Willem Robert van Hage

Contents

Chapter 1

Introduction

1.1 The Research Field of Ontology Alignment

This thesis is about ontology alignment, the linking of structured vocabularies. To understand why this apparently abstruse subject is currently an important subject of research in computer science it is necessary to know what ontologies are, what they are used for, and why they should be aligned.

An ontology[1] is a set of concepts and their mutual relationships. Ontologies capture part of the semantics (*i.e.* meaning) of these concepts using some formalism, such as model theory or description logics. Their strictness varies greatly, *cf.* Obrst (2006); McGuinness (2003). In practice, ontologies are often simple vocabularies, thesauri, or classification schemas, with relatively weak semantics. Examples of ontologies are taxonomies that describe the hierarchy of species, upper-level ontologies that define concepts such as events and agents, and thesauri that describe relations between words, like synonymy and hyponymy. Which aspects of concepts are described, for example, their super and subconcepts, what names they are given, if you can count them, what they are not, depends on what the ontology is used for. Ontologies can be used to reveal differences or even inconsistencies in the way people speak about the world, to classify objects by their properties, or to define exactly to which products a trade agreement applies. Ontologies used for sense disambiguation primarily contain information about which words refer to which things, while ontologies used for classification mainly contain hierarchical subclass relations, for example, a human is-a mammal.

The main goal of the ontologies that are discussed in this thesis is information disclosure, finding documents. These ontologies are usually thesauri that consist of a big hierarchy of concepts that stand for subjects of documents. They are used to overcome language barriers when searching for a document in a library. For example, by offering synonyms to help people find the right words (*e.g.* 'Mystery swine disease', see 'PRRS'), by listing translations into other languages (*e.g.* english: 'pet care', german: 'Heimtierpflege', czech: 'péče o domácí zvířata'), by pointing at related concepts (*e.g.* 'Tuber (truffles)' related to 'edible fungi'), and by ordering the subjects hierarchically (*e.g.* generic: 'potatoes', 'vegetables', 'plant products', or partitive: 'Liechtenstein', 'Western Europe', 'Europe').

Even though ontologies can clearly capture the meaning of concepts for one group of people, other groups may disagree about this meaning and have ontologies with conflicting specifications. People can use describe things in different ways. The differences can exists on many levels. People can use different languages, they can use different words for the

[1] The thing with an article in front of it, not the topic in philosophy. For a, see Gruber (1993).

same things (*e.g.* different jargon), they can describe things from different points of view (*e.g.* insulin as a hormone or as a protein), and they can disagree on the meaning of things (*e.g.* when they have different experiences or different idiom).

When people want to tap into each other's knowledge bases, they have to work out for which concepts in their ontology there is a corresponding concept in the other ontologies. The process of finding these correspondences between ontologies is called *ontology alignment*. These correspondences can be of many different kinds. The most common types are: equivalence, generic (is-a, subClassOf, *e.g.* cats are mammals), partitive (part-whole, like parts of a complex object, or the material something is made of, *cf.* Winston et al. (1987), and instantive (the type something has, *e.g.* Mona Lisa is a painting, Texel is an island). In this thesis we investigate the three most widely used relations, equivalence, subclass, and partitive.

To explain why ontology alignment has become an important issue we have to start with the influence of the internet and especially the World Wide Web on the flow of information. The World Wide Web has democratized information access. Anybody can provide information to anyone on the web and anybody can consume information from anyone. This has made it possible to easily draw information from various sources, for example, libraries, weather services, or bookstores, without the effort of having to physically visit all these places. Everybody can add to the growing network of documents. It is an open world. An interesting possibility this brings is searching in many different collections at the same time, federated search. This makes it possible to find a much greater number and greater variety of resources than before.

Simple federated search over collections is possible by simply distributing queries over the indexing systems of the collections. This way the interpretation of the query is left to the each system separately. This can lead to situations where one system interprets the query differently than the other system, which often leads to undesirable effects. To solve this problem the indexing systems have to 'understand each other' so that the meaning of a query is interpreted in the same way on each system.

In the early 1990s this was done by merging the ontologies. When two organizations wanted to cooperate, knowledge engineers and domain experts of both organizations would work together to create one new ontology that unified the old ontologies. Such unification initiatives were often large projects, for example, the Unified Medical Language System (UMLS) project (Lindberg et al., 1993; Bodenreider, 2004) or the Unified Agricultural Thesaurus (UAT) project (Hood and Ebermann, 1990; Friis et al., 1993; Clarke, 1996). Merging the ontologies of different organizations requires a high degree of cooperation, standardization, and commitment. Everybody has to agree about, for example, what the resulting ontology will look like, how to deal with different points of view, who is responsible for its maintenance, and under which conditions it can be used (*e.g.* licensing). This can lead to conflicts and, hence, many of these projects, like the UAT project, were not considered a success by all participating parties.

Currently, the preferred method of knowledge integration is ontology alignment. Alignment allows access to other ontologies via mediating mappings, while the original ontologies remain unchanged. On the one hand, this means that a network of aligned ontologies is not necessarily a consistent whole. Consequently, information systems that use the mappings

have to be tolerant to inconsistencies. On the other hand, this means the organizations only have to agree on issues pertaining to the alignment. Each organization can decide for itself how to deal with modeling, maintenance, and licensing issues for their own ontology.

1.2 Ontology Alignment and the Semantic Web

The Semantic Web is a W3C project conceived by Tim Berners-Lee, the inventor of the World Wide Web, that aims at publishing interlinked data on the web in a machine-understandable form (Berners-Lee et al., 2001; Antoniou and van Harmelen, 2004). The purpose of the Semantic Web is to create a web of data that computers can use, as opposed to the current web, which is meant to be used by humans. A number of conferences, such as the International Semantic Web Conference[2] (ISWC) and its regional counterparts in Europe and Asia, bring together researchers from different communities, like the knowledge representation and reasoning, databases, and hypermedia communities, to develop technology to make this web of data possible. The main elements of the Semantic Web are the Universal Resource Identifier (URI) and ontologies formulated in web languages, such as RDF(S) and the Web Ontology Language (OWL).[3] The URI is to things and concepts[4] what the URL is to locations on the web. They follow the same syntax as URL's, which are alco considered URI's. By representing idea's on the web, people can share them in the same fashion as content on the World Wide Web, by linking to them by using each other's URI's. The idea of using web technology for sharing vocabularies has made decentralized information integration and ontology alignment significantly easier, both from the technological and social perspective.

The idea of decentralized information integration by means of mediators–in the spirit of the Web–formed in the early 1990s, *e.g.* Wiederhold (1991, 1994), but it took a decade to become widely accepted. Similar developments happened in the meantime in different guises: peer-to-peer systems (*e.g.* Napster, Gnutella), the Service-Oriented Architecture[5], GRID computing (Foster and Kesselman, 1999), large-scale wiki projects like Wikipedia[6], and the development model underlying open-source software (Raymond, 1999). The common principle underlying these developments is unity in diversity. They allow different degrees of consensus to survive. When there is not one single solution that everybody can live with it is not all or nothing. Instead, many partial solutions that can coexist, so that stepwise improvement of cooperation becomes possible.

1.3 Ontology Alignment Techniques and Standards

The ontologies on the web are of many different kinds, ranging from heavy-weight ontologies formulated in logic, for example, using the OWL dialects OWL Full or OWL DL, to folksonomies and controlled vocabularies in ad-hoc formats. Most of the vocabularies on

[2] perhaps more accurately named the International Semantic Web Technology Conference

[3] Respectively, see `http://www.w3.org/RDF`, `http://www.w3.org/TR/rdf-schema`, and `http://www.w3.org/TR/owl-guide`.

[4] *i.e.* **everything** you can think of. See `http://www.w3.org/Addressing`.

[5] `http://www.oasis-open.org/committees/soa-rm`

[6] `http://www.wikipedia.org`

the web are simple hierarchies with little semantic commitment, *cf.* d'Aquin et al. (2007); Wang et al. (2006). Knowledge acquisition has been widely acknowledged as a bottle neck for semantic-web applications. A survey that shows this is the analysis performed in the Knowledge Web network of excellence (Nixon and Mochol, 2004). Ontologies simply do not come for free. The adaptation of legacy vocabularies is an obvious partial solution to this problem. In the past years, the research field of ontology alignment has rapidly developed into maturity. Many organizations, especially libraries and museums have undertaken alignment projects, *e.g.* Aleksovski et al. (2007). Since 2004 the Ontology Alignment Evaluation Initiative[7] has hosted an increasing number of widely varying alignment tasks ranging from the alignment of web directory structures to rich medical ontologies.

Outside of the web, the most common type of vocabularies are thesauri. In this thesis we will mainly focus on thesauri. One of the main topics of the Semantic Web Best Practices and Deployment Working Group (SWBPD) organized by the W3C is publishing thesauri on the semantic web.[8] The SWBPD advocates the use of Simple Knowledge Organization System (SKOS)[9] for the representation of thesauri on the semantic web, as opposed to OWL. Most thesauri follow the ANSI/NISO Z39.19 standard (ANSI/NISO, 2005) and use BT and NT relations that do not specify whether the relations are generic, instantive, or partitive. The Web Ontolology Language (OWL) standard only predefines the rdfs:subClassOf relation, a generic relation, and rdf:type, the instantive relation. This means converting a Z39.19 thesaurus to OWL can require serious knowledge engineering to determine the relation subtype of the BT/NT relations. Simply translating all BT/NT relations to rdfs:subClassOf will lead to incorrect inferences.[10] SKOS is a semantic web language specified in OWL that stays close to the meaning of the constructs in the Z39.19 standard. For example, it contains the skos:broader and skos:narrower relations that have the same semantics as the Z39.19 BT and NT relations.

The alignment relations that are used most frequently are the equivalence relation, followed at some distance by the instantive, subsumption and incompatibility relations. For the alignment of most OWL Lite or OWL DL ontologies one can simply use owl:equivalentClass for equivalence alignment of classes, owl:sameAs for equivalence alignment of individuals and rdfs:subClassOf for subsumption alignment of classes, and rdf:type for instantive alignment. The representation of incompatibility is more complex. The assertion of a owl:disjointWith relation is often an overstatement. For example, if the classes named 'males' and 'females' are aligned using owl:disjointWith then this would be in conflict with species of animals that are hermophrodites, like the garden snail,[11]. A weaker statement was probably intended. SKOS contains alignment relations analogous to the BT/NT hyponymy relations and to express synonymy, respectively skos:broadMatch, skos:narrowMatch, and skos:exactMatch. For the alignment of thesauri written down in SKOS (which itself is specified in OWL), one can use either RDFS/OWL properties, like rdfs:subClassOf, or SKOS alignment properties. A practical advantage of using SKOS relations for the alignment of thesauri as opposed to

[7] http://oaei.ontologymatching.org

[8] http://www.w3.org/2004/03/thes-tf/mission

[9] http://www.w3.org/2004/02/skos

[10] The article by Winston et al. (1987) about partitive relations contains a systematic analysis of the composition of partitive and generic relations.

[11] http://en.wikipedia.org/wiki/Helix_aspersa

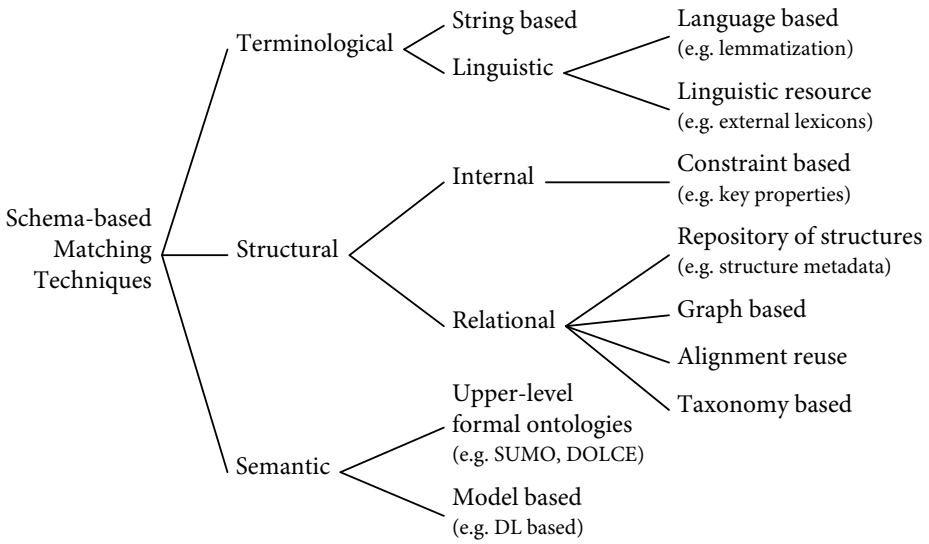


Figure 1.1: A taxonomy of ontology-alignment techniques, grouped by the type of input used. An adapted excerpt from Shvaiko and Euzenat (2005).

OWL properties is that, without additional rules, they do not make strong logical commitments. Since rdfs:subClassOf is stricter than skos:broader, alignment using rdfs:subClassOf can quickly lead to overcommitment.

There is a wide variety of techniques for the discovery of alignment relations. Most alignment techniques focus on finding equivalence relations between the concepts of the two vocabularies that are to be aligned. This is generally done by computing similarity scores between pairs of concepts from the vocabularies. These scores are based on any number of different features, such as: similar terms (*e.g.* 'Horse' and 'horses'), similar related concepts (*e.g.* 'sofa' and 'couch' are both 'furniture' and related to 'chaise longue'), or similar constraints (*e.g.* 'Russia' and 'Russian Federation' both have exactly 46 oblasts and 9 krais). An elaborate overview is given in Shvaiko and Euzenat (2005). An excerpt from that article is shown in figure 1.1, which shows a taxonomy of alignment techniques, based on the kind of input they use.

Only a few techniques use sophisticated reasoning, such as Description Logic (DL) reasoning, for example, Meilicke and Stuckenschmidt (2007), or background knowledge, for example, Aleksovski et al. (2006a,b); Sabou et al. (2007). The primary matching strategy of most systems is lexical comparison of the labels of concepts. Even the techniques that use, for example, property constraints or background knowledge from lexicons or upper-level ontologies, use lexical comparison first, before additional reasoning is employed. For instance, for a system to base a match on the fact that both concepts 'have four legs' it will have to conclude that the 'legs' are the same sense of the word leg (compare legs of animals, pants, furniture, journeys, races, etc.).

In the first chapters of this thesis we introduce two alignment techniques. Both are

linguistic techniques that use external linguistic resources. In some cases we apply part-of-speech tagging, which can be considered a language-based method. However, the main focus of this thesis is the evaluation of alignment approaches, as opposed to the techniques themselves. Therefore, we do not go into much detail about the techniques. Detailed descriptions of the other alignment techniques mentioned in this thesis can be found in Euzenat et al. (2006, 2007).

I.4 Problem Statement and Research Questions

The alignment of vocabularies can be time-consuming. For libraries, the most useful vocabularies to align are very large, containing tens or hundreds of thousands of concepts and various relations between these concepts. If two of these vocabularies have a significant topical overlap, then the number of correspondences between these vocabularies can also be very large. Manual ontology alignment can take years. For instance, aligning the Chinese Agricultural Thesaurus (CAT) of around 60,000 concepts to the Food and Agriculture Organization of the United Nations thesaurus (FAO), AGROVOC, of around 25,000 concepts took *seven man-years* of manual labor.[12] Therefore, automated ontology alignment could be valuable.

Automatic ontology-alignment systems are imperfect, significantly more so than human beings at this moment, but in completely different aspects than humans. Some tasks that are easy for humans, like word-sense disambiguation, are much more difficult for computers, while some tasks that are hard for humans, like meticulously performing repetitive tasks on large amounts of data, or logical reasoning are trivial for computers.[13] *It is unclear how ontology-alignment systems measure up to humans and how useful they are in practice.* The goal of this thesis is to discover when and how it is feasible to automate the ontology-alignment process. We make this goal more concrete by formulating four research questions.

I. *What is the quality of current alignment techniques?*

Ontology-alignment systems are imperfect. Some of the mappings they produce are erroneous. Some are simply understatements, like 'Amsterdam' skos:broadMatch 'Earth' if the concept 'Netherlands' is also an option, some are debatable, like near-synonyms[14] 'meteors' ≈ 'meteorites', and some are simply wrong, like 'Apples' ≠ 'Oranges'. In order to know when it is feasible to automate ontology alignment in practice it is vital to know what the performance is, with respect to the quality of the mappings, of current state-of-the-art ontology alignment techniques. An answer to this question requires measurements and hence also measurement methods. We perform separate investigations for equivalence, subsumption, and partitive alignment relations, because the evidence from which these relations can be derived differ, as well as their criteria for validity.

[12] This is based on personal communication with the project leader, Dr. Chang Chun, of the Chinese Academy of Agricultural Sciences.

[13] *cf.* Moravec's paradox, http://en.wikipedia.org/wiki/Moravec's_paradox

[14] This is especially an issue when dealing with multilingual thesauri.

II. *Which domain-specific factors influence this quality?*

Some mappings are easier to discover than others. One reason for this is that some concepts are simply more alike in their outward appearance and hence easier to match. For example, some have exactly the same name, or almost exactly the same name, for instance, the singular and plural form of the same word, like 'Buildings' and 'building'. While others referring to the same thing using very different words, like 'tremor' and 'quake', are much harder to match. In some domains of discourse, it is harder to find alignments than in other domains. This can depend on many factors, like the general level of consensus on the terminology that is used, the level of consensus on the structure of the domain (*i.e.* how well the domain is understood), or which type of alignment relations are more common. The answer to this question should partially explain the quantitative results of the first research question. We want to know which properties of the domain pose particular constraints on the performance of alignment techniques. For example, what kind of knowledge is available in the domain for the alignment techniques to use, like reference data to use as background knowledge, or well-known naming schemes that can be exploited. Within the scope of this thesis we can only investigate a small selection of domains, so the answers to this question also are limited.

III. *Which application-specific factors influence this quality?*

Another type of factors that influence the quality of alignment techniques is application-specific factors. These factors have to do with the task (prediction, diagnosis, assessment, etc. *cf.* Schreiber et al. (1994, 1999)) an application has to perform for which is needs the alignment. Different applications need different mappings to work. For example, an application about traceability of meat products might require geographical part-whole relations to categorize of locations, and equivalence relations between foodstuffs to match various synonyms of products in news feeds, while a patient classification application might need subclass relations between classes of symptoms to assist with a diagnosis task. Furthermore, some applications have higher performance requirements than others. For example, a television program recommendation system might already be considered good when there is a good recommendation amongst every four suggestions, while for another application this level of performance is completely unacceptable. Incorrect mappings can lead to bad results in an application. For example, information retrieval applications might suffer from topic drift and classification applications might suffer from inconsistency. Each application has its own minimal quality requirements to yield satisfactory results.

IV. *How can human preference be incorporated in the evaluation of automatic alignment methods?*

The reference by which we judge automatic alignment is manual alignment. The best automatic alignment system produces alignments that are indistinguishable from the work of a human expert.[15] In this thesis we only consider properties of the resulting alignment, not other performance properties of the systems, like user friendliness, speed, or hardware requirements. There are many methods and measures to compare the work of a human to that of a computer. For example, a possible evaluation method is to have human experts

[15] *cf.* the Turing test `http://en.wikipedia.org/wiki/Turing_test`.

create the entire alignment by hand, so that it becomes clear exactly which mappings are made by humans, but not by computers, or vice versa. However, this can be too costly. Sample-based evaluation methods can solve this problem at the cost of some uncertainty. In this thesis we investigate how sampling techniques can be applied to the evaluation of ontology alignment.

I.5 Contributions and Guide to the Reader

This thesis consists of two types of contributions to the field of ontology alignment. Chapter II and III contribute alignment techniques for relation types for which there are no satisfactory alternatives. Chapter VI-VIII contribute to the evaluation of alignment techniques, either by elaborating on an evaluation task or by working out an evaluation method. The following two lists give a short description of these contributions in the order by which they appear in this thesis, and how these contributions pertain to the research questions.

I.5.1 Outline by Contribution

Part-Whole and Subclass Relation Learning Techniques We extend the range of ontology alignment techniques. Most of the existing alignment techniques are designed to find equivalence relations using only the ontologies as input. We introduce complementary methods to find subclass and part-whole relations from textual background knowledge, for example, web pages or dictionaries. This is described in chapter II, which is based on van Hage et al. (2005); chapter III, which is based on van Hage et al. (2006); and chapter IV, which is based on van Hage and Katrenko (2007).

Comparative Evaluation of Alignment Systems We perform a comparative evaluation of alignment techniques. Most alignment projects have to do with library collections indexed with large thesauri. We introduce two tasks of the OAEI, the food and environment tasks, to measure the quality of alignment techniques on matching the thesauri of the United States Department of Agriculture (the NAL Agricultural Thesaurus), the Food and Agriculture Organization of the United Nations (AGROVOC), and the European Environment Agency (GEMET). We measure and compare the performance of the seven alignment systems that participated in these tasks. We investigate factors that influence how well the various techniques implemented by the systems perform. This is described in chapter V, which is based on van Hage et al. (2008b), and the report Euzenat et al. (2007).

Alignment Sampling and End-to-End Evaluation Methods One aspect in which the thesauri of the OAEI food and environment tasks are typical is that they are large. Hence, the alignments between them are also large. We describe a method for sample-based evaluation of ontology alignment to make the comparison of alignment techniques on large ontologies feasible. This is described in chapter VI, which is based on (van Hage et al., 2007).

Analysis of Alignment by Humans versus Computers In the OAEI automatic ontology alignment systems are compared to each other. This allows us to conclude which techniques outperform others in certain cases. However, it does not tell us how these techniques relate to human experts. Most of the current alignment work is carried out by human experts. To conclude how the automatic techniques would fare in practice we analyze the difference between a manually-created alignment (the alignment between the AGROVOC thesaurus and the German national library's Slagwortnormdatei) to the automatically-generated alignments of the OAEI food task. This is described in chapter viii, which is based on Lauser et al. (2008).

Relevance-Based Evaluation Method The evaluation tasks of the OAEI do not take into account which part of the alignment is most useful in practice. Most tasks assume that every mapping is equally valuable. In practice, this is not true. The typical case is not equal to the average case. We describe a method to draw samples to represent typical usage. We apply this method to the OAEI food task to complement its average-case estimates. This is described in chapter vii, which is based on van Hage et al. (2008a).

1.5.2 Outline by Research Question

I. Quality Measurement of Alignment Approaches Chapter ii to v contribute to our understanding of the quality of current alignment approaches. In chapter ii we measure the performance of various techniques to find subclass relations between the USDA Nutrient Database for Standard Reference and the FAO AGROVOC thesaurus. In chapter iii we measure the performance of a technique to find part-whole relations between a list of carcinogenic substances from the International Agency for Research on Cancer (IARC) and possible carriers of these substances the USDA National Agricultural Library's Agricultural Thesaurus (NALT) and AGROVOC. In chapter iv we measure the performance of the same technique to find part-whole relations without a restricted domain. In chapter v we measure the performance of seven different ontology alignment systems that find mainly equivalence relations by setting up a comparative evaluation task about the alignment of the NALT and AGROVOC thesauri.

II. Determination of Domain-Dependent Factors In chapter ii we investigate how the domain of food products, like rice and mozzarella, constrains the discovery of subclass relations. Also, we look at the influence of the type of text used for learning subclass relations. We compare learning from a cooking dictionary (domain specific) to learning from web pages indexed by the web search engine Google (domain neutral). In chapter iii we investigate the domain of carcinogenic substances, like benzene and asbestos, and that of substances that can contain these carcinogens by which they can reach humans, for example, dyes, pesticides, and animal fat, for the discovery of part-whole relations. As with the learning of subclass relations, we investigate the effect of language in web pages indexed by Google on the alignments that we find. In chapter iv we also learn part-whole relations from text, but we investigate this in the domain of generic english sentences to determine the effect of domain-specific language on the learning of part-whole relations. Consider a typical

sentence that mentions carcinogens from chapter III (specific domain): "*Benzene* is used in the *dehydration process*"; as opposed to a generic domain sentence: "John opened the *door* of the *car* with difficulty". In chapter V we apply stratification by domain to the alignments we evaluate. This way we can see how the alignment systems handle the specific properties of various domains, like taxonomy (plants, animals, etc.), biology and chemistry (genes, proteins, etc.), and geography. In chapter VIII we compare mappings from alignments in the agricultural domain. One alignment is between two thesauri that are specifically tailored to the agricultural domain, and the other an alignment between an agricultural thesaurus and a general domain thesaurus with a part about agriculture. Respectively, the automatically-generated alignments from the OAEI task described in chapter V between NALT and AGROVOC, and the manually created alignment between the German national library's Schlagwortnormdatei thesaurus (SWD) and AGROVOC. This allows us to investigate the influence of german-english multilinguality and the domain-specificity of the thesauri on the difficulty to find alignments.

III. DETERMINATION OF APPLICATION-DEPENDENT FACTORS In chapter V we investigate which relations can be found by current state-of-the-art alignment systems in the OAEI 2006 and 2007 food task. Also, we analyze common mistakes and omissions in the automatically-generated alignments. In chapter VI we suggest two methods to incorporate application requirements in the evaluation of alignment techniques. One proposes to measure alignment quality on samples of mappings that are typically required by the application, the other to measure the effect of alignments on an application by looking at the behavior of the entire application. The sample-based evaluation method used in chapter V is based on the former of the two methods. In chapter VII we describe a third method to incorporate application demands in the evaluation of alignment techniques. We start from typical usage scenarios of an application and deduce which mappings are necessary in these cases for the application to operate in a satisfactory manner. These mappings, as opposed to randomly drawn mappings, are used for the evaluation of the alignment. In chapter VIII we investigate which alignment relation type humans prefer to use and compare that to which alignment relation types automatic alignment techniques yield. This gives some insight into which mappings are *not* found by the systems participating in the OAEI food task described in chapter V. In chapter III and VII we test alignment methods in the context of two specific application scenarios, respectively a fact-finding and a metadata-based retrieval scenario.

IV. EVALUATION AND HUMAN PREFERENCE In chapter VI and chapter VII we discuss three different sampling techniques: *Alignment sampling*, *end-to-end evaluation*, and *relevance-based evaluation*. *Alignment sampling* and *relevance-based evaluation* measure which fraction of a number of sample mappings is found by automatic techniques (Recall) and which of the mappings that are found by automatic techniques would also be suggested by human experts (Precision). The way these samples are drawn differs. *End-to-end evaluation* is also a sampling technique, but the samples do not consist of mappings, but of application scenarios. Which alignment technique works better is deduced from which technique leads to more successful sample scenarios.

Some differences between the alignment behavior of humans and computers are hard

to deduce from the aggregated numbers that result from the techniques described above. Therefore, in chapter VIII, we perform an in-depth analysis of this difference by manually classifying a sample set of mappings by difficulty, *i.e.* how much knowledge is required to discover each mapping.

1.6 The Virtual Laboratories for e-Science Project

The context in which the work in this thesis was done is the Adaptive Information Disclosure (AID) subprogram of the Virtual Laboratories for e-Science project[16] (VL-e), a project funded by the Dutch government aimed at improving and extending the role of computers in scientific research (Herzberger, 2006). The overall theme of research in the VL-e project is GRID and service-oriented computing. AID focusses on information retrieval with background knowledge, learning semantic relations from text, and metadata-based information-disclosure techniques.

Using the GRID as the underlying architecture of a virtual laboratory allows researchers to share both their data and their computing power. GRID API's allow transparent access to file storage as well as processors. However, these API's currently only support low level functionality like resource allocation and job scheduling. Sharing research results happens on the level of file paths, access rights, and sending e-mail to tell people where your files are. The goal of the AID group in VL-e is to add high-level functionality for cooperation to the GRID. For example, search engines, web-service-based workflows to share your experiments, and metadata repositories to share the descriptions of your data, workflows, people, etc.

The applications we developed to provide these high-level functions were all implemented as SOAP web services that run on a GRID access point. This allows them to use GRID computers for computation or storage-intensive tasks like machine learning, search-engine indexing, or hosting the search index. At the same time it allows people to access the power of the GRID over the web. The web services are composed into using workflow tools like Taverna,[17] possibly incorporating web services hosted at other places on the web. The services set up by the AID group, together the AIDA toolkit, are illustrated in figure 1.2. The arrows indicate typical useful connections that a workflow could make. With the services in figure 1.2 AID set up two exemplar workflows in cooperation with the Bioinformatics and Foodinformatics groups of VL-e. Respectively for the discovery of gene-disease association from medical texts, and for metadata-based access to a collection of research questions about food and sensory experiments for literature research on the crunchyness and creamyness of food products.

Apart from setting up these workflows, AID cooperated with with the Bioinformatics group on specialized information retrieval in TREC Genomics, and text mining to learn protein interactions; and with the Foodinformatics groups on the development of a thesaurus-based information retrieval tool, format conversion of existing vocabularies, text mining for health risk assessment, and the alignment of agricultural thesauri.

[16] http://www.vl-e.nl
[17] http://taverna.sourceforge.net

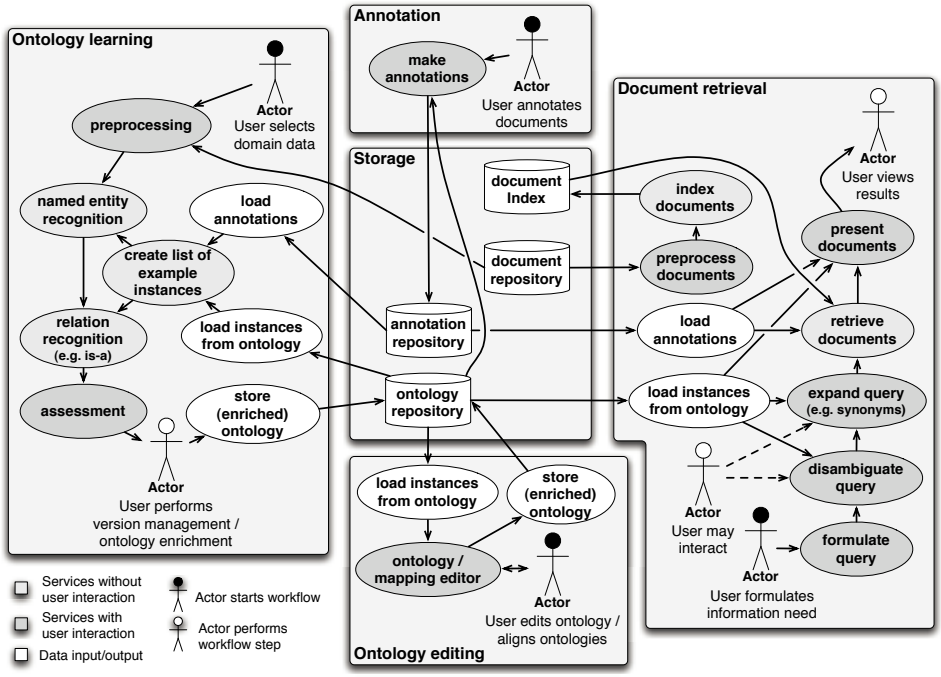


Figure 1.2: Web-service architecture of the AIDA toolkit.

This thesis is mainly about the last topic, the alignment of agricultural thesauri for the advancement of food and environmental research, like food product development and food safety research. Specifically, to develop a suitable evaluation methodology for existing alignment techniques, and to develop new alignment techniques whenever no suitable techniques exist for an alignment task. Hence, the data sets used in this thesis all deal with the food domain. The Foodinformatics group consists both of academic (Wageningen University and Research Centre)and corporate researchers (Unilever, Friesland Foods, TNO Quality of Life), so we only considered data sets that are available for use without restrictions.[18]

1.7 A Note on Terminology in this Thesis

An interesting (and ironic) fact is that, even within the ontology alignment community, there are many different words people use to refer to ontology alignment (and the various things that have to do with it, such as concepts and mappings). It seems almost every research group uses its own terms and definitions, each with its own subtle differences in meaning. In most of the articles that this thesis is made up of, we choose to adhere to the words used

[18]This eliminated, for example, the CAB thesaurus.

in Euzenat and Shvaiko (2007). That is, generally, we call the process 'ontology alignment', a match between two concepts a 'mapping' or 'correspondence', and the set of mappings that results from the ontology alignment process an 'alignment'. Sometimes mappings are considered to be directional correspondences, but for thesaurus alignment this distinction is often not important, as most alignment relations have an obvious inverse relation (*e.g.* skos:broadMatch is the inverse of skos:narrowMatch).

To properly interpret our work, please take a liberal stance with respect to which words we use and, in the scope of this thesis, consider the following words to be synonymous: '(ontology | schema | vocabulary) (alignment | mapping | matching)' all refer to ontology alignment; '(mapping | correspondence | cross-concordance)' all refer to a single mapping between two concepts; '(alignment | crosswalk | mapping)' all refer to a set of mappings; '(term | concept | class)' all refer to a concept; '(label | concept name | word)' all refer to words used to name and describe a concept; a 'preferred (label | term)' is the name of a 'descriptor (term)?' and an 'alternative (label | term)' is the name of a 'non-descriptor (term)?'.

Furthermore, besides the topic of this thesis, we consistently use the pronoun 'we' when we refer to the author, to acknowledge the influence of the co-authors and the scientific community.

Chapter II

Finding Subclass Relations

In this chapter we study the task of learning subclass relations from text for the purpose of aligning ontologies. We discuss four linguistic techniques that use either a web search engine or a dictionary as text source. We evaluate these techniques by aligning two vocabularies in the food domain. We measure Precision and Recall of the alignment methods on samples. This evaluation method was used as an input for the generic alignment sample evaluation *method described in chapter VI.*

This chapter is based on a paper coauthored by Sophia Katrenko and Guus Schreiber, "A Method to Combine Linguistic Ontology-Mapping Techniques, Willem Robert van Hage, Sophia Katrenko, Guus Schreiber" *(van Hage et al., 2005), which was presented at the fourth International Semantic Web Conference (ISWC 2005).*

Abstract We discuss four linguistic ontology-mapping techniques and evaluate them on real-life ontologies in the domain of food. Furthermore we propose a method to combine ontology-mapping techniques with high Precision and Recall to reduce the necessary amount of manual labor and computation.

II.1 Introduction

Ontologies are widely used to provide access to the semantics of data. To provide integrated access to data annotated with different, yet related, ontologies, one has to relate these ontologies in some way. This is commonly done by cross-referencing concepts from these ontologies. In different contexts this practice is called ontology mapping, schema matching, or meaning negotiation. In the literature one can find surveys of the widely varying methods of automated ontology mapping. For instance, in the surveys done by Kalfoglou and Schorlemmer (2003); and Rahm and Bernstein (2001). The latter organized the methods hierarchically. The ontology-mapping methods we develop in this chapter fall in the categories *schema-only based*, which means they work on the conceptual part of the ontology and not on the annotated individuals and *linguistic*, since we use the labels of the concepts. The techniques we use come from the field of *information retrieval* (IR).

The work in this chapter is done within the scope of the Adaptive Information Disclosure (AID) project, which is part of the greater effort of the Dutch *Virtual Labs for e-Science* project (VL-e)[1]. The AID project focusses on facilitating access to domain-specific text corpora,

[1] http://www.vl-e.nl

in particular articles about food. When the semantics of data sources or the information needs are of increasing complexity old-fashioned information-retrieval systems can fail to deliver due to the following reasons:

- Domain-specific terms can have homonyms in a different domain. For instance, 'PGA' stands for 'Polyglandular Autoimmune Syndrome' and the 'Professional Golfers' Association'.
- Synonyms used by different communities can be difficult to relate to each other. For instance, some refer to 'stomach acid' with 'Betaine HCl', others use 'Hydrochloric Acid'.
- Skewed term-frequency distributions can lead to failing weighting schemes. For instance, the term 'cancer' occurs as frequently as some stop words in the medical MedLine corpus, but it is an important term.

Ontologies pave the way for new techniques to facilitate access to domain-specific data. Semantic annotation of text resources can help to subdue jargon. (Kamps, 2004; Stuckenschmidt et al., 2004) Obviously accessing annotated data sources is not without problems of its own. In practice different data sources are often annotated with different ontologies.[2] In order to provide integrated access using multiple ontologies, some form of ontology mapping needs to be done.

