappings between the Royal Library’s (KB) GTT thesaurus and Brinkman thesaurus. You will be asked to evaluate 70 mappings. Each mapping has to be sorted into one of the following categories:

**Exact Match:** The source and target concepts both mean the same thing. They can be used interchangeably in most applications. E.g.: Orange (fruit) and orange (type of fruit)
Close Match: The source and target concepts almost mean the same thing; they are synonyms and can be used interchangeably in some applications. An example situation could be that the concepts appear in completely different type of hierarchy. For example: the concept “blowgun” in one thesaurus has the parent concept “weapon”, while it has the parent concept “conduit” in the other thesaurus. One thesaurus takes a functional view on the concept, the other uses the form of the object to categorize it.

Broader: You think the target concept should be a broader term than the source concept. E.g.: “milk product” (target) is broader term than “cheese” (source).

Narrower: You think the target concept should be a more specific/narrower term than the source concept. “Artist” in one vocabulary refers only to visual arts (painter, sculptor), while in the other vocabulary it also refers to musicians and performing artists. E.g.: a “waist-jacket” (target) is a narrower term than “jacket” (source).

Related: The two concepts have an associative relationship and are of two different (ontological) types. For example: a material and object made from it such as “milk” and “cheese”, an activity and an object such as “volleyball” and the “game volleyball”. Generic examples of such relationships are: process and agent, action and property (e.g.: “environmental cleanup” and “pollution”), action and product (e.g.: “cloth” and “weaving”), cause and effect, concept or object and origin, material and object, discipline and object or practitioner.

Not related: The two concepts are definitely not the same, nor do they have any other semantic relationship of the sort listed above.

Not sure: One or both concepts are used in the thesauri in a confusing or contradictory fashion making the exact relationship unclear.

Note: Sibling and part-of relations are not SKOS relations. If you encounter this instance please select not related.
### Examples of Disagreements Between Raters

#### B.1 Disagreements from Chapter 6

<table>
<thead>
<tr>
<th>AAT Concept</th>
<th>WordNet Concept</th>
<th>Selected Mapping Categories</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bhutanese</td>
<td>Bhutanese</td>
<td>close (2), exact, broad, related</td>
<td>The AAT concept refers to the style. The WordNet concept is the adjective indicating a characteristic of Bhutan.</td>
</tr>
<tr>
<td>Cavaliers  (soldiers)</td>
<td>Cavalier</td>
<td>narrow (2), exact, related, unrelated</td>
<td>The AAT concept refers to fighting men on horseback, in particular Royalists. The WordNet concept refers to Royalist supporters</td>
</tr>
<tr>
<td>Flow  (physics)</td>
<td>Flow</td>
<td>broad (2), exact, close, unrelated</td>
<td>The AAT concept refers to the movement of liquids, gases, electrical charges and other materials. The WordNet concept refers to the act of streaming.</td>
</tr>
<tr>
<td>Gnostics</td>
<td>Gnostic</td>
<td>unrelated (2), exact, close, related</td>
<td>The AAT concept refers to those who practice gnosticism. The WordNet concept refers to those possessing intellectual or esoteric knowledge.</td>
</tr>
<tr>
<td>Restorers</td>
<td>Refinisher</td>
<td>exact (2), close, related, unrelated</td>
<td>The AAT concept refers to those making changes in objects to approximate their original state. Separated from preservationists and conservers. The WordNet concept refers to skilled workers employed to restore or refinish buildings or furniture.</td>
</tr>
</tbody>
</table>

**Table B.1** Examples of disagreements between at least three raters in the AATWordNet experiment. The number in parentheses behind the mapping category indicates how many raters chose that category. When no number is present only one rater selected the category.
<table>
<thead>
<tr>
<th>GTT Concept</th>
<th>Brinkman Concept</th>
<th>Selected Mapping Categories</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doop, trouw-en begrafboekenregisters (Baptism, marriage and death certificates)</td>
<td>Familieregisters (family registry)</td>
<td>close(2), broad, narrow, related</td>
<td>Both concepts lack additional context</td>
</tr>
<tr>
<td>Horeca (hospitality industry)</td>
<td>Horecabeerdrijf (catering company)</td>
<td>exact (2), close (2), related</td>
<td>The GTT concept has hospitality industry as English label and has cafes, hotels, restaurants and other similar concepts as narrower concepts. The Brinkman concept has companies as broader concept.</td>
</tr>
<tr>
<td>Milieuverontreining (environmental pollution)</td>
<td>Afvalstoffen (waste materials)</td>
<td>narrow (2), related (2), unsure</td>
<td>The Brinkman concept has environmental pollution as broader concept.</td>
</tr>
<tr>
<td>Personeelsmanagement (personnel management)</td>
<td>Personeelsbeleid (personnel policy)</td>
<td>exact (2), related (2), close</td>
<td>The GTT concept has personnel policy as alternative label.</td>
</tr>
<tr>
<td>Ouder-kind-relaties (parent-child-relationships)</td>
<td>Ouderschap (parenthood)</td>
<td>related (2), close (2), broad</td>
<td>The GTT concept has family relations as broader concept. The Brinkman concept has fatherhood, motherhood and grandparents as narrower concepts.</td>
</tr>
<tr>
<td>Kerkschatten (church treasures)</td>
<td>Religieuze kunst (religious art)</td>
<td>related (3), broad, close</td>
<td>The GTT concept has no context. The Brinkman concept has art as parent term.</td>
</tr>
<tr>
<td>Vetten (Fats)</td>
<td>Voedingsleer (dietetics)</td>
<td>related (3), broad, unrelated</td>
<td>The GTT concept is related to the concept lipids. The Brinkman concept has various substances such as minerals and vitamins as narrower concepts, but not fats.</td>
</tr>
<tr>
<td>Routeren (routing)</td>
<td>Logistiek (logistics)</td>
<td>unrelated (3), related, unsure</td>
<td>The GTT concept has no additional context. The Brinkman term has infrastructure as narrower concept.</td>
</tr>
<tr>
<td>Chourase</td>
<td>Taalkunde, overig (Linguistics, other)</td>
<td>unrelated (3), broad, unsure</td>
<td>The GTT concept Chourase is a language of Nepal and has Kiranti languages as broader concept.</td>
</tr>
</tbody>
</table>

Table B.2 Examples of disagreements between at least three raters in the GTTInstance experiment. The translation of the label is in parentheses. The number in parentheses behind the mapping category indicates how many raters chose that category. When no number is present only one rater selected the category.
<table>
<thead>
<tr>
<th>GTT Concept</th>
<th>Brinkman Concept</th>
<th>Selected Categories</th>
<th>Mapping Categories</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Merken (brands)</td>
<td>Merken</td>
<td>exact (2), close (2), unrelated</td>
<td></td>
<td>The GTT concept has logos, trademarks, silvermarks and manufacturer’s trade mark as narrower terms. The Brinkman concept has brand-law as broader concept.</td>
</tr>
<tr>
<td>Menselijk lichaam (human body)</td>
<td>Menskunde (antropobiology)</td>
<td>related (3), exact, close</td>
<td></td>
<td>The GTT concept has male body and female body as narrower concepts. The Brinkman concept has biology as broader term and human body as alternative label.</td>
</tr>
<tr>
<td>Arbeids- en organisatiepsychologie (industrial psychology)</td>
<td>Arbeidspsychi</td>
<td>exact (2), close (2), narrow</td>
<td></td>
<td>The GTT concept has management science and psychology as broader concepts. The Brinkman concept has psychology as broader concept. In Dutch the GTT concept may appear to encompass more than the Brinkman concept.</td>
</tr>
<tr>
<td>Amsterdam Zuid (Amsterdam South)</td>
<td>Amsterdam Nieuw Zuid (Amsterdam New South)</td>
<td>close (2), broad (2), unrelated</td>
<td></td>
<td>The GTT concept has Amsterdam New South and Amsterdam Old South as alternative labels. The Brinkman thesaurus separates Amsterdam New South and Amsterdam Old South, and has Amsterdam South as broader concept for the two.</td>
</tr>
<tr>
<td>Amsterdam Zuid (Amsterdam South)</td>
<td>Amsterdam Oud Zuid (Amsterdam Old South)</td>
<td>exact (2), close, broad (2)</td>
<td></td>
<td>See comment above.</td>
</tr>
<tr>
<td>Gespleten gehemelte (cleft palate)</td>
<td>schisis (cleft lip and palate)</td>
<td>narrow (3), close, unrelated</td>
<td></td>
<td>The GTT concept is a separate concept from cleft lip to which it is related. The Brinkman concept has cleft palate, cleft lip and cleft jaw as alternative labels.</td>
</tr>
<tr>
<td>Hazenlip (cleft lip)</td>
<td>schisis (cleft lip and palate)</td>
<td>narrow (3), exact, close</td>
<td></td>
<td>See comment above</td>
</tr>
</tbody>
</table>

**Table B.3** Examples of disagreements between at least three raters in the GTTlexical experiment. The translation of the label is in parentheses. The number in parentheses behind the mapping category indicates how many raters chose that category. When no number is present only one rater selected the category.

### B.2 Disagreements from Chapter 7
<table>
<thead>
<tr>
<th>Agrovoc Concept</th>
<th>NALT concept</th>
<th>Rater A</th>
<th>Rater B</th>
<th>Rater C</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nucleosidases</td>
<td>Nucleotidases</td>
<td>unrelated</td>
<td>exact</td>
<td>exact</td>
<td>Both concepts are chemical compounds, sibling terms. Raters B and C made a mistake.</td>
</tr>
<tr>
<td>Morphine, Thebaine</td>
<td>Thebaine</td>
<td>narrow</td>
<td>unrelated</td>
<td>exact</td>
<td>The Agrovoc concept unites two sibling concepts as one, the NALT concept is an opium alkaloid and morphine is a sibling concept in NALT. The evaluation requires domain specific knowledge.</td>
</tr>
<tr>
<td>Placenta</td>
<td>Chorion (vertebrates)</td>
<td>unrelated</td>
<td>unrelated</td>
<td>broad</td>
<td>In Agrovoc chorion is an alternative term. However, because placenta seems to be restricted to mammals, and the NALT term indicates vertebrates, the two concepts do not seem to be equivalent.</td>
</tr>
<tr>
<td>Digestible cellulose</td>
<td>Cellulosic materials</td>
<td>unrelated</td>
<td>broad</td>
<td>broad</td>
<td>The source concept context is nutritive value, whereas the target concept is a material.</td>
</tr>
<tr>
<td>Metamorphosis</td>
<td>Metamorphosis</td>
<td>unrelated</td>
<td>exact</td>
<td>exact</td>
<td>Both terms refer to metamorphosis in insects. Rater A made a mistake.</td>
</tr>
<tr>
<td>Fenland soils</td>
<td>Wetland soils</td>
<td>unrelated</td>
<td>broad</td>
<td>broad</td>
<td>Fenland is a marshy region in England and a type of wetland. Evaluating these concepts requires domain specific knowledge.</td>
</tr>
<tr>
<td>Globulin</td>
<td>Legumin</td>
<td>unrelated</td>
<td>exact</td>
<td>narrow</td>
<td>Globulin is an animal protein and legumin is a plant protein. The concepts are sibling terms and require domain specific knowledge.</td>
</tr>
<tr>
<td>Kyrgyz SSR</td>
<td>Kyrgyzstan</td>
<td>broad</td>
<td>broad</td>
<td>exact</td>
<td>The source term is specific for Kyrgyzstan as a member of the USSR, whereas the target term encompasses both USSR and current time Kyrgyzstan.</td>
</tr>
<tr>
<td>Daylight, sunlight</td>
<td>Solar radiation, sunshine</td>
<td>unrelated</td>
<td>exact</td>
<td>unrelated</td>
<td>The source context is radiation and visible light, whereas the target context is a meteorological parameter.</td>
</tr>
<tr>
<td>Infection</td>
<td>Infection</td>
<td>exact</td>
<td>exact</td>
<td>unrelated</td>
<td>The Agrovoc scope note states the term is to be used for the process of becoming infected and not the resulting disease. The NALT parent term is “Pathological processes and conditions”, and the child concepts are processes and not diseases. This term requires domain specific knowledge.</td>
</tr>
<tr>
<td>Malnutrition</td>
<td>Nutrition</td>
<td>unrelated</td>
<td>broad</td>
<td>unrelated</td>
<td>The two concepts are antithetic. The target concept does have a child concept “Malnutrition”.</td>
</tr>
</tbody>
</table>

Table B.4 Examples of disagreements between raters A, B and C in the OAEI Food task.
<table>
<thead>
<tr>
<th>RKD Concept</th>
<th>Cornetto Concept</th>
<th>Rater A</th>
<th>Rater B</th>
<th>Rater D</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aardbei (strawberry)</td>
<td>Aardbei (strawberry plant)</td>
<td>exact</td>
<td>related</td>
<td>related</td>
<td>The RKD term is used to annotate depictions of the fruit and sometimes the plant.</td>
</tr>
<tr>
<td>Priester (priest)</td>
<td>Priester, priesteres (female or male priest)</td>
<td>exact</td>
<td>broad</td>
<td>broad</td>
<td>The RKD term is used for male priest(s). The Cornetto term includes male and female priest.</td>
</tr>
<tr>
<td>Priesteres (female priest)</td>
<td>Priester, priesteres (female or male priest)</td>
<td>exact</td>
<td>broad</td>
<td>broad</td>
<td>The RKD term is used for female priest(s). The Cornetto term includes male and female priest.</td>
</tr>
<tr>
<td>Ridderspoor (lark-spur)</td>
<td>Ridderspoor (knight’s spur)</td>
<td>unrelated</td>
<td>exact</td>
<td>unrelated</td>
<td>The two concepts are homonyms. Rater B made a mistake.</td>
</tr>
<tr>
<td>Zwaan (swan)</td>
<td>Zwaan (Swan constellation)</td>
<td>unrelated</td>
<td>exact</td>
<td>related</td>
<td>The bird and constellation have a metaphorical relation. Rater B made a mistake.</td>
</tr>
<tr>
<td>Trompe l’œil</td>
<td>Grisaille</td>
<td>unrelated</td>
<td>narrow</td>
<td>unrelated</td>
<td>In some cases trompe l’œil’s can be executed in grey, monochrome color which is known as a grisaille.</td>
</tr>
<tr>
<td>Hoeden (herding)</td>
<td>Oppassen (to watch out)</td>
<td>exact</td>
<td>broad</td>
<td>exact</td>
<td>The RKD term has livestock farming as broader term. The Cornetto term is in the context of being watchful.</td>
</tr>
<tr>
<td>Schrijver (writer)</td>
<td>Schrijver (someone who writes)</td>
<td>exact</td>
<td>broad</td>
<td>broad</td>
<td>The RKD broader term is a profession/role. The Cornetto term is used in the context of being literate.</td>
</tr>
<tr>
<td>Steen (stone)</td>
<td>Steen (stone as building material)</td>
<td>exact</td>
<td>unrelated</td>
<td>close</td>
<td>The RKD term is used to annotate stone as building material.</td>
</tr>
<tr>
<td>Steen (stone)</td>
<td>Steen (piece of rock)</td>
<td>unrelated</td>
<td>exact</td>
<td>close</td>
<td>The RKD term is also used to annotate images depicting a piece of rock as part of the scene.</td>
</tr>
<tr>
<td>Steen (stone)</td>
<td>Steen (rocky substance)</td>
<td>exact</td>
<td>unrelated</td>
<td>close</td>
<td>The Cornetto term refers to substance from the earth, for example a mountain of rock.</td>
</tr>
</tbody>
</table>

Table B.5 Examples of disagreement between raters A, D and B in the RKD-Cornetto experiment. The English translation of the concept label is in parentheses.
<table>
<thead>
<tr>
<th>Agrovoc Concept</th>
<th>WordNet Concept</th>
<th>Rater A</th>
<th>Rater B</th>
<th>Rater D</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shipbuilding</td>
<td>Naval engineering</td>
<td>narrow</td>
<td>exact</td>
<td>related</td>
<td>The Agrovoc term unifies shipbuilding with naval engineering under the concept of “Industry”. WordNet refers the concept as a branch of engineering concerning itself with design, building and operation of ships</td>
</tr>
<tr>
<td>Shipbuilding</td>
<td>Shipbuilding</td>
<td>narrow</td>
<td>close</td>
<td>exact</td>
<td>Same Agrovoc term as above. The WordNet concept refers to the building of ships as a more specific term than “building”.</td>
</tr>
<tr>
<td>Mauritius</td>
<td>Mauritius</td>
<td>related</td>
<td>close</td>
<td>close</td>
<td>The Agrovoc term has East Africa as parent concept an no other information. The WordNet concept is the country Mauritius</td>
</tr>
<tr>
<td>Spinach</td>
<td>Spinach (plant)</td>
<td>exact</td>
<td>related</td>
<td>narrow</td>
<td>The Agrovoc concept is in the “vegetable products”, “vegetables” hierarchy. The WordNet concept refers to the spinach plant</td>
</tr>
<tr>
<td>Spinach</td>
<td>Spinach (edible leaves)</td>
<td>related</td>
<td>exact</td>
<td>exact</td>
<td>The same Agrovoc concept as above. The WordNet concept refers to the edible spinach leaves used as food</td>
</tr>
<tr>
<td>Health</td>
<td>Fitness</td>
<td>related</td>
<td>related</td>
<td>narrow</td>
<td>The Agrovoc concept is used for human health in general terms. The WordNet term refers to fitness in the sense of being in good physical condition.</td>
</tr>
<tr>
<td>Respiration</td>
<td>Breathing (adjective)</td>
<td>unrelated</td>
<td>close</td>
<td>related</td>
<td>The Agrovoc term is in the context of a physiological function, whereas the WordNet term is the adjective for the ability to pass air through the lungs.</td>
</tr>
<tr>
<td>Respiration</td>
<td>Respiration</td>
<td>related</td>
<td>related</td>
<td>narrow</td>
<td>Same Agrovoc term as above. The WordNet term refers to a single act of breathing in and out.</td>
</tr>
<tr>
<td>Lubrication</td>
<td>Lubrication</td>
<td>close</td>
<td>close</td>
<td>related</td>
<td>The Agrovoc term refers to the act of lubrication whereas the WordNet term refers to the condition of being made slippery.</td>
</tr>
<tr>
<td>Vinegar</td>
<td>Vinegar</td>
<td>related</td>
<td>close</td>
<td>close</td>
<td>The Agrovoc term refers to the condiment and the WordNet term refers to a diluted form of acetic acid.</td>
</tr>
</tbody>
</table>

Table B.6 Examples of disagreements between raters A, D and B in the Agrovoc WordNet experiment.


on ontology matching (OM), pages 73–126.