Within AID we focus on food information corpora. This domain–like the medical domain–struggles with an information overload and jargon issues. For instance, everyday household terms are intermingled with names of proteins and other chemical compounds. This complicates the formulation of good search queries. In this chapter we test the applicability of four automated ontology-mapping techniques on real-life ontologies in the domain of food and assess their practical use. Specifically we try to map the USDA Nutrient Database for Standard Reference, release 16 (SR-16)[3] onto the UN FAO AGROVOC thesaurus (AGROVOC)[4] using that yield RDFS (Brickley and Guha, 2000) subClassOf relations. The four techniques we discuss are listed below.

1. Learn subclass relations between concepts from AGROVOC and SR-16 by querying Google for Hearst patterns. (Hearst, 1992)
2. Learn subclass relations by extracting them from Google snippets returned by the same queries with the help of shallow parsing using the TreeTagger part-of-speech tagger. (Schmid, 1994)
3. Learn subclass relations by extracting them from a semi-structured data source, the CooksRecipes.com Cooking Dictionary, with MINIPAR (Lin, 1998).
4. Use the Google hits method as a sanity check to filter the dictionary mining results.

[2] We use the term ontologies to include light-weight ontologies such as vocabularies and thesauri
[3] http://www.nal.usda.gov/fnic/foodcomp/Data/SR16/sr16.html
[4] http://www.fao.org/agrovoc

In section II.2 we discuss some related work to give an impression of current practice in relation extraction. In section II.3 we describe the experimental set-up we used in which we tested the four mapping techniques. In section II.4 we describe the four techniques in great detail and discuss the acquired results. In section II.5 we propose a method for applying the techniques in practice and we show how much manual labor can be saved.

II.2 Related Work

Brin proposed a method called Dual Iterative Pattern Relation Extraction (DIPRE) in his paper from 1998 (Brin, 1998). He tested the method on part of his Google corpus–which at the time consisted of about 24 million web pages–to learn patterns that link authors to titles of their books. These patterns were then used to retrieve author-title relation instances from the same corpus. An example of such a pattern is the HTML bit: "<li>title by author".

In 1992 Hearst devised a set of lexico-syntactic patterns for domain aspecific hyponym extraction (Hearst, 1992). Her patterns found entrance in many applications such as Cimiano and Staab's PANKOW system. (Cimiano and Staab, 2004) The first method we discuss in this chapter is similar to their work.

In their 2004 paper Cimiano and Staab try to accomplish two things. The first is a instance classification task: to classify geographical entities such as Amsterdam (City), Atlantic (Ocean), etc. The second is a subclass learning task: to reconstruct a subclass hierarchy of travel destinations mentioned in the LonelyPlanet website[5]. The method they use is the same for both tasks. They send Hearst patterns describing the relation they want to test to the Google API and depending on the number of hits Google returns they accept of reject the relation. For instance, the query "cities such as Amsterdam" yields 992 hits. Depending on which threshold they put on the number of hits they achieved Precision between 0.20 and 0.35 and Recall somewhere between 0.15 and 0.08. The higher the threshold, the higher the Precision and the lower Recall.

What we want to accomplish is a bit more complicated than either of Cimiano and Staab's tasks for two reasons. The food domain is less well-defined than the geographical domain, in which there are exhaustive thesauri such as TGN. The relations between the concepts are clearly defined. Countries have exactly one capital. Countries can border each other, etc. In the food domain such consensus does not exist. This means the evidence for relations that can be found in Google can be expected to be more ambiguous in the food domain than in the geographical domain.

II.3 Experimental Set-up

Our set-up consists of the two thesauri we want to connect, the auxiliary sources of knowledge we use to learn the mappings from, and a gold-standard mapping to assess the quality of the learnt relations. In section II.3.3 we discuss the gold standard and the evaluation measures we use.

[5] http://lonelyplanet.com/destinations

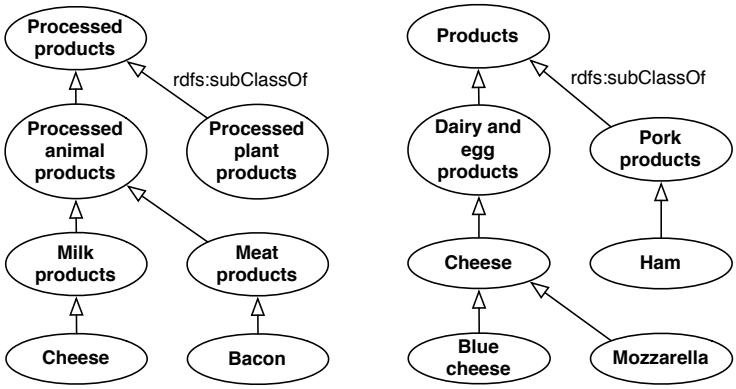


Figure II.1: excerpts from AGROVOC (left) and SR-16 (right).

II.3.1 THESAURI

AGROVOC This is a multi-lingual thesaurus made by the Food and Agriculture Organization of the United Nations (FAO). It consists of roughly 17,000 concepts and three types of relations derived from the ISO thesaurus standard: USE (preferred term), RT (related term) and BT (broader term). We use a RDFS version of this thesaurus where the broader term relation is represented with the RDFS subClassOf relation. The maximum depth of AGROVOC's subclass hierarchy is eight. Figure II.1 shows an excerpt from AGROVOC. The text boxes are classes with their names and the arrows stand for subclass relations.

SR-16 This is the Nutrient Database for Standard Reference version 16 (SR-16) made by the United States Department of Agriculture (USDA), converted to RDFS and OWL by the AID group. It consists of roughly 6,500 concepts and one relation, RDFS subClassOf. The maximum depth of the subclass hierarchy of SR-16 is four. Figure II.1 shows an excerpt from SR-16.

II.3.2 AUXILIARY KNOWLEDGE SOURCES

We used one general and one domain-specific source. The general source is Google and the domain-specific source is the CooksRecipes.com's Cooking Dictionary.

GOOGLE Google[6] is an open domain search engine. At the moment (mid 2005) Google indexes more than 8 billion pages. The large size of Google allows makes it possible to use it for statistical comparison of words. Google has a programming interface called the Google API, that at the moment allows researchers to pose 1,000 queries per day.

[6] http://www.google.com

CooksRecipes.com's Cooking Dictionary The CooksRecipes.com Cooking Dictionary provides definitions for ingredients, culinary terms and cooking techniques. It contains 1,076 definitions. An example entry is: "**Basmati** an aged, aromatic long-grain rice grown in the Himalayan foothills; has a creamy yellow color, distinctive sweet, nutty aroma and delicate flavor…"

II.3.3 Evaluation Method

In order to do full evaluation of the quality of a mapping between AGROVOC and SR-16 one would have to assess all possible subclass relations between a thesaurus of roughly 17,000 and one of around 6,500 classes. This sums up to something of the order of hundreds of millions of possible mapping relations. With smart pruning of the possible mapping this still would have left us with more work than time allowed. Therefor we took samples from both thesauri on a common topic. From SR-16 we took one set of concepts about meats, containing the parts about beef, pork and poultry (chicken, turkey bacon, ham, etc.). From AGROVOC we took two sets of concepts, one containing the part about animal products (minced meat, cheese, leather, etc.), and one containing the part about food categories (processed foods, instant foods, snack foods, etc.).

For the experiments with Google we created a gold standard mapping by hand from the set of SR-16 concepts to both sets of AGROVOC concepts. The size of the mapping from meats to animal products is 31 relations out of 3,696 possible relations. The size of the mapping from meats to food categories is 32 relations out of 792 possible relations.

The experiments with the CooksRecipes.com Dictionary yielded few results, distributed evenly over the thesauri, which made it hard to choose a subset of the thesaurus that contained a reasonable number of mapping relations. Therefor, we evaluated only the returned results. This means we are unable to say anything about Recall of the techniques using the CooksRecipes.com Dictionary.

The measures we used are Precision, Recall and F-Measure as used throughout the literature.[7] The F-Measure we use gives Precision and Recall an equal weight.

Protocol The protocol we used can be summarized as follows: All concepts are to be interpreted in their original context. For instance, in AGROVOC chicken is a subclass of product, which means none of the individuals of the chicken class are live chickens. Taking this into account chicken is not a subclass of frozen foods, because some chicken products are never frozen, but chicken is a subclass of poultry, because all chicken products qualify as poulty.

II.4 Experiments

II.4.1 Hearst patterns and Google hits

The mapping technique described in this section is approximately the same as Cimiano and Staab's *Learning by Googling* method. It derives relations from Google hit counts on certain

[7] http://en.wikipedia.org/wiki/Information_Retrieval

queries.

METHOD

1. **Create hypothetical relations between pairs of concepts from both thesauri.** For this experiment we chose to investigate all possible relations from any of the concepts in the predefined set of SR-16 concepts to any of the concepts in both of the predefined sets of AGROVOC concepts (see section II.3.3).

2. **Construct Google queries containing Hearst patterns for each pair of concepts.** We chose to use the same Hearst patterns as Cimiano and Staab (Cimiano and Staab, 2004) except the apposition and copula patters, to reduce the number of Google queries, because these patterns did not yield enough results to be useful. The patterns are listed in the table II.1.

	$concept_1$	such as	$concept_2$
such	$concept_1$	as	$concept_2$
	$concept_1$	including	$concept_2$
	$concept_1$	especially	$concept_2$
	$concept_1$	and other	$concept_2$
	$concept_1$	or other	$concept_2$

Table II.1: Hearst patterns used in this chapter.

   Since we are only interested in the combined result of all the patterns we can further reduce the number of queries by putting the patterns in a disjunction. We chose the disjunction to be as long as possible given the limit Google imposes on the number of terms in a query (which was 32 at the time).

3. **Send the queries to the Google API.**

4. **Collect the hit counts for all Heart patterns that give evidence for the existence of a relation.** For instance, add the hits on the queries "milk products such as cheese", "milk products including cheese", etc. Since all these hits give a bit of evidence that cheese is a subclass of milk products.

5. **Accept all hypothetical relations that get more hits than a certain threshold value.** Reject all others.

RESULTS The average number of hits for the mapping to food categories is about 2.5 and to animal products it is about 1.3. Only about 2.5% of the patterns had one or more hits. The maximum number of hits we found was in the order of 1,000, while Cimiano and Staab find hit counts in the order of 100,000. We suspect that this is the case because people rarely discuss the ontological aspects of food, because it is assumed to be common knowledge–everybody knows beef is a kind of meat–and hence can be left out. Since the total number of hits is so low we chose not to use a threshold, but to accept all relations that had one or more hits instead. Precision and Recall are shown in table II.2.

	Precision	Recall	F-Measure
to animal products	0.17 (10/58)	0.32 (10/31)	0.22
to food categories	0.30 (17/56)	0.53 (17/32)	0.38

Table II.2: Results of the Google hits experiment.

Discussion The performance of the PANKOW system of Cimiano and Staab on geographical data is 0.40 Precision at around 0.20 Recall for instance classification and 0.22 Precision at 0.16 Recall for subclass extraction.

Overall Recall seems to be less of a problem in the food domain than in the geographical domain. The decent Recall values can be explained by the large size of the current Google corpus. On simple matters it is quite exhaustive. Even though the total hit counts in the food domain are lower than in the geographical domain it seems that a greater percentage of the relations is mentioned in Google. Apparently not all LonelyPlanet destinations have been discovered by the general web public. If you are interested in really high Recall in the field of geography you can simply look up your relations in the Getty Thesaurus of Geographic Names (TGN)[8].

Precision of the mapping to animal products seems to be comparable to the subclass learning task Cimiano and Staab set for themselves. The overall low Precision can be explained by the fact that when you use Google as a source of mappings between two thesauri you turn it from one into two mapping problems: from the thesaurus to Google; and then from Google to the other thesaurus. That means you have to bridge a vocabulary gap twice and hence introduce errors twice.

Precision of mapping to food categories using Google hits seems to be comparable to that of instance classification. Mapping to animal products, *i.e.* mapping between concepts of similar specificity, appears to be more difficult.

II.4.2 Hearst patterns and Google Snippets

The second mapping technique is a modification of the previous technique. Instead of deriving relations from Google hit counts we analyze the snippets presented by Google that summarize the returned documents. We try to improve performance by shallow parsing the context of the occurrence of the Hearst pattern and remove false hits.

Method

1. **Follow step 1 through 3 from the *Hearst patterns and Google hits* method.**
2. **Collect all the snippets Google returns.** Snippets are the short exerpts from the web pages that show a bit of the context of the query terms.
3. **Extract the patterns.** To accomplish this we part-of-speech tag the snippets with TreeTagger and recognize sequences of adjectives and nouns as concept names. Then we try find all Hearst patterns over the concept names in the snippets.

[8] http://www.getty.edu/research/conducting_research/vocabularies/tgn

4. **Discard all patterns that contain concept names that do not exactly match the original concept names.** For instance, if the original pattern looked like "soup such as chicken", discard the matches on "soup such as chicken soup", because these give false evidence for the relation chicken subClassOf soup. We ignore prefixes to the concept names from the following list: 'other', 'various', 'varied', 'quality', 'high quality', 'fine', 'some', and 'many'. This unifies concept names such as 'meat products' and 'high quality meat products'.

5. **Count every remaining occurrence of the pattern as evidence that the relation holds.**

6. **Follow step 4 and 5 from the *Hearst patterns and Google hits* method.**

RESULTS Analysis of the snippets improves Precision while sacrificing Recall. Overall performance indicated by the F-Measure does not chance much. Shallow parsing the snippets removed many false hits. For instance, "salads such as chicken salad" does not lead to chicken subClassOf salad anymore. The exact Precision and Recall are shown in table II.3.

	Precision	Recall	F-Measure
to animal products	0.38 (7/18)	0.22 (7/31)	0.27
to food categories	0.50 (12/24)	0.37 (12/32)	0.42

Table II.3: Results of the Google snippets experiment.

DISCUSSION Even the Precision achieved with mapping to concepts of similar specificity (to animal products) is comparable to the level PANKOW achieves for instance classification. The mapping to food categories, which is closer to the instance classification task, now achieves a higher Precision and Recall than PANKOW.

As Cimiano and Staab noted downloading the whole documents for analysis could further improve the results. This might even improve Recall a bit if these documents contain more good Hearst patterns than those that caused them to appear in Google's result set.

II.4.3 EXTRACTION FROM A DICTIONARY

With the third mapping technique we try to exploit the implicit editor's guidelines of a dictionary to achieve an even higher grade of Precision than the Google Snippets technique described in the previous section. As an example we took a dictionary that includes terms from both thesauri, the CooksRecipes.com Cooking Dictionary. This dictionary is relatively small compared to the thesauri, but it covers about the same field as SR-16.

METHOD

Find regularities in the dictionary that highly correlate with subclass relations. We found that the editor of the dictionary often starts a definition with the superclass of the described concept. The following steps are tailored to exploit this regularity.

1. **Select all entries that describe a concept that literally matches a concept from AGROVOC or SR-16.**
2. **Parse the entry with MINIPAR.**
3. **Extract the first head from the parse tree.** For instance, the entry of the concept basmati starts with "an aged, aromatic long-grain rice grown in …" The first head in this sentence is 'rice'.
4. **Check if the first head corresponds to a concept in the other thesaurus** If basmati is a concept from AGROVOC, try to find the concept rice in SR-16 and vice versa.
5. **Construct a subclass relation between the concept matching the entry name and the one matching the first head.**

Results More than half of all the returned relations, even those failing the check in step 4, are correct subclass relations according to our strict evaluation protocol. As expected, given the relatively wide scope of the dictionary, step 4 eliminates most of the results. However the mapping relations that are left are of high quality. The exact results are shown in table II.4.

	Precision
relations not forming a mapping	0.53 (477/905)
mapping entire AGROVOC–SR-16	0.75 (16/21)

Table II.4: Results of the dictionary extraction experiment.

Discussion We exploited a regularity in the syntax of the data. This yields high Precision results. Clearly, Recall of this method is dependent on the size of the dictionary and the overlap between the dictionary and the thesauri.

We noticed that most of the errors could have been filtered out by looking for evidence on Google. For instance, the entry: "**leek** a member of the lily family *(Allium porrum)*; …" would cause our technique to suggest the relation leek subClassOf member. One query could have removed this false relation from the result list, because "member such as leek" gives no hits on Google.

II.4.4 Combination of Google hits & Dictionary extraction

The fourth technique is an improvement to the dictionary extraction technique. We use the Google hits technique to filter false relations out of the list of results provided by extraction.

Method

1. **Follow all the steps of the Dictionary Extraction method.** This yields a list of relations.
2. **For each extracted relation follow step 2–5 from the Google hits method.** This filters out all relations for which no evidence can be found on Google using Hearst patterns.

RESULTS Applying the Google hits technique as a sanity check on the extraction results greatly reduces the number of relations. Precision of this smaller result set is higher than with both the Google hits and dictionary extraction technique. Around 63% of the correct results were removed versus 92% of the incorrect results. The results are shown in table II.5.

	Precision
relations not forming a mapping	0.53 (477/905)
after Google hits sanity check	0.84 (178/210)
mapping entire AGROVOC to SR-16	0.75 (16/21)
after Google hits sanity check	0.94 (15/16)

Table II.5: Results of combining dictionary extraction and Google hits.

DISCUSSION The combination of Google hits and a dictionary gave the best Precision of the four techniques. Most of the mismatches caused by definitions that did not exactly fit the regularity that we exploited with the dictionary extraction technique were removed by applying the Google hits technique. On the other hand, a substantial portion of the correct results was also removed.

We noticed that most of the incorrect relations that were not removed are easily recognizable by hand. If the superclass is not directly food related the relation is usually false. For instance, mayonnaise subClassOf cold. Most relations to latin names of plants were inverted. For instance, rosemary subClassOf rosmarinus officinalis . There is another member of the rosemary family, 'Rosmarinus eriocalix', so rosmarinus officinalis should be a subclass.

II.5 METHOD PROPOSAL

As we discussed in section II.3.3 simply checking all possible relations between two ontologies is task of quadratic complexity. In theoretical computer science this might qualify as a polynomial with a low degree, but for a mapping technique that uses the Google API (which only allows 1,000 queries per account per day) this means it does not scale well. Furthermore, assessing a quadratic number of relations by hand is often not feasible. Therefor we propose to combine high Precision techniques and techniques that achieve a high Recall per human assessment. The method we propose is as follows:

1. **Find a small set of high Precision mapping relation as starting points, preferably distributed evenly over the ontologies.** This could be done with the last two techniques we described or with tools such as PROMPT[9]. Which technique works best depends largely on the naming conventions used in the ontologies.

2. **Manually remove all the incorrect relations.** Assessing the results of the dictionary extraction technique took about one man hour.

[9] http://protege.stanford.edu/plugins/prompt/prompt.html

3. **For each correct relation select the concepts surrounding the subject and object concepts.** For instance, if the SR-16 concept cheese (see figure II.1) was correctly mapped as a subclass of the AGROVOC concept Milk products, one would select a subclass tree from SR-16 that contains cheese and a subclass tree from AGROVOC that contains Milk products. This can be accomplished in the following two steps:

   (a) **Travel up the subclass hierarchy from the starting point.** Go as far as possible as long as it is still clear what is subsumed by the examined concept, without having to examine the subtrees of the sibling concepts. A suitable top concept from SR-16 could be Dairy and egg products because it is immediate clear to us what is subsumed by this concept without having to look at the Pork products concepts. A suitable top concept from AGROVOC could be Processed animal products.

   (b) **Select all subclasses of the two top concepts.** Collect the concepts as two sets.

   This could be done using tools such as Triple20[10] or Sesame[11].

4. **Find relations between the two sets of concepts returned in the previous step.** This could be done with the Google snippets technique.

5. **Manually remove all incorrect relations.** The evaluation of the mapping between the AGROVOC animal product concepts and the SR-16 meat concepts took us four man hours. Assessing all the mappings returned by the previous steps could take days. The higher the applied mapping techniques' Precision, the less time this step takes.

6. **Manually add all omissions.** Creating a list of omissions during the assessments of the previous step reduces the amount of work in this step. The higher the applied mapping techniques' Recall, the less time this step takes.

This method reduces the search space by eliminating cross-references between concepts in unrelated parts of the ontologies. For instance, possible relations between concepts in the part of AGROVOC about legumes and in the part of SR-16 about poultry would be ignored if step 1 did not yield any relations between those parts. Hence the number of queries we have to send to Google is reduced along with the number of necessary manual assessments low.

II.6 Discussion

We discussed four ontology mapping techniques and evaluated their performance. There is a clear trade-off between Precision and Recall. The more assumptions we make the higher Precision gets and the lower Recall. We showed that exploiting syntactic information by using a part-of-speech tagger can improve Precision of ontology-mapping methods based on Google hits such as our Google hits method and possibly PANKOW.

[10] http://www.swi-prolog.org/packages/Triple20
[11] http://www.openrdf.org

We showed that in our experiments finding subclass relations to generic concepts such as food categories is easier than mapping concepts that are roughly equal in specificity. We hypothesize that this is because the former discriminate more clearly between different interpretations of concepts and are therefor used more often. For instance, the phrase "chickens such as roosters" is less discriminating about the meaning of the word 'rooster' than "poultry such as roosters" or "birds such as roosters".

Furthermore, we introduced a method that extends the PANKOW two-step method by Cimiano and Staab to decrease the number of necessary Google queries and the amount of manual work.

II.7 ACKNOWLEDGEMENTS

This chapter has benefitted from input from the other AID group members: Pieter Adriaans, Jan van Eijck, Leonie IJzereef, Machiel Jansen, Maarten de Rijke. Furthermore we want to thank Marco Roos and Scott Marshall from the Micro Array Department of the University of Amsterdam, Michel Klein at the Computer Science department of the Free University Amsterdam for valuable discussions, Victor de Boer who organized the Ontology Learning and Population Workshop at the Human-Computer Studies Laboratory of the University of Amsterdam and and everybody who attended. Last we want to thank Thijs de Graaf, Wessel Kraaij and Dolf Trieschnigg at the Signal Processing group at TNO Science & Industry.

Chapter III

Finding Part-whole Relations

In this chapter we study the task of learning part-whole relations from text for the purpose of aligning ontologies. For this we use the first method described in chapter II. However, we do not use predefined patterns (as in chapter II) to learn relation instances, but we learn the patterns based on a set of seed relation instances and a web search engine. We evaluate this techniques by aligning a controlled vocabulary of known carcinogens to the AGROVOC and NALT thesauri. We measure Precision for each learnt pattern on a sample of produced relation instances, following the alignment sample evaluation *method described in chapter VI. To measure Recall we use a form of application-centered evaluation which used as an input for the generic* end-to-end evaluation *method described in chapter VI. This evaluation was centered around the task to reproduce lists of known media by which humans are exposed to carcinogens.*

This chapter is based on a paper coauthored by Hap Kolb and Guus Schreiber, "A Method for Learning Part-Whole Relations, Willem Robert van Hage, Hap Kolb, Guus Schreiber" *(van Hage et al., 2006), which was presented at the fifth International Semantic Web Conference (ISWC 2006).*

Abstract Part-whole relations are important in many domains, but typically receive less attention than subsumption relation. In this chapter we describe a method for finding part-whole relations. The method consists of two steps: (i) finding phrase patterns for both explicit and implicit part-whole relations, and (ii) applying these patterns to find part-whole relation instances. We show results of applying this method to a domain of finding sources of carcinogens.

III.1 Introduction

A plethora of existing vocabularies, terminologies and thesauri provide key knowledge needed to make the Semantic Web work. However, in using these sources witinin one context, a process of alignment is needed. This has already been identified as a central problem in semantic-web research. Most aligment approaches focus on finding equivalence and or subclass relations between concepts in diffeent sources. The objective of this chapter is to identifying alignment relations of the part-whole type. Part-whole relations play a key role in many application domains. For example, part-whole is a central structuring principle in artefact design (ships, cars), in chemistry (structure of a substance) and medicine (anatomy). The nature of part-whole has been studied in the area of formal ontology (*e.g.*, Artale et al.,

1996). Traditionally, part-whole receives much less attention than the subclass/subsumption relation.

The main objective of this chapter is to develop a method for learning part-whole relations from existing vocabularies and text sources. Our sample domain is concerned with food ingredients. We discuss a method to learn part-whole relations by first learning phrase patterns that connect parts to wholes from a training set of known part-whole pairs using a search engine, and then applying the patterns to find new part-whole relations, again using a search engine. We apply this method in a use case of assisting safety and health researchers in finding sources of carcinogenic substances using Google. We evaluate the performance of the pattern-learning and the relation-learning steps, with special attention to the performance of patterns that implicitly mention part-whole relations. Furthermore we perform an end-to-end task evaluation to establish whether our method accomplishes the task.

In section III.2 we describe the use case on which we evaluate end-to-end performance and pose performance criteria. In section III.3 we discuss the experimental set-up we use to learn part-whole relations. In section III.4 and III.5 we describe the learning and application of patterns to find part-whole relations and evaluate the performance of the patterns in terms of Precision. In section III.6 we evaluate Recall on four sample carcinogens. Section III.7 discusses related work. We conclude with a discussion of the results and open research questions in section III.8.

III.2 USE CASE

An important application area of part-whole learning is health and safety research. Experts in this field are faced with hard information retrieval tasks on a regular bases. News of a benzene spill in a river, for example, will trigger questions like "Is the general public's health in danger?", "Are there any foodstuffs we should avoid?", and "Are there any occupational risks, fishermen perhaps?". The first task the health and safety researchers are faced with is to find out via which pathways the substance in question can reach humans. Only then can they investigate if any of these pathways apply to the current situation. A sizable part of this problem can be reduced to finding all part-whole relations between the substance and initially unknown wholes in scientific literature and reports from authorities in the field such as the United States Food and Drugs Administration[1] (FDA) and Environmental Protection Agency[2] (EPA), and the World Health Organization[3] (WHO).

The wholes should be possible routes through which humans can be exposed to the substance. For example, tap water, exhaust fumes, or fish. We will not go into detail discussing the roles these concepts play that leads to the actual exposure. For example, when humans are exposed to benzene in fish by eating the fish, fish assumes the role of food. Relevant part-whole relations can be of any of the types described by Winston, Chaffin, and Herrmann (Winston et al., 1987).

[1] http://www.fda.gov
[2] http://www.epa.gov
[3] http://www.who.int

component-integral object "Residents might have been exposed to *benzene* in their *drinking water*."

member-collection "*Benzene* belongs in the group of *BTX-aromatics*."

portion-mass "*3 tons* of the *benzene emissions* can be attributed to the dehydrator."

stuff-object "*Aftershave* used to contain *benzene*."

feature-activity "*Benzene* is used in the *dehydration process*." The part in this case is not benzene itself, but the application of benzene, which is abstracted over with the word 'used'.

place-area "*Benzene* was found in the *river*." The part in this case is the location where the benzene was found, which is left anonymous.

The automation of the knowledge discovery task described above is a success if and only if the following criteria are met:

1. The key concepts of each important pathway through with a carcinogen can reach humans should be found. (*i.e.,* Recall should be very high.)
2. The researchers should not be distracted by too many red herrings. (*i.e.,* Precision should be sufficient.)

Precision can be evaluated in a straightforward manner by counting how many of the returned part-whole relations are valid. The evaluation of Recall however poses a greater problem. We are attempting to learn unknown facts. How can one measure which percentage of the unknown facts has been learnt when the facts are unknown? For this use case we will solve this problem by looking at exposure crises for four substances (acrylamide, asbestos, benzene, and dioxins) that have been documented in the past. We know now which pathways led to the exposure in the past. This means we can construct sets of pathways we should have known at the time of these crises and use these sets to evaluate Recall.

III.3 Experimental Set-up

In this chapter we will use two-step method to learn part-whole relations. First we learn lexical patterns from known part-whole pairs, using search engine queries. Then we apply these patterns to a set of parts to find wholes that are related to these parts, also using search engine queries. To constrain the size of the search space we will constrain both the set of parts and the set of wholes to controlled vocabularies. In more detail, the method works as follows:

1. **Learning part-whole patterns.**
   (a) Construct a search query for each part-whole pair in a training set.
   (b) Collect phrases from the search results that contain the part-whole pair.

(c) Abstract over the parts and wholes in the phrases to get patterns.

(d) Sort the patterns by frequency of occurrence. Discard the bottom of the list.

2. **Learning wholes by applying the patterns.**

(a) Fill in each pattern with all parts from a set of part instances, while keeping the wholes free.

(b) Construct search queries for each filled in pattern.

(c) Collect phrases from the search result that contain the filled in pattern.

(d) Extract the part-whole pairs from the phrases.

(e) Constrain the pairs to those with wholes from a controlled vocabulary.

(f) Sort the pairs by frequency of occurrence. Discard the bottom of the list.

In the following two sections we will describe the details of the data sets we used and we will motivate the decisions we made.

III.4 Learning Part-Whole Patterns

In this section we will describe the details of step 1 in our part-whole learning method, described in the previous section. We will describe the training set we used and the details of the application of step 1 on this training set, and analyze the resulting patterns.

Our training set consists of 503 part-whole pairs, derived from a list of various kinds of food additives and food product types they can occur in created by the International Food Information Council[4] (IFIC) and the FDA.[5] The list contains 58 additives (parts) and 113 food products (wholes), grouped together in 18 classes of additives such as sweeteners and preservatives. An example is shown in Fig. III.1. It is not specified which additives occur in which food products. To discover this, we took the cartesian product of the additives and the food products and filtered out the pairs that yielded no hits on Google[6] when put together in a wildcard query. For example, the pair ⟨table-top sugar, aspartame⟩ is filtered out, because the query `"table-top sugar * aspartame" or "aspartame * table-top sugar"` yields no hits.

For all 503 part-whole pairs that did yield results we collected the first 1,000 snippets (or as many snippets as were available). We attempted to part-of-speech tag these snippets. This did not produce good results, because nearly all snippets were incomplete sentences and many were lists of substances. For example, "…Water)*, Xanthan Gum, Brassica Campestris (Rapeseed), Essential Oils [+/- CI 77491,CI …". None of the part-of-speech taggers we tried were able to deal with this. Therefore we used the untagged snippets and looked up all consistent phrases that connected the part and whole from the query. In these phrases we substituted all parts and wholes by the variables "part and whole". This yielded 4,502 unique

[4] http://www.ific.org

[5] http://www.cfsan.fda.gov/~dms/foodic.html

[6] http://www.google.com

Type	Sweeteners
What They Do	Add sweetness with or without the extra calories.
Examples of Uses	Beverages, baked goods, confections, table-top sugar, substitutes, many processed foods.
Product Label Names	Sucrose (sugar), glucose, fructose, sorbitol, mannitol, corn syrup, high fructose corn syrup, saccharin, aspartame, sucralose, acesulfame potassium (acesulfame-K), neotame

Figure III.1: An excerpt from the IFIC and FDA list of food additives.

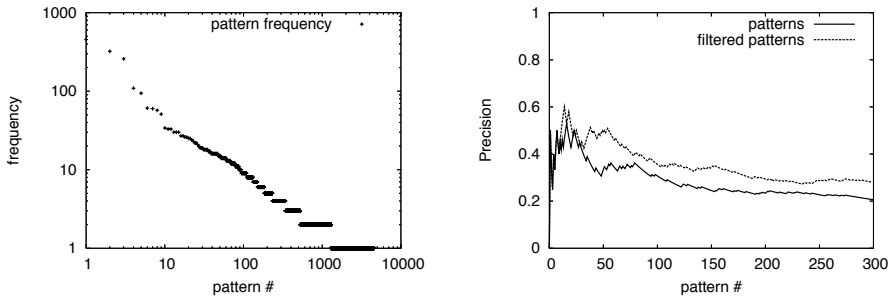


Figure III.2: (left) Frequency distribution in the training set of the learnt patterns. Referred to as T in table III.3. (right) Precision@n (*i.e.*, # correct part of patterns in the top-n / n) graph over the top-300 most frequent patterns, before and after filtering out patterns that contain labels of AGROVOC or NALT concepts.

patterns, which we sorted by frequency of occurrence. The frequencies of the patterns are shown in Fig. III.2.

Due to the fact that there were many lists of substances in our data there were also many patterns that did not describe a part-whole relation, but that were merely part of a list of substances containing the part and the whole. These patterns can be easily recognized, because they contain names of substances. For example, for the pair ⟨cheese, enzymes⟩ the following snippet was returned: "*cheese* (pasteurized milk, cheese cultures, salt, *enzymes*)". An example of a good snippet is: "All *cheese* contains *enzymes*.". To exclude lists we removed all patterns that contain, apart from the part and whole, labels of concepts in agricultural thesauri. The thesauri we used are the NAL Agricultural Thesaurus[7] and the AGROVOC Thesaurus[8]. (We used the SKOS[9] version of these thesauri.) This filtered out 1,491 patterns, of which only 12 were correct part-whole patterns. Fig. III.2 shows a Precision graph of the list of patterns before and after the filtering step.

To restrict the number of Google queries needed to find wholes for parts we decided

[7] http://agclass.nal.usda.gov/agt
[8] http://www.fao.org/agrovoc
[9] http://www.w3.org/2004/02/skos

not to use all of the remaining 3,011 patterns, but to select the most productive patterns. We analyzed the 300 patterns that produce the most results. For each pattern we looked at the snippets it returned. If the majority of the occurrences of the pattern described a proper part-whole relation (*i.e.*, Precision ≥ 0.5) we classified the pattern as part-whole. Otherwise we classified it as not part-whole.

We distinguished the following groups of patterns, based on the most common types of errors that led to the classification of the pattern as not part-whole. A pattern can yield more than one type of false relations, but the classification is based on the most common of the error types.

too specific Too training-set specific to be useful. Either the pattern contains adjectives or it yields no hits due to over-training.

too generic The pattern matches part-whole relations, but also too many non-part-whole relations to be useful. For example, the pattern "whole part", as in 'barn door', can match any type of collocation.

is a The pattern primarily matches hyponyms. The language used to describe member/collection relations is also used for hyponyms.

conjunction/disjunction The pattern primarily matches conjunctions / disjunctions.

related The pattern connects terms that are related, but not part-whole related.

wrong Not a proper pattern for any other reason. Most of the errors in the wrong category can be attributed to the lack of sophisticated linguistic analysis of the phrases.

Table III.2 shows the build-up of the different error types.

"part to whole"	→	"add part to whole", "added part to whole"
"part to the whole"	→	"add part to the whole", "added part to the whole"
"part gives the whole"	→	"part gives the whole its"
"part containing whole"	→	"part-containing whole"
"part reduced whole"	→	"part-reduced whole"
"part increased whole"	→	"part-increased whole"

Table III.1: Manually corrected patterns.

We corrected 6 patterns that were classified as not part-whole, and added them to the part-whole patterns. These patterns are not counted in table III.2. They are listed in table III.1. Notice that in the English grammar, hyphenation turns a part-whole relation into its inverse. For example, 'sugar-containing cake' and 'cake containing sugar'.

While analyzing the correct part-whole patterns we noticed that the phrases that deal with part-whole relations do not always explicitly state that relation. Often, the part-whole relation has to be inferred from the description of a process that led to the inclusion of the part in the whole or the extraction of the part from the whole. For example, from the

pattern class	example pattern	# patterns in class
part-whole		83
part of	whole containing part	40
made with	part added to whole	36
source of	part found in whole	7
not part-whole		217
wrong	part these whole, part organic whole	186
too specific	part in commercial whole	10
too generic	part of whole	7
is a	whole such as part	5
related	part as well as whole	4
conjunction	part and whole, whole and part	3
disjunction	part or whole, whole or part	2

Table III.2: Analysis of the top-300 most frequently occurring patterns.

sentence "I add *honey* to my *tea*." we can infer that honey is part of the tea, even though the sentence only mentions the process of adding it. In addition to explicit descriptions of part-whole relations we distinguish two types of phrases that mention part-whole relations implicitly.

part of The phrase explicitly describes a part-whole relation. For example, "There's alcohol in beer.".

source of The phrase implicitly describes a part-whole relation by describing the action of acquiring the part from the whole. For example, "Go get some *water* from the *well*.".

made with The phrase implicitly describes a part-whole relation by describing a (construction) process that leads to a part-whole relation. For example, "I add *honey* to my *tea*".

Table III.2 shows that together, the implicit patterns account for a third of the total number of part-whole pairs.