Cultural collections are composed (of images) of cultural heritage objects, their description in metadata, and associated vocabularies with terms used in the metadata. Collections vary not only in content, but as collection metadata was created independently by each institution, they also employ a variety of formats making the integration of several collections into a single virtual collection difficult. The metadata and vocabularies first need to be converted to a common format and then linked to each other.

In this thesis we start by investigating the steps necessary for the integration of a collection into a large virtual collection. We identify four distinct steps and illustrate them in a case study. First, we convert all vocabularies accompanying the collection metadata into a common format (SKOS). Second, we convert the metadata schema, which is the set of elements describing the metadata, to more generic schemas (Dublin Core and VRA). Third, we convert the values of the metadata. In this step we identify meaningful values that either come from vocabularies or may be identifiable objects in their own right, such as a painter whose works are contained in several collections. As a last option, we keep values as text. In the fourth and last step we align collection vocabularies to vocabularies that are already part of the virtual collection, by creating mappings between similar concepts. As a result, the vocabularies that form the indexes of the collections also become integrated. The vocabularies may be aligned manually, which is a time-intensive task, or automatically. There is a multitude of alignment tools available that implement one or more alignment techniques. Although these tools are evaluated and compared in the Ontology Alignment Evaluation Initiative (OAEI), there is no methodological advice on how to align vocabularies from start to finish.

In the remainder of the thesis we focus on vocabulary alignment. Our main research question is as follows: How can we combine vocabulary alignment techniques, assess their performance and evaluate alignments?

In Chapters 3 and 4 we study how we can combine vocabulary alignment techniques and assess their performance. In our experiment in Chapter 3, two vocabularies are aligned using string matching and off-the-shelf alignment tools. We find that although the individual tools perform relatively poorly, by selecting subsets of mappings we are able to boost the quality of the mappings in comparison to the original set.

In Chapter 4 we perform a similar experiment on large vocabularies. Due to the size of the vocabularies we were unable to use the off-the-shelf tools from Chapter 3 and resort to simple string matching techniques to generate mappings. As the number of generated mappings was too high to evaluate in their entirety, we took random samples in order to be able to assess the value of the alignment techniques and compare them. In a step-by-step method we select subsets of mappings of higher quality than the original set of mappings.
In the Chapter 5 we focus on reusing existing equivalence mappings between multiple vocabularies and generate composed mappings. We investigate whether composed mappings are transitive by setting up experiments with vocabularies from different domains: the medicine and cultural heritage library domain. We generate composed mappings and examine the number of composed mappings and their quality in comparison to the mappings used to create them. Our findings matched our expectations about mapping composition. For example, we found that vocabularies that overlap in the domain they describe are more likely to have high quality composed mappings. One aspect that deserves more study is the fact that the quality of composed mappings was higher than expected, almost as high as the quality of mappings used to create the composed mappings.

In order to evaluate the quality of an alignment at least part of the mappings need to be evaluated manually. For manual evaluation to be reliable, at least a subset of mappings needs to be evaluated by independent human raters. Then their agreement level needs to be calculated, and if it is sufficiently high the result of the manual evaluation can be used as a test set or reference alignment. In our experiments we found that the agreement level was in some cases rather low. In Chapters 6 and 7 we study the level of agreement between multiple raters in several manual evaluation experiments, and attempt to identify the causes for low agreement levels. We find that improved guidelines with clear descriptions of the task and evaluation categories improves the level of agreement between raters. Nevertheless there are also aspects that make evaluation difficult, such as the inherent ambiguity of concepts and their descriptions, and the background knowledge of raters which influences their interpretation of concepts. Additionally, characteristics of vocabularies, such as their representation and their domain and the selection of suitable matching categories also influence the outcome of an evaluation.

In this thesis we show that combining alignment techniques through an interactive process is an effective and transparent method for generating high quality mappings. We also show that a large number of factors influence manual evaluation. These factors include the characteristics of vocabularies, the matching categories and the inherent vagueness of concepts.

We conclude with a method for evaluating alignments which includes reporting agreement levels between raters, reporting on how disagreements were dealt with, and emphasizes that the evaluation needs to be made public. Second, we propose a method for aligning vocabularies, where we keep track of the source of mappings. As the alignment is performed in explicit steps it is possible to select mappings of a specific quality, and thus the alignment can be tailored to any kind of application.
Karakteristiek voor collecties van instellingen voor cultureel erfgoed is dat ze bestaan uit (afbeeldingen van) culturele objecten, hun beschrijving in de vorm van metadata, en bijbehorende woordenlijsten of vocabulaires. Deze laatste worden gebruikt voor het indexeren van metadata, waardoor het zoeken naar objecten in de collectie eenvoudiger wordt. Collecties variëren niet alleen in hun inhoud, maar maken ook gebruik van vele verschillende metadataformaten doordat de metadata onafhankelijk worden samengesteld door elke instelling. Dit bemoeilijkt de integratie van verschillende collecties in een enkele virtuele collectie. De metadata en vocabulaires moeten eerst omgezet worden in een gemeenschappelijk formaat en vervolgens aan elkaar gekoppeld worden.

In dit proefschrift beginnen wij met het onderzoeken van de stappen die nodig zijn voor de integratie van een collectie in een meeromvattende virtuele collectie. We identificeren vier verschillende stappen en illustreren ze met behulp van een case study. Ten eerste zetten we alle vocabulaires die bij de metadata horen om in een gemeenschappelijk formaat (SKOS). Ten tweede zetten we de metadata schemata (de verzameling elementen die de metadata beschrijven) om in generieke schema’s (Dublin Core en VRA). Ten derde zetten we de waarden van de metadata om. In deze stap proberen we zinvolle waarden te identificeren die ofwel afkomstig zijn uit vocabulaires ofwel zelf identificeerbare objecten zijn, zoals een schilder wiens werken opgenomen zijn in verschillende collecties. Als beide niet mogelijk zijn, houden we als laatste optie waarden aan als tekst. In de vierde en laatste stap beelden we collectievocabulaires af op vocabulaires die al deel uitmaken van de virtuele collectie; met andere woorden, we maken verbindingen tussen vergelijkbare concepten. Hierdoor wordt de collectie die geïndexeerd is met het afgebeelde vocabulaire geïntegreerd met de andere collecties. De vocabulaires kunnen handmatig (een tijdrovende taak) of automatisch afgebeeld worden op andere vocabulaires. Er zijn vele tools beschikbaar die gebruik maken van één of meerdere technieken om een afbeelding te maken. Hoewel deze tools geëvalueerd en vergeleken worden in de jaarlijkse Ontology Alignment Evaluation Initiative (OAEI) is er geen methodologisch advies over hoe vocabulaires van begin tot eind verbonden kunnen worden.

In het vervolg van het proefschrift richten we ons op vocabulaire-afbeelding. Onze belangrijkste onderzoeksvraag is als volgt: Hoe kunnen we afbeeldingstechnieken combineren, hun prestatie beoordelen en de resulterende afbeelding evalueren?

In de hoofdstukken 3 en 4 richten we ons op het eerste deel van de onderzoeksvraag, en bestuderen we het combineren van afbeeldingstechnieken voor vocabulaires en het beoordelen van hun prestaties. In ons experiment in hoofdstuk 3 worden twee vocabulaires afbeeld met behulp van tekstvergelijking (“string matching”) en kant en klare afbeeldingstools. We concluderen dat de
afzonderlijke tools relatief slecht presteren; echter, door het uitkiezen van specifieke deelverzamelingen is de kwaliteit van gegenereerde verbindingen in de deelverzameling hoger dan die in de oorspronkelijke verzameling.