When applying patterns to learn part-whole relations it is useful to make this distinction into three types, because it turns out that these three types have rather different Precision and Recall properties, listed in table III.3. The patterns in the part of class yield the most results with high Precision. The patterns in the made with class also yield many results, but—somewhat surprisingly—with much lower Precision, while the patterns in the source of class yield few results, but with high Precision.

The 91 patterns we used for the discovery for wholes are the 83 classified as part-whole in table III.2 and the 8 listed in table III.1 on the right side. They are listed in table III.6.

III.5 Finding Wholes

In this section we will describe the details of step 2 in our part-whole learning method, described in the previous section. We will describe the sets of part and whole instances we used, and analyze the resulting part-whole relations.

In the use case we focus on finding wholes that contain a specific substance. Initially, any concept name is a valid candidate for a whole. We tackle this problem by first reducing the set of valid wholes to those that occur in a phrase that matches one of the patterns learnt in step 1 of our method. This corresponds to step 2C and 2D of our method. Then we prune this set of potential wholes using two large, agricultural, and environmental thesauri that are geared to indexing documents relevant to our use case. We remove all wholes that do not match a concept label in either thesaurus. This corresponds to step 2E of our method. The former reduction step asserts that there is a part-whole relation. The latter that the whole is on topic.

We select the possible part instances from a list of carcinogens provided by the International Agency for Research on Cancer[10] (IARC). In the IARC Monographs on the Evaluation of Carcinogenic Risks to Humans[11] carcinogenic agents, mixtures and exposures are classified into four groups: positively carcinogenic to humans, probably or possibly carcinogenic to humans, not classifiable as carcinogenic to humans, and probably not carcinogenic to humans. We took the agents and mixtures from the group of positively carcinogenic factors. We interpreted each line in the list as a description of a concept. We removed the references and expanded the conjunctions, interpreting each conjunct as a label of the concept. *i.e.*, For example, we transform the list entry "Arsenic [7440-38-2] and arsenic compounds (Vol. 23, Suppl. 7;1987)" into a concept arsenic with the labels 'Arsenic' and 'arsenic compounds'. The resulting list contains 73 concepts, with 109 labels in total. We applied the 91 patterns that resulted from the process described section III.4 on these 109 labels to discover wholes. We allow for words—generally articles and adjectives—to appear in between the whole and the rest of the pattern. For example, the pattern "part in whole" can be interpreted as "part in $*$ whole", and hence will match "part in deep-sea whole" and "part in the whole". This also means there can be overlap between the sets of part-whole pairs retrieved by patterns. From the resulting filled-in patterns we extracted the wholes. We filtered out all wholes from this list that do not appear in the UN FAO AGROVOC Thesaurus and the USDA NAL Agricultural Thesaurus. When put together, these thesauri contain 69,746 concepts with 87,357 labels in total. Thus limiting the set of discoverable wholes to 69,746 concepts. For each remaining whole in the list we construct a part-whole relation.

An assessment of the part-whole results is shown in table III.6. We approximated Precision for the 91 patterns we used to find wholes based on a random sample of 25 discovered pairs. The results are shown under 'Precision'. The number of hits per pattern are listed under *D*. This number includes duplicate phrases and multiple phrases describing the same part-whole pair. Table III.4 in section III.6 shows how many unique wholes are found for four example parts.

III.6 Analysis

In section III.2 we stated two criteria that have to be met for the application of our part-whole learning method to be a success. Precision has to be sufficient, and Recall has to be very

[10] http://www.iarc.fr
[11] http://monographs.iarc.fr/ENG/Classification

pattern class	# patterns in class	T	D	avg. Precision
part of	40	744	84,852	0.81
made with	36	525	33,408	0.69
source of	7	111	8,497	0.83

Table III.3: Average pattern performance per pattern class. T is the number of times patterns in the class occur in the training set. D is the number of discovered part-whole phrases.

concept (part)	# of wholes found	Recall
acrylamide	350	13/15 (.86)
asbestos	402	11/15 (.73)
benzene	479	13/15 (.86)
dioxins	439	12/15 (.80)

Table III.4: Recall on four sample substances.

high. In section III.4 and III.5 we analyzed the results in terms of frequency and Precision. We achieved an average Precision of 0.74. In this section we will assess Recall.

Since even the knowledge of experts of whether or not a substance is contained in some whole is far from complete we can not create a complete gold standard to measure Recall. It is simply infeasible. We can, however, approximate Recall by computing it on samples.

We set up four test cases centered towards discovering possible causes of exposure to a specific carcinogenic agent. The agents we chose are acrylamide, asbestos, benzene, and dioxins. These substances have all caused health safety crises in the past and possible exposure to them has been extensively documented. For each case we decided on 15 important concepts that contain the carcinogen and define a possible exposure route. For example, you can be exposed to acrylamide by eating fried food such as french fries, because acrylamide can be formed in the frying process. The selection of the wholes was based on reports from the United States Environmental Protection Agency (EPA) and the Netherlands Organization for Applied Scientific Research (TNO) Quality of Life. The cases were set up without knowledge of the data set and the learning system, to minimize the hindsight bias, but with knowledge of the concepts in the AGROVOC and NALT thesauri. The sets of wholes are shown in table III.5, along with the rank at which the whole occurs in the list of discovered wholes. Recall and the total number of discovered wholes are shown in table III.4.

For all of the cases we found a large majority of the important concepts. For half of the missed concepts we found concepts that are very closely related. For example, we did not find the concept 'cement pipes', but we did find 'cement' and 'pipes', and we did not find 'air', but we did find 'air pollution' and 'atmosphere'.

The data sets and the results can be found at the following web location: `http://www.few.vu.nl/~wrvhage/carcinogens`.

Acrylamide

concept (whole)	rank
coffee	18
fried food	22
plastics industry	39
smoke	42
drinking water	43
olives	103
paper	109
dyes	114
soil	144
fish	158
herbicide	181
water treatment	195
textiles	275
air	not found
baked food	not found

Benzene

concept (whole)	rank
leaded gasoline	1
water	4
solvents	9
smoke	10
dyes	32
pesticides	68
soil	69
detergents	76
cola	84[12]
rubber	161
bottled water	191
rivers	228
lubricants	340
air	not found[13]
fats	not found

[12]soft drinks appear at rank 5
[13]found air pollution and atmosphere

Asbestos

concept (whole)	rank
insulation	5
vermiculite	9
roofing	12
building materials	16
flooring	23
rocks	37
water	47
brakes	67
adhesives	127
cars	160
mucus	211
cement pipes	not found[14]
sewage	not found[15]
air	not found
feces	not found

[14]found cement and pipes
[15]found refuse and wastewater

Dioxins

concept (whole)	rank
fish	2[16]
paper	3
soil	7
herbicides	8
defoliants	17[17]
water	32
smoke	38
bleach	39
chickens	75
animal fat	106
animal feed	138
waste incineration	142
pigs	not found[18]
air	not found[19]
diesel trucks	not found[20]

[16]also found fishermen
[17]also found vietnam
[18]found cattle and livestock
[19]found air quality
[20]found exhaust gases

Table III.5: Recall bases for four sample substances.

Prec.	*D*	pattern	Prec.	*D*	pattern
0.84	26,799	part in whole	0.76	980	part content in the whole
0.68	8,787	whole with part	0.96	745	part-treated whole
0.84	5,266	part in the whole	0.84	786	part derived from whole
0.96	4,249	part from whole	0.76	852	whole rich in part
0.68	5,917	part for whole	0.28	2,306	whole high part
0.60	5,794	part content whole	0.88	617	part-containing whole
0.88	3,949	whole contain part	0.20	2,571	whole add part
1.0	2,934	whole containing part	0.72	700	part in most whole
0.64	4,415	part based whole	0.80	623	part for use in whole
0.72	3,558	whole using part	0.40	1,169	part to make whole
0.92	2,591	part levels in whole	0.72	630	add part to the whole
1.0	2,336	part-laden whole	0.72	580	part enriched whole
0.84	2,327	part content in whole	0.56	703	part in many whole
1.0	1,945	whole contains part	0.96	404	part-enriched whole
0.76	2,536	whole have part	0.72	527	part contents in whole
0.72	2,622	part into whole	0.52	608	added part to whole
0.88	2,035	part is used in whole	0.92	314	part occurs naturally in whole
1.0	1,760	part found in whole	0.84	288	part extracted from whole
0.52	3,217	part free whole	0.96	226	whole enriched with part
1.0	1,672	part is found in whole	0.68	310	part to our whole
0.88	1,834	part-rich whole	0.16	1,160	whole provide part
0.80	1,994	part used in whole	0.68	247	added part to the whole
0.92	1,680	part content of whole	0.72	220	whole with added part
0.20	7,711	whole for part	0.96	137	part found in many whole
0.96	1,497	part is present in whole	1.0	124	whole containing high part
0.84	1,600	add part to whole	0.76	134	part replacement in whole
0.88	1,496	part added to whole	0.60	133	part for making whole
0.80	1,597	part in their whole	0.88	64	whole fortified with part
0.92	1,372	part-based whole	0.76	74	whole have part added
0.88	1,421	part in these whole	0.96	54	part-fortified whole
1.0	1,218	whole that contain part	0.36	120	part compound for whole
1.0	1,203	part levels in the whole	0.36	120	part fortified whole
0.84	1,361	part in all whole	1.0	24	whole sweetened with part
1.0	1,112	part contained in whole	0.16	89	whole preserves part
0.76	1,455	part in some whole	0.91	11	part-reduced whole
0.84	1,301	part in your whole	0.90	10	part gives the whole its
1.0	1,058	part present in whole	0.04	85	part sweetened whole
0.76	1,350	part in our whole	0.27	11	part-increased whole
1.0	985	part laden whole	0.67	3	part-added whole
0.32	3,052	whole use part	1.0	1	part-sweetened whole
0.52	1,648	whole mit part	1.0	1	part to sweeten their whole
0.84	930	whole made with part	1.0	1	part fortification of whole
0.88	885	part-free whole	0.0	0	part additions in various whole
0.52	1,477	part is in whole	0.0	0	part used in making whole
0.80	945	part is added to whole	0.0	242	part hydrogenated whole
0.92	811	whole high in part			

Table III.6: The 91 patterns used for the learning of wholes, ordered by the number of correct pairs it yielded. Prec. is Precision approximated on a sample of 25 occurrences (or less if freq. < 25). *D* is the number of discovered part-whole phrases.

III.7 Related Work

The method of automatic learning of relations by first learning patterns and then applying these patterns on a large corpus is widely used. An example in the domain of business mergers and production is described in the 1999 article by Finkelstein-Landau and Morin (1999). Their work on extracting companies-product relations touches lightly upon the subject of this chapter. Another example of pattern-based relation learning on the web is the KnowItAll system of Etzioni et al. (2004). The learning of part-whole relations however is quite rare. Two examples, are Berland and Charniak (1999) and Girju et al. (2003).

Berland and Charniak learn part-whole patterns from a part-of-speech tagged corpus, the Linguistic Data Consortium's (LDC) North American News Corpus (NANC). To illustrate the pattern learning phase they mention five example patterns. "whole's part", "part of {the|a} whole", "part in {the|a} whole", "parts of wholes", and "parts in wholes". The domain they used for evaluation is component/integral object relations between artifacts such as cars and windshields. Even though our domain is quite different, we found all five of their example patterns using our training data, respectively at rank 294, 290, 12, 128, and 2 (of 4,502 learnt patterns).

Girju, Badulescu, and Moldovan, used the SemCor 1.7 corpus and the LA Times corpus from the Ninth Text Retrieval Conference (TREC-9). They used the meronyms from WordNet (Miller, 1995), mainly component/integral object and member/collection relations. Girju, Badulescu, and Moldovan also make the distinction between explicit and implicit part-whole constructions, but the implicit constructions they focus on are mainly possessive forms like 'the girl's mouth', 'eyes of the baby', 'oxygen-rich water', and 'high heel shoes'. They list the three most frequent patterns, which also contain part-of-speech tags. "part of whole", "whole's part", and "part *Verb* whole". We found the first two patterns, as mentioned above, and many instances of the third pattern, such as "part fortified whole" at rank 4.

Other applications of part-whole relations than discovering sources of substances are query expansion for image retrieval (Hollink, 2006, Ch. 6), and geographical retrieval (Buscaldi et al., 2005).

III.8 Discussion

Our experimental setup assumes that all interesting information pertaining to some carcinogenic substance can be obtained in one single retrieval step. The construction of complex paths from the substance to the eventual exposure has to happen in the mind of the user—and depends solely on his expertise and ingenuity. This is a severe limitation that leaves room for considerable improvement. A relatively straightforward extension would be to iterate the retrieval step using suitable wholes found in retrieval step $n-1$ in the part slot in retrieval step n. Separation of roles, classes, etc. amongst the wholes by means of classification (*cf., e.g.,* Guarino and Welty, 2004) might be necessary to limit the inevitable loss of precision. For example, if step $n-1$ yielded that there is benzene in some fish, then proceeding to investigate in step n whether these fish are part of people's diet. If, however, step $n-1$ yielded that benzene is part of a group of carbon-based chemicals, then proceeding to investigate these chemicals might lead to excessive topic drift.

The usefulness of such an extension depends to a large extent on the validity of some sort of transitive reasoning over the paths. Yet, the transitivity characteristics of part-whole expressions are notoriously quirky. Existing accounts actually either take the classical route set out by Stanislaw Lesniewski in the 1920's, defining the relations in question axiomatically and with little consideration for actual usage, or they formulate reasoning patterns for specific application domains and expressions (*cf., e.g.,* Schulz and Hahn, 2005). Neither approach is applicable to the mixed bags of 'interesting' token relations our setup derives from natural language usage. A rare attempt to ground reasoning patterns in the general usage of part-whole expressions is contained in Winston et al. (1987). Even though our layout is orthogonal (and not even coextensive) to their influential classification of part-whole relations, their basic intuition w.r.t. transitivity does carry over to our case. In short:

1. The part-whole relations, P, expressed in natural language form a partial order $\mathcal{P} = \langle P, \geq \rangle$;

2. The weakest link determines the interpretation of a chain of part-whole pairs w.r.t. transitivity;

3. Transitivity fails if the chain contains uncomparable relation instances (w.r.t. $\geq$).

Contrary to Winston et al. (1987) we assume that there is some weakest mereological relation, *i.e.*, the poset $\mathcal{P}$ has a minimum element. (2) can then be generalized as follows:

2'. Any element of $\mathcal{P}$ which is compatible with (*i.e.*, as least as weak as) every relation used to form a chain of part-whole pairs determines a transitive interpretation of that chain.

This means that for every chain of part-whole pairs there is a meaningful, albeit sometimes rather weak, transitive interpretation available. It depends solely on the intended utilization whether the information obtained in this way is specific enough to be useful. What has its merits in a task with a strong element of exploration and novelty detection like our use case, may well be a show stopper for tasks such as diagnosis in a process control environment. Refinements, especially concerning the classification of relation types and the properties of the poset of relations are necessary to extend the general applicability of this approach.

This is especially true when our work is placed in the more general context of vocabulary and ontology alignment. Most ontology-alignment systems aim at finding equivalence relations. Yet, many real-world alignment cases have to deal with vocabularies that have a different level of aggregation. (*cf.*, van Hage et al., 2005) In such cases equivalent concepts are quite rare, while aggregation relations, such as broader/narrower term, subclass and part-whole, are common. The carcinogen-source discovery case can be seen as an ontology-alignment problem where the alignment relation is the part-whole relation and the vocabularies are the controlled vocabulary of IARC group 1 carcinogens, and the AGROVOC and NALT thesauri. Under this perspective our work describes a first step towards a novel approach to ontology alignment. The influence part-whole alignment relations have on the consistency of the resulting aligned ontologies is unknown.

Acknowledgements

Margherita Sini and Johannes Keizer (FAO), Lori Finch (NAL), Fred van de Brug (TNO), Dean Allemang (BU), Alistair Miles (CCLRC) and Dan Brickley (W3C), the IARC, EPA, IFIC, and FDA, Vera Hollink (UvA), Sophia Katrenko (UvA), Mark van Assem (VUA), Laura Hollink (VUA), Véronique Malaisé (VUA). This work is part of the Virtual Lab e-Science project[21].

[21] http://www.vl-e.org

Chapter iv

Finding Relations in Generic-Domain Text

In this chapter we study the classification of relation instances as true or false for seven different semantic relations. The relation-learning method discussed in chapter ii and iii can be subdivided into two phases: relation candidate discovery, and relation candidate verification (i.e. classification as true or false). The first phase yields sentences that possibly mention a semantic relation between two terms, the second phase classifies these relation instances are true or false. Whereas the relation-learning methods described in the previous two chapters deals with both phases, the task described in this chapter only deals with relation candidate verification, and hence can be seen as part of the methods described in chapter ii and iii. Two-phased relation learning can be applied to ontology alignment by finding sentences that mention a concept from both ontologies. The incorrect relation instances are filtered out of this set of candidates by classifying them and discarding the false relation instances.

This chapter is based on a paper coauthored by Sophia Katrenko, "UVAVU: WordNet Similarity and Lexical Patterns for Semantic Relation Classification, Willem Robert van Hage, Sophia Katrenko" *(van Hage and Katrenko, 2007), which was presented as a poster at the fourth International Workshop on Semantic Evaluations (SemEval-2007).*

iv.1 Introduction

This chapter describes the entry of the University of Amsterdam and the Vrije Universiteit Amsterdam in the comparative evaluation task *Classification of Semantic Relations between Nominals*[1], task 4 of the fourth International Workshop on Semantic Evaluation (SemEval-2007). All participants were requested to write a concise report about their system without introducing the task. The introduction to the task can be found in Girju et al. (2007), which is printed in the same proceedings, preceding the reports of the participants. An excerpt from this summary paper, describing the task and related work follows.

> The theme of Task 4 is the classification of semantic relations between simple nominals (nouns or base noun phrases) other than named entities–honey bee, for example, shows an instance of the Product- Producer relation. The classification occurs in the

[1] Information about the task and the data sets can be found at `http://www.apperceptual.com/semeval.html`

> context of a sentence in a written English text. Algorithms for classifying semantic relations can be applied in information retrieval, information extraction, text summarization, question answering and so on. The recognition of textual entailment (Tatu and Moldovan, 2005) is an example of successful use of this type of deeper analysis in high-end NLP applications. The literature shows a wide variety of methods of nominal relation classification. They depend as much on the training data as on the domain of application and the available resources. Rosario and Hearst (2001) classify noun compounds from the domain of medicine, using 13 classes that describe the semantic relation between the head noun and the modifier in a given noun compound. Rosario et al. (2002) classify noun compounds using the MeSH hierarchy and a multi-level hierarchy of semantic relations, with 15 classes at the top level. Nastase and Szpakowicz (2003) present a two-level hierarchy for classifying noun-modifier relations in base noun phrases from general text, with 5 classes at the top and 30 classes at the bottom; other researchers (Turney and Littman, 2005; Turney, 2005; Nastase et al., 2006) have used their class scheme and data set. Moldovan et al. (2004) propose a 35-class scheme to classify relations in various phrases; the same scheme has been applied to noun compounds and other noun phrases (Girju et al., 2005). Chklovski and Pantel (2004) introduce a 5-class set, designed specifically for characterizing verb-verb semantic relations. Stephens et al. (2001) propose 17 classes targeted to relations between genes. Lapata (2002) presents a binary classification of relations in nominalizations. There is little consensus on the relation sets and algorithms for analyzing semantic relations, and it seems unlikely that any single scheme could work for all applications. For example, the gene-gene relation scheme of Stephens et al. (2001), with relations like *X* phosphorylates *Y*, is unlikely to be transferred easily to general text. We have created a benchmark data set to allow the evaluation of different semantic relation classification algorithms. We do not presume to propose a single classification scheme, however alluring it would be to try to design a unified standard–it would be likely to have shortcomings just as any of the others we have just reviewed. Instead, we have decided to focus on separate semantic relations that many researchers list in their relation sets. We have built annotated data sets for seven such relations. Every data set supports a separate binary classification task.

The goal of task 4 is to classify instances of semantic relations as true or false, depending whether the relation holds in a sentence that describes the relation, as opposed to relation learning tasks this presumes a given set of candidate relation instances. For example, the part-whole relation instance macadamia nuts-cake in the sentence "The *macadamia nuts* in the *cake* also make it necessary to have a very sharp knife to cut through the cake neatly." would be given and should be classified as true, because the macadamia nuts are part of the cake in the context of this sentence.

The semantic relations considered in this task are Cause-Effect (*e.g.* virus and flu), Instrument-Agency (*e.g.* knife and surgeon), Product-Producer (*e.g.* honey and bee), Origin-Entity (*e.g.* grapes and wine), Theme-Tool (*e.g.* soup and pot), Part-Whole (*e.g.* wheel and car), and Content-Container (*e.g.* wine and bottle). For each of these seven semantic relations the participants were given 140 training sentences, 70 testing sentences (both consisting of approximately 50% true and false relation instances). The negative instances are all 'near misses', as opposed to pairs of completely unrelated concepts.

The features used by the classifier are, for example, typical hypernyms of the subject and object of the semantic relation. In the case of part-whole relations this can be, for example,

objects and materials (*e.g.* pottery and clay), or groups and elements. In the example of the macadamia nuts and the cake this means the learning algorithm would have to learn that a cake can be seen as a set of ingredients and that macadamia nuts are ingredients and based on that it should conclude that part-whole relations between a set of ingredients (a group) and an ingredient (an element) are common and hence probably true. If the relation instance would have been, for example, between an object and an event, it probably would have been false.

IV.2 Experimental Set-up

The system we used to classify the semantic relations consists of two parallel binary classifiers. We ran this system for each of the seven semantic relations separately. Each classifier predicts for each instance of the relation whether it holds or not. The predictions of all the classifiers are aggregated for each instance by disjunction. That is to say, each instance is predicted to be false by default unless any of the classifiers gives evidence against this.

To generate the submitted predictions we used two parallel classifiers: (1) a classifier that combines 11 WordNet-based similarity measures, see section IV.3.1, and (2) a classifier that learns lexical patterns from Google and the Waterloo Multi-Text System (WMTS) (Turney, 2004) snippets and applies these on the same corpora, see section IV.3.2.

Three other classifiers we experimented with, but that were not used to generate the submitted predictions: (3) a classifier that uses string kernel methods on the dependency paths of the training sentences, see section IV.4.1, (4) a classifier that uses string kernels on the local context of the subject and object nominals in the training sentences, see section IV.4.2 and (5) a classifier that uses hand-made lexical patterns on Google and WMTS, see section IV.4.3.

IV.3 Submitted Run

IV.3.1 WordNet-based Similarity Measures

WordNet 3.0 (Fellbaum, 1998) is the most frequently used lexical database of English. As this resource consists of lexical and semantic relations, its use constitutes an appealing option to learning relations. In particular, we believe that given two mentions of the same semantic relation, their arguments should also be similar. Or, in analogy learning terms, if $R_1(X_1, Y_1)$ and $R_2(X_2, Y_2)$ are relation mentions of the same type, then $X_1 :: Y_1$ as $X_2 :: Y_2$. Our preliminary experiments with WordNet suggested that few arguments of each relation are connected by immediate hyperonymy or meronymy relations. As a result, we decided to use similarity measures defined over WordNet (Pedersen et al., 2004). The WordNet::Similarity package includes 11 different measures, which mostly use either the WordNet glosses (*lesk* or *vector* measures) or the paths between a pair of concepts (Leacock & Chodorow; Palmer) to determine their relatedness .

To be able to use WordNet::Similarity, we mapped all WordNet sense keys from the training and test sets to the earlier WordNet version (2.1). Given a relation $R(X, Y)$, we

computed the relatedness scores for each pair of arguments X and Y. The scores together with the sense keys of arguments were further used as features for the machine learning method. As there is no a priori knowledge on what measures are the most important for each relation, all of them were used and no feature selection step has been taken.

We experimented with a number of machine learning methods such as k-nearest neighbor algorithm, logistic regression, bayesian networks and others. For each relation the best performing method on the training set was selected (using 5-fold cross-validation).

IV.3.2 Learnt Lexical Patterns

This classifier models the intuition that when a pair of nominals is used in similar phrases as another pair they share at least one relation, and when no such phrases can be found they do not share any relation. Applied to the semantic relation classification problem this means that when a pair in the test set can be found in the same patterns as pairs from the training set, the classification for the pair will be true.

To find the patterns we followed step 1 to 6 described in Turney (2006), with the exception that we used both Google and the WMTS to compute pattern frequency.

First we extracted the pairs of nominals $\langle X, Y \rangle$ from the training sentences and created one Google query and a set of WMTS queries for each pair. The Google queries were of the form "X * Y" OR "Y * X". Currently, Google performs morphological normalization on every query, so we did not make separate queries for various endings of the nominals. For the WMTS we did make separate queries for various morphological variations. We used the following set of suffixes: '-tion(s—al)', '-ly', '-ist', '-ical', '-y', '-ing', '-ed', '-ies', and '-s'. For this we used Peter Turney's `pairs` Perl package. The WMTS queries looked like [n]>([5].."X"..[i].."Y"..[5]) and [n]>([5].."Y"..[i].."X"..[5]) for $i = 1, 2, 3$ and $n = i + 12$, and for each variation of X and Y. Then we extracted sentences from the Google snippets and cut out a context of size 5, so that we were left with similar text segments as those returned by the WMTS queries. We merged the lists of text segments and counted all n-grams that contained both nominals for $n = 1$ to 6. We substituted the nominals by variables in the n-grams with a count greater than 10 and used these as patterns for the classifier. An example of such a pattern for the Cause-Effect relation is "generation of Y by X". After this we followed step 3 to 6 of Turney (2006), which left us with a matrix for each of the seven semantic relations, where each row represented a pair of nominals and each column represented the frequency of a pattern, and where each pair was classified as either true or false. The straightforward way to find pattern frequencies for the pairs in the test set would be to fill in these patterns with the pairs of nominals from the test set. This was not feasible given the time limitation on the task. So instead, for each pair of nominals in the test set we gathered the top-1000 snippets and computed pattern frequencies by counting how often the nominals occur in every pattern on this set of text segments. We constructed a matrix from these frequencies in the same way as for the training set, but without classifications for the pairs. We experimented with various machine learning algorithms to predict the classes of the pairs. We chose to use k-nearest neighbors, because it was the only algorithm that gave more subtle predictions than true for every pair or false for every pair. For each semantic relation we used the value of k that produced the highest

F_1 score on 5-fold cross validation on the training data.

IV.4 Additional Runs

IV.4.1 String Kernels on Dependency Paths

It has been a long tradition to use syntactic structures for relation extraction task. Some of the methods as in Katrenko and Adriaans (2007) have used information extracted from the dependency trees. We followed similar approach by considering the paths between each pair of arguments X and Y. Ideally, if each syntactic structure is a tree, there is only one path from one node to the other. After we have extracted paths, we used them as input for the string kernel methods (Daumé, 2004). The advantage of using string kernels is that they can handle sequences of different lengths and already proved to be efficient for a number of tasks.

All sentences in the training data were parsed using MINIPAR (Lin, 1998). From each dependency tree we extracted a dependency path (if any) between the arguments by collecting all lemmas (nodes) and syntactic functions (edges). The sequences we obtained were fed into string kernel. To assess the results, we carried out 5-fold cross-validation. Even by optimizing the parameters of the kernel (such as the length of subsequences) for each relation, the highest accuracy we obtained was equal 61.54% (on Origin-Entity relation) and the lowest was accuracy for the Instrument-Agency relation (50.48%).

IV.4.2 String Kernels on Local Context

Alternatively to syntactic information, we also extracted the snippets of the fixed length from each sentence. For each relation mention of $R(X, Y)$, all tokens between the relation arguments X and Y were collected along with at most three tokens to the left and to the right. Unfortunately, the results we received on the training set were comparable to those obtained by string kernels on dependency paths and less accurate than the results provided by WordNet similarity measures or patterns extracted from the Web and WMTS. As a consequence, string kernel methods were not used for the final submission.

IV.4.3 Manually-created Lexical Patterns

The results of the method described in section IV.3.2 are quite far below what we expected given earlier results in the literature (Turney, 2006; van Hage et al., 2005, 2006; Berland and Charniak, 1999; Etzioni et al., 2004). We think this is caused by the fact that many pairs in the training set are non-stereotypical examples. So often the most commonly described relation of such a pair is not the relation we try to classify with the pair. For example, common associations with the pair ⟨body,parents⟩ are that it is the parents' body, or that the parents are member of some organizing body, while it is a positive example for the Product-Producer relation. We wanted to see if this could be the case by testing whether more intuitive patterns give better results on the test set. The patterns we manually created for each relation are shown in table IV.1. If a pair gives any results for these patterns on

Google or WMTS, we classify the pair as true, otherwise we classify it as false. The results are shown in table IV.2. We did not use these results for the submitted run, because only automatic runs were permitted. The manual patterns did not yield many useful results at all. Apparently intuitive patterns do not capture what is required to classify the relations in the test set. The patterns we used for the Part-Whole (6) relation had an average Precision of 0.50, which is much lower than the average Precision found in van Hage et al. (2006), which was around 0.88. We conclude that both the sets of training and test examples capture different semantics of the relations than the intuitive ones, which causes common sense background knowledge, such as Google to produce bad results.

rel.	patterns
1.	X causes Y, X caused by Y, X * cause Y
2.	X used Y, X uses Y, X * with a Y
3.	X made by Y, X produced by Y, Y makes X, Y produces X
4.	Y comes from X, X * source of Y, Y * from * X
5.	Y * to * X, Y * for * X, used Y for * X
6.	X in Y, Y contains X, X from Y
7.	Y contains X, X in Y, X containing Y, X into Y

Table IV.1: Hand-written patterns.

relation	N	Prec.	Recall	F_1	Acc.
1. Cause-Effect	6	1	0.15	0.25	0.56
2. Instr.-Agency	2	1	0.05	0.10	0.54
3. Prod.-Prod.	4	0.75	0.05	0.09	0.35
4. Origin-Ent.	6	0.33	0.05	0.09	0.35
5. Theme-Tool	2	0	0	0	0.56
6. Part-Whole	16	0.50	0.31	0.38	0.64
7. Cont.-Cont.	11	0.54	0.16	0.24	0.50

Table IV.2: Results for hand-written lexical patterns on Google and WMTS.

IV.5 RESULTS

IV.5.1 WORDNET-BASED SIMILARITY MEASURES

table IV.3 shows the results of the WordNet-based similarity measure method. In the 'methods' column, the abbreviation LR stands for logistic regression, K-NN stands for k-nearest neighbor, and DT stands for decision trees.

IV.5.2 LEARNT LEXICAL PATTERNS

table IV.4 shows the results of the learnt lexical patterns method. For all relations we used the k-nearest neighbor method.

relation	method	Prec.	Recall	F_1	Acc.
1. Cause-Effect	LR	0.48	0.51	0.49	0.45
2. Instr.-Agency	DT	0.65	0.63	0.64	0.62
3. Prod.-Prod.	DT	0.67	0.50	0.57	0.46
4. Origin-Ent.	LR	0.50	0.47	0.49	0.49
5. Theme-Tool	LR	0.54	0.52	0.53	0.62
6. Part-Whole	DT	0.54	0.73	0.62	0.67
7. Cont.-Cont.	2-NN	0.66	0.55	0.60	0.62

Table IV.3: Results for similarity-measure methods.

relation	method	Prec.	Recall	F_1	Acc.
1. Cause-Effect	3-NN	0.53	0.76	0.63	0.54
2. Instr.-Agency	2-NN	0.47	0.89	0.62	0.46
3. Prod.-Prod.	2-NN	0	0	0	0.33
4. Origin-Ent.	2-NN	0.47	0.22	0.30	0.54
5. Theme-Tool	3-NN	0.39	0.93	0.55	0.38
6. Part-Whole	2-NN	0.36	1	0.53	0.36
7. Cont.-Cont.	2-NN	0.51	0.97	0.67	0.51

Table IV.4: Results for learnt lexical patterns on Google and WMTS.

IV.6 Discussion

Our methods had the most difficulty with classifying relation 1, 3 and 4. We wanted to see if human assessors perform less consistent for those relations. If so, then those relations would simply be harder to classify. Otherwise, our system performed worse for those relations. We manually assessed 30 sample sentences from the test set, 15 of which were positive examples and 15 were false examples. The result of a comparison with the test set is shown in table IV.5. The numbers listed there represent the fraction of examples on which we agreed with the judges of the test set. There was quite a large variation in the inter-judge agreement,

	inter-judge agreement	
relation	judge 1 vs. reference	judge 2 vs. reference
1. Cause-Effect	0.93 (28/30)	0.93 (28/30)
2. Instrument-Agency	0.77 (23/30)	0.77 (23/30)
3. Product-Producer	0.87 (26/30)	0.80 (24/30)
4. Origin-Entity	0.80 (24/30)	0.77 (23/30)
5. Theme-Tool	0.80 (24/30)	0.77 (23/30)
6. Part-Whole	0.97 (29/30)	1.00 (30/30)
7. Content-Container	0.77 (23/30)	0.77 (23/30)

Table IV.5: Inter-judge agreement.

but for relation 1 and 6 the consensus was high. We conclude that the reason for our low performance on those relations is not caused by the difficulty of the sentences, but due to other reasons. Our intuition is that the sentences, especially those of relation 1 and 3, are

easily decidable by humans, but that they are non-stereotypical examples of the relation, and thus hard to learn. The following example sentence breaks common-sense domain and range restrictions: Product-Producer #142 *"And, of course, everyone wants to prove the truth of their beliefs through experience, but the <e1>belief</e1> begets the <e2>experience</e2>."* The common-sense domain and range restriction of the Product-Producer relation are respectively something like 'Entity' and 'Agent'. However, 'belief' is generally not considered to be an entity, and 'experience' not an agent. The definition of Product-Producer relation used for the Challenge is more flexible and allows therefore many examples which are difficult to find by such common-sense resources as Google or WordNet.

Chapter v

Evaluating Ontology-Alignment Techniques

The OAEI Food & Environment Tasks

In this chapter we discuss a comparative evaluation of ontology alignment techniques. The evaluation is based on the task to align the AGROVOC and NALT thesauri. We perform a quantitative evaluation using the alignment sample evaluation *method described in chapter vi, and a qualitative analysis of typical mistakes and omissions in the alignments submitted by the participants.*

This chapter is based on a paper coauthored by Margherita Sini, Lori Finch, Hap Kolb and Guus Schreiber, "The OAEI food task: An Analysis of a Thesaurus Mapping Task, Willem Robert van Hage, Margherita Sini, Lori Finch, Hap Kolb, Guus Schreiber" *(van Hage et al., 2008b), which has been submitted for publication. Parts of this chapter have been published in reports about the results of the OAEI 2006 and 2007 (Euzenat et al., 2006, 2007) that were presented at the first and second International Workshop on Ontology Matching (OM-2006 and OM-2007).*

The appendix to this chapter contains a short description of a related task, the alignment of the AGROVOC, NALT, and GEMET thesauri, in the context of the OAEI 2007 environment task. *This was published in the report about the results of the OAEI 2007 (Euzenat et al., 2007).*

Abstract This chapter describes the *food task* of the Ontology Alignment Evaluation Initiative (OAEI) 2006 and 2007. The OAEI[1] is a comparative evaluation effort to measure the quality of automatic ontology-alignment systems. The *food task* focuses on the alignment of thesauri in the agricultural domain. It aims at providing a realistic task for ontology-alignment systems by which the relative performance of the alignment systems can be evaluated. Research groups from around the world signed up their ontology-alignment system for the task. Each system automatically constructed an alignment. The alignments were then compared by means of statistical performance measures to get clues about which techniques work best for automatic ontology alignment. To complement this quantitative evaluation we performed an in-depth qualitative analysis of the results to draw conclusions about the strengths and weaknesses of the various alignment approaches and the specific challenges of thesaurus alignment and its evaluation.

[1] http://oaei.ontologymatching.org

V.1 INTRODUCTION

Ontology alignment has become a major research focus in the area of distributed Web applications. The Web has has made it possible to access multiple libraries at the same time. Different libraries have different indexing schema's. This makes federated access difficult. In the past, this was solved by unifying the schema's. This can fail when there are non-reconcilable differences between the schema's or conflicts of interest. Alignment can be seen as an alternative to schema unification, *cf.* Clarke (1996). The schema's stay unchanged; instead cross-links between the the schema's are added. Differences between the schema's are allowed to persist. (Huang et al., 2005, 2006) The alignment has to be maintained, but this is a smaller issue to solve than to arrange joint maintenance of a unified schema.

Initially, OAEI focused on alignment of heavy-weight OWL-based ontologies. However, in practice the domains in which alignment is needed are typically information retrieval tasks where documents (including multimedia documents such as images and video) have been indexed with different thesauri. Such concept schemes can best be viewed as light-weight ontologies. Many thesauri follow the ANSI/NISO and ISO standards for thesauri, such as ANSI/NISO Z39.19, ISO 2788 (for monolingual thesauri), and ISO 5964 (for multi-lingual thesauri), see Hodge (2000). Within the Semantic Web community SKOS (Simple Knowledge Organization System) has been developed for the purpose of providing a format for publishing such thesauri on the Web. SKOS (Miles and Bechhofer, 2004) allows one to define a concept scheme with a URI for each concept so that we can create unambiguous alignments between the thesauri. SKOS provides a special alignment vocabulary, the SKOS Mapping Vocabulary (discussed in more detail in section v.2.3).

The main research objective of this chapter concerns the *evaluation methodology for ontology alignments*. We use the OAEI 2006/2007 results as a case study to get insight into evaluation issues such as the way in which recall and precision should be assessed. In real-life alignment cases (of which the *food task* is an example) there is often no gold standard for the alignment available. We are also interested in characteristics of thesaurus alignment in comparison with general ontology alignment.

We start by explaining the data involved in the OAEI 2006 and 2007 *food task*. Section v.2 describes the vocabularies that were used and section v.3 describes the alignments submitted by the participants. Subsequently, we discuss in section v.4 the general evaluation method that we followed. In sections v.4.1 and v.4.2 we elaborate on the specific details of the OAEI 2006 and 2007 food task evaluation. In section v.5 we quantitatively compare the performance of the participating systems. Finally, in section v.6 we perform a qualitative analysis of the results, where we discuss in some detail typical issues with respect to alignment of thesauri.

V.2 VOCABULARIES

The thesauri used for this task are the United Nations Food and Agriculture Organization AGROVOC thesaurus, and the United States National Agricultural Library Agricultural Thesaurus. We selected these thesauri because they are both large and widely used. The thesauri were supplied unaltered to the participants in their native SKOS format and a

simplified OWL-Lite version. The 2006 OWL-Lite version was made by Wei Hu. The 2007 OWL-Lite version follows the same rules as those used by Antoine Isaac for the OAEI 2007 library track.[2] The versions used for the OAEI 2006 and 2007 *food task* can be downloaded at `http://www.few.vu.nl/~wrvhage/oaei2006` and `http://www.few.vu.nl/~wrvhage/oaei2007/food.html` respectively.

v.2.1 AGROVOC

The UN FAO AGROVOC thesaurus was developed by agriculture domain experts at the FAO and the Commission of the European Communities, in the early 1980s. It is updated by FAO roughly every three months. AGROVOC is used to index a multitude of data sources all over the world, one of which is the AGRIS/CARIS[3] literature reference database. Many international organizations use localized excerpts of the thesaurus. Information about these projects and links to the respective web pages can be found at `http://www.fao.org/aims`. There are manually created alignments from AGROVOC to the Chinese Agricultural Thesaurus and the German National Library's Schlagwortnormdatei, and an automatically generated alignment to the European Environment Agency's GEMET.[4] AGROVOC is available in many different formats including ISO 2709 (format for bibliographic information interchange), SKOS, OWL,[5] and TermBase eXchange (TBX).[6] All formats are generated from a native custom MySQL form. The current version of AGROVOC thesaurus can be browsed online at `http://www.fao.org/agrovoc`. An online collaborative maintenance system for AGROVOC, called the AGROVOC Concept Server Workbench, is under development.[7] Future versions of the thesaurus will also be made available through a web service.

For the OAEI 2006 *food task* we used the May 2006 version which consists of 28,174 descriptor terms (*i.e.* prefered terms) and 10,028 non-descriptor terms (*i.e.* alternative terms). It is multilingual in ten languages (English, French, Spanish, Arabic, Chinese, Portugese, Czech, Japanese, Thai, and Slovak). For the OAEI 2007 *food task* we used the February 2007 version which consists of 28,445 descriptor terms and 12,531 non-descriptor terms and is multilingual in eleven languages (the same as listed before, plus German). Strictly speaking, AGROVOC is a translated thesaurus and not a multilingual thesaurus. It started with an English version and was later translated into other languages by domain experts from the respective countries. The terms are grouped into categories from the AGRIS/CARIS Classification Scheme.[8]

The SKOS format has exactly one skos:Concept per descriptor term. The term itself is a skos:prefLabel of the skos:Concept. Non-descriptors (USE) are modelled as skos:altLabels. USE+ is downgraded to multiple unrelated skos:altLabel relations. BT, NT, and RT relations are modelled as skos:broader, skos:narrower, and skos:related relations between the respective skos:Concepts. AGRIS/CARIS Classification Scheme categories are modelled as skos:Con-

[2] `http://www.few.vu.nl/~aisaac/oaei2007`
[3] `http://www.fao.org/agris`
[4] `http://www.few.vu.nl/~wrvhage/oaei2007/environment.html`
[5] `http://www.w3.org/2004/OWL`
[6] `http://www.lisa.org/standards/tbx`
[7] `http://www.fao.org/aims/aos.jsp`
[8] `http://www.fao.org/aims/ag_classifschemes.jsp`

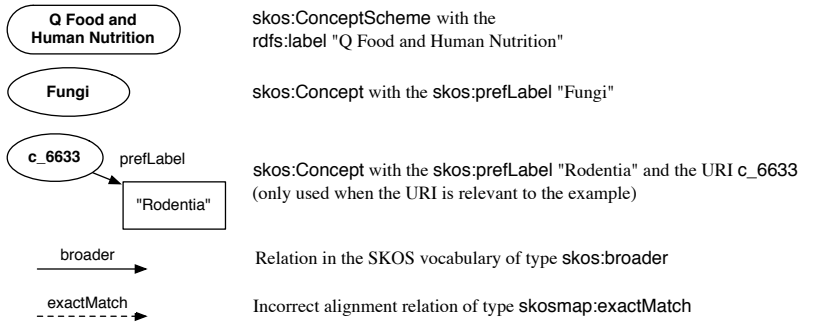


Figure V.1: Legend to the visual symbols used in this chapter.

ceptSchemes. The broadest concept that has a AGRIS/CARIS classification is modelled as a top concept of that skos:ConceptScheme using skos:hasTopConcept. Whenever scope notes exist they are attached to the skos:Concept as strings using the skos:scopeNote property.

An excerpt of AGROVOC is shown in figure V.2 on the left side. In all figures in this chapter we will depict skos:Concepts as an oval filled with the skos:prefLabel text. In cases where we explicitly want to show skos:altLabel and skos:prefLabel we depict the skos:Concept as an oval filled with its URI, connected to boxes that represent its various labels. skos:ConceptShemes are depicted as boxes with round sides. An overview of these visual symbols is shown in figure V.1.

V.2.2 NAL Agricultural Thesaurus

The USDA NAL Agricultural Thesaurus (NALT) was created by the National Agricultural Library to disclose information of the Agricultural Research Service of the United States Department of Agriculture. In 2002 the first English edition was published. In 2007 the first Spanish version of the NALT was published. Both are updated annually. The NALT is used to index the AGRICOLA[9] literature reference database of the USDA, the Food Safety Research Information Office[10] (FSRIO) research projects database, the NAL Digital Repository[11] (NALDR), and various data sources of the Agriculture Network Information Center[12] (AgNIC). There is an automatically generated alignment to the European Environment Agency's GEMET thesaurus. NALT is available in SKOS, and MARC, and a custom ASCII and XML format. The SKOS format is generated from the XML format. This transformation follows the same rules as described above for the SKOS version of AGROVOC. The current English version of the NALT thesaurus can be browsed online at `http://agclass.nal.usda.gov/agt/agt.shtml`. More information about the Spanish version can be found online at `http://agclass.nal.usda.gov/agt_Espanol/agt_es.shtml`.

[9] `http://agricola.nal.usda.gov`
[10] `http://fsrio.nal.usda.gov`
[11] `http://naldr.nal.usda.gov`
[12] `http://www.agnic.org`

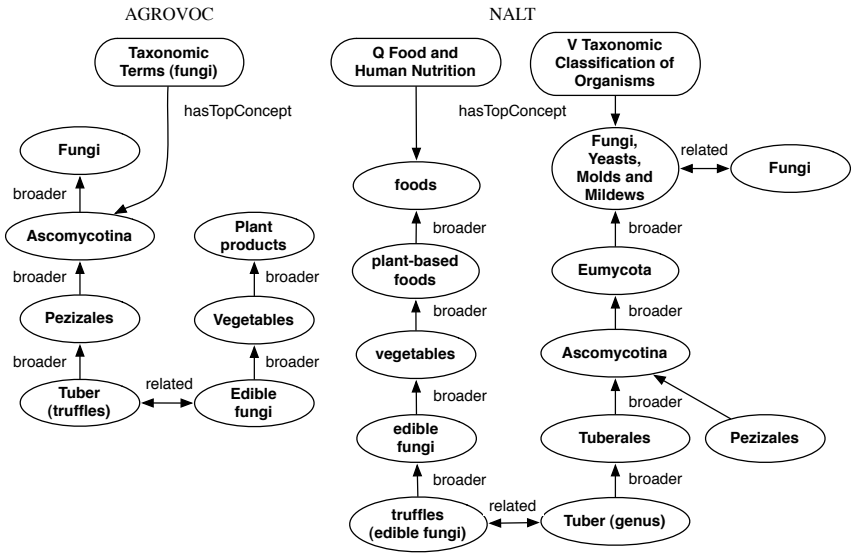


Figure v.2: The concept representing truffles in AGROVOC and NALT.

For the OAEI 2006 *food task* we used the 2006 version of the NALT which consists of 41,577 descriptor terms and 24,525 non-descriptor terms and is English monolingual. For the OAEI 2007 *food task* we used the 2007 version, which consists of 42,326 descriptor terms and 25,985 non-descriptor terms. We only use the English version.

An excerpt of NALT is shown in figure v.2 on the right side.

v.2.3 SKOS Mapping Vocabulary

For the alignment we use relations from the SKOS Mapping Vocabulary.[13] The participants were allowed to use the following relations: skos:narrowMatch, skos:exactMatch, and skos:broadMatch. The other relations and boolean combinators (skos:minorMatch, skos:majorMatch, skos:AND, skos:OR, skos:NOT) of the SKOS Mapping Vocabulary were not used in the evaluation. The participants were requested to hand in an RDF file in alignment format[14] (Euzenat, 2004) that contains information about the properties of the alignment, like which ontologies are involved, and properties of each relation in the alignment, like which concepts are aligned and the confidence the participant's ontology alignment system gave to the relation. An example of such an RDF file is shown in the code listing in figure v.3. The example shows two alignment relations, nalt:osteomyeliti skos:exactMatch agrovoc:c_12988 (Osteomyelitis), and favism skos:exactMatch agrovoc:c_6051 (Poisoning). The relations get a confidence of respectively 1.0 and 0.89.

[13] http://www.w3.org/2004/02/skos/mapping/spec
[14] http://alignapi.gforge.inria.fr

```
<?xml version='1.0' encoding='utf-8'?>
<rdf:RDF xmlns='http://knowledgeweb.semanticweb.org/heterogeneity/
    alignment'
         xmlns:rdf='http://www.w3.org/1999/02/22-rdf-syntax-ns#'
         xmlns:xsd='http://www.w3.org/2001/XMLSchema#'>

  <Alignment>
    <xml>yes</xml>
    <level>0</level>
    <type>11</type>
    <onto1>http://agclass.nal.usda.gov/nalt/2007.xml</uri1>
    <onto2>http://www.fao.org/aos/agrovoc</uri2>
    <map>
      <Cell>
        <entity1 rdf:resource='http://agclass.nal.usda.gov/nalt/2007.xml#
            osteomyelitis'/>
        <entity2 rdf:resource='http://www.fao.org/aos/agrovoc#c_12988'/>
        <measure rdf:datatype='http://www.w3.org/2001/XMLSchema#float
            '>1.0</measure>
        <relation>http://www.w3.org/2004/02/skos/mapping#exactMatch</
            relation>
      </Cell>
    </map>
    <map>
      <Cell>
        <entity1 rdf:resource='http://agclass.nal.usda.gov/nalt/2007.xml#
            favism'/>
        <entity2 rdf:resource='http://www.fao.org/aos/agrovoc#c_6051'/>
        <measure rdf:datatype='http://www.w3.org/2001/XMLSchema#float
            '>0.89</measure>
        <relation>http://www.w3.org/2004/02/skos/mapping#exactMatch</
            relation>
      </Cell>
    </map>
    ...
  </Alignment>
</rdf:RDF>
```

Figure v.3: The RDF format used for the submission of alignments. This example shows two skos:exactMatch relations with a confidence of respectively 1.0 and 0.89.

v.3 Participants and Submitted Alignments

The OAEI 2006 *food task* had five participants: South East University with the Falcon-AO 0.6 system (Hu et al., 2006); University of Pittsburgh with the Prior system (Mao and Peng, 2006); Tsinghua University with the RiMOM system (Li et al., 2006); University of Leipzig with the COMA++ system (Massmann et al., 2006); and Universitá degli Studi di Milano with the HMatch system (Castano et al., 2006). Each team provided between 10,000 and 20,000 alignment relations. This amounted to 31,112 unique alignment relations in total. All of these mappings were of the type skos:exactMatch. None of the systems was able to discover skos:broadMatch or skos:narrowMatch mappings. There was a high agreement between the best three systems, RiMOM, Falcon-AO, and HMatch. Details are shown in table v.1. From this table we can also deduce that there is a relatively large set of 'easy' mappings that are recognized by all systems.

The OAEI 2007 *food task* also had five participants: South East University with the Falcon-AO 0.7 system (Hu et al., 2007); Tsinghua University with the RiMOM system (Li et al., 2007); Politecnico di Milano with the X-SOM system (Curino et al., 2007); and the Knowledge Media Institute with two systems, DSSim (Nagy et al., 2007) and SCARLET (Sabou et al., 2007). Each team provided between 6583 (X-SOM) and 18,420 (RiMOM) alignment relations. This amounted to 37,384 unique alignment relations in total. The SCARLET system discovered skos:exactMatch, skos:broadMatch, and skos:narrowMatch relations. The other systems only discovered skos:exactMatch relations. There was a slightly lower agreement between RiMOM and Falcon-AO (the Jaccard similarity coefficient, $|A \cap B|/|A \cup B|$, was $11{,}203/22{,}517 = 0.50$ as opposed to $11{,}585/26{,}984 = 0.75$ in 2006). The other systems found much more different sets of alignment relations than the other systems in 2006. The SCARLET system is a complete outlier compared to the other systems.

v.4 Evaluation Procedure

In this section we will describe the evaluation process we used to compare the various submissions. The main two statistics we used to compare the alignments are Precision and Recall. If we call the set of all alignment relations that were submitted by a participant *Found* and the set of all alignment relations we would like to receive (*i.e.* all correct alignment relations) *Correct*, Precision and Recall can be defined as follows:

$$\text{Precision} = \frac{|Found \cap Correct|}{|Found|} \tag{v.1}$$

$$\text{Recall} = \frac{|Found \cap Correct|}{|Correct|} \tag{v.2}$$

Figure v.4 illustrates these definitions. In practice, the computation of Precision and Recall require the assessment of all relations in the set of *Found* relations and the determination of the cardinality of the set of all *Correct* relations.

The assessment of all *Found* relations requires human assessors to decide whether tens of thousands of alignment relations are correct or incorrect. The experience of the OAEI

2006

system	# mappings returned	# mappings shared with *n* other systems 0	1	2	3	4
RiMOM	13,975	868	1,042	2,121	4,389	5,555
Falcon-AO	13,009	642	419	1,939	4,400	5,555
Prior	11,511	1,543	1,106	676	2,631	5,555
COMA++	15,496	11,610	1,636	629	2,028	5,555
HMatch	20,001	7,000	981	2,045	4,420	5,555
all systems	31,112	21,663	2,592	2,470	4,467	5,555

2007

system	# mappings returned	# mappings shared with *n* other systems 0	1	2	3	4
RiMOM	18,419	7,052	6,131	3,774	1,462	0
Falcon-AO	15,300	2,964	6,933	3,941	1,462	0
X-SOM	6,583	4,083	317	725	1,458	0
DSSim	14,962	9,273	876	3,353	1,460	0
SCARLET exactMatch	81	9	27	39	6	0
broadMatch & narrowMatch	6,038	6,038	0	0	0	0
all systems	41,967	29,419	7,142	3,944	1,462	0

Table V.1: Distribution of the systems' results. Shown are the number of mappings returned by each system and how many mappings are also returned by *n* of the other systems.

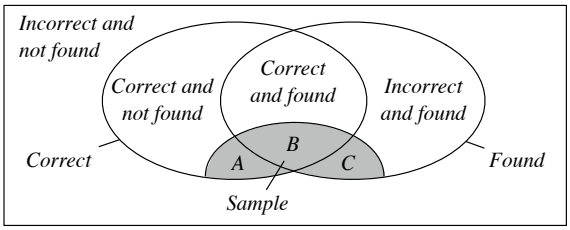


Figure V.4: Venn diagram to illustrate the sets of relations that are relevant to the sample evaluation. $A \cup B$ is a sample of the population of Correct alignment relations. $B \cup C$ is a sample of the population of Found alignment relations.

has shown that a voluntary human assessor can judge around 250 alignment relations per hour for at most a few hours. That means 10,000 alignments cost around 40 man-hours. For most large organizations that want to know the quality of an ontology alignment system this is a feasible investment. For evaluation fora such as the OAEI, this is not feasible. For the comparative evaluation of multiple systems we even have to assess multiple sets of *Found* relations.

The assessment of all *Correct* requires the manual construction of the entire desired alignment. Manual construction of the entire alignment is even more costly than the assessment of all *Found* relations, because it involves searching for good alignment relations, which is more difficult than simply judging the validity of a set of given relations. To illustrate this we can look at the manual construction of the alignment between the Chinese Agricultural Thesaurus (CAT), which consists of 64,638 concepts, and AGROVOC. This alignment is directional from CAT to AGROVOC and hence not complete, and consists of 24,686 alignment relations. Chang Chun of the Chinese Academy of Agricultural Sciences (CAAS) revealed at the Eighth Agricultural Ontology Service (AOS) meeting[15] that the construction took 15 PhD students (in relevant fields of research, like biology) 24 man-hours each during 6 months. The students were paid per alignment and followed a strict protocol. They made at most around 150 alignment relations per hour. (Liang et al., 2005)

If you are not interested in the evaluation as such, but in a complete alignment, automatic ontology alignment might not be necessary, because the total investment for the manual construction of an alignment is, for many purposes, not significantly larger than that of verifying an automatically constructed alignment. Provided that time, money, and access to adequately educated people are not an issue. In these cases manual ontology alignment might be worth the investment.

To make the computation of Precision and Recall feasible for the OAEI *food task*, we performed sample evaluation. Sample evaluation assumes that measurements on a randomly drawn sample can be extrapolated to the entire population. The larger the sample, the less the estimation based on the sample will deviate from the true value on the entire population. In our case, that means that we can extrapolate the performance of a system on a small set of alignment relations to all relevant alignments. We work with small subsets of all *Found* and *Correct* relations from which we generalize to the entire set of *Found* or *Correct* relations. The grey areas $B \cap C$ and $A \cap B$ in figure v.4 illustrate the samples used for the evaluation of respectively Precision and Recall. In section v.4.1 and v.4.2 we will go into detail on how these samples were constructed and how the human judges operated exactly.

Sample evaluation comes with a price. It introduces sampling error, bias due to the accidental inclusion and omission of certain elements from the population in the sample. The smaller the sample is, the more likely it is that important features of the population are accidentally overlooked. For instance, we know that the automatic alignment of concepts that represent the animal species is quite simple compared to the alignment of concepts that represent socio-economic phenomena. If a random sample of alignments by accident overrepresents animal species then the performance estimate based on this sample will be too optimistic. The fact that there are many potential alignment relations between animal species and few between socio-economic phenomenae even makes it quite likely that a

[15] http://www.fao.org/aims/pub_aos8.jsp

random sample from all alignment relations contains no socio-economic relations, but quite a few animal species relations. To minimize this kind of bias, we did a separate evaluation for sets of alignment relations that we know in advance to require different alignment strategies. The separate results are combined into a weighted average to give a fair overall performance indication. The statistical technique we used to accomplish this for Precision and Recall were different. For Precision we use stratified sampling, while for Recall sampling we use cluster sampling. (Cochran, 1977) The main reason for this difference is that the set of all *Found* alignment relations, as opposed to all *Correct* alignments, is predetermined. Hence we can easily draw samples from it.

In order to draw samples from the set of all *Correct* alignments we have to draw from the set of *all* alignment relations and filter out the incorrect alignments. Clearly, some parts of the cartesian product of the sets of terms from the two thesauri will contain more correct alignment relations than others (*e.g.* there are bound to be matches between the parts about plants of both thesauri, but not between the part about plants of one thesaurus and the area about countries of the other). So if we want to use our time optimally–which we have to do to make the evaluation feasible–we will look for correct alignment relations in the areas that are likely to contain some and not in the areas that are unlikely to contain any. This concession breaks one of the assumptions of stratified sampling, the assumption that the *entire* population is partitioned and that all elements get an equal chance to be selected for a sample.

The closest thing to stratified sampling that does not make assumptions we can not meet is cluster sampling where the clusters are not selected randomly. The price we pay for the reduction in assessment time is that we have no indication how large the error margin is when we generalize from the samples to the entire population of *Correct* alignment relations.

V.4.1 PRECISION

We estimate Precision using stratified sampling from the set of all *Found* alignment relations. This set is different for the 2006 and 2007 *food task* and different for each participating system. We discounted the effect of two kinds of features in the evaluation: how many systems submitted a certain relation, and the topic of the relation. The intuition behind this is the following. It can be expected that the quality of alignment relations that are submitted by all systems and relations that are submitted by, for instance, only one system will be different. It can also be expected that some topics are easier than others, as we explained in the beginning of section V.4 about terms representing animal species.

We first partitioned the set of all *Found* alignment relations into strata with a different topic. All relations between concepts that fall into these topics were grouped together. In 2006 we distinguished three categories of topics in the thesauri that each required a different level of domain knowledge of the assessors: *taxonomical* concepts (plants, animals, bacteria, etc.) that can be aligned with a few simple rules and lexical matching, *biological and chemical* concepts (structure formulas, terms from generics, etc.) that contain many synonyms and lexical variants, and *miscellaneous*, the remaining concepts (geography, agricultural processes, etc.) that can be expected to require a diverse set of techniques to match. In 2007 we distinguished four categories op topics. The same as used in 2006

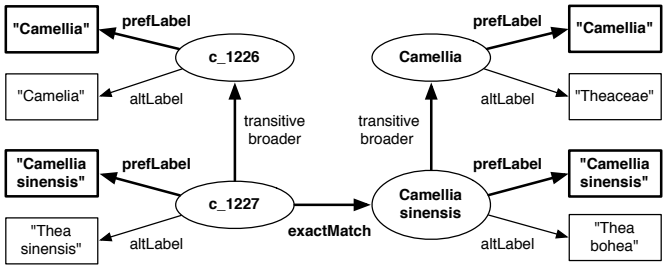


Figure v.5: Automatic assessment of taxonomical terms. The two concepts representing Camellia sinensis are considered equivalent, because they have a matching label and some of their ancestors also have matching labels.

plus *geographical* concepts (countries, provinces, etc.). We chose to separate these from the *miscellaneous* set, because there is much consensus about the naming of geographical locations. This makes the alignment of geographical concepts much easier than other topics in the *miscellaneous* set.

From each of the sets shown in table v.1 we took a random sample from each of the topic strata, such that both commonly and rarely returned alignment relations would be represented in each topic. Together, this led to the samples shown in table v.2, that had to be assessed.

Under the authority of taxonomists at the USDA the taxonomical stratum was automatically assessed completely using the strict rules that apply to the naming scheme of taxonomy. These rules are that if the preferred term of concept *A* is literally the same as either the preferred or the alternative term of concept *B* then the concepts are considered to be equivalent, provided that the same goes for an ancestor of *A* and *B*. This is illustrated in figure v.5. This assumes that the same taxonomical names always signify the same species, group, kingdom, or the like. In general, this is not true for taxonomical names, but only for names that are disambiguated by the last name of the author that first published the classification and year of the publication, *e.g.* "Passer domesticus (Linnaeus, 1758)". An example of homonymy in species names is 'Cereus', which can be either a cactus or sea anemone. In the case of NALT and AGROVOC, however, this ambiguation is not necessary, because the species names were based on the same literature and many of the concepts were copied from the same sources. Therefore, if the terms match it is extremely likely that they refer to the same species.

The other strata were all manually assessed by a group of domain experts. In 2006 this was done by domain experts of the NAL and the FAO, and a group of computer scientists at the EKAW workshop. In 2007 it was done by domain experts of the NAL, FAO, TNO Quality of life, Unilever, Wageningen Agricultural University, and the European Environment Agency. The assessed samples can be downloaded from `http://www.few.vu.nl/~wrvhage/oaei2006/gold_standard` and `http://www.few.vu.nl/~wrvhage/oaei2007/gold_standard`.

2006

stratum topic	stratum size (N_h)	sample size (n_h)	stratum weight
taxonomical	18,399	18,399	0.59
bio/chem	2,403	250	0.08
miscellaneous	10,310	650	0.33
all topics	31,112		

2007

stratum topic	stratum size (N_h)	sample size (n_h)	stratum weight
taxonomical	23,023	23,023	0.55
bio/chem	3,965	200	0.09
geographical	1,354	86	0.03
miscellaneous	13,625	476	0.32
all topics	41,967		

Table V.2: Sizes of the strata and of the samples from those strata that were assessed to evaluate Precision. The last column shows how much the stratum weighed in the calculation of a system's mean Precision.

ASSESSMENT TOOL FOR PRECISION For the assessment we used an alignment assessment tool developed at TNO by Willem Robert van Hage. An adaptation of this tool was also used for the assessment of the OAEI 2007 library task. A screengrab is shown in figure V.6. This tool reads a set of mappings in the common format for alignments and outputs a web form that is used by judges to assess the mappings. The results of the form are submitted to the organizer of the *food task*. The assessment process of a mapping follows three steps.

1. The judge decides if the relation specified above the arrow (between the two green boxes) holds between the two bold concepts. If the relation holds he skips to point 2 and goes straight to point 3. If it does not hold he goes to point 2;

2. The judge tries to specify an alternative relation, either by changing the relation type, or the concepts. If possible he select exactMatch and specifies the proper concepts between which the exactMatch relation holds. Otherwise he selects broadMatch or narrowMatch and specifies the proper concepts between which that relation holds.

3. The judge changes the default value of the assessment, 'unknown', into either 'true' or 'false'. If the relation holds and he arrived here from point 1 he selects 'true'. If the relation does not hold, but if he successfully selected an alternative relation (at point 2) that does hold, he also select 'true'. If the relation does not hold and no correct alternative could be found at point 2, select 'false'.

Finally, if the judge wishes to document his decision he fills out the entry box at the bottom of the assessment form.

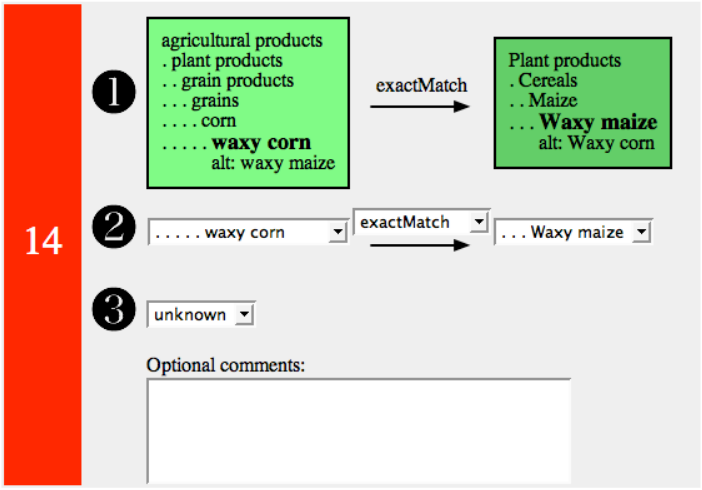


Figure v.6: Screengrab of the assessment tool used to evaluate Precision. Shown is the 14th mapping relation from a sample set of mappings, nalt:'waxy corn' skos:exactMatch agrovoc:'Waxy maize'.

2006		NAL & FAO (KR and food experts)		
		true	**false**	**unknown**
computer science	**true**	253	13	0
researchers at EKAW	**false**	6	52	0
(KR experts, agriculture laymen)	**unknown**	4	8	0

Table v.3: Comparison between the assessments by judges from the NAL and FAO and computer scientist judges. Shown is a confusion matrix of the 336 alignments from the OAEI 2006 *food task* that were judged by both groups. Each alignment was judged once by someone from each group.

Inter-Judge Agreement Both in 2006 and 2007 all samples were assessed by domain experts, but to find out how important it is to involve domain experts in the assessment part of the work was repeated by laymen, computer scientists at the EKAW workshop (mainly knowledge representation experts). The agreement between the group of domain experts and the group of computer scientists was 72%. The computer scientists were less likely to judge a mapping to be correct than the domain experts. They judged 78% of the sample mappings to be 'true', while the domain experts judged 85% to be 'true'. A more exact analysis is shown in table v.3, which shows the judgements of the overlapping set of alignment relations. From this data we can compute Cohen's kappa to show how similar the judgements of the two parties are. We use Cohen's kappa as opposed to, for example, Fleiss' kappa, because we only have one judgement for each alignment per group and thus only

two parties that can agree or disagree. Cohen's kappa is defined as follows:

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}$$

Where $Pr(a)$ is the relative observed agreement among raters, and $Pr(e)$ is the probability that agreement is due to chance. If the raters are in complete agreement then $\kappa = 1$. If there is no agreement among the raters (other than what would be expected by chance) then $\kappa \leq 0$. A detailed description can be found in Cohen (1960). The κ among the two groups of judges was 0.73, which signifies a substantial agreement, which is higher than we expected. A κ of around 0.65 is not unusual between domain experts that are supposed to agree. Apparently, most alignments that are true are clearly true and slightly less, but still many of the false alignments are clearly false. We questioned some of the judges from both groups. The laymen tended to be more sceptical about the correctness of an alignment relation, because they felt it was worse to make an inappropriate generalization than an inappropriate discrimination, whenever they were not familiar with the kind of generalizations that are common in agricultural library systems. If we would have used the assessments made by the laymen for this evaluation instead of those made by the domain experts the estimated Precision scores would have been slightly lower, but it is unlikely that the ranking of the participants would have changed.

SIGNIFICANCE TESTING As a significance test on Precision scores of the systems we used the Bernoulli distribution (van Hage et al., 2007). Precision of system A, P_A, can be considered to be significantly greater than Precision of system B, P_B, if their estimated values, $\hat{P}_A$ and $\hat{P}_B$ are far enough apart. In general, based the Bernoulli distribution, this is the case when the following equation holds:

$$|\hat{P}_A - \hat{P}_B| > 2\sqrt{\frac{\hat{P}_A(1-\hat{P}_A)}{n_A} + \frac{\hat{P}_B(1-\hat{P}_B)}{n_B}} \tag{V.3}$$

where n_A and n_B are the size of the set of assessed alignment relations that were returned by respectively system A or B. This number is always less or equal to the numbers in table V.2, which shows the total number of assessed relations for all systems. The significance test in Equation V.3 was used to determine which of the systems performs best on each of the three or four strata. The greatest error margin occurs when both systems have a Precision close to 0.5, when it is at most $1/\sqrt{n}$. When the results of the strata are combined, we are able to distinguish smaller differences in the results than for each of the strata alone. The upper bound of the error is equal to the error of simple random sampling (Cochran, 1977). The significance test we used for the combined result is as follows. We denote the estimated Precision of system A on stratum h as $\hat{P}_{A,h}$, the size of stratum h as N_h, and the size of the sample from stratum h as n_h (see table V.2). We can conclude that system A performs significantly better than system B when the following equation holds:

$$|\hat{P}_A - \hat{P}_B| > 2\sqrt{\sum_{h=1}^{L} \frac{\hat{P}_{A,h}(1-\hat{P}_{A,h})}{N_A}\left(\frac{N_h}{n_h} - 1\right) + \sum_{h=1}^{L} \frac{\hat{P}_{B,h}(1-\hat{P}_{B,h})}{N_B}\left(\frac{N_h}{n_h} - 1\right)} \tag{V.4}$$

The greatest error margin still occurs when both systems have a Precision close to 0.5, but it is as most $2/\sqrt{2n}$. Again, the overall significantly best performance is indicated with a ○ and if the best result was not significantly higher than the runner up this is indicated with a ●.

v.4.2 Recall

We estimate Recall using cluster sampling from the set of all *Correct* alignment relations. These are exactly all the relations that would be in an ideal finished alignment. This set is the same for all systems. It is the same for 2006 and 2007 with the exception of changed or added concepts. Therefore, we can use the same samples for the estimation of Recall for all systems. We also chose to reuse the samples we used in 2006 for 2007 with some updates. An advantage of this is that the results of 2006 and 2007 are easily comparable. A disadvantage is that there is the possibility that participants will train their systems on the samples and thus achieve better performance on the samples than on the rest of the alignment. This can cause a positive bias in the results. We were not concerned about this, because each participating system was only allowed one configuration for all the OAEI tasks. The *food task* is only one of all the tasks. The others focus on anatomy, directories, jobs, conferences, and dutch libraries. Each task has a different optimal setting for the ontology alignment systems. Therefore, specific optimization on the food Recall samples is unlikely, because it is probably counter productive for the participants.

Like the samples we used for Precision, the samples we used for the evaluation of Recall are on a certain topic. We chose several sub-hierarchies of the two thesauri and manually created the full alignment between the concepts in these selections. The topics we used in 2006 are: all *oak trees* (everything under the concept representing the Quercus genus), all *rodents* (everything under Rodentia), geographical concepts of *Europe* (countries), and everything under the NALT concept animal health and all AGROVOC concepts that have alignment relations to these concepts and their sub-concepts. The sizes of these samples are shown in table v.4, along with the percentage of the alignment relations that was of type exactMatch, as opposed to broadMatch and narrowMatch. The average percentage of exactMatch in the 2006 sample was 70%. In 2007 we chose to add an additional geographical sample, *topography* below country level, because the 2006 geographical sample gave the impression that the percentage of exactMatch relations in the geographical domain is much higher than it really is. This is especially the case for concepts below country level, like provinces, which often do not have an exact counterpart in the other thesaurus and thus require a broadMatch or narrowMatch relation to be aligned.

Mapping Tool for Recall To create these samples we used the AIDA Thesaurus Browser, a SKOS browser that supports parallel browsing of two thesauri, concept search, mapping traversal, and the addition, change and removal of mappings of the SKOS Mapping Vocabulary. This tool was developed at TNO by Willem Robert van Hage in the context of the VL-e project.[16] It is an AJAX application that accesses a SOAP service wrapper of Sesame 1.2[17] through Java servlets. The service wrapper is part of the AIDA web service toolkit, which

[16] http://www.vl-e.nl
[17] http://openrdf.org

topic	size	% exactMatch	used in year 2006	2007
animal health	34	57%	✓	✓
oak trees (taxonomical)	41	84%	✓	✓
rodents (vernacular)	42	32%	✓	✓
Europe (country level)	74	93%	✓	✓
topography (below country level)	164	35%		✓

Table V.4: Sizes of the sets of manually created alignments used to evaluate Recall.

also includes wrappers for the Lucene search engine and several machine learning tools.[18] A screengrab of the tool is shown in figure V.7. A preliminary version of the Recall samples was made at the Vrije Universiteit Amsterdam and was verified and extended by domain experts at the the FAO and USDA to produce the final Recall samples. The samples can be downloaded from `http://www.few.vu.nl/~wrvhage/oaei2006/gold_standard` and `http://www.few.vu.nl/~wrvhage/oaei2007/gold_standard`.