In hoofdstuk 4 voeren we een vergelijkbaar experiment uit, maar dan op grotere vocabulaires. Vanwege de omvang van de vocabulaires waren we niet in staat om de kant en klare tools uit hoofdstuk 3 te gebruiken. Daarom kozen we voor eenvoudige tekstvergelijkingstechnieken om verbindingen te maken. Doordat het aantal gegenereerde verbindingen te groot is konden wij geen handmatige evaluatie toepassen op de gehele verzameling van verbindingen. Om de waarde van elke techniek te kunnen beoordelen namen wij steekproeven van de verbindingen. In een stapvoor-stap methode selecteren wij deelverzamelingen van verbindingen die onderling van hogere kwaliteit zijn dan de oorspronkelijke afbeelding.

In hoofdstuk 5 richten we ons op het hergebruik van verbindingen tussen gelijkwaardige concepten om verbindingen tussen meerdere vocabulaires samen te stellen. Wij testen of gelijkwaardige verbindingen overdraagbaar zijn in de praktijk door experimenten uit te voeren op vocabulaires uit meerdere domeinen: geneeskunde, cultureel erfgoed en het bibliotheekdomein. We genereren samengestelde verbindingen en onderzoeken ze op hun aantal en hun kwaliteit. Onze bevindingen kwamen overeen met onze verwachtingen, met name dat samengestelde verbindingen tussen vocabulaires die hetzelfde domein beschrijven van hogere kwaliteit zijn dan verbindingen tussen vocabulaires uit verschillende domeinen. Een verrassende bevinding die meer onderzoek verdient is dat de kwaliteit van samengestelde verbindingen hoger is dan verwacht, namelijk vergelijkbaar met de kwaliteit van de gebruikte verbindingen.

Voor het beoordelen van de kwaliteit van een afbeelding is het nodig om tenminste een gedeelte handmatig te beoordelen. Om zeker te zijn dat de handmatige evaluatie op zijn beurt betrouwbaar is, dient tenminste één subset van verbindingen door onafhankelijke beoordelaars te worden geëvalueerd. Hierna wordt de interbeoordelaarsbetrouwbaarheid berekend. Als deze voldoende is kunnen de handmatig geëvalueerde verbindingen gebruikt worden als testverzameling of reference alignment?. In onze voorgaande experimenten vonden we dat de interbeoordelaarsbetrouwbaarheid vrij laag uitvalt. In hoofdstuk 6 en 7 bestuderen we de mate van overeenstemming tussen meerdere beoordelaars in verschillende experimenten, en we proberen de oorzaken van lage niveaus van overeenkomst te identificeren. Wij concluderen dat betere richtlijnen met een duidelijke beschrijving van de evaluatietaak en de evaluatiecategorieën de mate van overeenstemming tussen beoordelaars aanzienlijk verbeteren. Toch zorgen bepaalde eigenschappen van handmatige evaluatie ervoor dat dit een moeilijke taak blijft, met name de intrinsieke dubbelzinnigheid van begrippen en hun beschrijvingen, en het feit dat de achtergrondkennis van beoordelaars hun interpretatie van begrippen beïnvloedt. Daarnaast hebben kenmerken van vocabulaires, zoals hun representatie en hun domein, en de gekozen evaluatiecategorieën ook invloed op de uitkomst van een evaluatie.

In dit proefschrift laten we zien dat het combineren van afbeeldingstechnieken door middel van een interactief proces een effectieve en transparante methode is voor het genereren van verbindingen van hoge kwaliteit. We tonen ook aan dat een groot aantal factoren handmatige beoordeling
beïnvloedt. Deze factoren zijn onder andere de kenmerken van vocabulaires, de gekozen evaluatiecategorieën en de inherente vaagheid van concepten. Er moet rekening gehouden worden met deze factoren bij het verbinden van vocabulaires.

We sluiten af met een methode voor het evalueren van verbindingen waarbij onder andere de overeenstemmingsniveaus tussen de beoordelaars gerapporteerd worden, samen met een rapport over de manier waarop meningsverschillen werden behandeld. Hierbij wordt het belang van het openbaar maken van de evaluatie en het rapporteren daarover benadrukt. Daarnaast stellen we een methode voor voor het verbinden van vocabulaires. Deze methode maakt het mogelijk om de oorsprong van verbindingen te traceren doordat het koppelen in expliciete stappen gebeurt, waardoor verbindingen van bepaalde kwaliteit geselecteerd kunnen worden. Hierdoor kan het verbindingproces aangepast worden aan verschillende toepassingen.
<table>
<thead>
<tr>
<th>Year</th>
<th>Author(s)</th>
<th>Title/Focus</th>
</tr>
</thead>
<tbody>
<tr>
<td>1998</td>
<td>Johan van den Akker (CWI)</td>
<td>Degas - An Active, Temporal Database of Autonomous Objects</td>
</tr>
<tr>
<td></td>
<td>Floris Wiesman (UM)</td>
<td>Information Retrieval by Graphically Browsing Meta-Information</td>
</tr>
<tr>
<td></td>
<td>Ans Steuten (TUD)</td>
<td>A Contribution to the Linguistic Analysis of Business Conversations within the Language/Action Perspective</td>
</tr>
<tr>
<td></td>
<td>Dennis Breuker (UM)</td>
<td>Memory versus Search in Games</td>
</tr>
<tr>
<td></td>
<td>Eduard W. Oskamp (RUL)</td>
<td>Computerondersteuning bij Straftoemeting</td>
</tr>
<tr>
<td>1999</td>
<td>Mark Sloof (VU)</td>
<td>Physiology of Quality Change Modelling; Automated Modelling of Quality Change of Agricultural Products</td>
</tr>
<tr>
<td></td>
<td>Rob Potharst (EUR)</td>
<td>Classification using Decision Trees and Neural Nets</td>
</tr>
<tr>
<td></td>
<td>Don Beal (UM)</td>
<td>The Nature of Minimax Search</td>
</tr>
<tr>
<td></td>
<td>Jacques Penders (UM)</td>
<td>The Practical Art of Moving Physical Objects</td>
</tr>
<tr>
<td></td>
<td>Aldo de Moor (KUB)</td>
<td>Empowering Communities: A Method for the Legitimate User-Driven Specification of Network Information Systems</td>
</tr>
<tr>
<td></td>
<td>Niek J.E. Wijngaards (VU)</td>
<td>Re-Design of Compositional Systems</td>
</tr>
<tr>
<td></td>
<td>David Spelt (UT)</td>
<td>Verification Support for Object Database Design</td>
</tr>
<tr>
<td></td>
<td>Jacques H.J. Lenting (UM)</td>
<td>Informed Gambling: Conception and Analysis of a Multi-Agent Mechanism for Discrete Reallocation</td>
</tr>
<tr>
<td>2000</td>
<td>Frank Niessink (VU)</td>
<td>Perspectives on Improving Software Maintenance</td>
</tr>
<tr>
<td></td>
<td>Koen Holtman (TU/e)</td>
<td>Prototyping of CMS Storage Management</td>
</tr>
<tr>
<td></td>
<td>Carolien M.T. Metselaar (UvA)</td>
<td>Sociaal-organisatorische Gevolgen van Kennistechnologie; een Procesbenadering en Actorperspectief</td>
</tr>
<tr>
<td></td>
<td>Geert de Haan (VU)</td>
<td>ETAG, A Formal Model of Competence Knowledge for User Interface Design</td>
</tr>
<tr>
<td></td>
<td>Ruud van der Pol (UM)</td>
<td>Knowledge-Based Query Formulation in Information Retrieval</td>
</tr>
<tr>
<td></td>
<td>Rogier van Eijk (UU)</td>
<td>Programming Languages for Agent Communication</td>
</tr>
<tr>
<td></td>
<td>Niels Peek (UU)</td>
<td>Decision-Theoretic Planning of Clinical Patient Management</td>
</tr>
<tr>
<td></td>
<td>Veerle Coupé (EUR)</td>
<td>Sensitivity Analysis of Decision-Theoretic Networks</td>
</tr>
<tr>
<td></td>
<td>Florian Waas (CWI)</td>
<td>Principles of Probabilistic Query Optimization</td>
</tr>
<tr>
<td></td>
<td>Niels Nes (CWI)</td>
<td>Image Database Management System Design Considerations, Algorithms and Architecture</td>
</tr>
<tr>
<td></td>
<td>Jonas Karlsson (CWI)</td>
<td>Scalable Distributed Data Structures for Database Management</td>
</tr>
<tr>
<td>2001</td>
<td>Silja Renooij (UU)</td>
<td>Qualitative Approaches to Quantifying Probabilistic Networks</td>
</tr>
<tr>
<td></td>
<td>Koen Hindriks (UU)</td>
<td>Agent Programming Languages: Programming with Mental Models</td>
</tr>
<tr>
<td></td>
<td>Maarten van Someren (UvA)</td>
<td>Learning as Problem Solving</td>
</tr>
</tbody>
</table>