The guidelines used to make the mapping were the following:

1. Starting from the side of AGROVOC, try to find a skos:exactMatch for every concept in the sample. If this is impossible, try to find a skos:narrowMatch or skos:broadMatch. Always choose the broader concept of these alignments as narrow as possible and the narrower concept as broad as possible.

2. Investigate the surrounding concepts of the target concept in NALT. If the surrounding concepts are still on the topic for the sample, try to map this concept back to AGROVOC using skos:exactMatch. If this is impossible, try to find a skos:narrowMatch or skos:broadMatch.

SIGNIFICANCE TESTS The sample selection procedure we chose, which is based on completely aligning sub-hierarchies where we expect many alignment relations, saved us a lot of time. This made it feasible to construct Recall samples. The downside of this is that the results are not fully generalizable to a greater population, because an assumption for generalization to the sample frame is random selection where each element gets an equal non-zero probability to be selected. Therefore, the application of significance measures would produce meaningless results.

Recall for all the systems is shown in table V.6.

V.5 EVALUATION OUTCOME

In this section we will examine the quantitative evaluation results. We will first discuss the performance of the participating systems per year. Then we will look at the results of the systems that participated in both years (RiMOM and Falcon-AO) and investigate the difference. Finally, we will look into the performance of the systems' aggregated results.

[18] `http://www.adaptivedisclosure.org/aida-toolkit`

Figure v.7: Screengrab of the AIDA Thesaurus Browser, which was used to create alignment samples for the evaluation of Recall.

2006					
Precision for	RiMOM	Falcon-AO	Prior	COMA++	HMatch
taxonomical	0.82	0.83 ○	0.68	0.43	0.48
bio/chem	0.85 ●	0.80	0.81	0.76	0.83
miscellaneous	0.78	0.83 ○	0.74	0.70	0.80
all topics	**0.81**	**0.83 ●**	**0.71**	**0.54**	**0.61**

2007					SCARLET	
Precision for	RiMOM	Falcon-AO	X-SOM	DSSim	exact	broad & narrow
taxonomical	0.54	0.81 ○	0.26	0.37	0.60	0.13
bio/chem	0.84	0.91	0.92	0.86	1.00 ●	0.17
geographical	0.97	0.95	1.00 ●	0.94	0.00	1.00
miscellaneous	0.69	0.86 ○	0.62	0.57	0.75	0.44
all topics	**0.62**	**0.84 ○**	**0.45**	**0.49**	**0.66**	**0.25**

Table V.5: Precision results based on sample evaluation.

V.5.1 RESULTS 2006

The Precision and Recall outcomes of the 2006 evaluation are shown at the top of table V.5 and table V.6. Overall, RiMOM and Falcon-AO were the best systems and COMA++ performed significantly worse than the other systems, mainly due to bad results for the taxonomical part of the task.

PRECISION The taxonomical parts of the thesauri accounted for by far the largest part of the mappings, 59% of all submitted mappings. The more difficult mappings that required lexical normalization, such as structure formulas, and relations that required background knowledge, such as many of the relations in the miscellaneous domain, accounted for a smaller part of the alignment. This caused systems that did well at the taxonomical mappings to have a great advantage over the other systems.

The RiMOM and Falcon-AO systems performed well at the largest two strata, taxonomical and miscellaneous, and thus achieved high Precision. What set them apart from the rest was mainly their strict acceptance criterion for alignments.

The COMA++ system lagged behind due to liberal lexical matching. Terms with a high edit distance were accepted as matches, for example, 'Buttiauxella noackiae' and 'Arca noae', because both contain the substring "␣noa". This was particularly harmful in the taxonomical stratum, because of three reasons. (1) Many latinized names have similar prefixes and suffixes and have a tendency to start with a 'c' or 'p'. This decreases the edit distance amongst unrelated terms. (2) Different species from the same genus always share the same first name, for example 'Camellia sinensis' and 'Camellia japonica'. This greatly decreases the edit distance between different species. (3) It is not uncommon for species from completely different kingdoms, for example, plants and animals, to have the same specific name.[19] An example is 'caerula', which means blue and is contained in the scientific name of the blue

[19] http://en.wikipedia.org/wiki/List_of_Latin_and_Greek_words_commonly_used_in_systematic_names

2006					
Recall for	RiMOM	Falcon-AO	Prior	COMA++	HMatch
animal health	0.18 (0.55)	0.09 (0.27)	0.06 (0.18)	0.12 (0.36)	0.15 (0.45)
oak trees	0.85 (0.92)	0.83 (0.89)	0.85 (0.92)	0.54 (0.58)	0.80 (0.87)
rodents	0.07 (0.12)	0.00 (0.00)	0.05 (0.08)	0.02 (0.04)	0.00 (0.00)
Europe	0.70 (0.84)	0.69 (0.82)	0.65 (0.77)	0.24 (0.29)	0.68 (0.81)
all topics	**0.50 (0.71)**	**0.46 (0.65)**	**0.45 (0.64)**	**0.23 (0.33)**	**0.46 (0.65)**

2007					SCARLET
Recall for	RiMOM	Falcon-AO	X-SOM	DSSim	all relation types
animal health	0.21 (0.64)	0.21 (0.64)	0.00 (0.00)	0.06 (0.18)	0.00 (0.00)
oak trees	0.93 (1.00)	0.93 (1.00)	0.10 (0.12)	0.22 (0.24)	0.00 (0.00)
rodents	0.24 (0.42)	0.40 (0.71)	0.07 (0.10)	0.17 (0.29)	0.00 (0.00)
Europe	0.70 (0.84)	0.81 (0.97)	0.08 (0.10)	0.34 (0.40)	0.00 (0.00)
geography	0.26 (0.74)	0.32 (0.90)	0.05 (0.14)	0.18 (0.50)	0.01 (0.02)
all topics	**0.42 (0.78)**	**0.49 (0.90)**	**0.06 (0.11)**	**0.20 (0.37)**	**0.00 (0.00)**

Table v.6: Tentative estimation of Recall based on sample evaluation. The numbers between parentheses show Recall when only the exactMatch alignments of the reference alignments are considered.

tit (a bird), 'Parus caeruleus', and the blue passion flower (a flowering plant), 'Passiflora caerulea'.

The HMatch system performed as well as the RiMOM and Falcon-AO systems, except in the taxonomical domain. This was due to the same reasons as those described previously for the COMA++ system, but on a smaller scale. Most of the mistakes for taxonomical alignment relations were due to point 2. Also, terms from completely different parts of the thesauri were matched when there was only lexical overlap. For example, 'Jordan' (a river) and 'Triglops jordani' (a fish).

Recall All systems only returned skos:exactMatch mappings. This means Recall of all systems was limited to 71%. For example, RiMOM achieved 50% where it could achieve 71% and 71% where it could achieve 100% in table v.6.

The RiMOM system managed to discover more good results than the Falcon-AO system on the four small sample Recall bases, at the cost of some Precision. These were mainly results where the preferred labels were different and had to be matched to an alternative label. For example, 'Entomopathogenic fungi' and 'Entomogenous fungi'. RiMOM was less strict in these cases.

In general, performance on the rodents and animal health samples was bad. This was due to a large number of alignment relations in these sets that required some reasoning or background knowledge to find and a high number of broadMatch and narrowMatch relations. An example from the animal health set is the deduction that is required to conclude that 'bee viruses' have a broadMatch 'invertebrate viruses'. A system will have to reason that bees are invertebrates. None of the systems was able to accomplish this. Many of the alignment relations from the rodents set required background knowledge, or reasoning over related

term relations. In the NALT thesaurus the colloquial names of animals are linked to the scientific names with a related term relation. That means in order to match 'Geomyidae' to 'Gophers' it is necessary to recognize that this is a pattern in NALT.

The other sets, oak trees and Europe, were relatively easy for the systems. All systems except COMA++ were able to find around 70% to 80% of these alignment relations. There was no particular reason why the COMA++ system was unable to find a similar number of relations. The system simply returned only part of the results to boost Precision and selected the wrong part. For example, the match 'Italy' and 'Italy' was returned, but the match 'Bulgaria' and 'Bulgaria', which would have gotten at least the same confidence score, was not.

V.5.2 RESULTS 2007

The Precision and Recall outcomes of the 2007 evaluation are shown at the bottom of table V.5 and table V.6. The RiMOM and Falcon-AO systems are still in the lead, but the RiMOM system showed a large drop in performance, while the Falcon-AO system seems to have made a small improvement over last year's results. The performance indications of SCARLET on the biological and chemical set looks higher than that of the other systems, but the total number of exactMatch relations SCARLET returned was only marginal. The number of relations returned in the biological and chemical set was only 2 and they were both correct. That means the best two systems on that set were X-SOM and Falcon-AO.

PRECISION The Falcon-AO system was clearly the best system in 2007. This was mainly due to its consistent behavior on the taxonomical set, but also the miscellaneous set. Other systems could match Falcon-AO on the biological and chemical, and geographical sets, but performed worse on the other two sets.

The X-SOM and DSSim systems show the largest difference in performance. The large majority of the taxonomical results can be attributed to extremely liberal use of edit distance matching without disambiguation using the structure of the thesauri. Many of these matches link concepts from completely unrelated parts of the thesauri. For example, 'crushers' (equipment) has exactMatch 'Mares' (animal). The only connection is that 'crushers' has an alternative label 'mashers', which also starts with 'ma' and ends with an 's'. Another similar example is 'housing' has exactMatch 'Fomes' (a bracket fungus). The former concept has an alternative label 'homes', which also ends in 'omes'. This phenomenon was the strongest in the taxonomical part, due to regularities in latin names described before.

RECALL In 2007, the Falcon-AO system performed particularly well at the rodents set. There is an absolute difference of about 20% with the runner up, the RiMOM system. It is clear from the results that the context of the concepts, such as labels of related terms in NALT, are exploited whenever there is a lack of information. An example of a relation that was found is the 'Geomyidae' to 'Gophers' example, described before.

The X-SOM system had an unexpectedly low Recall on the Europe set. The X-SOM system has a string similarity module and the country names of the Europe set are lexically similar. However, it struggled with the large size of the food thesauri. Therefore, we expect

that the low Recall score is due to unfortunate partitioning of the thesauri, which precluded many matches from the result set.

The SCARLET system found almost none of the relations in the manually constructed alignments. Yet, a significant part of the relations that were returned were judged to be correct. The explanation for this paradoxical situation has to do with the evaluation method we used. The Recall samples consisted only of those relations that a human expert would create. These relations are all as strict as possible. Whenever an equivalent concept is available, an exactMatch relation is created. Only when no equivalent concept is available, a broadMatch or narrowMatch is created. These hierarchical relations are chosen as flat as possible, as explained in section v.4.2. All more diagonal relations can be inferred from these relations. For example, if there is no equivalent for the concept 'car', it would be aligned to 'motorized vehicle' and not to 'vehicle'. If 'motorized vehicle' is a narrow term of 'vehicle' then we can already deduce from that broader/narrower relation and the alignment relation that 'car' also has a broadMatch 'vehicle'. Most of the relations that were found by the SCARLET system were very diagonal while a much flatter correct alignment relation was available. By our strict evaluation method, which measures how close the system's output is to human output and not how close their logical consequences are, nearly no correct relations were found. This is a pessimistic outcome. A more optimistic, and for some use cases perhaps a more realistic, outcome could have been calculated using the Semantic precision and Semantic recall metrics (Euzenat, 2007) instead of the Precision and Recall metrics we used.

v.5.3 Comparison 2006–2007

There were two systems that participated in 2006 and 2007, the RiMOM and Falcon-AO systems. The RiMOM system was changed considerably in the meantime, while the 2007 Falcon-AO system was simply an improved version of the 2006 Falcon-AO system.

Precision The RiMOM system had an unexpected fall in Precision from 2006 to 2007. This was due to bad performance in the taxonomical and miscellaneous sets. The main reason is that in 2007 the RiMOM system also returned many alignment relations that are based on partial lexical matching. Even though many of these partial matches are long or even complete words, for example, 'fat substitutes' has exactMatch 'Caviar substitutes', they are still more often incorrect than correct.

The Falcon-AO system showed a small drop in performance on the taxonomical test set, but a big improvement on the biological and chemical set. This was due to the decision to use edit distance instead of I-Sub for lexical similarity on the *food task*. I-Sub works better for short terms, while edit distance works better for long terms. Most terms in AGROVOC and NALT are quite long. Edit distance is more tolerant to small differences between terms than I-Sub. This allowed matches between chemical terms that only differed by a hyphen or a set of parentheses, like 'parathion-methyl' to 'Parathion methyl', which are common in chemical names. It also allowed matches between species names that are only subtly different, yet refer to completely different species, like 'Helostomatidae' (a fish) to 'Belostomatidae' (a beetle). In general, the Falcon-AO system performed better in 2007 than in 2006 due to improvements in the matching strategy. Apart from bug fixes, a big difference is that the

	RiMOM 2006	Falcon-AO 2006	RiMOM 2007	Falcon-AO 2007
RiMOM 2006	1	0.75	0.48	**0.91**
Falcon-AO 2006		1	0.46	0.74
RiMOM 2007			1	0.50
Falcon-AO 2007				1

Table V.7: Jaccard similarity ($|A \cap B|/|A \cup B|$) between the sets of submitted alignment relations of RiMOM and Falcon-AO in 2006 and 2007. The results of RiMOM 2006 and Falcon-AO 2007 are remarkably similar.

more correspondences based on lexical matches with a high confidence are found the less hard the system try to find additional matches using less dependable matchers, such as its context matcher. This precluded many bad matches to be included in the result set when better lexical matches were already included. The results of this strategy are very similar to those of RiMOM's risk minimization strategy in 2006.

RECALL There was a large Recall improvement by both RiMOM and Falcon-AO. Especially in the animal health and rodents sets. These were the harder sets to produce. Both systems employed a more tolerant lexical matching technique, which led to more matches. The Falcon-AO system was better capable of making the final decision which alignment relations to include in the result set than RiMOM. For example, the simple mapping of 'Rats' with alternative label 'Rattus' to 'Rats', fell outside the final selection of results by RiMOM, but was returned by Falcon-AO.

SYSTEM SIMILARITY The results of the Falcon-AO 2007 system are very similar to those of the RiMOM 2006 system. They are even more similar to the results of the RiMOM 2006 system than to the Falcon-AO 2006 results. Table V.7 shows the similarity between the sets of RiMOM and Falcon-AO for the years 2006 and 2007. The reason for this similarity is an improvement in Falcon-AO's lexical matching algorithm, which makes it very similar to that used by the RiMOM 2006 system. Most of the matches are derived mainly from evidence provided by lexical clues. The other matching strategies, such as Falcon-AO's GMO (structural similarity) or RiMOM's path similarity strategy, are minor sources of evidence. The RiMOM 2007 system focussed on adding extra sources of evidence, which hurt their performance, while the Falcon-AO system learnt of RiMOM's 2006 results and simply fixed the bugs in their lexical matching algorithm.

We expect that the overlapping part of the results of the Falcon-AO 2007 and RiMOM 2006 systems is close to the part of the alignment that can be acquired by means of lexical matching techniques and that the rest of the alignment can not be found using lexical matching techniques. Therefore, without the application of completely different sources of evidence, such as background knowledge in the form of third party ontologies or text mining, the performance of the Falcon-AO 2007 system is representative of the maximum performance one can expect for ontology alignment systems on thesaurus alignment tasks such as the *food task*.

2006					
mapping found by # systems	1	2	3	4	5
average Precision	0.06	0.35	0.67	0.86	0.99
# mappings	21,663	2,592	2,470	4,467	5,555

2007					
mapping found by # systems	1	2	3	4	5
average Precision	0.19	0.81	0.88	0.91	–
# mappings	29,419	7,142	3,944	1,462	0

Table v.8: Consensus: average Precision of the mappings returned by a number of systems.

v.5.4 Consensus

It has to be noted that a potential user of ontology-alignment systems does not necessarily have to limit himself to only one alignment system. Simple ensemble methods such as majority voting can improve Precision. To give an impression of this we list the average Precision of the alignment relations submitted by n systems in table v.8. For $n = 4$ and 5 (*i.e.* the mappings that were returned by 4 out of 5 systems or all of the systems) Precision is significantly higher than for the best system by itself, Falcon-AO in this case. In 2006, nearly all of the 5,555 mappings found by majority voting are correct. Obviously, these are the 'easy' mappings. Whether they are useful or not useful depends on the application of the mappings–if high Precision is more important than high Recall–and remains a topic for future research. In 2007 the result sets of the systems varied much more and thus majority voting worked less well, but still the quality of the alignment relations returned by 3 or more systems is well over that of the best system.

v.6 Analysis

In this section we will discuss a number of issues that limit the performance of alignment systems. Some of these issues are technical and are easy to solve. Others are more fundamental problems that cannot be solved soon if at all.

Inappropriate 'Spelling Correction' Incorrect matches such as nalt:patients skos:exactMatch agrovoc:Patents and nalt:aesthetics skos:exactMatch agrovoc:anaesthetics are caused by inappropriate spelling correction code. In general, tolerating spelling differences in thesauri is not an effective technique, but if it is applied nonetheless it should only be applied when there is no exact literal match. For example, there is a concept representing 'patients' in both thesauri. Recognizing this should trigger a alignment system to refrain from suggesting a mapping to 'patents'. The problem is greater for short terms than for long terms, because edit-distance based measures can be tuned better on long terms than on short terms, because the impact of changing a letter is greater in a short term than in a long term. Changing one letter in a short term changes its lexical shape more and is more likely to cause a difference in meaning than changing one letter in a long term.

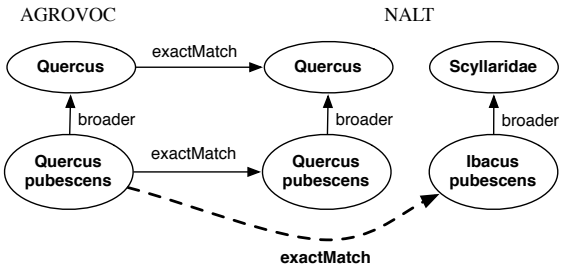


Figure V.8: Failing to recognize the naming scheme can lead to wrong mappings.

Apart from incorrect partial phrase matches, like 'disease reservoirs' to 'water reservoirs', where a partial word overlap is assumed to indicate equivalence, the most common source of mistakes is inappropriate spelling correction. However, especially in the chemical domain, spelling correction also causes a great Recall gain.

Spelling correction should only be applied when the resulting term does not have a distinctly different meaning. A heuristic that could possibly be used to predict this is the comparison of word frequency distributions of the local textual context of the terms in some suitable large corpus of text. Currently, none of the ontology alignment systems implement this technique.

LABELS FOLLOWING NAMING SCHEMES Labels often follow naming schemes. Real-life ontologies often use more than one naming scheme. Both AGROVOC and NALT have a large section on biological taxonomy. The labels of these concepts follow the Linnaeic system of species names. Concepts in other sections of the thesauri (e.g. the sections on geography) follow different schemes. It is vital that lexical matchers recognize that different naming schemes require different matching rules. Perhaps the most common matching rule is postfix matching. This rule states that terms that end in the same word have similar meaning. For instance, 'lime stone' and 'sand stone' are similar. They are both kinds of 'stone'. Two terms from the Linnaeic system that end in the same word, such as 'Quercus pubescens' (a tree) and 'Ibacus pubescens' (a crustacean) are completely dissimilar. Failing to recognize that the Linnaeic system needs prefix matching and not postfix matching can lead to many wrong mappings. The bold arrow in figure V.8 indicates this wrong mapping.

USE AND USE FOR MODELED WITH ALTLABEL When USE is modeled using skos:altLabel the difference between synonyms, obsolete terms, and acknowledgment of lack of detail disappears. In figure V.9 AGROVOC does not include detailed descriptors for the concept nalt:Sigmodon. In fact, a few levels of taxonomical distinctions are left out. The skos:altLabel 'Sigmodon' is added to indicate this omission. It indicates that users that desire to refer to sigmodons should use the agrovoc:c_6633 concept, that symbolizes all rodents. A computer without prior knowledge about this modeling decision cannot distinguish this from synonymy represented with skos:altLabel. This will cause most systems to conclude there is a skos:exactMatch between agrovoc:c_6633 and nalt:Sigmodon, while the proper mapping

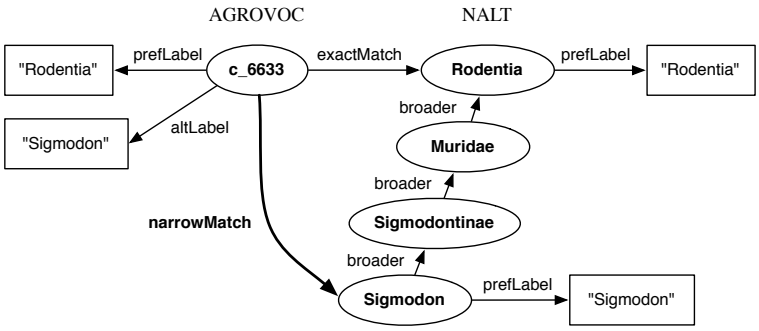


Figure v.9: USE modeled with skos:altLabel in AGROVOC.

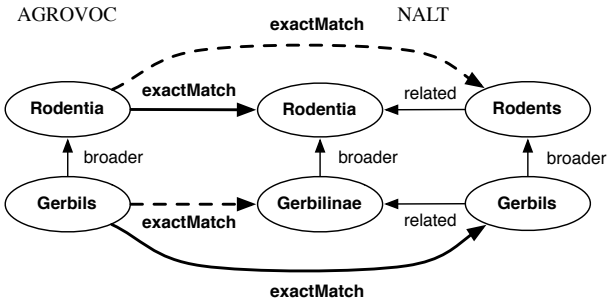


Figure v.10: Separate hierarchies for colloquial names and scientific names.

between these concepts is a skos:narrowMatch.

Vernecular Names and Scientific Names A delicate problem is that of vernacular versus scientific names for the same species. Take the example illustrated in figure v.10 of gerbils with the scientific name 'Gerbilinae'. In NALT, the two types of names each have their own hierarchy, because vernacular names often do not exactly correspond to scientific names. There are Gerbilinae that are not Gerbils (*e.g.* sand rats and jirds), but there is no scientific name for Gerbils. It is also common to have scientific groups that have no vernacular name (*e.g.* nearly all taxonomical terms about bacteria). In AGROVOC the two are combined, because in the indexed documents they both refer to the same actual species. For example, 'Roe deer' BT 'Cervidae' BT 'Ruminants'. A complicating factor is indexing rules. In the AGRIS and AGRICOLA literature reference databases documents are indexed with scientific names whenever the animal or plant in the wild is meant, but the colloquial name is used when the domesticated animal or the product derived from the plant is. For example, 'cows' are domesticated cows, while 'Bos taurus' are wild cows, and 'Zea mays' is the corn plant, while 'maize' or 'corn' is used for the product. The separation of vernacular and scientific names has the advantage that it enables more specific querying of the database, but that query expansion is necessary to find everything about cows or corn. Unification

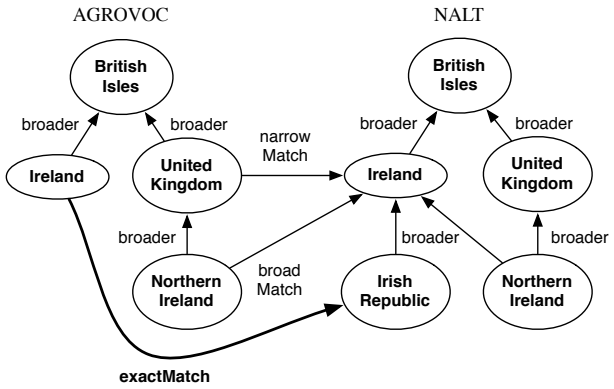


Figure V.11: Concepts representing different senses of a term.

of vernacular and scientific names makes that easier, but makes finding things specifically about the product harder.

Whenever alignment relations are traversed, it is clear that one enters another party's view of the world. The main reason to apply alignment relations is liberal query expansion. This considered, we are lead to believe that in the case of Gerbils there should be skos:exactMatch mappings to both hierarchies in NALT. We created the evaluation samples for Recall based on this assumption. Whether it is the proper treatment depends on the application of the mappings. Depending on the specific indexing rules of the collections, terms can symbolize different views of the concepts or refer to the same extension. This is not limited to species names, but also occurs with, for example, structural formulas of chemicals.

In AGROVOC and NALT this problem is extremely common, because the largest part of the thesauri deals with species names.

CLASHING SENSES Sometimes all is not what it seems. Seemingly obvious mappings can be wrong. Consider 'Ireland' and the 'British Isles' in figure V.11. The British Isles can be partitioned in two ways, administrative and geographical. Respectively, the Irish Republic and the United Kingdom; or Ireland and the other islands of the British Isles, which all belong to the United Kingdom.

A natural intuition of people is the assumption that sibling concepts are disjoint. Therefore, if the distinction is made between Ireland and the United Kingdom, the most obvious interpretation is the administrative case. Even though in itself the word 'Ireland' is more likely to refer to the island that to the nation, which is officially named the 'Irish republic', people will immediately default to the nation. The lack of a broader relation between agrovoc:Northern Ireland and agrovoc:Ireland further supports their choice. Another natural intuition is that narrower concepts are strictly narrower than (*i.e.* not equivalent to) their parents. This means that the existence of the concept nalt:Irish Republic makes people assume that nalt:Ireland refers to the entire island. The narrower concept nalt:Northern Ireland confirms

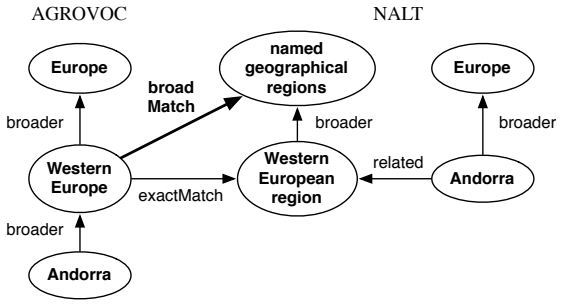


Figure v.12: The is no evidence in the thesauri for this skos:broadMatch.

this. In the example this means that agrovoc:Ireland should be equivalent to nalt:Irish Republic.

In this case, a computer could solve this problem if a few OWL statements were added that proclaim siblings to be disjoint and broader concept to be not equivalent to narrower concepts. This kind of approach, however, is likely to cause more harm than good in the entire thesaurus. Thesaurus concepts are inherently vague and such a strict interpretation often causes unintentional inconsistencies. A technique that uses the added axioms as heuristics might be more suitable.

Obviously, the *Colloquial names and scientific names* issue, described previously, is also an example of clashing senses. Hence, this issue is a common phenomenon. There might not be as many geographical concepts as taxonomical concepts, but in applications geographical concepts are amongst the most commonly used concepts. Many fielded or facetted search clients support geographical selection of resources. Some data sets are better separated by nation (*e.g* legal documents), others are better served by a geographical separation (*e.g.* weather data).

No Evidence in the Thesauri for Some Correct Mappings In many cases it is simply impossible to find certain mappings without resorting to external knowledge sources, such as a third ontology, concrete domain reasoning, text mining, or traditional knowledge acquisition. An example of a mapping that is impossible to find is shown in figure v.12. Western Europe is clearly a named geographical region, but the skos:broader relation between nalt:Western European region and nalt:named geographical regions alone is not enough evidence to suggest this. AGROVOC contains no concepts that are lexically similar to the latter NALT concept.

An other example is: nalt:cytoplasmic polyhedrosis virus skos:broadMatch agrovoc:Reoviridae. None of the broader or narrower concepts have any lexical similarities, yet the mapping is sound. A search query on the MedLine collection with PubMed Central[20] reveals many articles that mention the relation. An excerpt from one of these articles that gives evidence for the mapping is: *"Cytoplasmic polyhedrosis viruses (CPVs) belong to the genus Cypovirus in the family Reoviridae (13, 36)."* (Ikeda et al., 2001)

[20] http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=113995

This situation is common outside of areas where there is high consensus on the jargon (*e.g.* the taxonomical, geographical, or anotomical domains) and in the more general areas of the thesauri, *i.e.* near to the top concepts. In some areas (*cf.* the animal health Recall sample) alignments that require some degree of background knowledge are even the majority. The current ontology alignment systems, and even humans for that matter, have great difficulty to find these hard alignment relations. Therefore, the true magnitude of the problem is hard to quantify.

USEFUL BROADMATCH AND NARROWMATCH RELATIONS ARE HARD TO FIND The SCARLET system found thousands of hierarchical relations. A large part of these relations was correct, yet Recall scores on our samples are extremely low. This means that these relations are not the kind of relations domain experts would assert, even if many of them are not strictly false. Most of these relations are undercommitments. An example is the relation nalt:technology skos:narrowMatch agrovoc:Diesel engines. It appeared in the Precision sample for the SCARLET system and was judged to be true, but it would never appear in a manually constructed Recall sample. A thesaurus editor would always try to find the strictest relation that does not overcommit. AGROVOC has a concept agrovoc:Technology and NALT has a concept nalt:diesel engines. These two concepts would provide stricter matches and hence the alignment between nalt:technology and agrovoc:Diesel engines would never be asserted.

Whether undercommitments are a big issue depends on the application. If the only thing that matters for an application are the top concepts (*e.g.* for a rough separation of documents into topic categories) then undercommitments are no problem. If the alignment is used for delicate query expansion then undercommitments are nearly useless.

V.7 DISCUSSION

From this work we can draw conclusions on various levels: The specific challenges of thesaurus alignment in the agricultural domain, the importance of certain features for the quality of alignment systems in such tasks, the particularities of the evaluation of thesaurus alignment relations of various types, and in which cases thesaurus alignment can be automated with good results.

SPECIFIC CHALLENGES OF THE FOOD TASK The main challenges for alignment systems in the OAEI 2006 and 2007 *food task* were the following:

- Compared to the data sets of the other OAEI tasks, AGROVOC and NALT are large. Only systems that could deal with the size of the AGROVOC and NAL thesauri (*e.g.* by correctly partitioning the data sets) could achieve satisfactory results. Some groups did not participate in the *food task*, because their systems were unable to load the thesauri. Most systems attempted some kind of partitioning. The quality of the partitioning turned out to be one of the decisive factors for overall system performance, for example the difference between Falcon-AO and X-SOM in table V.5 is partially caused by the different partitioning strategies of the systems.

- Only systems that were able to deal with the relatively weak semantic structure of thesauri could do well. Whereas most OWL ontologies have one label per class and a number of property restrictions, most SKOS thesauri have many labels, but lack property restrictions. This means systems could not rely on description-logic reasoning and were required to do term disambiguation. The systems that had the best lexical matching strategies (RiMOM 2006 and Falcon 2006 and 2007) performed significantly better than systems that focussed more on other facets.

- Both thesauri contain concepts from many different domains. Only systems that were able to do proper lexical analysis of labels that use various naming conventions could avoid large numbers of mistakes. Some systems did very well in some domains, but very poorly in other domains, for example, X-SOM did very well in the geographical and biochemical domain, but very poorly in the taxonomical domain. Systems that performed well overall were the clear winners.

Conclusions of the Qualitative Analysis The two most important conclusions of the qualitative analysis of the OAEI 2006 and 2007 *food task* results are:

- Within one thesaurus there can be many different kinds of labels (*e.g.* scientific names of species, structure formula's of molecules, named entities, medical terminology of various kinds, diverse types of jargon, etc.) Being able to deal with various naming schemes used in labels is, by far, the most important quality of a thesaurus alignment system.

- There is idiom in thesauri, 'abuse' of the semantic features. For example, USE / skos:altLabel is sometimes used to indicate missing detail (see section v.6, page 72), that RT / skos:related usually also implies disjointness, and that BT / skos:broader should usually be considered as strictly broader. Alignment systems can gain by exploiting these rules.

Strict and Relaxed Evaluation Methods For the evaluation of the alignments in the OAEI *food task* we chose to draw samples. Each sample alignment relation was either verified individually or constructed individually for the measurement of respectively Precision and Recall.

For the evaluation of broadMatch and narrowMatch relations there is a discrepancy between how we measured Precision and Recall. The correctness criterion for Precision could be summarized as: "Is the relation valid?", while the criterion for Recall could be summarized as: "Is the relation valid and as strict as possible?". The intuition behind the current Precision assessment criterion corresponds to that of Semantic precision, while the intuition behind the Recall criterion corresponds to strict Recall. We could have assessed Precision in the same strict way as we used for Recall or Recall in the same relaxed way as we used for Precision to get respectively a lower bound or an upper bound on the performance scores. This could have been accomplished by using either the current evaluation method for Precision and Semantic recall for Recall (relaxed), or a stricter criterion for Precision and the current evaluation method for Recall (strict).

The reason why we did not do this is a pragmatic one. We wanted to perform the exact same evaluation procedure for the *food task* in 2007 as we did in 2006. All of the systems in 2006 were only able to find exactMatch relations and for the evaluation of exactMatch relations there is no discrepancy between the current evaluation methods for Precision and Recall, because these relations are never an element of any other alignment relation's logical consequence. There are no equivalence relations in the logical consequence of an equivalence relation, only hierarchical relations.

A similar problem occurs in the evaluation of XML retrieval systems that are allowed to return nested parts of documents. These systems have to decide whether they should return the entire section, or only the most relevant paragraphs. A strict evaluation method states that only the most relevant elements (the smallest element containing all relevant information) should be returned. A relaxed evaluation method states that all enveloping elements or even contained elements can also be returned. The INEX evaluation (Kazai et al., 2004) initiative has experimented quite extensively with different combinations of strict and relaxed evaluation methods.

APPLICATION OF THESAURUS ALIGNMENT Two important factors that determine how useful automatic ontology alignment can be in practice are the domains covered by the thesauri and the desired reliability of the results.

As we can see in table V.5 and V.6, some domains are more easily aligned automatically than others. The geography domain, for instance, is an easy domain. The Falcon-AO 2007 system was able to find more than 90% of all exactMatch relations. Domains concerning roles of objects where there are different perspectives on the same objects are hard. An example is the category animal health (see table V.6) where you have mappings between, for instance, flukes as a species of worms and flukes as a kind of parasites. Or in the category rodents there are mappings between mice as a species and mice as a pest. The best systems were only able to find about 60% of the exactMatch relations and around 20% of all relations (see table V.6).

The fact that 90% of all equivalence relations between geographical terms can be found automatically by itself does not mean that it is always a wise decision to automate the alignment process for geographical terms. If you are dealing with an application where subtle differences are important, like the status of Northern Ireland or Montenegro, it is probably a better idea to construct the entire geographical alignment by hand. In many cases, this is feasible, considering the relatively small number of alignment relations in the geographical domain (as compared to, for example, the taxonomical domain). Judging by our experience with the OAEI 2006 and 2007 *food task*, we estimate that the verification of alignment relations consumes roughly 5 times less time than searching for the alignment relations by hand without suggested relations. So in some cases where Recall is high complete manual verification of an automatically-created alignment can potentially save time.

We presented a quantitative and qualitative evaluation of thesaurus-alignment techniques. Thesauri might be relatively weak semantic structures, yet they are widespread and used for a multitude of tasks in various contexts. This very versatility is what makes the evaluation of thesaurus alignment complicated. Ideally, every task gets its own evaluation method that takes into account its specific properties. For example, the evaluation of a

classification task would use stricter measures than that of a browsing or recommendation task. As opposed to picking a number of different measures for different tasks we chose to pick a neutral evaluation measure. We complemented this quantitative evaluation with an in-depth qualitative analysis discussing the inherent strengths and weaknesses of the various alignment techniques employed by the systems.

Acknowledgements

We would like to thank the NAL and FAO for allowing us to use their thesauri and for the time and resources they committed to this work. Our special gratitude goes to everybody that helped with the assessment of the alignment samples: Gudrun Johannsen and Caterina Caracciolo at the FAO, Nicole Koenderink, Hajo Rijgersberg, and Lars Hulzebos at the Wageningen Agricultural University, Fred van de Brug and Marco Bouman at TNO Quality of life, and Evangelos Alexopoulos at Unilever, and everybody at the EKAW 2006 workshop who took part in the inter-judge agreement experiment. Furthermore, we would like to thank the participants of the ECOTERM 2007 workshop for valuable discussions. This work was partly supported by the Dutch BSIK project Virtual Laboratory for e-science (`http://www.vl-e.nl`).

v.8 Impendix – The OAEI 2007 Environment Task

The *environment task* is comprised of three alignments between three thesauri: the two thesauri of the *food task* (AGROVOC and NALT), and the European Environment Agency thesaurus, GEMET. The participants were allowed to the third thesaurus as background knowledge to align the other two for the construction of any of the three alignment.

v.8.1 Test Set

The task of this case consists of matching three thesauri formulated in SKOS:

GEMET The European Environment Agency (EEA) GEneral Multilingual Environmental Thesaurus, version July 2007. This thesaurus consists of 5,298 concepts, each with descriptor terms in all of its 22 languages (bg, cs, da, de, el, en, en-US, es, et, eu, fi, fr, hu, it, nl, no, pl, pt, ru, sk, sl, sv).

AGROVOC The United Nations Food and Agriculture Organization (FAO) AGROVOC thesaurus, version February 2007. This thesaurus consists of 28,445 descriptor terms, i.e., prefered terms, and 12,531 non-descriptor terms, i.e., alternative terms. It is multilingual in eleven languages (en, fr, de, es, ar, zh, pt, cs, ja, th, sk).

NALT The United States National Agricultural Library (NAL) Agricultural thesaurus, version 2007. This thesaurus consists of 42,326 descriptor terms and 25,985 non-descriptor terms. NALT is monolingual, English.

Participants had to match these SKOS versions of GEMET, AGROVOC and NAL using the exactMatch, narrowMatch, and broadMatch relations from the SKOS Mapping Vocabulary.

v.8.2 Evaluation Procedure

The evaluation procedure used is the same as for the *food task* with the exception that we used slightly different categories of sample topics.

Precision

For the evaluation of Precision for the GEMET-AGROVOC and GEMET-NALT alignments we distinguished six categories of topics in the thesauri that each required a different level of domain knowledge of the assessors: Taxonomical concepts (plants, animals, bacteria, etc.), biological and chemical terms (structure formulas, terms from generics, etc.), geographical terms (countries, regions, etc.), natural resources (fishery, forestry, agriculture, mining, etc.), health risk management (pollution, food, air, water, disasters, etc.), and the remaining concepts (administration, materials, military aspects, etc.). The results for the NALT-AGROVOC are shown in the section about the food alignment task. The sizes of the categories and the part that was assessed are shown in table v.9.

	GEMET-AGROVOC		GEMET-NALT	
topic	# alignments	# assessed	# alignments	# assessed
taxonomical	500	39	802	33
biological / chemical	541	43	841	51
geographical	167	40	164	39
natural resources	412	51	450	39
health risk management	602	38	738	52
miscellaneous	1884	48	1988	51

Table v.9: Categories of alignments that were separately assessed for the estimation of Precision.

Recall

For the evaluation of recall we used a set of sub-hierarchies of the thesauri about: Concepts from agriculture in the broad sense of the word including fishery (fishing equipment, aquaculture methods, etc.) and animal husbandry (animal diseases, animal housing, etc.), and geological concepts like countries and place types (the Baltic states, alluvial plains, etc.), The sizes of the samples are shown in table v.10, along with the percentage of exactMatch alignments in each sample.

	GEMET-AGROVOC		GEMET-NALT	
topic	# alignments	% exactMatch	# alignments	% exactMatch
agriculture	89	69%	92	66%
geology	136	64%	138	56%

Table v.10: Reference alignments that were used for the estimation of Recall.

v.8.3 Results

Two systems took part in the OAEI 2007 environment alignment task, South East University (Falcon-AO 0.7) and the Knowledge Media Institute (DSSim). Both systems returned only exactMatch alignments. Table v.11 shows the number of alignments the two systems returned for each of the three alignments.

	# alignments		
system	NALT-AGROVOC	GEMET-AGROVOC	GEMET-NALT
Falcon-AO	15,300	1384	1374
DSSim	14,962	3030	4278

Table v.11: Number of alignments that were returned by the participating systems

Best Precision The GEMET thesaurus is very shallow compared to the AGROVOC and NALT thesauri, but it does offer definitions and labels in many languages. This means that there is very little information for the matching systems to reason with. That means

lexical comparison is usually the only thing that the alignment system can exploit. The Falcon-AO system performed best at both tasks, achieving a similar Precision as with the easier NALT-AGROVOC alignment. An overview of all the Precision results is shown in table V.12.

	GEMET-AGROVOC		GEMET-NALT	
Precision for	Falcon-AO	DSSim	Falcon-AO	DSSim
taxonomical	0.95	0.27	0.87	0.16
bio/chem	0.54	0.00	0.88	0.53
geographical	1.00	0.30	0.77	0.29
natural resources	1.00	0.53	0.95	0.32
health risk man.	0.95	0.38	0.88	0.50
miscellaneous	0.90	0.39	0.82	0.53
overall	**0.88**	**0.33**	**0.86**	**0.44**

Table V.12: Precision results based on sample evaluation.

BEST RECALL The Falcon-AO system performs significantly better than the DSSim system on the GEMET-AGROVOC and GEMET-NALT alignments. However, it does not achieve similar Recall scores as for the NALT-AGROVOC alignment.

	GEMET-AGROVOC		GEMET-NALT	
Recall for	Falcon-AO	DSSim	Falcon-AO	DSSim
agriculture	0.43 (0.62)	0.11 (0.16)	0.36 (0.54)	0.16 (0.25)
geology	0.37 (0.59)	0.18 (0.29)	0.26 (0.47)	0.17 (0.30)
overall	**0.39 (0.60)**	**0.15 (0.24)**	**0.30 (0.50)**	**0.16 (0.27)**

Table V.13: Recall results based on sample evaluation. The numbers between parentheses show Recall when only the exactMatch alignments of the reference alignments are considered.

Chapter VI

Evaluation Methods for Ontology Alignment

In this chapter we describe two methods for the evaluation of alignment approaches, alignment sample evaluation *and* end-to-end evaluation. *The former measures the quality of the alignment itself, the latter measures its effect on an application.* Alignment sample evaluation *is applied in chapter II, III, and V, and it is used as part of the* relevance-based evaluation *method described in chapter VII.* End-to-end evaluation *is applied in chapter III to determine Recall.*

This chapter is based on a paper coauthored by Antoine Isaac and Zharko Aleksovski, "Sample Evaluation of Ontology-Matching Systems, Willem Robert van Hage, Antoine Isaac, Zharko Aleksovski" *(van Hage et al., 2007), which was presented at the fifth International Workshop on the Evaluation of Ontologies and Ontology-based tools (EON 2007). The order of the sections was adapted to better suit the content of this thesis.*

Abstract Ontology matching exists to solve practical problems. Hence, methodologies to find and evaluate solutions for ontology matching should be centered on practical problems. In this chapter we propose two statistically-founded evaluation techniques to assess ontology-matching performance that are based on the application of the alignment. Both are based on sampling. One examines samples of mappings, the other the behavior of an alignment in use. We show the assumptions underlying these techniques and describe their limitations.

VI.1 Introduction

In the context of the Semantic Web project an overwhelming number[1] of ontologies have been published on the web. Cross-referencing between these ontologies by means of ontology matching is now necessary. Ontology matching has thus been acknowledged as one of the most urgent problems for the community, and also as one of the most scientifically challenging tasks in semantic-web research.

Consequently, many matching tools have been proposed, which is a mixed blessing: *comparative* evaluation of these tools is now required to guide both ontology-matching research and application developers in search of a solution. One such effort, the Ontology Alignment Evaluation Initiative[2] (OAEI) provides a collaborative comparison of state-of-

[1] `http://swoogle.umbc.edu` indexes over 10,000 ontologies by 2007.
[2] `http://oaei.ontologymatching.org`

the-art mapping systems which has greatly accelerated the development of high-quality techniques. The focus of the OAEI has been mainly on comparing mapping techniques for research.

Good evaluation of ontology-matching systems takes into account the purpose of the alignment.[3] Every application has different requirements for a matching system. Some applications use rich ontologies, others use simple taxonomies. Some require equivalence correspondences, others subsumption or even very specific correspondences such as artist-style or gene-enzyme. Also, the scope of concepts and relations is often determined by unwritten application-specific rules (*cf.* Šváb et al. (2007)). For example, consider the subclass correspondence between the concepts Gold and Jewelry. This correspondence holds if the scope of Gold is limited to the domain of jewelry. Otherwise the two would just be related terms. In either case, application determines relevance.

The best way to evaluate the quality of an alignment is trough extensive practical use in real-world applications. This, however, is usually not feasible. The main reason for this is usually lack of time (*i.e.* money). Benchmarks and experiments using synthesized ontologies can reveal the strengths and weaknesses of ontology-matching techniques, but disregard application-specific requirements. Therefore, the second best option is to perform an evaluation that mimics actual usage. Either by performing a number of typical usage scenarios or by specifying the requirements an application has for the alignment and then testing whether these requirements are met. The final measure for system performance in practice is user satisfaction. For the evaluation of matching systems, this means that a set of correspondences is good if users are satisfied with the effect the correspondences have in an application.

Most current matching evaluation metrics simulate user satisfaction by looking at a set of assessed correspondences. For example, Recall expresses how many of the assessed correspondences are found by a system. This has two major problems. (1) Some correspondences have a larger logical consequence than others. That is to say, some correspondences subsume many other correspondences, while some only subsume themselves. This problem is addressed quite extensively by Euzenat (2007) and Ehrig and Euzenat (2005). (2) Correct correspondences do not automatically imply happy users. The impact of a correspondence on system performance is determined not only by its logical consequence, but also by its relevance to the user's information need. A correspondence can be correct and have many logical implications, but be irrelevant to the reasoning that is required to satisfy the user. Also, some correspondences have more impact than others.

In the following sections we propose two alternative approaches to include relevance into matching evaluation, one based on *alignment sample evaluation* (Sec. VI.2), and one based on *end-to-end evaluation* (Sec. VI.4). Both approaches use sample evaluation, but both what is sampled and the sample selection criteria are different. The former method uses sample sets of correspondences which are selected in such a way that they represent different requirements of the alignment. The latter uses sample queries, disregarding the alignment itself, and hence providing objectivity. We investigate the limitations of these statistical techniques and the assumptions underlying them. Furthermore, we calculate

[3] In this chapter we use the definitions as presented in Euzenat and Shvaiko (2007): An ontology matching system produces a set of correspondences called an alignment.

upper bounds to the errors caused by the sampling. Finally, in Sec. VI.3 we will demonstrate the workings of the latter of the two evaluation methods in the context of the OAEI 2006 *food task*.

VI.2 Alignment Sample Evaluation

This evaluation approach is based on the assessment of the alignment itself. In practice, it is often too costly to manually assess all the correspondences. A solution to this problem is to take a small *sample* from the whole set of correspondences (Cochran, 1977). This set is manually assessed and the results are generalized to estimate system performance on the whole set of correspondences. As opposed to the elegant abstract way of evaluating system behavior provided by *end-to-end evaluation*, *alignment sample evaluation* has many hidden pitfalls. In this section we will only investigate the caveats that are inherent to sample evaluation. We will not consider errors based on non-sampling factors such as judgement biases, peculiarities of the ontology-matching systems or ontologies, and other unforeseen sources of evaluation bias.

Simple Random Sampling

p	true proportion of the samples produced that is correct (unknown)
n	number of sample correspondences used to approximate p
$\hat{P}$	approximation of p based on a sample of size n
δ	margin of error of $\hat{P}$ with 95% confidence

The most common way to deal with this problem is to take a small *simple random sample* from the whole set of correspondences. Assessing a set of correspondences can be seen as classifying the correspondences as *Correct* or *Incorrect*. We can see the output of a matching system as a *Bernoulli random variable* if we assign '1' to every *Correct* correspondence and '0' to each *Incorrect* correspondence it produces. The true Precision of a system is the probability with which this random variable produces a '1', p. We can approximate this p by the proportion of ones in a *simple random sample* of size n. With a confidence of 95% this approximation, $\hat{P}$, lies in the interval:

$$\hat{P} \in [p - \delta, p + \delta] \quad \text{where} \quad \delta = \frac{1}{\sqrt{n}} \tag{VI.1}$$

The variance of $\hat{P}$ can be approximated with:

$$VAR(\hat{P}) \approx \frac{\hat{P}(1-\hat{P})}{n}$$

Both Precision and Recall can be estimated using samples. In the case of Precision we take a random sample from the output of the matching system, *Found* in Fig. VI.1. In this figure the sample for Precision is illustrated as $B \cup C$. The results for this sample can be generalized to results for the set of all *Found* correspondences. In the case of Recall we

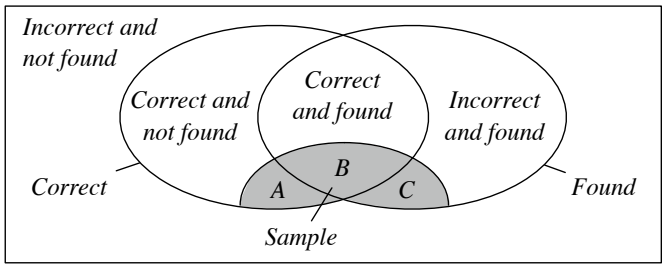
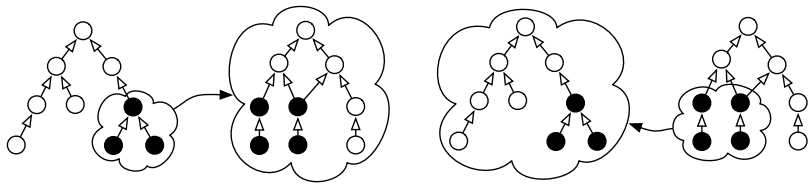


Figure VI.1: Venn diagram to illustrate sample evaluation. $A \cup B$ is a sample of the population of Correct correspondences. $B \cup C$ is a sample of the population of Found correspondences.



mapping from ontology X to ontology Y mapping from ontology Y to ontology X

Figure VI.2: Concepts to consider when creating a sample for Recall evaluation based on a topic. Black concepts are 'on topic', white concepts 'off topic'. For example, the black concepts have something to do with steam engines and the white concepts do not. Concepts to consider for sample correspondences are marked by clouds. This avoids bias against cross-topic correspondences.

take a random sample from the set of all correct correspondences, *Correct* in Fig. VI.1. The sample for Recall is illustrated as $A \cup B$. The results for this sample can be generalized to results for the set of all *Correct* correspondences.

A problem with taking a random sample from all *Correct* correspondences is it is unknown which correspondences are correct and which are incorrect a priori. A proper random sample can be taken by randomly selecting correspondences between all possible correspondences between concepts from the two aligned ontologies, *i.e.* a subset of the cartesian product of the sets of concepts from both ontologies. Each correspondence has to be judged to filter out all incorrect correspondences. This can be very time-consuming if there are relatively few valid correspondences in the cartesian product. The construction time of the sample of correct correspondences can be reduced by only judging parts of the ontologies that have a high topical overlap. For example, one can only consider all correct mappings between concepts having to do with steam engines. (*cf. e.g.* Wang et al. (2007)) It is important to always match concepts about a certain topic in ontology X to *all* concepts in ontology Y, and all concepts about the same topic in ontology Y to *all* concepts in ontology X. This is illustrated in Fig. VI.2. This avoids a bias against correspondences to concepts outside the sample topic.

There are two caveats when applying this approximation method. (1) A sample of correct mappings constructed in this way is arbitrary, but not completely random. Correspondences in the semantic vicinity of other correspondences have a higher probability of being selected than 'loners'. This means ontology matching techniques that employ structural aspects of the ontologies are slightly advantaged in the evaluation. (2) The method works under the assumption that correspondences inside a topic are equally hard to derive as correspondences across topics.

Stratified Random Sampling

N	size of the entire population, *e.g.* the set of all correct correspondences
h	one stratum of the entire population
N_h	size of stratum h
n_h	number of sample correspondences used to approximate p of stratum h
$\hat{P}_h$	approximation of p for the correspondences in stratum h

A better way than *simple random sampling* to perform sample evaluation is *stratified random sampling*. In stratified sampling, the population (*i.e.* the entire set of correspondences used in the evaluation) is first divided into subpopulations, called *strata*. These strata are selected in such a way that they represent parts of the population with a common property. Useful distinctions to make when stratifying a set of correspondences are: different alignment relations (*e.g.* equivalence, subsumption), correspondences in different domains (*e.g* cats, automobiles), different expected performance of the matching system (*e.g.* hard and easy parts of the alignment), or different levels of importance to the use case (*e.g* mission critical versus nice-to-have). The strata form a partition of the entire population, so that every correspondence has a non-zero probability to end up in a sample. Then a sample is drawn from each stratum by *simple random sampling*. These samples are assessed and used to score each stratum, treating the stratum as if it were an entire population. The approximated proportion and margin of error can be calculated with *simple random sampling*.

Stratified random sampling for the evaluation of alignments has two major advantages over simple random sampling. (1) The separate evaluation of subpopulations makes it easier to investigate the conditions for the behavior of matching techniques. If the strata are chosen in such a way that they distinguish between different usages of the correspondences, we can draw conclusions about the behavior of the correspondences in a use case. For example, if a certain matching technique works very well on chemical concepts, but not on anatomical concepts, then this will only come up if this division is made through stratification. (2) Evaluation results for the entire population acquired by combining the results from stratified random sampling are more precise than those of simple random sampling. With simple random sampling there is always a chance that the sample is coincidentally biased against an important property. While every property that is distinguished in the stratification process will be represented in the sample.

The results of all the strata can be combined to one result for the entire population by weighing the results by the relative sizes of the strata. Let N be the size of the entire population and $N_1, \ldots, N_L$ the sizes of strata 1 to L, so that $N_1 + \cdots + N_L = N$. Then the

weight of stratum h is N_h/N. Let n_h be the size of the *simple random sample* in stratum h and $\hat{P}_h$ be the approximation of proportion p in stratum h by the sample of size n_h. We do not require the sample sizes $n_1, \ldots, n_L$ to be equal, or proportional to the size of the stratum. The approximated proportion in the entire population, $\hat{P}$, can be calculated from the approximated proportions of the strata, $\hat{P}_h$, as follows:

$$\hat{P} = \frac{1}{N} \sum_{h=1}^{L} N_h \hat{P}_h$$

The variance of $\hat{P}$ can be approximated by

$$VAR(\hat{P}) \approx \sum_{h=1}^{L} \frac{\hat{P}(1-\hat{P})}{n_h} \cdot \frac{N_h - n_h}{N}$$

Due to the fact that the variance of the binomial distribution is greatest at $p = 0.5$, we know that the greatest margin-of-error occurs when $\hat{P} = 0.5$. That means that with a confidence of 95% the approximation of $\hat{P}$ lies in the interval:

$$\hat{P} \in [p - \delta, p + \delta] \quad \text{where} \quad \delta = \frac{1}{\sqrt{N}} \sqrt{\sum_{h=1}^{L} \left(\frac{N_h}{n_h} - 1\right)} \tag{VI.2}$$

Comparative Alignment Sample Evaluation

p_A	true proportion of the correspondences produced by system A that is correct (unknown)
$\hat{P}_A$	sample approximation of p_A
$\hat{P}_{A,h}$	$\hat{P}_A$ in stratum h

To compare the performance of two systems, A and B, using sample evaluation, we calculate their respective $\hat{P}_A$ and $\hat{P}_B$ and check if their margins of error overlap. If this is not the case, we can assume with a certain confidence that p_A and p_B are different, and hence that one system is significantly better than the other. For *simple random sampling* this can be calculated as follows:

$$|\hat{P}_A - \hat{P}_B| > 2\sqrt{\frac{\hat{P}_A(1-\hat{P}_A)}{n} + \frac{\hat{P}_B(1-\hat{P}_B)}{n}} \tag{VI.3}$$

For *stratified random sampling* this can be calculated as follows:

$$|\hat{P}_A - \hat{P}_B| > 2\sqrt{\sum_{h=1}^{L} \frac{\hat{P}_{A,h}(1-\hat{P}_{A,h})}{N}\left(\frac{N_h}{n_h} - 1\right) + \sum_{h=1}^{L} \frac{\hat{P}_{B,h}(1-\hat{P}_{B,h})}{N}\left(\frac{N_h}{n_h} - 1\right)} \tag{VI.4}$$

For both methods the maximum difference needed to distinguish P_A from P_B with a confidence of 95% is $2/\sqrt{2n}$. So if, depending on the type of sampling performed, equation (VI.3) or (VI.4) holds, there is a significant difference between the performance of system A and B.

VI.3 Alignment Sample Evaluation in Practice

In this section we will demonstrate the effects of *alignment sample evaluation* in practice by applying *stratified random sampling* on the results of the OAEI 2006 *food task*[4] for the estimation of Precision and we will calculate the margin of error caused by the sampling process.

The OAEI 2006 *food task* is a thesaurus alignment task between the Food and Agriculture Organisation of the United Nations (FAO) AGROVOC thesaurus and the thesaurus of the United States Department of Agriculture (USDA) National Agricultural Library (NAL). Both thesauri are supplied to participants in SKOS and OWL Lite[5]. The alignment had to be formulated in SKOS Mapping Vocabulary[6] and submitted in the common format for alignments[7]. A detailed description of the OAEI 2006 *food task* can be found in Euzenat et al. (2006); Shvaiko et al. (2007).

Five teams submitted an alignment: Falcon-AO, COMA++, HMatch, PRIOR, and RiMOM. Each alignment consisted only of one-to-one semantic equivalence correspondences. The size of the five alignments is shown below. The number of unique *Found* correspondences was 31,112. The number of *Correct* correspondences can be estimated in the same order of magnitude. In our experience, voluntary judges can only reliably assess a few hundred correspondences per day. That means this means assessing all the *Found* correspondences in the alignments would already take many judges a few weeks of full-time work. This is only feasible with significant funding. Thus, we performed a sample evaluation.

system	RiMOM	Falcon-AO	Prior	COMA++	HMatch	all systems
# *Found*	13,975	13,009	11,511	15,496	20,001	31,112

During a preliminary analysis of the results we noticed that the performance of the different systems was quite consistent for most topics, except correspondences between taxonomical concepts (*i.e.* names of living organisms such as 'Bos Taurus') with latin names where some systems performed noticeably worse than others. This was very surprising given that there was a straightforward rule to decide the validity of a taxonomical correspondence, due to similar editorial guidelines for taxonomical concepts in the two thesauri. Two concepts with the same preferred label and some ancestors with the same preferred label are equivalent. Also, when the preferred label of one concept is literally the same as the alternative label of the other and some of their ancestors have the same preferred label they are equivalent. For example, the African elephant in AGROVOC has a preferred label 'African elephant' and an alternative label 'Loxodonta africana'. In NALT it is the other way around.

These rules allowed us to semi-automatically assess the taxonomical correspondences. This was not possible for the other correspondences. So we decided to separately evaluate correspondences from and to taxonomical concepts. We also noticed that most other correspondences were very easy to judge, except correspondences between biochemical

[4] http://www.few.vu.nl/~wrvhage/oaei2006

[5] The conversion from SKOS to OWL Lite was provided by Wei Hu.

[6] http://www.w3.org/2004/02/skos/mapping/spec

[7] http://oaei.ontologymatching.org/2006/align.html

concepts (*e.g.* 'protein kinases') and substance names (*e.g.* 'tryptophan 2,3-dioxygenase'). These required more than a layman's knowledge of biology or chemistry. So we decided to also evaluate biological and chemical concepts separately, with different judges. This led to three strata: taxonomical correspondences, biological and chemical correspondences, and the remaining correspondences. The sizes of the strata, along with the size of the evaluated part of the stratum and the corresponding stratum weights are shown below. Precision

stratum topic	stratum size (N_h)	sample size (n_h)	stratum weight (N_h/N)
taxonomical	18,399	18,399	0.59
biological and chemical	2,403	250	0.08
miscellaneous	10,310	650	0.33
all strata	31,112	21,452	

estimates using these strata have a maximum margin of error of:

$$2 \cdot \sqrt{\frac{0.5 \cdot (1-0.5)}{31112} \cdot \left(\left(\frac{18399}{18399} - 1 \right) + \left(\frac{2403}{250} - 1 \right) + \left(\frac{10310}{650} - 1 \right) \right) \cdot 2} \approx 3.8\%$$

at a confidence level of 95%. That means that, under the assumption that there are no further biases in the experiment, a system with 82% Precision outperforms a system with 78% Precision with more than 95% confidence.

If, for example, we are interested in the performance of a system for the alignment of biological and chemical concepts and use the sample of 250 correspondences to derive the performance on the entire set of 2,403 correspondences our margin of error would be $1/\sqrt{250} \approx 6.3\%$. Comparison of two systems based on only these 250 sample biological and chemical correspondences gives results with a margin of error of $2/\sqrt{2 \cdot 250} \approx 8.9\%$. That means with a confidence level of 95% we can distinguish a system with 50% Precision from a system with 59% Precision, but not from a system with 55% Precision.

VI.4 End-to-end Evaluation

An alternative approach is *end-to-end evaluation*. This approach is completely system-performance driven, based on a sample set of representative information needs. The performance is determined for each trial information need, using a measure for user satisfaction. For example, such an information need could be *"I would like to read a good book about the history of steam engines."* and one could use F-score or the Mean-Reciprocal Rank[8] of the best book in the result list, or the time users spent to find an answer. The set of trials is selected such that it fairly represents different kinds of usage, *i.e.* more common cases receive more trials. Real-life topics should get adequate representation in the set of trials. In practice the trials are best constructed from existing usage data, such as log files of a baseline system. Another option is to construct the trials in cooperation with domain experts. A concrete example of an end-to-end evaluation is described by Voorhees and Tice (2000). In their paper, Voorhees and Tice explicitly describe the topic construction method and the

[8]One over the rank of the best possible result, *e.g.* 1/4 if the best result is the fourth in the list.

measure of satisfaction they used for the end-to-end evaluation of the TREC-9 question-answering track. The size and construction methods of test sets for end-to-end retrieval have been investigated extensively in the context of information retrieval evaluation initiatives such as TREC (Voorhees, 1998), CLEF, and INEX[9]. When all typical kinds of usage are fairly represented in the sample set, the total system performance can be acquired by averaging the scores.[10] To evaluate the effect of an ontology alignment, one usually compares it to a baseline alignment in the context of the same information system. By changing the alignment while keeping all other factors the same, the only thing that influences the results is the alignment. The baseline alignment can be any alignment, but a sensible choice is a trivial alignment based only on simple lexical matching.

Comparative End-to-end Evaluation

n	number of test trials (*e.g.* information system queries) in the evaluation sample
A, B	two ontology-matching systems
A_i	outcome of the evaluation metric (*e.g.* Semantic precision (Euzenat, 2007)) for the i-th test trial for system A
$I[A_i > B_i] = \begin{cases} 1 & A_i > B_i \\ 0 & A_i \leq B_i \end{cases}$	interpretation function that tests outperformance
$S_+ = \sum I[A_i > B_i]$	number of trials for which system A outperforms system B

To compare end-to-end system performances we determine whether one system performs better over a significant number of trials. There are many tests for statistical significance that use pairwise comparisons. Each test can be used under different assumptions. A common assumption is the normal distribution of performance differences: small differences between the performance of two systems are more likely than large differences, and positive differences are equally likely as negative differences. However, this is not very probable in the context of comparative evaluation of matching systems. The performance differences between techniques are usually of a much greater magnitude than estimation errors. There are many techniques that improve performance on some queries while not hurting performance on other queries. This causes a skewed distribution of the performance differences. Therefore, the most reliable test is the Sign-test (Hull, 2000; van Rijsbergen, 1979). This significance test only assumes that two systems with an equal performance are equally likely to outperform each other for any trial. It does not take into account how much better a system is, only in how many cases a system is better. The test gives reliable results for at least 25 trials. It needs relatively large differences to proclaim statistical significance, compared to other statistical tests. This means statistical significance calculated in this way is *very* strong evidence.

[9] respectively `http://trec.nist.gov`, `http://www.clef-campaign.org`, and `http://inex.is.informatik.uni-duisburg.de`

[10] A more reliable method for weighted combination of the scores that uses the variance of each performance measurement is described by Meier (1953).

To perform the Sign-test on the results of systems A and B on a set of n trials, we compare their scores for each trial, $A_1, \ldots, A_n$ and $B_1, \ldots, B_n$. Based on these outcomes we compute S_+, the total the number of times A has a better score than B. For example, the number of search queries for which A retrieves better documents than B. The null-hypothesis is that the performance of A is equal to that of B. This hypothesis can be rejected at a confidence level of 95%† if

$$\frac{2 \cdot S_+ - n}{\sqrt{n}} > 1.96$$

For example, in the case of 36 trials, system A performs significantly better than system B when it outperforms system B in at least 23 of the 36 trials.

VI.5 CONCLUSION

We presented two alternative techniques for the evaluation of ontology-matching systems and showed the margin of error that comes with these techniques. We also showed how sample evaluation can be applied and what the statistical results mean in practice in the context of the OAEI 2006. Both techniques allow a more application-centered evaluation approach than current practice.

Apart from sampling errors we investigated in this chapter, there are many other possible types of errors that can occur in an evaluation setting. (Some of which are discussed by Avesani et al. (2005).) Other sources of errors remain a subject for future work. Also, this chapter leaves open the question of which technique to choose for a certain evaluation effort. For example, when you want to apply evaluation to find the best ontology matching system for a certain application. The right choice depends on which technique is more cost effective. In practice, there is a trade-off between cheap and reliable evaluation: With limited resources there is no such thing as absolute reliability. Yet, all the questions we have about the behavior of matching systems will have to be answered with the available evaluation results. The nature of the use case for which the evaluation is performed determines which of the two approaches is more cost effective. Depending on the nature of the final application, evaluation of end-to-end performance will sometimes turn out to be more cost effective than investigating the alignment, and sometimes the latter option will be a better choice. We will apply the techniques presented in this chapter to the food, environment, and library tasks of the forthcoming OAEI 2007.[11] This should give us the opportunity to further study this subject.

†About 95% of the cases fall within 1.96 times the standard deviation from the mean of the normal or binomial distribution. In the derivations we use 2 instead of 1.96 for the sake of simplicity. This guarantees a confidence level of more than 95%.

[11] http://oaei.ontologymatching.org/2007/

Acknowledgments

We would like to thank Frank van Harmelen, Guus Schreiber, Lourens van der Meij, Stefan Slobach (VU), Hap Kolb, Erik Schoen, Jan Telman, and Giljam Derksen (TNO), Margherita Sini (FAO), Lori Finch (NAL).

Chapter vii

Relevance-Based Evaluation of Ontology Alignment

In this chapter we describe an ontology-alignment evaluation method that uses a form of biased sampling, relevance-based evaluation*. This method is an extension of* alignment sample evaluation*, described in chapter vi. As opposed to* alignment sample evaluation*, which draws samples disregarding their value to users,* relevance-based evaluation *draws samples based on their utility to a selection of prototypical usage scenarios. We apply* relevance-based evaluation *on the alignment between AGROVOC and NALT of the OAEI 2007* food task*, described in chapter v.*

This chapter is based on a paper coauthored by Hap Kolb and Guus Schreiber, "Relevance-Based Evaluation of Ontology Alignment, Willem Robert van Hage, Hap Kolb, Guus Schreiber" *(van Hage et al., 2008a).*

Abstract Current state-of-the-art ontology-alignment evaluation methods are based on the assumption that alignment relations come in two flavors: correct and incorrect. Some alignment systems find more correct mappings than others and hence, by this assumption, they perform better. In practical applications however, it does not only matter *how many* correct mappings you find, but also *which* correct mappings you find. This means that, apart from correctness, relevance should also be included in the evaluation procedure. In this chapter we demonstrate how to incorporate relevance in sample evaluation of alignment approaches by using high relevancy as a selection criterion when drawing sample mappings. We expand the sample-based evaluation of the OAEI 2007 *food task* with relevance-based evaluation and compare the results of this new evaluation method to the existing results. This leads to new insights on the performance of the participating ontology-alignment systems in practice.

vii.1 Introduction

In recent years ontology alignment has become a major field of research (Kalfoglou and Schorlemmer, 2003; Shvaiko and Euzenat, 2005; Euzenat and Shvaiko, 2007). Especially in the field of digital libraries it has had a great impact. Many libraries have made the transition to offer access to their resources through the web. This has made it possible to access multiple collections at the same time. Different libraries have different indexing schema's

and protocols. This complicates federated access. Alignment offers a way to bridge the semantic gap between the indexing schema's so that users can profit from their joint coverage.

Good evaluation of alignment approaches is important. In past decades, research communities that focus on other complex computer-science subjects, such as natural-language processing and information retrieval, have developed suitable evaluation methods. Some of their methods in these communities are applicable to ontology alignment and have been adopted in recent years by evaluation efforts such as the Ontology Alignment Evaluation Initiative (OAEI). The main contribution of this work is to improve the evaluation methodology of alignment to better capture the performance of alignment approaches in actual applications. We introduce a simple evaluation method, *relevance-based evaluation*, that remedies some of the shortcomings of existing methods by using a sampling technique that takes the needs of users into account. We apply this method to the data of the OAEI 2007 *food task* (Euzenat et al., 2007).

In section VII.2 we discuss existing evaluation methods. In section VII.3 we describe our new method, relevance-based evaluation. In section VII.4 we describe the procedure we followed to apply relevance-based evaluation to alignment in the agricultural domain. In section VII.5 we go into detail on every step of this procedure and the data sets that were involved. In section VII.6 we show the results of relevance-based evaluation on the OAEI 2007 *food task* data and compare them to the existing results. We test how these new results, based on our "second opinion", differ from the old results and we draw conclusions about the validity of the OAEI 2007 *food task* results.

VII.2 Alignment Evaluation

Nearly all existing evaluation measures used to determine the quality of alignment approaches are based on counting mappings (Euzenat et al., 2007; Euzenat, 2007). For instance, in the context of ontology alignment, the definition of Recall is defined as the number of correct mappings a system produces divided by the total number of correct mappings that can possibly be found (*i.e.* that are desired to be part of the result). Regardless of their differences, most of these measures have one thing in common: They do not favor one mapping over the other in order to give an objective impression of system performance. Any mapping could prove to be important to some application. Therefore, they can only tell us *how many* mappings are found on average by a system, but not *which* mappings are found and whether the mappings that are found are those that are useful for a certain application. Whenever someone wants to decide which alignment approach is best suited for his application (*e.g.* Mochol et al. (2006)) he will have to reinterpret average expected performance in the light of his own needs. This can be a serious obstacle for users.

A solution to this problem is to incorporate the importance of mappings (*i.e.* relevance) into the evaluation result. This solution immediately raises two new problems:

1. How to come up with suitable importance weights
2. How to define a simple and intuitive way to use these weights

With respect to problem 1, there are many sensible ways to weigh the importance of mappings. One possibility is to assign weights that are independent of how often the mappings

are used, but dependent on the size of the logical implication of a mapping, *cf.* Semantic Precision and Semantic Recall (Euzenat, 2007). The intuition underlying this method is that mappings with a greater logical consequence have more benefit to users, because more implications can be made using these mappings. However, a mapping might have a large logical consequence while it is never used in a specific application. In this chapter we do not account for logical implications. Another possibility is to assign weights according to how often a mapping can be expected to be used (Hollink et al., 2008). This method makes the assumption that each concept has an equal probability of being used as a query in an application. Under this assumption, Hollink et al. estimate the frequency a mapping will be used based on the distance of a mapping the query concept. In this chapter we do not assume a uniform query distribution.

Likewise, with respect to problem 2, there are many sensible ways to incorporate mapping importance into an evaluation method. They can, for example, be used as coëfficients in a lineair equation. (*cf.* Kekäläinen's approach to including varying degrees of relevance in information-retrieval evaluation in Kekäläinen (2005)) Or, in the case of sample evaluation, they can be used to weigh sample sets as a whole (*cf.* the *Alignment Sample Evaluation* method described in van Hage et al. (2007)).

Another related evaluation approach is described in Porzel and Malaka (2004), where the ontology construction process is guided by its effect on end-to-end task performance.