1 Abbreviations: SIKS - Dutch Research School for Information and Knowledge Systems; CWI - Centrum voor Wiskunde en Informatica, Amsterdam; EUR - Erasmus Universiteit, Rotterdam; KUB - Katholieke Universiteit Brabant, Tilburg; KUN - Katholieke Universiteit Nijmegen; OU - Open Universiteit; RUL - Rijksuniversiteit Leiden; RUN - Radboud Universiteit Nijmegen; TUD - Technische Universiteit Delft; TU/e - Technische Universiteit Eindhoven; UL - Universiteit Leiden; UM - Universiteit Maastricht; UT - Universiteit Twente, Enschede; UU - Universiteit Utrecht; UvA - Universiteit van Amsterdam; UvT - Universiteit van Tilburg; VU - Vrije Universiteit, Amsterdam.
4 Evgueni Smirnov (UM) Conjunctive and Disjunctive Version Spaces with Instance-Based Boundary Sets
5 Jacco van Ossenbruggen (VU) Processing Structured Hypermedia: A Matter of Style
6 Martijn van Welie (VU) Task-Based User Interface Design
7 Bastiaan Schonhage (VU) Diva: Architectural Perspectives on Information Visualization
8 Pascal van Eck (VU) A Compositional Semantic Structure for Multi-Agent Systems Dynamics
9 Pieter Jan ’t Hoen (RUL) Towards Distributed Development of Large Object-Oriented Models, Views of Packages as Classes
10 Maarten Sierhuis (UvA) Modeling and Simulating Work Practice BRAHMS: a Multiagent Modeling and Simulation Language for Work Practice Analysis and Design
11 Tom M. van Engers (VU) Knowledge Management: The Role of Mental Models in Business Systems Design

2002
1 Nico Lassing (VU) Architecture-Level Modifiability Analysis
2 Roelof van Zwol (UT) Modelling and Searching Web-based Document Collections
3 Henk Ernst Blok (UT) Database Optimization Aspects for Information Retrieval
4 Juan Roberto Castelo Valdueza (UU) The Discrete Acyclic Digraph Markov Model in Data Mining
5 Radu Serban (VU) The Private Cyberspace Modeling Electronic Environments Inhabited by Privacy-Concerned Agents
6 Laurens Mommers (UL) Applied Legal Epistemology: Building a Knowledge-based Ontology of the Legal Domain
7 Peter Boncz (CWI) Monet: A Next-Generation DBMS Kernel For Query-Intensive Applications
8 Jaap Gordijn (VU) Value Based Requirements Engineering: Exploring Innovative E-Commerce Ideas
9 Willem-Jan van den Heuvel (KUB) Integrating Modern Business Applications with Objectified Legacy Systems
10 Brian Sheppard (UM) Towards Perfect Play of Scrabble
11 Wouter C.A. Wijngaards (VU) Agent Based Modelling of Dynamics: Biological and Organisational Applications
12 Albrecht Schmidt (UvA) Processing XML in Database Systems
13 Hongjing Wu (TU/e) A Reference Architecture for Adaptive Hypermedia Applications
14 Wieke de Vries (UU) Agent Interaction: Abstract Approaches to Modelling, Programming and Verifying Multi-Agent Systems
15 Rik Eshuis (UT) Semantics and Verification of UML Activity Diagrams for Workflow Modelling
17 Stefan Manegold (UvA) Understanding, Modeling, and Improving Main-Memory Database Performance

2003
1 Heiner Stuckenschmidt (VU) Ontology-Based Information Sharing in Weakly Structured Environments
2 Jan Broersen (VU) Modal Action Logics for Reasoning About Reactive Systems
3 Martijn Schuemie (TUD) Human-Computer Interaction and Presence in Virtual Reality Exposure Therapy
4 Milan Petkovic (UT) Content-Based Video Retrieval Supported by Database Technology
5 Jos Lehmann (UvA) Causation in Artificial Intelligence and Law – A Modelling Approach
6 Boris van Schooten (UT) Development and Specification of Virtual Environments
7 Machiel Jansen (UvA) Formal Explorations of Knowledge Intensive Tasks
8 Yong-Ping Ran (UM) Repair-Based Scheduling
9 Rens Kortmann (UM) The Resolution of Visually Guided Behaviour
10 Andreas Lincke (UT) Electronic Business Negotiation: Some Experimental Studies on the Interaction between Medium, Innovation Context and Cult
11 Simon Keizer (UT) Reasoning under Uncertainty in Natural Language Dialogue using Bayesian Networks
12 Roeland Ordelman (UT) Dutch Speech Recognition in Multimedia Information Retrieval
13 Jeroen Donkers (UM) Nosce Hostem – Searching with Opponent Models
<table>
<thead>
<tr>
<th>2004</th>
<th>2005</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Stijn Hoppenbrouwers (KUN)</td>
<td>1 Floor Verdenius (UvA)</td>
</tr>
<tr>
<td>Freezing Language:</td>
<td>Methodological Aspects of</td>
</tr>
<tr>
<td>Conceptualisation Processes across</td>
<td>Designing Induction-Based</td>
</tr>
<tr>
<td>ICT-Supported</td>
<td>Applications</td>
</tr>
<tr>
<td>Organisations</td>
<td>2 Erik van der Werf (UM) AI techniques for the game</td>
</tr>
<tr>
<td></td>
<td>of Go</td>
</tr>
<tr>
<td>15 Mathijs de Weerdt (TUD)</td>
<td>3 Franc Grootjen (RUN) A Pragmatic Approach to the</td>
</tr>
<tr>
<td>Plan Merging in Multi-Agent</td>
<td>Conceptualisation of Language</td>
</tr>
<tr>
<td>Systems</td>
<td>4 Nirvana Meratnia (UT) Towards Database Support</td>
</tr>
<tr>
<td></td>
<td>for Moving Object data</td>
</tr>
<tr>
<td>16 Menzo Windhouwer (CWI)</td>
<td>5 Gabriel Infante-Lopez (UvA) Two-Level Probabilistic</td>
</tr>
<tr>
<td>Feature Grammar Systems</td>
<td>Grammars for Natural Language Parsing</td>
</tr>
<tr>
<td>- Incremental Maintenance of Indexes to Digital</td>
<td>6 Pieter Spronck (UM) Adaptive Game AI</td>
</tr>
<tr>
<td>Media Warehouse</td>
<td>7 Flavius Frasincar (TU/e) Hypermedia Presentation</td>
</tr>
<tr>
<td></td>
<td>Generation for Semantic Web Information Systems</td>
</tr>
<tr>
<td>17 David Jansen (UT)</td>
<td>8 Richard Vdovjak (TU/e) A Model-driven Approach</td>
</tr>
<tr>
<td>Extensions of Statecharts with</td>
<td>for Building Distributed Ontology-based Web Applications</td>
</tr>
<tr>
<td>Probability, Time, and Stochastic Timing</td>
<td>9 Jeen Broekstra (VU) Storage, Querying and Inferencing</td>
</tr>
<tr>
<td></td>
<td>for Semantic Web Languages</td>
</tr>
<tr>
<td>18 Levente Kocsis (UM)</td>
<td>10 Anders Bouwer (UvA) Explaining Behaviour: Using</td>
</tr>
<tr>
<td>Learning Search Decisions</td>
<td>Qualitative Simulation in Interactive Learning Environments</td>
</tr>
<tr>
<td>2005</td>
<td>2005</td>
</tr>
<tr>
<td>1 Floor Verdenius (UvA)</td>
<td>Methodological Aspects of</td>
</tr>
<tr>
<td></td>
<td>Designing Induction-Based Applications</td>
</tr>
<tr>
<td>2 Erik van der Werf (UM)</td>
<td>AI techniques for the game</td>
</tr>
<tr>
<td></td>
<td>of Go</td>
</tr>
<tr>
<td>3 Franc Grootjen (RUN)</td>
<td>A Pragmatic Approach to the Conceptualisation of Language</td>
</tr>
<tr>
<td></td>
<td>Towards Database Support for Moving Object data</td>
</tr>
<tr>
<td>4 Nirvana Meratnia (UT)</td>
<td>Two-Level Probabilistic Grammars for Natural Language Parsing</td>
</tr>
<tr>
<td>5 Gabriel Infante-Lopez (UvA)</td>
<td>Adaptive Game AI</td>
</tr>
<tr>
<td>6 Pieter Spronck (UM)</td>
<td>A Model-driven Approach for Building Distributed Ontology-based Web Applications</td>
</tr>
<tr>
<td>7 Flavius Frasincar (TU/e)</td>
<td>Storage, Querying and Inferencing for Semantic Web Languages</td>
</tr>
<tr>
<td>8 Richard Vdovjak (TU/e)</td>
<td>Explaining Behaviour: Using Qualitative Simulation in Interactive Learning Environments</td>
</tr>
<tr>
<td>9 Jeen Broekstra (VU)</td>
<td>Agent Based Matchmaking and Clustering - A Decentralized Approach to Search</td>
</tr>
<tr>
<td></td>
<td>Distributed Simulation in Industry</td>
</tr>
<tr>
<td>10 Anders Bouwer (UvA)</td>
<td>Een Computermodel voor het Ondersteunen van Euthanasiebeslissingen</td>
</tr>
<tr>
<td>11 Elth Ogston (VU)</td>
<td>Web-Service configuration on the Semantic Web; Exploring how semantics meets pragmatics</td>
</tr>
<tr>
<td>12 Csaba Boer (EUR)</td>
<td>Analysis of the Dynamics of Cognitive Processes</td>
</tr>
<tr>
<td>13 Fred Hamburg (UL)</td>
<td>Usability of XML Query Languages</td>
</tr>
<tr>
<td>14 Borys Omelayenko (VU)</td>
<td>Software Specification Based on Re-usable Business Components</td>
</tr>
<tr>
<td>15 Tibor Bosse (VU)</td>
<td>Test-selection strategies for probabilistic networks</td>
</tr>
</tbody>
</table>
19 Michel van Dartel (UM) Situated Representation
20 Cristina Coteanu (UL) Cyber Consumer Law, State of the Art and Perspectives
21 Wijnand Derks (UT) Improving Concurrency and Recovery in Database Systems by Exploiting Application Semantics

2006
1 Samuil Angelov (TU/e) Foundations of B2B Electronic Contracting
2 Cristina Chisalita (VU) Contextual issues in the design and use of information technology in organizations
3 Noor Christoph (UvA) The role of metacognitive skills in learning to solve problems
4 Marta Sabou (VU) Building Web Service Ontologies
5 Cees Pierik (UU) Validation Techniques for Object-Oriented Proof Outlines
6 Ziv Baida (VU) Software-aided Service Bundling - Intelligent Methods & Tools for Graphical Service Modeling
7 Marko Smiljanic (UT) XML schema matching – balancing efficiency and effectiveness by means of clustering
8 Eelco Herder (UT) Forward, Back and Home Again - Analyzing User Behavior on the Web
9 Mohamed Wahdan (UM) Automatic Formulation of the Auditor’s Opinion
10 Ronny Siebes (VU) Semantic Routing in Peer-to-Peer Systems
11 Joeri van Ruth (UT) Flattening Queries over Nested Data Types
12 Bert Bongers (VU) Interactivation - Towards an ecology of people, our technological environment, and the arts
13 Henk-Jan Lebbink (UU) Dialogue and Decision Games for Information Exchanging Agents
14 Johan Hoon (VU) Software Requirements: Update, Upgrade, Redesign - towards a Theory of Requirements Change
15 Rainer Malik (UU) CONAN: Text Mining in the Biomedical Domain
16 Carsten Riggelsen (UU) Approximation Methods for Efficient Learning of Bayesian Networks
17 Stacey Nagata (UU) User Assistance for Multitasking with Interruptions on a Mobile Device
18 Valentin Zhizhkun (UvA) Graph transformation for Natural Language Processing
19 Birna van Riemsdijk (UU) Cognitive Agent Programming: A Semantic Approach
20 Marina Velikova (UvT) Monotone models for prediction in data mining
21 Bas van Gils (RUN) Aptness on the Web
22 Paul de Vrieze (RUN) Fundaments of Adaptive Personalisation
23 Ion Juvina (UU) Development of Cognitive Model for Navigating on the Web
24 Laura Hollink (VU) Semantic Annotation for Retrieval of Visual Resources
25 Madalina Drugan (UU) Conditional log-likelihood MDL and Evolutionary MCMC
26 Vojkan Mihajlovic (UT) Score Region Algebra: A Flexible Framework for Structured Information Retrieval
27 Stefano Bocconi (CWI) Vox Populi: generating video documentaries from semantically annotated media repositories
28 Borkur Sigurbjornsson (UvA) Focused Information Access using XML Element Retrieval

2007
1 Kees Leune (UvT) Access Control and Service-Oriented Architectures
2 Wouter Teepe (RUG) Reconciling Information Exchange and Confidentiality: A Formal Approach
3 Peter Mika (VU) Social Networks and the Semantic Web
5 Bart Schermer (UL) Software Agents, Surveillance, and the Right to Privacy: a Legislative Framework for Agent-enabled Surveillance
6 Gilad Mishne (UvA) Applied Text Analytics for Blogs
7 Natasa Jovanovic’ (UT) To Whom It May Concern - Addressee Identification in Face-to-Face Meetings
8 Mark Hoogendoorn (VU) Modeling of Change in Multi-Agent Organizations
9 David Mobach (VU) Agent-Based Mediated Service Negotiation
10 Huib Aldewereld (UU) Autonomy vs. Conformity: an Institutional Perspective on Norms and Protocols
11 Natalia Stash (TU/e) Incorporating Cognitive/Learning Styles in a General-Purpose Adaptive Hypermedia System

12 Marcel van Gerven (RUN) Bayesian Networks for Clinical Decision Support: A Rational Approach to Dynamic Decision-Making under Uncertainty

13 Rutger Rienks (UT) Meetings in Smart Environments; Implications of Progressing Technology

14 Niek Bergboer (UM) Context-Based Image Analysis

15 Joyca Lacroix (UM) NIM: a Situated Computational Memory Model

16 Davide Grossi (UU) Designing Invisible Handcuffs. Formal investigations in Institutions and Organizations for Multi-agent Systems

17 Theodore Charitos (UU) Reasoning with Dynamic Networks in Practice

18 Bart Orriens (UvT) On the development and management of adaptive business collaborations

19 David Levy (UM) Intimate relationships with artificial partners

20 Slinger Jansen (UU) Customer Configuration Updating in a Software Supply Network

21 Karianne Vermaas (UU) Fast diffusion and broadening use: A research on residential adoption and usage of broadband internet in the Netherlands between 2001 and 2005

22 Zlatko Zlatev (UT) Goal-oriented design of value and process models from patterns

23 Peter Barna (TU/e) Specification of Application Logic in Web Information Systems

24 Georgina Ramirez Camps (CWI) Structural Features in XML Retrieval


2008

1 Katalin Boer-Sorbán (EUR) Agent-Based Simulation of Financial Markets: A modular, continuous-time approach


3 Vera Hollink (UvA) Optimizing hierarchical menus: a usage-based approach

4 Ander de Keijzer (UT) Management of Uncertain Data - towards unattended integration

5 Bela Mutschler (UT) Modeling and simulating causal dependencies on process-aware information systems from a cost perspective

6 Arjen Hommersom (RUN) On the Application of Formal Methods to Clinical Guidelines, an Artificial Intelligence Perspective

7 Peter van Rosmalen (OU) Supporting the tutor in the design and support of adaptive e-learning

8 Janneke Bolt (UU) Bayesian Networks: Aspects of Approximate Inference

9 Christof van Nimwegen (UU) The paradox of the guided user: assistance can be counter-effective

10 Wauter Bosma (UT) Discourse oriented Summarization

11 Vera Kartseva (VU) Designing Controls for Network Organizations: a Value-Based Approach

12 Jozsef Farkas (RUN) A Semiotically oriented Cognitive Model of Knowledge Representation

13 Caterina Carraciolo (UvA) Topic Driven Access to Scientific Handbooks

14 Arthur van Bunningen (UT) Context-Aware Querying; Better Answers with Less Effort


16 Henriette van Vugt (VU) Embodied Agents from a User's Perspective

17 Martin Op't Land (TUD) Applying Architecture and Ontology to the Splitting and Allying of Enterprises

18 Guido de Croon (UM) Adaptive Active Vision

19 Henning Rode (UT) From document to entity retrieval: improving precision and performance of focused text search

20 Rex Arendsen (UvA) Geen bericht, goed bericht. Een onderzoek naar de effecten van de introductie van elektronisch berichtenverkeer met een overheid op de administratieve lasten van bedrijven

21 Krisztian Balog (UvA) People search in the enterprise

22 Henk Koning (UU) Communication of IT-architecture

23 Stefan Visscher (UU) Bayesian network models for the management of ventilator-associated pneumonia

24 Zharko Aleksovski (VU) Using background knowledge in ontology matching
25 Geert Jonker (UU) Efficient and Equitable exchange in air traffic management plan repair using spender-signed currency
26 Marijn Huijbregts (UT) Segmentation, diarization and speech transcription: surprise data unraveled
27 Hubert Vogten (OU) Design and implementation strategies for IMS learning design
28 Ildiko Flesh (RUN) On the use of independence relations in Bayesian networks
29 Dennis Reidsma (UT) Annotations and subjective machines- Of annotators, embodied agents, users, and other humans
30 Wouter van Atteveldt (VU) Semantic network analysis: techniques for extracting, representing and querying media content
31 Loes Braun (UM) Pro-active medical information retrieval
32 Trung B. Hui (UT) Toward affective dialogue management using partially observable markov decision processes
33 Frank Terpstra (UvA) Scientific workflow design; theoretical and practical issues
34 Jeroen de Knijf (UU) Studies in Frequent Tree Mining
35 Benjamin Torben-Nielsen (UvT) Dendritic morphology: function shapes structure