VII.3 Relevance-Based Evaluation

The evaluation method we propose in this chapter consists of two steps:

1. **Gather Relevant Mappings** Depending on the application, we determine which mappings are directly involved in achieving the user's goals (*e.g.* finding documents of special importance). These mappings are considered relevant, the rest is considered irrelevant. We gather a set of relevant mappings that reflects typical usage scenarios.

2. **Apply Sample Evaluation of Relevant Mappings** We assume that the selected relevant mappings are representative of all mappings that are useful to the application. We calculate performance scores on the sample of relevant mappings using existing sample-evaluation methods (*e.g.* van Hage et al. (2007)).

As opposed to existing methods to account for the relevance of mappings that include it as a variable in an evaluation measure, we use relevance to steer the sample-selection process. Instead of randomly selecting mappings for the evaluation of alignment approaches (*cf.* the *food* and *environment tasks* described in Euzenat et al. (2007)) we select *only* those that are relevant to an application. This way we can use existing and well-understood evaluation metrics, like Precision and Recall, to measure performance on important tasks as opposed to fictive average-case performance. The advantages of not adapting the evaluation measure, but influencing the drawing of samples are the following:

- The evaluation measure can remain simple. This makes it easier to interpret what scores mean.

- Using the same evaluation measure for relevance-based as for non-relevance-based evaluation allows us to easily explore how performance in specific applications differs from average-case performance, because only the samples differ.
- Existing experiments can be easily extended to account for new use cases. Additional samples can be added to compensate for the underrepresentation of certain usage scenarios.
- Different sources of relevance estimates can be used besides each other, because the estimation is not part of the evaluation measure.

VII.4 EXPERIMENTAL SET-UP

We demonstrate how relevance-based evaluation works by applying it to the existing results of the OAEI 2007 *food task*, which did not take relevance into account. We determine relevance for the mappings based on hot topics related to this task, like global warming and increasing food prices, which we obtain by means of query-log analysis, expert interviews, and news feeds. For the original OAEI 2007 *food task*, Recall was measured on samples that represent the frequency of topics in the vocabularies. For example, if 60% of the concepts in the vocabularies were animal or plant species names, then sample mappings from and to animal and plant species names determined 60% of the end result. In this chapter we will repeat the measurement of Recall on samples that represent the relevance of mappings to finding documents on hot topics. For example, most species names, except 'Oryza sativa' (the rice plant) are probably irrelevant to the hot topic of rising rice prices. On the other hand, topics that are covered by few concepts in the vocabularies[1] might prove to be vital to hot topics. We implemented the two steps described in section VII.3 as follows:

GATHER RELEVANT MAPPINGS

1. **Gather topics that represent important use cases.** In this step we research which topics are currently "hot" in agriculture. We gather topics from the query log files of the FAO AGRIS/CARIS search engine, the FAO newsroom website, and interviews with experts from the FAO's David Lubin library and the TNO Quality of Life food-safety group. We manually construct search-engine queries for each topic. Further elaboration can be found in section VII.5.1.
2. **Gather documents that are highly relevant to the topics.** In this step we ascertain which documents would be sufficient for the hot topics. We gather suitable candidate documents from the part of the FAO AGRIS/CARIS and USDA AGRICOLA reference databases that overlaps. We use a free-text search engine[2] and manually filter out all irrelevant documents, see section VII.5.2.

[1] This refers to the number of concepts in the thesaurus on a given subject, not the number of times they are used to index a document.

[2] `http://www.fao.org/agris/search`

3. **Collect the meta-data describing the subject of these documents and align the concepts that describe the subject of the documents to concepts in the other thesaurus.** In this step we determine which mappings are necessary to find these documents. We collect values of the Dublin Core subject field from the AGRIS/CARIS and AGRICOLA reference databases. These values come from subject vocabularies, respectively AGROVOC and the NAL Agricultural Thesaurus. We manually align each concept to the most similar concept in the other vocabulary, see section VII.5.3. The resulting mappings make up our sample set of relevant mappings.

Apply Sample Evaluation on Relevant Mappings

4. **Count how many of these mappings have been found by ontology alignment systems and compare system performance based on these counts.** We re-calculate Recall for the top-4 systems of the OAEI 2007 *food task*, following the same procedure as described in Euzenat et al. (2007); van Hage et al. (2007), but use the new set of relevant mappings. The details and results can be found in section VII.6.

VII.5 Sample Construction

VII.5.1 Topics

In order to get a broad overview of current affairs in the agricultural domain we gathered topics from three sources: AGRIS/CARIS search log analysis, topics in the "Focus on the issues" section of the FAO Newsroom, and interviews with a food-safety expert at TNO Quality of Life and a reference librarian at the David Lubin Memorial Library of the FAO. A long description of the topics that resulted from these three sources can be found in the appendix of this chapter, section VII.8.

Log analysis

The FAO AGRIS/CARIS search engine is used by a broad range of people all around the world: Information scientists at agricultural research facilities, farmers in search of new techniques for their profession, internal FAO information officers, people involved in development and education, and the occasional data mining bot. This means the query log is very heterogeneous. After simple syntactic preprocessing of the queries we sorted them by frequency and selected four topics that were represented by multiple, easily interpretable queries amongst the most frequent queries of the log. Amongst the top queries are many query-syntax mistakes, single-letter queries (*e.g.* `M`, perhaps the initial of an author), stray boolean operators (*e.g* `AND` without actual terms), and spelling mistakes (*e.g.* `babanas`). Many of the most frequent terms are clearly not related to hot topics, like `University` or `title`. For most queries it is impossible to reconstruct the original meaning without unreasonable guessing. For example, the query `rice` does not reveal which aspect of rice was intended. In such cases we searched for queries that contained `rice`, like `paddy rice and fertilizers` or `rice fish system`. When this yielded a connection to current affairs

we added it to the list of hot topics. An important reason to practice rice/fish cultivation, for example, is the great reduction of pesticide that it permits.[3]

The hot topics that were selected based on evidence mainly from query log analysis were the following: [4] Avian influenza, Malaria in Africa, Genetic modification of soy, Cattle traceability.

THE FAO NEWSROOM

One of the main tasks of the FAO is to disseminate information about agriculture (*i.e.* agronomy, forestry, and fishery) to the world. The Newsroom[5] is one of the channels the FAO uses to reach people around the world. The Newsroom has a section about current events.[6] We used this section as a source of hot topics and to verify evidence from query-log analysis and interviews.

The hot topics that were selected based on evidence mainly from the FAO newsroom were the following: [4] Rice and pesticides, The role forestry can play in climate change, Plants and advancing desertification, Biofuels and their effect on corn prices, Biofuels and their effect on water supply.

EXPERT OPINIONS

Information officers and reference librarians at the FAO in Rome and food-safery researchers at the Netherlands Organisation for Applied Scientific Research (TNO) deal with questions from journalists on a daily basis. Apart from consulting tangible sources of topics we have also consulted these domain experts. Besides confirming the topics we obtained from the other two sources they mentioned these additional issues: [4] Acrylamide found in fried foods, Benzene found in food or drink, Dioxins found in food or drink, The effect of bee extinction on pollination, The effect of fish farming and antibiotics use on wild fish.

VII.5.2 DOCUMENTS

Per topic we retrieved the top-100 hits of a full-text search on the AGRIS/CARIS search engine limited to the set of documents that is shared between the AGRICOLA and AGRIS/-CARIS collections.[7] From these 1500 documents we selected only the ones that are relevant to our topics and that have been assigned Dublin Core subject terms in both collections. This left 52 documents. How many suitable double-annotated documents we found per topic and how many of these were also relevant is shown in table VII.1. For four of the topics we found no documents that were both relevant and indexed in both collections: Cattle traceability, both topics about biofuels, and the effect of antibiotics in fish farming on wild fish. The reason for this is that these topics are all very new issues. The greatest overlap between the AGRIS/CARIS and AGRICOLA collections exists for documents published

[3] see: http://www.fao.org/newsroom/en/news/2005/102401

[4] Detailed descriptions of the topics can be found in appendix section VII.8.

[5] `http://www.fao.org/newsroom`

[6] `http://www.fao.org/newsroom/en/focus`

[7] This can be accomplished by limiting the search to data from the USDA data center by adding `+center:(US)` to the search query.

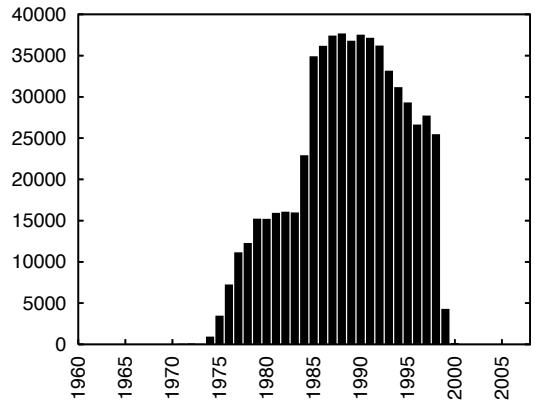


Figure VII.1: The number of documents imported by the FAO from the USDA AGRICOLA collection into the AGRIS/CARIS collection per year.

between 1985 and 1995. The total number of documents that was imported from AGRICOLA to AGRIS/CARIS per year is shown in figure VII.1. After the year 2000 no documents have been imported and thus it is hard to find relevant documents for new issues. We assume that the 52 double-annotated relevant documents are representative of the set of all relevant documents with subject meta-data, *i.e.* also the documents with only annotations in one of the two collections. These are the documents for which alignment could make the biggest difference. This is a reasonable assumption, because the indexing process of both collections is regulated by a protocol. The indexing protocol of both libraries differ quite a lot, but within each collection annotations are quite stable. For both libraries it goes that not all documents are indexed, but those that are were indexed by the same protocol.

VII.5.3 Mappings

Now that we have established which documents are potentially important to find, we will decide which mappings will be of most benefit to someone who wants to find them. This can be done with a search engine that employs mappings. There are many possible ways in which such a search engine works. Each retrieval method has strong and weak points. Some methods that apply mappings during retrieval work well with an incomplete set of mappings, others do not. Some make use of the extra synonyms that are made available through mappings, others are geared towards exploiting extra hierarchical relations. To maximize the generalizability of our work, we avoid having to choose a specific retrieval method by making two assumptions about the retrieval methods that will be used.

1. We assume that retrieval methods work best if each concept (in both vocabularies) is aligned to the most similar concept in the other vocabulary.[8]

2. We assume that relevant mappings are all equally important during the retrieval process and that irrelevant mappings are all equally unimportant for retrieval.

[8] As opposed to alignments consisting mainly of, for example, rdf:type or partitive relations.

topic	suitable documents	suitable and relevant	indexed with # concepts NALT	AGROVOC	mappings
avian influenza	48	9	52	35	72
malaria in africa	51	12	75	67	112
genetic modification of soy	40	1	8	10	12
cattle traceability	34	0	-	-	-
rice and pesticides	41	3	10	5	12
climate change and forestry	21	4	29	25	36
desertification	47	8	21	23	33
biofuels and corn price	35	0	-	-	-
biofuels and water	5	0	-	-	-
acrylamide in fried foods	31	5	20	13	25
benzene in food	13	5	28	24	38
dioxins in food	9	4	26	15	31
bee extinction	62	2	15	5	18
fish farming and antibiotics	1	0	-	-	-

Table VII.1: Statistics per topic. Shown are the number of double-annotated documents in the top-100 of the AGRIS/CARIS search engine, the number of relevant documents amongst these, the number of indexing terms used for these documents, and the number of mappings this led to for the relevance-based reference alignment.

The first assumption corresponds exactly to the protocol that was used by human experts to create the reference alignments for the OAEI *food* and *environment tasks*. The second merely states that we use boolean weights for relevancy or, specifically, that we will create a sample set of only relevant mappings.

Given this, the set of mappings that works best for finding the 53 relevant documents is the set that aligns each of the describing concepts with its most similar counterpart. For example, if a document is indexed with the concepts agrovoc:chickens and agrovoc:frying in AGRIS/CARIS and with nalt:chickens and nalt:fried_foods in AGRICOLA then the ideal set of mappings for this document is:

```
agrovoc:chickens  skos:exactMatch  nalt:chicken  .
agrovoc:frying  skos:exactMatch  nalt:frying  .
agrovoc:foods  skos:narrowMatch  nalt:fried_foods  .
```

In this way we manually mapped the 266 NALT concepts and 212 AGROVOC concepts, see table VII.1, to their counterpart in the other thesaurus with the help of thesaurus experts at the FAO and USDA, Gudrun Johannsen and Lori Finch. This led to a sample reference alignment consisting of 347 mappings[9]: 74 broadMatch / narrowMatch and 273 exactMatch (79%). 11 concepts had no exact, broader or narrower counterpart. This is a higher percentage of exactMatch mappings than we expected based on our experiences with the OAEI *food task*. For the *food task*, arbitrary subhierarchies of the AGROVOC and NAL

[9]Adding up the number of mappings per topic leads to a total of 373 mappings. The lower total is due to overlap between the topics.

	submitted to OAEI 2007 food	required for hot topics
taxonomical	55%	14%
biological/chemical	9%	20%
geographical	3%	8%
miscellaneous	33%	58%

Table VII.2: The relative size of topics in the sets of mappings found by the participants of the OAEI 2007 *food task* and in the set of mappings that is necessary to find documents on hot topics.

thesaurus were drawn and manually aligned with the other thesaurus. Most of the resulting mappings were equivalence relations. The sample sets, the percentage of equivalence mappings in the reference alignment (*i.e.* the desired equivalence relations) varied between 54% and 71%.

Table VII.2 gives an overview of the kinds of mappings in the new reference alignment and the kinds submitted to the OAEI 2007 *food task* by the participants. The sample reference alignments of the OAEI 2007 *food task* focussed much more on taxonomical terms and less on biological and chemical terms. Common categories of mappings that were not recognized as such in the OAEI 2007 food evaluation were: scientific methods, anatomy, and production or processing techniques (*e.g.* for crops or natural resources).

VII.6 Sample Evaluation Results

Having constructed a new sample reference alignment we can use it to measure the performance of alignment approaches. We choose to reiterate the evaluation of Recall[10] on the OAEI 2007 *food task* for two reasons: It allows us to show the effect of relevance-based evaluation as opposed to non-relevance-based evaluation by referring to known results; and it offers a second opinion to test the validity of the evaluation method used for the OAEI *food tasks*. The latter is important, because the evaluation of Recall under the open-world assumption is inherently tricky business (*i.e.* an unsolved subject of research). If the results of relevance-based evaluation differ significantly from the results of independent evaluation then we should wonder whether non-relevance-based evaluation as it is performed in all OAEI tasks is a suitable evaluation method.

For the sake of simplicity we calculate Recall scores of the top-4 of the systems that participated in the OAEI 2007 *food task*. The results are shown in table VII.3. There are a number of striking points to note about these results.

If we look at the difference between rows labeled "OAEI 2007 food" and those labeled "hot topics" in table VII.3 we can see that for most systems there is a significant positive or negative difference. Falcon-AO performs 6% better on only exactMatch mappings for hot topics than it did in the OAEI 2007 *food task*, while DSSim performs 21% worse on hot topics, a very large relative difference.

[10] "the whole truth" as opposed to "nothing but the truth", $|Correct \cap Found| / |Correct|$.

	Falcon-AO	RiMOM	DSSim	X-SOM
OAEI 2007 food, only exactMatch (54% of total)	0.90	0.77	0.37	0.11
hot topics, only exactMatch (79% of total)	0.96 ↑	0.60 ↓	0.16 ↓	0.07 ↓
OAEI 2007 food, exact, broad, narrowMatch	0.49	0.42	0.20	0.06
hot topics, exact, broad, narrowMatch	0.75 ↑	0.47 ↑	0.12 ↓	0.05 ≈

Table VII.3: Recall of alignment approaches measured on sample mappings biased towards relevance to hot topics in agriculture and on impartial, non-relevance-based sample mappings from the OAEI 2007 *food task*.

Overall, the difference with non-relevance-based evaluation is quite great. In the second row of table VII.3, we can see that for exactMatch relations performance in general is lower for relevance-based evaluation than for non-relevance-based evaluation, with the exception of Falcon-AO, although the relative difference is small. However, even though there is a clear difference, the ranking of the alignment approaches is left unchanged. The results of relevance-based evaluation seem to exaggerate the differences between the performance of the approaches. This can be explained by the relatively high number of obvious matches (93%) in the set of mappings on hot topics. None of the approaches was able to find a substantial number of difficult mappings, but the best approaches were good at finding all obvious mappings before resorting to speculation about the harder mappings. The relatively high number of easy matches significantly boosts the scores of approaches that find the obvious matches. We expect that the reason why so many of the relevant mappings are easy is that the indexers at the USDA and FAO attempt to help users by using the most obvious words. (*cf.* the debated *basic level* described by Eleanor Rosch et al. in Rosch et al. (1976))

Another thing we can note is that the best two systems, Falcon-AO and RiMOM performed relatively good for all relation types, the last row of table VII.3. This has nothing to do with their ability to find particular relation types, because they found no broadMatch and narrowMatch relations. It is due to the kind of exactMatch relations they *did*, which were mostly of the obvious kind (*i.e.* literal matches), which was exactly the kind that was needed most for the hot topics. The high percentage of exactMatch relations in the set on hot topics accentuates their behavior. The converse goes for DSSim, which found a relatively low number of obvious mappings.

Fewer broadMatch and narrowMatch mappings seem to be needed than one would expect from the non-relevance-based evaluation method. Compare the percentage in the OAEI 2007 Recall set, 54%, to the percentage based on hot topics, 78.6%. Although there is a large part of the AGROVOC and NALT vocabularies that does not have a counterpart in the other vocabulary, the portion that is actually used suffers less than one would expect from this mismatch. Apparently, indexers mainly pick their terms from a limited set, which shows a greater overlap. (After all, why needlessly complicate things?) On one hand this means that approaches that can only find equivalence mappings perform better in practice than was expected. On the other hand it confirms the expectation that a large part (*more than 20%*) of the mappings that are needed for federated search over AGRIS/CARIS and AGRICOLA consists of other relations than equivalence relations. Also, one can conclude

that systems that are incapable of finding a substantial number of equivalence relations can only play a marginal role.

VII.7 Discussion

By using relevance as a sample criterion we avoid having to come up with an artificial approximation of importance. We can simply explore the performance difference on samples consisting of relevant mappings and samples consisting of irrelevant mappings. This has a few advantages. We can use existing evaluation measures without adaptation, therefore results using this method are easily comparable to existing results. Due to the simplicity of this method results are easy to interpret. Linear weighing of the mappings by some real value representing relevance as in Kekäläinen (2005), for example, can make it difficult to see whether an alignment approach found many marginally relevant mappings or a few reasonably relevant mappings. If you use the weights for the drawing of the samples you can save the sample for later use. We can easily extend existing experiments. For instance, to investigate a new use case.

Under minimal assumptions we avoid having to choose a specific retrieval method while retaining the the character of an end-to-end evaluation. (*cf.* the *End-to-end Evaluation* method described in van Hage et al. (2007)) This saves us the effort of extensive user studies while not ignoring the behavior of alignment approaches in real-life situations.

Considering the fact that AGROVOC and NALT are two of the most widely used agricultural ontologies, and that they are prototypical examples of domain thesauri in their design we conclude the following. From the point of view of a developer of a federated search engine in the agricultural domain that needs an alignment we can conclude that at the moment the Falcon-AO is a good starting point. For use cases similar to the prototypical set-up described in this chapter, Falcon-AO can be expected to find three quarters of the mappings. Demands change through time, and hence, current thesauri, current hot topics, and perhaps current alignment techniques will be outdated.

Another thing to note, which is besides the main message of this chapter, is that this empirical study has shown that at least 20% of the required mappings to solve the typical federated-search problem are hierarchical relations. Even though this is a smaller fraction than we initially expected it is still a large part.

Acknowledgments

We would like to thank the FAO and NAL for allowing us to use their collections and the participants of the OAEI 2007 *food task*. Specific thanks go to Edgar Meij for his thoughts on log analysis, Fred van de Brug and Patricia Merrikin for their input of topics, Andrew Bagdanov and Stefano Anibaldi for allowing us to access the AGRIS/CARIS search log files, Lori Finch and Gudrun Johannson for creating part of the reference alignment, Tuukka Ruotsalo and Ruud Stegers for their comments, and Laura Hollink, Mark van Assem, Mike Frame, and Gail Hodge for valuable discussions. This work has been partially funded by the Dutch BSIK project Virtual Laboratory for e-science (VL-e).

vii.8 Impendix – Detailed Topic Descriptions

Log Analysis

Avian influenza
query: `"avian influenza"`,
description: There have been numerous avian influenza epidemics, but especially the H5N1 outbreak of 1997-1998 in South East Asia was one of the big recent events in agriculture. It has had a great impact on farmers world-wide and international travel.
information need: where can avian influenza be found, who is susceptible, and what measures are or can be taken.

Malaria in Africa
query: `"malaria africa"`,
description: Malaria is one of the biggest problems in agricultural communities in Africa.
information need: Where can malaria be found in Africa, which agricultural processes influence the risk of exposure, and what measures are or can be taken to avoid the disease.

Genetic modification of soy
query: `"genetic modification soy"`,
description: Soy is one of the most important crops in the world. Genetic modification of soy is one of the topics that has never left the spotlight in recent years, due to its potential benefits for (and possible damage to) the world's food supply.
information need: the effects of genetic modification in soy or the effects of cross-pollination of modified soy and regular plants.

Cattle traceability
query: `"cattle traceability"`,
description: Knowledge about the entire food chain with respect to meat and other cattle products have become a hot topic since recent Foot-and-mouth disease and BSE epidemics.
information need: which methods can be used to trace cattle products.

The FAO Newsroom

Rice and pesticides
query: `"rice pesticides"`,
description: Rice is the most common staple food of the world. Hence, pesticides used with rice have a large impact on people.
information need: which pesticides are used with rice, how can pesticide use be decreased for rice.

The role forestry can play in climate change
query: `"climate change forestry"`,
description: Globally, forests trap one trillion tons of carbon. Forestry plays a large part in controlling CO_2.
information need: how does forestry influence greenhouse gasses and climate change, how great can this impact be.

Plants and advancing desertification

query: `"desertification plants"`,
description: Certain plants can be used to combat advancing deserts and hence to save threatened farmable soil.
information need: which plants can help under which circumstances.

Biofuels and their effect on corn prices
query: `"biofuels corn price"`,
description: The use of corn to produce the biofuel ethanol has led to an increased demand of corn, which has increased corn prices worldwide. This is a serious issue for poor regions.
information need: how large is the problem, who is affected, what measures can be taken.

Biofuels and their effect on water supply
query: `"biofuels water"`,
description: Farming crops for the production of biofuel is taxing for the world's water supply.
information need: how much water is needed for a liter of biofuel, are there sustainable ways to produce biofuels with respect to water consumption.

Expert Opinions

Acrylamide found in fried foods (TNO)
query: `"acrylamide fried foods"`,
description: Health risks caused by substances in food occur all the time. Sometimes they become big issues. One of such occasions was triggered by the discovery that the hazardous chemical acrylamide can be formed during the frying process of, for instance, french fries.
information need: how is acrylamide formed, how does it end up in food, who can be exposed.

Benzene found in food or drink (TNO)
query: `"benzene food"`,
description: The level of benzene in water are formally regulated internationally, but for soft drinks regulation only informally. Recently, high levels of benzene have been found in bottled water and orange flavored soda's.
information need: in which occasions has benzene been found in the food chain, who can be exposed.

Dioxins found in food or drink (TNO)
query: `"dioxins food"`,
description: Dioxins are a family of toxic chemicals that can accumulate in fat. If dioxin enters the food chain (*e.g.* as insecticide) it can end up in humans.
information need: how does dioxin enter the food chain, who can be exposed.

The effect of bee extinction on pollination (FAO)
query: `"bee extinction pollination"`,
description: Many species of bees and bumblebees are facing extinction, mainly due to degradation and destruction of their habitats. Bees pollinate many plants, like apples and tomatoes.

information need: what are the ecological consequences of bee extinction, what measures can be taken to avoid bee extinction.

The effect of fish farming and antibiotics use on wild fish (FAO)
query: `"fish farming wild fish antibiotics"`,
description: Fish farms are an important source of fish for consumption. Most farmed fish is given antibiotics. When such fish escape from a farm (*e.g.* during a storm) they can spread diseases that are carried by them, but that do not affect them, to wild fish.
information need: which diseases are spread, which fish are affected most, what can be done to decrease the use of antibiotics.

Chapter viii

Automatic versus Manual Ontology Alignment

In this chapter we compare an automatically constructed alignment to a manually constructed alignment. Specifically, we compare the joint alignments between AGROVOC and NALT, described in chapter v, to the alignment between AGROVOC and the Schlagwortnormdatei, constructed by GESIS/IZ (Mayr and Petras, 2008a).

This chapter is based on a paper coauthored with Boris Lauser, Gudrun Johannsen, Caterina Caracciolo, Johannes Keizer, and Philipp Mayr, "Comparing Human and Automatic Thesaurus Mapping Approaches in the Agricultural Domain, Boris Lauser, Gudrun Johannsen, Caterina Caracciolo, Johannes Keizer, Willem Robert van Hage, Philipp Mayr" *(Lauser et al., 2008), which will be presented at the International Conference on Dublin Core and Metadata Applications (DC 2008). Small adaptations were made to this paper.*

The work described in this chapter was done in close cooperation with the coauthors of this paper, during a working visit to the Food and Agriculture Organization of the United Nations in Rome. My contribution is the experimental design, part of the assessment work, the AGROVOC-NALT alignment, and part of the analysis. Gudrun Johannsen contributed to the assessment work and the analysis. Philipp Mayr contributed the AGROVOC-SWD alignment and added to the analysis. Caterina Caracciolo contributed to related work. Boris Lauser contributed most of the article. Johannes Keizer took the initiative to cooperate.

Abstract Knowledge organization systems (KOS), like thesauri and other controlled vocabularies, are used to provide subject access to information systems across the web. Due to the heterogeneity of these systems, alignment between vocabularies becomes crucial for retrieving relevant information. However, aligning thesauri is a laborious task, and thus the automation of the alignment process is a topic of research. This chapter examines two alignment approaches involving the agricultural thesaurus AGROVOC, one created by machines and one by humans. We are addressing the basic question "What are the pros and cons of human and automatic mapping and how can they complement each other?" By pointing out the difficulties in specific cases or groups of cases and grouping the sample into simple and difficult types of mappings, we show the limitations of current automatic methods and come up with some basic recommendations on what approach to use when.

VIII.1 INTRODUCTION

Information on the Internet is constantly growing and with it the number of digital libraries, databases and information-management systems. Each system uses different ways of describing their metadata, and different sets of keywords, thesauri and other knowledge organization systems (KOS) to describe its subject content. Accessing and synthesizing information by subject across distributed databases is a challenging task, and retrieving all information available on a specific subject in different information systems is nearly impossible. One of the reasons is the different vocabularies used for subject indexing. For example, one system might use the keyword 'snakes', whereas the other system uses the taxonomic name 'Serpentes' to classify information about the same subject. If users are not aware of the different 'languages' used by the systems, they might not be able to find all the relevant information. If, however, the system itself "knows", by means of mappings, that 'snakes' is equivalent to 'Serpentes', the system can appropriately translate the user's query and therefore retrieve the relevant information without the user having to know about all synonyms or variants used in the different databases.

Aligning thesauri and other knowledge organization systems in specific domains of interest can therefore enhance the access to information in these domains. System developers for library search applications can incorporate the use of alignments into the search applications. The mappings can hence be utilized at query time to translate a user query into the terminology used in the different systems of the available mappings and seamlessly retrieve consolidated information from various databases.[1]

Alignments are usually established by domain experts, but this is a labor intensive, time consuming and error-prone task (Doerr, 2001). For this reason, the possibility of creating mappings in an automatic or semi-automatic way is being investigated, *cf.* Vizine-Goetz et al. (2004); Euzenat and Shvaiko (2007); Kalfoglou and Schorlemmer (2003); Maedche et al. (2002). However, so far, research has focused mainly on the quantitative analysis of the automatically obtained mappings, *i.e.* purely in terms of Precision and Recall of either end-to-end document retrieval or of the quality of the sets of mappings produced by a system. Only little attention has been paid to a comparative study of manual and automatic alignment. A qualitative analysis is necessary to learn how and when automatic techniques are a suitable alternative to dependable but expensive manual alignment. The work described in this chapter aims to fill that gap. We will elaborate on mappings between three KOS in the agricultural domain: AGROVOC, NALT and SWD.

- AGROVOC[2] is a multilingual, structured and controlled vocabulary designed to cover the terminology of all subject fields in agriculture, forestry, fisheries, food and related domains (*e.g.* environment). The AGROVOC Thesaurus was developed by the Food and Agriculture Organization of the United Nations (FAO) and the European Commission, in the early 1980s. It is currently available online in 17 languages (more are under development) and contains 28,718 descriptors and 10,928 non-descriptors in the English version.

[1] See the implementation of such an automatic translation service in the German social sciences portal Sowiport, available at `http://www.sowiport.de`

[2] `http://www.fao.org/aims/ag_intro.htm`

- The NAL Thesaurus[3] (NALT) is a thesaurus developed by the National Agricultural Library (NAL) of the United States Department of Agriculture and was first released in 2002. It contains 42,326 descriptors and 25,985 non-descriptors organized into 17 subject categories and is currently available in two languages (English and Spanish). Its scope is very similar to that of AGROVOC. Some areas such as economical and social aspects of rural economies are described in more detail.

- The Schlagwortnormdatei[4] (SWD) is a subject authority file maintained by the German National Library and cooperating libraries. Its scope is that of a universal vocabulary. The SWD contains around 650,000 keywords and 160,000 relations between terms. The controlled terms cover all disciplines and are classified within 36 subject categories. The agricultural part of the SWD contains around 5,350 terms.

These controlled vocabularies (AGROVOC, NALT, and SWD) have been part of two mapping initiatives, conducted by the Ontology Alignment Evaluation Initiative (OAEI) and by the GESIS Social Science Information Centre (GESIS-IZ) in Bonn.

The Ontology Alignment Evaluation Initiative (OAEI) is an internationally-coordinated initiative to form consensus on the evaluation of ontology-alignment techniques. The goal of the OAEI is to help to improve the work on ontology alignment by organizing an annual comparative evaluation of ontology-alignment systems on various tasks. In 2006 and 2007 there was a task that consisted of aligning the AGROVOC and NALT thesauri, called the *food task*. A total of eight systems participated in this event. For this experiment we consider the results of the five best performing systems that participated in the OAEI 2007 food task: Falcon-AO, RiMOM, X-SOM, DSSim and SCARLET. Details about this task, the data sets used and the results obtained can be found on the website of the food task[5]. The alignment relations that participants could use were the SKOS Mapping Vocabulary relations exactMatch,broadMatch, and narrowMatch, because these correspond to the most commonly accepted thesaurus relations: USE, USE FOR, BT, and NT (ANSI/NISO, 2005).

In 2004, the German Federal Ministry for Education and Research funded a major terminology-alignment initiative called Competence Center Modeling and Treatment of Semantic Heterogeneity[6] at the GESIS-IZ, which published its conclusion at the end of 2007, see Mayr and Petras (2008a). The goal of this initiative was to organize, create and manage alignments between major controlled vocabularies (thesauri, classification systems, subject heading lists), initially centred around the social sciences but quickly extending to other subject areas. To date, 25 controlled vocabularies from 11 disciplines have been intellectually (manually) connected with vocabulary sizes ranging from 1,000-17,000 terms per vocabulary. More than 513,000 relations were constructed in 64 crosswalks. All terminology-alignment data is made available for research purposes. We also plan on using the alignments for user assistance during initial search query formulation as well as for ranking of retrieval results (Mayr et al., 2008). The evaluation of the value added by alignments

[3] http://agclass.nal.usda.gov/agt.shtml

[4] http://www.d-nb.de/standardisierung/normdateien/swd.htm

[5] http://www.few.vu.nl/ wrvhage/oaei2007/food.html. Both the results and gold standard samples are available in RDF format.

[6] The project was funded by BMBF, grant no. 01C5953. http://www.gesis.org/en/research/information_technology/komohe.htm.

and the results of an information retrieval experiment using human generated terminology alignments is described in Mayr and Petras (2008b). The AGROVOC-SWD alignment was created within this initiative in 2007.

VIII.2 Related Work

Many thesauri, amongst which AGROVOC and the Aquatic Sciences and Fisheries Abstracts Thesaurus[7] (ASFA) are being converted into ontologies, in order to enhance their expressiveness and take advantage of the tools made available by the semantic web community. In the Networked Ontologies project[8] (NeOn) an experiment was carried out to automatically align AGROVOC and ASFA. Since ASFA is a specialized thesaurus in the area of fisheries and aquaculture, the mapping with AGROVOC resulted in an alignment with the fisheries-related terms of AGROVOC. The mappings were extracted by means of the SCARLET system (*cf.* section 3) and were of three types: superclass/subclass, disjointness and equivalence. Evaluation was carried out manually by two FAO experts, in two runs: first with a sample of 200 randomly selected mappings, then with a second sample of 500 mappings. The experts were also supported in their evaluation by the graphical interface. The results obtained were rather poor (Precision was 0.16 in the first run of the evaluation and 0.28 in the second run), especially if compared with the high results obtained by the same system with the alignment of AGROVOC and NALT (*cf.* section 3). The hypothesis formulated to explain this low performance is related to the low degree of overlap between AGROVOC and ASFA,[9] and that the terms in ASFA may not be well covered by the Semantic Web, as required by SCARLET. Cases like this clearly show how beneficial it would be to gain a clear understanding of when manual alignment is more advisable than automatic alignment (as in the case of the AGROVOC-ASFA mapping) or the other way around (as in the case of the AGROVOC-NALT mapping analyzed in this chapter).

Another alignment exercise was carried out aligning AGROVOC to the Chinese Agricultural Thesaurus (CAT) described in (Liang et al., 2006). The alignment has been carried out using the SKOS Mapping Vocabulary[10] (version 2004) and addresses an important issue in aligning thesauri and other KOS: multilinguality. AGROVOC has been translated from English to Chinese, whereas CAT has been translated from Chinese to English. This creates potential problems as the following example illustrates: CAT '水稻'/'Oryza sativa' was originally aligned to AGROVOC 'Oryza sativa'. However, upon closer examination, the Chinese lexicalization in AGROVOC of 'Oryza sativa', which is '稻', appears to be the broader term of the CAT Chinese term. Moreover, a search in AGROVOC for the CAT Chinese term '水稻', shows the English translation as 'Paddy'. These discrepancies indicate the weakness of the above mentioned procedure and the necessity of cross checking all lexicalizations in both languages. Such cases pose hard problems for automatic alignment algorithms and can only be addressed with human support at the moment. Other related

[7] http://www4.fao.org/asfa/asfa.htm

[8] http://neon-project.org

[9] In particular, a problem could be the different level of detail of the two resources, as ASFA tends to be very specific on fisheries related terms.

[10] http://www.w3.org/2004/02/skos/mapping/spec/

System	Falcon-AO	RiMOM	X-SOM	DSSim	SCARLET		
Mapping type	=	=	=	=	=	< >	null(0)
# mappings	15,300	18,419	6583	14,962	81	6038	647
Precision	**0.84**	0.62	0.45	0.49	0.66	0.25	
Recall	**0.49**	0.42	0.06	0.20	0.00	0.00	

Table VIII.1: The OEAI 2007 *food task*. Results (in terms of Precision and Recall) of the 5 systems participating in the initiative. Best scores are in **bold face**. All systems found equivalence mappings only, except SCARLET that also found hierarchical mappings.

work on semantic interoperability can be found in Patel et al. (2005).

VIII.3 The AGROVOC-NALT Alignment within the OAEI

In the OAEI 2007 *food task*, five systems using distinct alignment techniques were compared on the basis of manual sample evaluation. Samples were drawn randomly from each of the sets of mappings supplied by the systems to measure Precision. Also, a number of small parts of the alignment were constructed manually to measure Recall. Details about the procedure can be found in chapter V and VI. Each participant documented their alignment method in a paper in the Ontology Matching 2007 workshop[11] (Sabou et al., 2007; Curino et al., 2007; Nagy et al., 2007; Hu et al., 2007; Li et al., 2007). Table VIII.1 summarizes, for each system, the type of mapping found, how many mappings were identified and the Precision and Recall scores measured on the set of returned mappings.