2009
1 Rasa Jurgelenaite (RUN) Symmetric Causal Independence Models
2 Willem Robert van Hage (VU) Evaluating Ontology-Aignment Techniques
3 Hans Stol (UvT) A Framework for Evidence-based Policy Making Using IT
4 Josephine Nabukunya (RUN) Improving the Quality of Organisational Policy Making using Collaboration Engineering
5 Sietske Overbeek (RUN) Bridging Supply and Demand for Knowledge Intensive Tasks - Based on Knowledge, Cognition, and Quality
6 Muhammad Subianto (UU) Understanding Classification
7 Ronald Poppe (UT) Discriminative Vision-Based Recovery and Recognition of Human Motion
8 Volker Nannen (VU) Evolutionary Agent-Based Policy Analysis in Dynamic Environments
9 Benjamin Kanagwa (RUN) Design, Discovery and Construction of Service-oriented Systems
10 Jan Wielemaker (UvA) Logic programming for knowledge-intensive interactive applications
11 Alexander Boer (UvA) Legal Theory, Sources of Law & the Semantic Web
12 Peter Massuthé (TU/e, Humboldt-Universität zu Berlin) Operating Guidelines for Services
13 Steven de Jong (UM) Fairness in Multi-Agent Systems
14 Maksym Korotkiy (VU) From ontology-enabled services to service-enabled ontologies (making ontologies work in e-science with ONTO-SOA)
15 Rinke Hoekstra (UvA) Ontology Representation - Design Patterns and Ontologies that Make Sense
16 Fritz Reul (UvT) New Architectures in Computer Chess
17 Laurens van der Maaten (UvT) Feature Extraction from Visual Data
18 Fabian Groffen (CWI) Armada, An Evolving Database System
19 Valentin Robu (CWI) Modeling Preferences, Strategic Reasoning and Collaboration in Agent-Mediated Electronic Markets
20 Bob van der Vecht (UU) Adjustable Autonomy: Controlling Influences on Decision Making
21 Stijn Vanderlooy (UM) Ranking and Reliable Classification
22 Pavel Serdyukov (UT) Search For Expertise: Going beyond direct evidence
23 Peter Hofgesang (VU) Modelling Web Usage in a Changing Environment
24 Annerieke Heuvelink (VU) Cognitive Models for Training Simulations
25 Alex van Ballegooij (CWI) “RAM: Array Database Management through Relational Mapping”
26 Fernando Koch (UvT) An Agent-Based Model for the Development of Intelligent Mobile Services
27 Christian Glahn (OU) Contextual Support of social Engagement and Reflection on the Web
28 Sander Evers (UT) Sensor Data Management with Probabilistic Models
29 Stanislav Pokraev (UT) Model-Driven Semantic Integration of Service-Oriented Applications
30 Marcin Zukowski (CWI) Balancing vectorized query execution with bandwidth-optimized storage
31 Sofiya Katrenko (UvA) A Closer Look at Learning Relations from Text
<table>
<thead>
<tr>
<th>Page</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>Rik Farenhorst and Remco de Boer (VU) Architectural Knowledge Management: Supporting Architects and Auditors</td>
</tr>
<tr>
<td>33</td>
<td>Khiet Truong (UT) How Does Real Affect Affect Recognition In Speech?</td>
</tr>
<tr>
<td>34</td>
<td>Inge van de Weerd (UU) Advancing in Software Product Management: An Incremental Method Engineering Approach</td>
</tr>
<tr>
<td>35</td>
<td>Wouter Koelewijn (UL) Privacy en Politiegegevens; Over geautomatiseerde normatieve informatieuitwisseling</td>
</tr>
<tr>
<td>36</td>
<td>Marco Kalz (OUN) Placement Support for Learners in Learning Networks</td>
</tr>
<tr>
<td>37</td>
<td>Hendrik Drachsler (OUN) Navigation Support for Learners in Informal Learning Networks</td>
</tr>
<tr>
<td>38</td>
<td>Riina Vuorikari (OU) Tags and self-organisation: a metadata ecology for learning resources in a multilingual context</td>
</tr>
<tr>
<td>39</td>
<td>Christian Stahl (TUE, Humboldt-Universitaet zu Berlin) Service Substitution – A Behavioral Approach Based on Petri Nets</td>
</tr>
<tr>
<td>40</td>
<td>Stephan Raaijmakers (UvT) Multinomial Language Learning: Investigations into the Geometry of Language</td>
</tr>
<tr>
<td>41</td>
<td>Igor Berezhnyy (UvT) Digital Analysis of Paintings</td>
</tr>
<tr>
<td>42</td>
<td>Toine Bogers (UvT) Recommender Systems for Social Bookmarking</td>
</tr>
<tr>
<td>43</td>
<td>Stephan Raaijmakers (UvT) Multinomial Language Learning: Investigations into the Geometry of Language</td>
</tr>
<tr>
<td>44</td>
<td>Igor Berezhnyy (UvT) Digital Analysis of Paintings</td>
</tr>
<tr>
<td>45</td>
<td>Toine Bogers (UvT) Recommender Systems for Social Bookmarking</td>
</tr>
<tr>
<td>46</td>
<td>Virginia Nunes Leal Franqueira (UT) Finding Multi-step Attacks in Computer Networks using Heuristic Search and Mobile Ambients</td>
</tr>
<tr>
<td>47</td>
<td>Roberto Santana Tapia (UT) Assessing Business-IT Alignment in Networked Organizations</td>
</tr>
<tr>
<td>48</td>
<td>Jilles Vreeken (UU) Making Pattern Mining Useful</td>
</tr>
<tr>
<td>49</td>
<td>Loredana Afanasiev (UvA) Querying XML: Benchmarks and Recursion</td>
</tr>
<tr>
<td>2010</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Matthijs van Leeuwen (UU) Patterns that Matter</td>
</tr>
<tr>
<td>2</td>
<td>Ingo Wassink (UT) Work flows in Life Science</td>
</tr>
</tbody>
</table>
24 Dmytro Tykhonov () Designing Generic and Efficient Negotiation Strategies
25 Zulfiqar Ali Memon (VU) Modelling Human-Awareness for Ambient Agents: A Human Mindreading Perspective
26 Ying Zhang (CWI) XRPC: Efficient Distributed Query Processing on Heterogeneous XQuery Engines
27 Marten Voulon (UL) Automatisch contracteren
28 Arne Koopman (UU) Characteristic Relational Patterns
29 Stratos Idreos (CWI) Database Cracking: Towards Auto-tuning Database Kernels
30 Marieke van Erp (UvT) Accessing Natural History - Discoveries in data cleaning, structuring, and retrieval
31 Victor de Boer (UvA) Ontology Enrichment from Heterogeneous Sources on the Web
32 Marcel Hiel (UvT) An Adaptive Service Oriented Architecture: Automatically solving Interoperability Problems
33 Robin Aly (UT) Modeling Representation Uncertainty in Concept-Based Multimedia Retrieval
34 Teduh Dirgahayu (UT) Interaction Design in Service Compositions
35 Dolf Trieschnigg (UT) Proof of Concept: Concept-based Biomedical Information Retrieval
36 Jose Janssen (OU) Paving the Way for Lifelong Learning: Facilitating competence development through a learning path specification
37 Niels Lohmann (TUE) Correctness of services and their composition
38 Dirk Fahland (TUE) From Scenarios to components
39 Ghazanfar Faroq Siddiqui (VU) Integrative modeling of emotions in virtual agents
40 Mark van Assem (VU) Converting and Integrating Vocabularies for the Semantic Web
41 Guillaume Chaslot (UM) Monte-Carlo Tree Search
42 Sybren de Kinderen (VU) Needs-driven service bundling in a multi-supplier setting - the computational e3-service approach
43 Peter van Kranenburg (UU) A Computational Approach to Content-Based Retrieval of Folk Song Melodies
44 Pieter Bellekens (TUE) An Approach towards Context-sensitive and User-adapted Access to Heterogeneous Data Sources, Illustrated in the Television Domain
45 Vasilios Andrîkopoulos (UvT) A theory and model for the evolution of software services
46 Vincent Pijpers (VU) e3alignment: Exploring Inter-Organizational Business-ICT Alignment
47 Chen Li (UT) Mining Process Model Variants: Challenges, Techniques, Examples
48 Milan Lovric (EUR) Behavioral Finance and Agent-Based Artificial Markets
49 Jahn-Takeshi Saito (UM) Solving difficult game positions
50 Bouke Huurnink (UVA) Search in Audiovisual Broadcast Archives
51 Alia Khairia Amin (CWI) Understanding and supporting information seeking tasks in multiple sources
52 Peter-Paul van Maanen (VU) Adaptive Support for Human-Computer Teams: Exploring the Use of Cognitive Models of Trust and Attention
53 Edgar Meij (UvA) Combining Concepts and Language Models for Information Access

2011

1 Botond Cseke (RUN) Variational Algorithms for Bayesian Inference in Latent Gaussian Models
2 Nick Tinnemeier (UU) Organizing Agent Organizations. Syntax and Operational Semantics of an Organization-Oriented Programming Language
3 Jan Martijn van der Werf (TUE) Compositional Design and Verification of Component-Based Information Systems
4 Hado van Hasselt (UU) Insights in Reinforcement Learning; Formal analysis and empirical evaluation of temporal-difference learning algorithms
5 Base van der Raadt (VU) Enterprise Architecture Coming of Age - Increasing the Performance of an Emerging Discipline.
6 Yiwen Wang (TUE) Semantically-Enhanced Recommendations in Cultural Heritage
7 Yujia Cao (UT) Multimodal Information Presentation for High Load Human Computer Interaction
8 Nieske Vergunst (UU) BDI-based Generation of Robust Task-Oriented Dialogues
9 Tim de Jong (OU) Contextualised Mobile Media for Learning
10 Bart Bogaert (UvT) Cloud Content Contention
11 Dhaval Vyas (UT) Designing for Awareness: An Experience-focused HCI Perspective
12 Carmen Bratosin (TUE) Grid Architecture for Distributed Process Mining
13 Xiaoyu Mao (UvT) Airport under Control. Multiagent Scheduling for Airport Ground Handling
14 Milan Lovric (EUR) Behavioral Finance and Agent-Based Artificial Markets
15 Marijn Koolen (UvA) The Meaning of Structure: the Value of Link Evidence for Information Retrieval
16 Maarten Schadd (UM) Selective Search in Games of Different Complexity
17 Jiujin He (UvA) Exploring Topic Structure: Coherence, Diversity and Relatedness
18 Mark Ponsen (UM) Strategic Decision-Making in complex games
19 Ellen Rusman (OU) The Mind’s Eye on Personal Profiles
20 Qing Gu (VU) Guiding service-oriented software engineering - A view-based approach
21 Linda Terlouw (TUD) Modularization and Specification of Service-Oriented Systems
22 Junte Zhang (UvA) System Evaluation of Archival Description and Access
23 Wouter Weerkamp (UVA) Finding People and their Utterances in Social Media
24 Herwin van Welbergen (UT) Behavior Generation for Interpersonal Coordination with Virtual Humans On Specifying, Scheduling and Realizing Multimodal Virtual Human Behavior
25 Syed Waqar ul Qounain Jaffry (VU) Analysis and Validation of Models for Trust Dynamics
26 Matthijs Aart Pontier (VU) Virtual Agents for Human Communication - Emotion Regulation and Involvement-Distance Trade-Offs in Embodied Conversational Agents and Robots
27 Aniel Bhulai (VU) Dynamic website optimization through autonomous management of design patterns
28 Rianne Kaptein (UvA) Effective Focused Retrieval by Exploiting Query Context and Document Structure
29 Faisal Kamiran (TUE) Discrimination-aware Classification
30 Egon van den Broek (UT) Affective Signal Processing (ASP): Unraveling the mystery of emotions
31 Ludo Waltman (EUR) Computational and Game-Theoretic Approaches for Modeling Bounded Rationality
32 Nees-Jan van Eck (EUR) Methodological Advances in Bibliometric Mapping of Science
33 Tom van der Weide (UU) Arguing to Motivate Decisions
34 Paolo Turrini (UU) Strategic Reasoning in Interdependence: Logical and Game-theoretical Investigations
35 Maaike Harbers (UU) Explaining Agent Behavior in Virtual Training
36 Erik van der Spek (UU) Experiments in serious game design: a cognitive approach
37 Adriana Burlutiu (RUN) Machine Learning for Pairwise Data, Applications for Preference Learning and Supervised Network Inference
38 Nyree Lemmens (UM) Bee-inspired Distributed Optimization
39 Joost Westra (UU) Organizing Adaptation using Agents in Serious Games
40 Viktor Clerc (VU) Architectural Knowledge Management in Global Software Development
41 Luan Ibraimi (UT) Cryptographically Enforced Distributed Data Access Control
42 Michal Sindlar (UU) Explaining Behavior through Mental State Attribution
43 Henk van der Schuur (UU) Process Improvement through Software Operation Knowledge
44 Boris Reuderink (UT) Robust Brain-Computer Interfaces
45 Herman Stehouwer (UVT) Statistical Language Models for Alternative Sequence Selection
46 Beibei Hu (TUD) Towards Contextualized Information Delivery: A Rule-based Architecture for the Domain of Mobile Police Work
47 Azizi Bin Ab Aziz (VU) Exploring Computational Models for Intelligent Support of Persons with Depression
48 Mark Ter Maat (UT) Response Selection and Turn-taking for a Sensitive Artificial Listening Agent
49 Andreea Niculescu (UT) Conversational interfaces for task-oriented spoken dialogues: design aspects influencing interaction quality

2012
1 Terry Kakeeto (UvT) Relationship Marketing for SMEs in Uganda
2 Muhammad Umair (VU) Adaptivity, emotion, and Rationality in Human and Ambient Agent Models
3 Adam Vanya (VU) Supporting Architecture Evolution by Mining Software Repositories
4 Jurriaan Souer (UU) Development of Content Management System-based Web Applications
5 Marijn Plomp (UU) Maturing Interorganisational Information Systems
6 Wolfgang Reinhardt (OU) Awareness Support for Knowledge Workers in Research Networks
7 Rianne van Lambalgen (VU) When the Going Gets Tough: Exploring Agent-based Models of Human Performance under Demanding Conditions
8 Gerben de Vries (UvA) Kernel Methods for Vessel Trajectories
9 Ricardo Neisse (UT) Trust and Privacy Management Support for Context-Aware Service Platforms
10 David Smits (TUE) Towards a Generic Distributed Adaptive Hypermedia Environment
11 J.C.B. Rantham Prabhakara (TUE) Process Mining in the Large: Preprocessing, Discovery, and Diagnostics
12 Kees van der Sluijs (TUE) Model Driven Design and Data Integration in Semantic Web Information Systems
13 Suleman Shahid (UvT) Fun and Face: Exploring non-verbal expressions of emotion during playful interactions
14 Evgeny Knutov (TUE) Generic Adaptation Framework for Unifying Adaptive Web-based Systems
15 Natalie van der Wal (VU) Social Agents. Agent-Based Modelling of Integrated Internal and Social Dynamics of Cognitive and Affective Processes
16 Fiemke Both (VU) Helping people by understanding them - Ambient Agents supporting task execution and depression treatment
17 Amal Elgammal (UvT) Towards a Comprehensive Framework for Business Process Compliance
18 Eltjo Poort (VU) Improving Solution Architecting Practices
19 Helen Schonenberg (TUE) What’s Next? Operational Support for Business Process Execution
20 Ali Bahramisharif (RUN) Covert Visual Spatial Attention, a Robust Paradigm for Brain-Computer Interfacing
21 Roberto Cornacchia (TUD) Querying Sparse Matrices for Information Retrieval
22 Thijs Vis (UvT) Intelligence, politie en veiligheidsdienst: verenigbare grootheden?
23 Christian Muehl (UT) Toward Affective Brain-Computer Interfaces: Exploring the Neuropsychology of Affect during Human Media Interaction
24 Laurens van der Werff (UT) Evaluation of Noisy Transcripts for Spoken Document Retrieval
25 Silja Eckartz (UT) Managing the Business Case Development in Inter-Organizational IT Projects: A Methodology and its Application
26 Emile de Maat (UVA) Making Sense of Legal Text
27 Hayrettin Gurkok (UT) Mind the Sheep! User Experience Evaluation & Brain-Computer Interface Games
28 Nancy Pascall (UvT) Engendering Technology Empowering Women
29 Almer Tigelaar (UT) Peer-to-Peer Information Retrieval
30 Alina Pommeranz (TUD) Designing Human-Centered Systems for Reflective Decision Making
31 Emily Bagarukayo (RUN) A Learning by Construction Approach for Higher Order Cognitive Skills Improvement, Building Capacity and Infrastructure
32 Wietse Visser (TUD) Qualitative multi-criteria preference representation and reasoning
33 Rory Sie (OUN) Coalitions in Cooperation Networks (COCOON)
34 Pavol Jancura (RUN) Evolutionary analysis in PPI networks and applications
35 Evert Haasdijk (VU) Never Too Old To Learn – Online Evolution of Controllers in Swarm- and Modular Robotics
36 Denis Ssebugwawo (RUN) Analysis and Evaluation of Collaborative Modeling Processes
37 Agnes Nakakawa (RUN) A Collaboration Process for Enterprise Architecture Creation
38 Selmar Smit (VU) Parameter Tuning and Scientific Testing in Evolutionary Algorithms
39 Hassan Fatemi (UT) Risk-aware design of value and coordination networks
40 Agus Gunawan (UvT) Information Access for SMEs in Indonesia
41 Sebastian Kelle (OU) Game Design Patterns for Learning
42 Dominique Verpoorten (OU) Reflection Amplifiers in self-regulated Learning
43 André van Cleeff (UT) Physical and Digital Security Mechanisms: Properties, Combinations and Trade-offs