The system that performed best at the OAEI 2007 *food task* was Falcon-AO. It found around 80% of all equivalence relations using lexical matching techniques. However, it was unable to find any hierarchical relations. Also, it did not find relations that required background knowledge to discover. This led to a recall score of around 50%. The SCARLET system was the only system that found hierarchical relations using the semantic web search engine Watson[12] (Sabou et al., 2007). Many of the mappings returned by SCARLET were objectively speaking valid, but more generic than any human would suggest. This led to a Recall score close to zero.

VIII.4 The AGROVOC-SWD Alignment by GESIS-IZ

The GESIS-IZ approach considers intellectually (manually) created relations that determine equivalence, hierarchy (*i.e.* broader or narrower terms), and association mappings (*i.e.* related terms) between terms from two controlled vocabularies. Typically, vocabularies will be related bilaterally, that means there is an alignment relating terms from vocabulary *A* (start terms in table VIII.2) to vocabulary *B* (end terms) as well an alignment relating terms from *B* to *A*. Bilateral relations are not necessarily symmetrical. For example, the term 'Computer' in *A* is aligned with term 'Information System' in *B*, but the same term

[11] http://www.om2007.ontologymatching.org/
[12] http://watson.kmi.open.ac.uk

direction	# mappings	=	<	>	∧	null(0)	start terms	end terms
AG.-SWD	6254	5500 (4557)	100	314	337	3	6119	6062
SWD-AG.	11,189	6462 (4454)	3202	145	1188	192	10,254	6171

Table VIII.2: The AGROVOC-SWD alignment. Numbers of established mappings by type and by direction. The numbers in parentheses are the number of equivalence relations between concepts with literally identical terms.

'Information System' in *B* is aligned with another term 'Data base' in *A*. Bilateral mappings are only one approach to treat semantic heterogeneity; compare Hellweg et al. (2001) and Zeng and Chan (2004). The approach allows the following 1:1 or 1:n mappings: Equivalence (=) means identity, synonym, quasi-synonym; Broader terms (<) from a narrower to a broad; Narrower terms (>) from a broad to a narrower; association (∧): mapping between related terms; and null (0) which means that a term can not be mapped to another term. The first three of these relations correspond to the exactMatch, broadMatch, and narrowMatch relations from the SKOS Mapping Vocabulary. The AGROVOC-SWD alignment is a completely manually-constructed bilateral alignmennt that involves large parts of the vocabularies (see table VIII.2). Both vocabularies were analysed in terms of topical and syntactical overlap before the aligning started. All alignments in the GESIS-IZ approach are established by researchers, terminology experts, domain experts, and postgraduates. Essential for a successful alignment is the complete understanding of the meaning and semantics of the terms and the intensive use of the internal relations of the vocabularies concerned. This includes performing lots of simple syntactic checks of word stems but also semantic knowledge, *i.e.* to lookup synonyms and other related or associated terms.

The establishment of the alignment is based on the following practical rules and guidelines:

1. During the mapping of the terms, all existing intra-thesaurus relations (including scope notes) have to be used.
2. The utility of the established relations has to be checked. This is especially important for combinations of terms (1:n relations).
3. 1:1 relations are preferred.
4. Word groups and relevance adjustments have to be made consistently. In the end the semantics of the mappings are reviewed by experts and samples are empirically tested for document Recall and Precision (definition from information retrieval). Some examples of the rules in the KoMoHe approach can be found in Mayr and Petras (2008a).

VIII.5 Experimental Set-up

Given these two approaches, one completely carried out by human subject experts and the other by machines trying to simulate the human task, the basic questions are: "Who

performs more efficiently in a certain domain?", "What are the differences?", and "Where are the limits?". We hypothesize that:

1. Machines are humans' equals in domains with clear naming schemes, like taxonomy and geography. For other domains, machines are, as yet, inferior.

2. Machines are incapable of finding mappings that require knowledge of the domain that is not explicitly encoded in the thesauri that are to be aligned.

In order to test these hypotheses and draw conclusions, a qualitative assessment is needed.

VIII.5.1 Matching the Mappings

We first matched the mappings for the overlapping AGROVOC terms that have been mapped both to NALT and to SWD. For this we matched the AGROVOC term with the aligned NALT terms (in English) and the aligned SWD term (in German): about 5,000 AGROVOC terms have been aligned in both approaches. For the AGROVOC-NALT alignment, we took the entire set of suggestions made by five systems participating in OAEI 2007. We also listed the number of systems that have suggested the mapping between the AGROVOC and the NALT term (between 1 and 5) and the specific mapping that has been assigned in the SWD alignment (equality, broader, narrower or related match). In case of several suggestions for a mapping (*e.g.*, the AGROVOC term 'Energy value' has been suggested to be mapped to 'energy' as well as 'digestible protein' in the NAL thesaurus; the latter being an obvious mistake made by one of the systems.) we left all the multiple suggestions to be evaluated later. We then grouped the matched mappings into the higher level subject categories of AGROVOC and finally into four major terminology groups: Taxonomic, Biological/Chemical, Geographic, and Miscellaneous. These categories are the same as those used in the OAEI *food task* evaluation. This was done in order to be able to draw more detailed conclusions on the difficulty of mappings based on the terminology group a particular mapping falls into. These groups were chosen in order to be more specific on whom to contact to evaluate the respective mappings. This will give an indication on what kind of knowledge is generally harder for computer systems to map and what kind of background knowledge might also be needed to solve the more difficult cases.

VIII.5.2 Rating a Sample of the Mappings

Out of the about 5,000 mappings, we chose a representative sample of 643 mappings to be manually assessed. The mappings for the sample have been randomly drawn, fairly representing each of the terminology groups. We then assigned one of the following 6 difficulty ratings once for each of the mappings, AGROVOC-NALT and AGROVOC-SWD respectively. The assessments were done by Gudrun Johannsen and Willem Robert van Hage. Table VIII.3 summarizes our rating. Due to the strict protocol, there were few disagreements between the assessors. On a randomly drawn sample of 97 of these assessments, the unweighted Kappa was 0.86, which indicates almost perfect agreement.

	Rating	Explanation
1.	Simple	the prefLabels are literally the same / exact match
2.	Alt Label	there is a literal match with an alternative label / synonym in the other thesaurus
3.	Easy Lexical	the labels are so close that any laymen can see that they are the same terms/concepts
4.	Hard Lexical	the labels are very close, but one would have to know a little about the naming scheme used in the thesaurus (*e.g.* some medical phrases have a different meaning when the order of the words is changed and doctors know that)
5.	Easy Background Knowledge	there are no clues as in point 1-4 for a match, but the average adult laymen knows enough to conclude that there is a mapping
6.	Hard Background Knowledge	there are no clues as in point 1-4 for a match and you have to be an expert in some field, *e.g.* agriculture, chemistry, or medicine, to deduce that there is a mapping

Table VIII.3: Scale used to rate the mapping based on their "difficulty". The scale goes from 1 (Simple) to 6 (Hard Background Knowledge).

VIII.6 RESULTS

The assessment of the sample selection of 643 mappings is summarized in table VIII.4. The table is grouped by major subject groups: Taxonomic, Biological/Chemical and Miscellaneous. For both alignments (AGROVOC-NALT and AGROVOC-SWD), the table shows, what percentage of the mappings in the respective group are Simple, Easy Lexical, etc. The numbers in brackets are the absolute numbers. For example, in the group Miscellaneous: 18% of the AGROVOC-SWD mappings in this subject group have been found to be of difficulty 6 (Hard Background Knowledge), whereas only 1.4% of the AGROVOC-NALT mappings have been given this rating.

Table VIII.5 shows the mappings that have been wrongly assigned with the automatic approach in the AGROVOC-NALT alignment. In the assessment, we have specified if these wrong mappings should have been broader mappings (>), narrower mappings (<), related term mappings (∧) or simply completely wrong, *i.e.* null (0) and should not have been suggested.

The Geographic group has been left out from the table, since the sample contained only very few mappings (20). In any case, we can make the rather trivial statement that the Geographic group turns out to be rather simple, *i.e.* there seems to be an overall consensus on country names and other geographic concepts (in our case, the geographic group consists basically of country names). However, we have to be careful with this statement, especially when it comes to geopolitics. Borders of countries and similarly sensitive concepts might be called the same in two systems (and therefore seem simple and would be suggested by an automatic mapping tool with high security), but actually defined differently and mapping the two could raise sensitive issues. Take, for example, 'Taiwan': In AGROVOC, the non-preferred term 'China (Taiwan)' refers to the preferred term 'Taiwan', which has the broader term (BT) 'China', whereas in NALT 'Taiwan' USE FOR 'China (Taiwan)' has the broader

Taxonomic	Simple	Alt Label	Easy Lexical	Easy Backgr.	Hard Lexical	Hard Backgr.
AG.-SWD	27% (70)	39% (102)	7% (18)	3.4% (9)	6.5% (17)	17% (45)
AG.-NALT	65% (170)	23% (59)	1.1% (3)	0% (0)	1.9% (5)	0% (0)

Biological /Chemical	Simple	Alt Label	Easy Lexical	Easy Backgr.	Hard Lexical	Hard Backgr.
AG.-SWD	62% (53)	21% (18)	1.2% (1)	2.3% (2)	1.2% (1)	12% (10)
AG.-NALT	65% (55)	13% (11)	3.5% (3)	0% (0)	3.5% (3)	1.2% (1)

Misc.	Simple	Alt Label	Easy Lexical	Easy Backgr.	Hard Lexical	Hard Backgr.
AG.-SWD	33% (92)	12% (33)	10% (28)	17% (46)	9.8% (27)	18% (50)
AG.-NALT	49% (136)	24% (67)	4.0% (11)	0.36% (1)	1.8% (5)	1.4% (4)

Table VIII.4: Rating of the mappings by terminology groups (taxonomic, biological, miscellaneous) and by rating of difficulty.

should be:	<	>	null (0)	∧	total wrong
Taxonomic	2.7% (7)	0.38% (1)	5.7% (15)	0.38% (1)	9.2% (24 of 262)
Biological / Chemical	2.3% (2)	1.2% (1)	11% (9)	0% (0)	14% (12 of 84)
Miscellaneous	1.4% (4)	0.36% (1)	14% (38)	3.3% (9)	19% (52 of 277)
all groups	2.0% (13)	0.0% (3)	9.6% (62)	1.5% (10)	14% (88 of 643)

Table VIII.5: Mapping of AGROVOC-NALT. Classification of wrong equivalence mappings. 21 geographical mappings were omitted.

term 'East Asia'. Another example, which is currently an issue, is the concept 'Macedonia'. It has been used in the Codex Alimentarius[13] to refer to the former Yugoslavian Republic of Macedonia. However, since there is also a region in Greece, which is called Macedonia, the Greek authorities have requested the Codex Alimentarius to use 'The former Yugoslavian Republic of' in the name of the concept. Moreover, country definitions are time dependent. How a user might best map geographical terms depends on the use case. For some purposes, where the alignment is not used in mission critical situations, automatic alignment can be a quick and good solution. For other purposes it might be better to map all geographical terms manually, which is generally feasible due to the relatively small number of countries in the world (as compared, for example, to plant species).

[13]The Codex Alimentarius Commission was created in 1963 by FAO and WHO to develop food standards, guidelines and related texts such as codes of practice under the Joint FAO/WHO Food Standards Programme. The main purposes of this Programme are protecting health of the consumers, ensuring fair trade practices in the food trade, and promoting coordination of all food standards work undertaken by international governmental and non-governmental organizations. It is available at: `http://www.codexalimentarius.net/web/index_en.jsp`.

VIII.7 ANALYSIS

Analyzing the other groups listed in the table leads to the few first statements: First of all, we can say that in general, Biological/Chemical like Geographical terminology is fairly easy to align (over 60% rated as Simple). This result makes sense, since like for geographical concepts there is probably a good consensus in the world on names of biological entities and chemicals.[14] Taking into account the alternative labels, this statement also holds for the group of taxonomic terminology alignment. Apparently, in the German language there are more discrepancies on the usage of preferred versus non-preferred labels and synonyms than in the English language. The Miscellaneous group (which includes the majority of mappings) appears to be the most difficult one. About 14% of the automatically suggested mappings were even wrong, and it shows the highest percentage of Hard Background Knowledge mappings.

Furthermore, the mappings, we found that the AGROVOC-SWD alignment has a considerable amount of broader (>) and narrower (<) mappings. These are in general more difficult to find than equivalence mappings (either very easy or very difficult, because Hard Background Knowledge may be required), and therefore pose a big problem to automatic alignment techniques. The SWD part on agriculture is also considerably smaller than the AGROVOC or NAL thesaurus and therefore many broader and narrower mappings are possible. Automatic alignment techniques have difficulty with such discrepancies. Apparently, subterms are often a good lexical clue for a < or > relation, but how does a computer decide which of the subterms is the superclass? Sometimes it is easy because one of the subterms is an adjective, while the other is a noun (*e.g.* 'mechanical damage' is a damage), but sometimes both are nouns (*e.g.* 'Bos taurus' is a Bos, not a taurus, but 'fruit harvester' is a harvester), and this is hard to parse. There are also cases where lexical inclusion can bring confusion, for example, 'Meerrettich' (horseradish is 'Armoracia rusticana') and 'Meerrettichbaum' (horseradish tree is 'Moringa oleifera'), as they refer to completely different concepts. Eventually, this problem might be solved by machine learning, but current alignment systems do not have any functionality to detect various common naming conventions.

It is remarkable that for the harder mappings (Hard Lexical, Easy Background, Hard Background), the percentage that has been found by the automatic approaches is overall very little (at most 3.5% for Hard Lexical biological/chemical terms), whereas the manual alignment approach can obviously identify these mappings. For example, in the Miscellaneous group, more than 40% of the manual AGROVOC-SWD mappings fall into one of the three hardest ratings. The automatic mappings with this rating accumulate to less than 4%. Table VIII.5 shows the numbers of wrong automatic mapping suggestions. The percentages in the three hardest ratings of the AGROVOC-NALT alignment are obviously cases of wrong suggestions, as listed in table VIII.5, which were either completely wrong mappings or should have been broader, narrower or related mappings.

It is not impossible, however, for automatic techniques to also detect even Hard Background Knowledge mappings, for example, by means of text mining. Some of these are eas-

[14] Organizations like The American Chemical Society (CAS, `http://www.cas.org/expertise/cascontent/registry`) maintains lists of unique identifiers for chemicals in various languages. Various resources are also available that relate various chemical names to their CAS identifiers.

ier to solve than others, because some background knowledge is simply easier to find. For instance, there are many web pages about taxonomy, but few about 'Lebensmittelanalyse' (food analysis). There are also many about chemicals, but few that state that a 'Heckstapler' (rear stapler) is some kind of 'Handhabungsgeraet' (handling equipment).

Some more concrete examples of mappings of varying difficuly:

1. Mapping rated Alt label. AGROVOC-NALT 'Marketing Strategies' = 'Marketing Techniques'. This mapping has been rated 'alt label', since, for example, in AGROVOC, 'Marketing Strategy' is the non-descriptor of 'Marketing Techniques'. This case makes it easy for an automatic classifier. However, this might also be misleading. In the agriculture domain, it might be correct to declare equivalence between these terms. However, in another domain there might actually be no mapping or at most a related term mapping. For example, in the business area, marketing strategies differ from marketing techniques substantially in that the strategies are long term objectives and roadmaps whereas the marketing techniques are operational techniques used in the marketing of certain products. For an automatic alignment system, this is difficult to detect and alternative labels as they are sometimes found in thesauri, might be misleading.

2. Mapping rated Hard Background Knowledge. Both in AGROVOC and the NAL Thesaurus there is the term 'falcons' (exact match, simple mapping) while in SWD the German term 'Falke' does not exist, and thus had to be mapped to the broader term 'Greifvögel' (predatory birds) which requires human background knowledge. However, in this case, the human knowledge could be found by a mapping system, if it would exploit the German Wikipedia. On the page about Falke,[15] it states: "Die Falken (Gattung Falco) sind Greifvögel...".

3. Mapping rated Hard Background Knowledge. In SWD the term 'Laubfresser' (folivore) which does not exist in AGROVOC or in NALT had to be mapped to the broader term 'Herbivore'. This is another example where Hard Background Knowledge is needed.

4. Sometimes terms which seem to match exactly are incorrectly machine-mapped, for example, when they are homonyms. Example: 'Viola' in AGROVOC it is the taxonomic name of a plant (violets) while in SWD it refers to a musical instrument. In this case the relationship is o. Sense disambiguation techniques such as the ontology partitioning performed by some of the current mapping systems, like Falcon-AO, should be able to solve most of these ambiguities by recognizing that none of the broader or narrower terms of 'Viola' and 'violet' are similar.

Some of the mappings of course will remain impossible for automatic techniques that do not exploit sources of background knowledge, for example, one of the AGROVOC-SWD mappings that found that 'Kater' (tomcat) is a 'männliches Individuum' (male individual).

[15] `http://de.wiktionary.org/wiki/Falke` or `http://de.wiktionary.org/wiki/Greifvogel`.

VIII.8 CONCLUSION

We conclude that for the alignments and automatic alignment systems considered in this study, hypothesis 1 does not hold as strictly as it was phrased. For the geographical mappings concidered in this study, automatic systems were indeed indistinguisable from humans, but for taxonomical alignment, they are slightly worse than humans. Furthermore, if we consider the percentage of simple taxonomical mappings in the AGROVOC-SWD and AGROVOC-NALT alignments (see the first column of the first two rows of table VIII.4), we can see that the AGROVOC-NALT taxonomical terms are much easier to align than the AGROVOC-SWD taxonomical terms. This is due to AGROVOC and NALT's common origins. The taxonomical terms are largely based on the same literature, while this is not the case with SWD and AGROVOC. Based on the first row of table VIII.4, we conclude that humans compensate for this higher difficulty by drawing on internalized background knowledge and linguistic understanding. The latter asset can currently be compensated for with Natural Language Processing techniques, but there is, as yet, no good technique to compensate for the former.

An assumption behind hypothesis 1 was that taxonomy and geography were two easy domains, while biological and chemical terms like genes (with their many synonyms) would be considerably more difficult for automatic techniques than other domains. Row three and four of table VIII.4 shows this to be a wrong assumption. Less background knowledge was needed for the alignment of these terms, than for the alignment of taxonomical terms, while lexical techniques were more useful.

We conclude that hypothesis 2 holds. Not taking into account the six mappings where either easy or hard background knowledge was needed by human aligners. We consider these six mappings, which are listed under easy and hard background knowledge in rows two and four of table VIII.4 as "noise", *i.e.* lucky guesses based on other features than background knowledge by the automatic alignment techniques.

We have seen that automatic alignment can definitely be very helpful and effective in case of Simple and Easy Lexical mappings. From our results, it appears that apart from Taxonomic vocabulary and Geographic concepts, also Biological and Chemical Terminology falls into this category. In general, there seems to be more consensus on how to name concepts in these domains than in the other domains covered by AGROVOC. However, we need to be careful in these areas, where often word similarity does not mean that this is a potential mapping. These can be serious traps for automatic alignment techniques (like in the case of geopolitical issues).

Things get potentially more difficult in the case of more diversified groups/categories (in our case just summarized as Miscellaneous). Here, often background knowledge is needed to infer the correct mapping, and automatic alignment systems are able to identify only very little of these correctly. Most of the automatic suggestions are simply wrong or should not be equivalence relationships but broader, narrower or related terms.

The bottom line is that for the moment, alignment should not be seen as a monolithic exercise, but we can take the best of both approaches and use automatic alignment approaches to get to the simple and easy lexical mappings and then use human knowledge to control the ambiguous cases.

Acknowledgments

We would like to thank Lori Finch at the NAL for her extensive help on the AGROVOC-NALT alignment and for many discussions that contributed to this work. Van Hage was supported by the Dutch BSIK project Virtual Laboratory for e-science (http://www.vl-e.nl). The project at GESIS-IZ was funded by the German Federal Ministry for Education and Research, grant no. 01C5953. Philipp Mayr wishes to thank all our project partners and my colleagues in Bonn for their collaboration.

Chapter IX

Conclusions and Discussion

In this chapter we conclude this thesis by reflecting on the four research questions posed in chapter I. For each of these questions we recall the results from the chapters that addressed it, draw conclusions, and discuss related issues. We close with a general discussion and recommendations for future work.

IX.1 Revisiting the Research Questions

I What is the quality of current alignment techniques?

The first research question captured our desire to know how well current state-of-the-art techniques can find alignment relations. We addressed this question in two ways. First, in chapter II, III, and IV, by developing new techniques where no satisfactory alternatives exist and measuring their performance in practical use cases. Second, in chapter V, by investigating evaluation methodology and by applying it to existing alignment systems. By the results from these chapters we get an overview of the quality of current techniques for finding equivalence, subclass, and partitive relations–the three most common alignment relations–in the scope of agricultural information-retrieval tasks.

Equivalence Alignment We have shown in chapter V and its appendix V.8 that the Falcon-AO system currently implements the best equivalence alignment technique for the alignment of agricultural thesauri. In table V.5 on page 66 and table V.6 on page 67 we can see that the 2007 version of the system (Falcon-AO 0.7) achieves 0.83 Precision and 0.90 Recall for the discovery of equivalence relations between AGROVOC and NALT. On the tasks of aligning AGROVOC and NALT to the less structured GEMET thesaurus, Falcon-AO achieves similar Precision, but lower Recall. In table V.12 on page 82 and table V.13 on page 82 we can see that, for aligning GEMET to AGROVOC and NALT, Falcon-AO achieves respectively 0.88 and 0.86 Precision, and 0.60 and 0.50 Recall. A summary of these results is shown in table IX.1.

The reason why Falcon-AO currently outperforms the other alignment systems can be summarized as: It is better at deciding when to ignore low-confidence matches. The most important part of schema-based matching (as opposed to instance-based matching) is matching labels. Nearly all of the correct mappings found by Falcon-AO can be attributed mainly to lexical matches. Other techniques, such as structural matching, can help to rank the results and to provide matches when no lexical matches exist, but they play a minor role in aligning thesauri. The most important part of lexical matching is to decide when to

alignment task	approach	type	Precision	Recall
OAEI 2006 AGROVOC-NALT	RiMOM	equivalence	0.81	0.71
OAEI 2007 AGROVOC-NALT	Falcon-AO	equivalence	**0.83**	**0.90**
OAEI 2007 AGROVOC-GEMET	Falcon-AO	equivalence	**0.88**	**0.60**
OAEI 2007 NALT-GEMET	Falcon-AO	equivalence	**0.86**	**0.50**

Table IX.1: The quality of approaches for the alignment of agricultural thesauri using equivalence relations.

alignment task	approach	type	Precision	Recall
AGROVOC-SR-16	*Hearst Patt. and Google Hits*	subclass	0.17–0.30	**0.32–0.53**
AGROVOC-SR-16	*Hearst Patt. and Google Snippets*	subclass	**0.38–0.50**	0.22–0.37

Table IX.2: The quality of approaches for the alignment of agricultural thesauri using subclass relations.

tolerate slight differences between the labels and when to be strict. When there is already an obvious match, adding low-confidence alternatives harms the quality of the result. A discussion of this issue can be found in section V.6 on page 71.

SUBCLASS ALIGNMENT We have shown in chapter II that it is possible to find subclass alignment relations using a web search engine. Finding subclass alignment relations is significantly harder than finding equivalence alignment relations. The most important reason for this is that there is less lexical evidence for them. The two primary causes of this are: As opposed to equivalent classes, subclasses have different names; and subclass relations are not mentioned often in natural language, because they are assumed to be common knowledge. In table IX.2 we summarize results from table II.2 on page 21 and table II.3 on page II.3. In this table we can see that the technique based on hit counts achieves better Recall (roughly between 30% and 50%), while the technique based on syntactic analysis of snippets achieves better Precision (roughly between 40% and 50%). Steeper subclass relations (*i.e.* where the level of abstraction between the subclass and superclass is larger) are easier to find than more even subclass relations (*i.e.* closer to equivalence relations). The quality of other techniques in the literature vary, depending on the domain, but subclass learning is far from a solved problem.

Two alternative approaches that also use background knowledge, *Extraction from a Dictionary* (see section II.4.3) and the SCARLET system (Sabou et al., 2007), achieve higher Precision, but find few or 'unintuitive' results. Extraction from the CooksRecipes.com Cooking Dictionary (see table II.4 on 23) yielded 477 correct relations, but only 16 of these were mappings between AGROVOC and SR-16, the rest led to concepts outside either ontology. SCARLET (see table V.1 on page 56) found 6,038 subclass relations between AGROVOC and NALT using the Watson semantic-web search engine, of which depending on the evaluation either 25% (see table V.5 on page 66) or 70% (Sabou et al., 2007) are correct.[1] All of

[1]This large difference can be attributed to the strong weight attached to the taxonomical stratum in the OAEI 2007 food task evaluation, see table V.2 on page 60. The evaluation in Sabou et al. (2007) randomly selected mappings from the result set, weighing each topic equally.

alignment task	approach	type	Precision	Recall
AG./NALT-IARC	*Lexical Patt. and Google Snippets*	part-whole	**0.74**	**0.82**
SemEval 2007	*WordNet Similarity Measures*	part-whole	**0.54**	0.73
SemEval 2007	*Lexical Patt. and Google Snippets*	part-whole	0.36	**1**
SemEval 2007	*WordNet Similarity Measures*	containment	**0.66**	0.55
SemEval 2007	*Lexical Patt. and Google Snippets*	containment	0.51	**0.97**

Table IX.3: The quality of approaches for the alignment of agricultural thesauri to a controlled vocabulary (AGROVOC/NALT-IARC Group 1 carcinogens) using partitive relations, and the classification of general-domain partitive relations (SemEval 2007).

these subclass relations were steeper than those that would have been suggested by human experts, see the paragraph about Recall in section V.5.2 on page 68.

Partitive Alignment We have shown in chapter III and IV that the text-mining techniques used for learning subclass relations can also be used for learning partitive relations. There are many kinds of partitive relations, like place-area, member-collection, and stuff-object. In chapter III we investigated learning all six types of partitive relations described by Winston et al. (1987) from the web in the context of a food-safety retrieval task. In chapter IV we investigated these six types of invariant part-whole relation as well as temporary containment, like eggs in a basket or people in a bus. In table IX.3 we summarize the results from table III.6 on page 37, table III.4 on page 35, and of table IV.4 on page 47 and table IV.3 on page 47. Partitive alignment is easier than subclass alignment, but harder than equivalence alignment.[2] As opposed to subclass relations, part-whole relations are not considered to be common knowledge. To the contrary, in fields like anatomy, topography, recipes and chemistry, they are cherished assets and commonly discussed (or they are well guarded secrets).

Evaluation Methodology Realistic alignment tasks are too large for exhaustive evaluation. It is possible to trade in some certainty about the results for a reduction in evaluation effort. In chapter VI we propose two methods for sample-based evaluation of alignment approaches: *Alignment Sample Evaluation* and *End-to-end Evaluation*. The former provides more insight in the strengths and weaknesses of alignment techniques with respect to which mappings are discovered and which are overlooked. The latter provides a better estimation of the performance of alignment techniques as part of an application. In chapter VII we introduce *Relevance-based Evaluation*, a method that combines the advantages of *Alignment Sample Evaluation* and *End-to-end Evaluation*. This approach can be summarized as: Performing *Alignment Sample Evaluation* on samples deduced from the requirements of *End-to-end Evaluation* usage scenarios.

Apart from the cost of aligning vocabularies once, there is a more fundamental reason that makes sample evaluation the only viable evaluation method for ontology alignment. This is the never ending dynamics of the world, ontologies, and application requirements.

[2] At least, this goes for finding partitive relations in general. The determination of the part-whole subtype is another problem.

Complete evaluation of an alignment makes no sense if the alignment and its use will be different tomorrow. Ongoing small-scale evaluation experiments that represent current demands are more suitable for the open-ended nature of networked ontologies.

II Which domain-specific factors influence the quality of alignment techniques?

The second research question aimed at partially explaining the differences in quality between the various alignment tasks undertaken to answer the first research question. The aspect we considered for this question was the domain of the aligned concepts. We approached this question by performing separate evaluations for a number of domains and alignment-relation types. In chapter ii we investigated subclass alignment in the domain of food products found in supermarkets and categories of these products, like dairy products and confectionery. In chapter iii we investigated part-whole alignment in the domain of food safety. Specifically, we look into chemicals and media that can contain these chemicals, such as food stuffs, air, water, and objects people come into physical contact with on a regular basis, like construction materials. In chapter iv we investigated seven relation types, amongst which part-whole and containment relations, without a domain restriction. As opposed to chapter iii this also includes, for example, metaphorical or intangible part-whole relations, like parts of a problem. In chapter v we investigated equivalence alignment between two agricultural thesauri with a broad scope. These thesauri themselves deal with many domains. We perform separate evaluations for a few clearly separable domains: geography, taxonomy, biochemistry, and the remaining domains as a whole. In chapter viii we investigated the difference between the alignment between two agricultural thesauri and a general domain thesaurus.

Variation in Alignment Relation Types Some alignment relation types are easier to learn than others. Of the relations discussed in chapter ii, iii, iv, and v, equivalence relations are the easiest to find, then partitive relations and then subclass relations. The work described in these chapters, combined with related work on learning relations of various types (*e.g.* subclass, author-work, person-profession) from web search engines (Cimiano and Staab, 2004; Geleijnse et al., 2006; Geleijnse and Korst, 2007a,b; Ruenes, 2007), leads us to conclude that there are two factors that influence the learning difficulty of relation types the most and which are tied to the domain. First, some relations have stricter domain (*the mathematical sense of the word!*) or range restrictions than others. For example, authorship relations always hold between a person and a work, while subclass relations can potentially hold between any classes with similar properties (*cf.* Guarino and Welty (2004)). The more restricted the domain and range of a relation are, the easier it is to learn, because it is easier to filter out incorrect relation instances. For example, member-collection partitive relations are restricted to subjects that are entities and objects that are groups of entities. Therefore, a relation between an entity and an event can never be a valid member-collection relation. Second, some relations are subject to more discussion on the web than others. For example, more people talk about Francis Ford Coppola being the author of The Godfather, than about bell peppers being a kind of fruit. The more evidence there is for a relation, the easier

it is to learn.

High and Low Consensus Domains In some domains, like anatomy and geography, there is more consensus about how to structure knowledge than in other domains, like farming, or economy. There are few of the former and many of the latter kind of domains. What matters the most to ontology alignment is not if *one* community has worked out a domain to high level of detail and understanding, but whether this understanding is shared amongst communities. Ontology alignment can be seen as an automation of the ongoing discussion that leads to understanding of each other's perspective on a domain. Therefore, the more discussion has happened before ontologies are aligned, the fewer decisions are left to the alignment system.

In chapter V we distinguished three domains with a relatively high level of consensus from the rest. These are, in order of decreasing level of consensus: geographical terms (countries, etc.), biochemical terms (proteins, chemicals, etc.), and taxonomical terms (both scientific and vernacular names of species). Other domains (farming techniques, rural economy, etc.) on average have a lower level of consensus. This shows itself in the alignment system performance results in table V.5 on page V.5.

Naming Schemes versus Generic Language Some domains use their own language to describe concepts. Two examples are taxonomy and chemistry. In taxonomy this specific 'language' actually consists of two languages: latin, and the binomial (*e.g.* 'Homo sapiens') and trinomial (*e.g.* 'Homo sapiens idaltu') nomenclature governed by Nomenclature Codes such as the International Code of Botanical Nomenclature (ICBN), International Code of Nomenclature for Cultivated Plants (ICNCP), and their counterparts for other kingdoms. In chemistry there are various regulated types of names, such as International Union of Pure and Applied Chemistry nomenclature formulae like 1,3,7-trimethyl-1H-purine-2,6(3H,7H)-dione for caffeine, chemical formulae like $C_8H_{10}N_4O_2$, and Chemical Abstracts Service (CAS) numbers like 58-08-2. Alignment of ontologies that both use such a standardized naming scheme is simple if the alignment technique recognizes the naming scheme. Otherwise, it can be a source of systematic error. For example, alignment systems that assume all labels follow general english grammar will parse 'Homo' the term 'Homo sapiens' as an adjective and 'sapiens' as a noun, which can lead the system to conclude a 'Homo sapiens' is a kind of 'sapiens', as opposed to a kind of 'Homo'. None of the alignment systems discussed in this thesis support the adaptive recognition of naming schemes.

III Which application-specific factors influence the quality of alignment techniques?

The perceived quality of alignment techniques is not only influenced by the domain of the alignment, but also how the alignment is used. With the third research question we wanted to reveal how applications dictate what is a good alignment. We approached this question in three ways: (1) We tested alignment methods in the context of specific application scenarios, a fact-finding scenario in chapter III and a metadata-based retrieval scenario in chapter VII; (2) We investigated omissions in automatically created alignments in chapter

V and VIII to see what kind of mappings are not available for prospective applications; and (3) we developed methods to incorporate application requirements into the evaluation of alignment techniques.

The two application-specific factors we recognized to be the most important in this thesis are reliability and relevance.

REQUIRED RELIABILITY OF RESULTS An alignment with a certain accurateness and completeness can work well for one application, but can fall short for another. Applications for exploratory browsing of resources require a lower level of reliability than search applications that offer a small high-quality list of results, which in turn require a lower level of reliability than decision-support systems. In this thesis we approach ontology alignment from the perspective of metadata-based information retrieval, an application type with intermediate quality requirements, where there is usually a human in the loop, for example, at the end of the retrieval process selecting the relevant hits from a list of resources. Compared to fully-automated tasks, such interactive tasks permit a lower level of reliability.

The nature of the task also influences the required reliability of relations. For example, if the task is classifying resources into broad subject categories using subclass mappings, then it does not matter much how strict (*i.e.* close to equivalence) the mappings are, as long as low-level concepts are subsumed by the correct high-level concepts. However, if the task demands more fine-grained reasoning, its reliability requirements will be higher.

APPLICATION RELEVANCE Not every mapping in an alignment is equally valuable to every application. For a specific application, some topics might be more relevant than others, or some mapping types might be more relevant than others. An alignment technique might be good at aligning taxonomical terms, and bad at aligning medical terms. If there are many more taxonomical terms than medical terms then the average case quality of technique might be very high, but it would not work well for an application that requires geographical mappings.

An evaluation based on the achievement of application goals can be used to determine alignment quality in the context of an application. In chapter III and VII we demonstrate two such evaluations, that account for application demands. The evaluations described in chapter II and V do not take application demands into account. In chapter V we do divide the evaluation into separate samples for different domains. This allows limited conclusions about how the alignment will perform in an application in a certain domain.

IV How can human preference be incorporated in the evaluation of automatic alignment methods?

The fourth and last research question dealt with evaluation methodology. We see ontology alignment as an automation of alignment by humans, and hence, as something that is only undertaken when it is worth a human's time and effort. Evaluation methods guide the development of new alignment techniques. They dictate the standard by which the quality of techniques is compared. It is important that the properties captured by evaluation methods coincide with what is deemed good in reality. In this thesis we distinguish two

properties of a good automatically-generated alignment: (1) It is indistinguishable from human work; and (2) It induces good results in an application. Hence, one alignment can be considered to be better than another alignment when it either emulates human work more accurately, or causes better results in an application, than the other alignment. Most current evaluation methods ignore the application demands, because these are inherently subjective. In this thesis we argue that, although such neutral evaluation methods are seemingly objective, they do not adequately capture the second property of a good alignment. That is, they do not measure whether an alignment will yield good results in an application. The fourth research question called for a remedy for this shortcoming. We addressed this question by proposing new evaluation methods that incorporate application demands, and by determining typical differences between manual and automatic alignments. In chapter VI we proposed *end-to-end evaluation*, which measures alignment quality only by application performance, ignoring the alignment itself altogether. In chapter VII we proposed an alternative method, *relevance-based evaluation*, an adaptation of *alignment sample evaluation*, discussed in chapter VI. *Relevance-based evaluation* measures alignment quality on a sample of mappings that are required for the successful completion of a number of prototypical application scenarios. This produces results that are biased towards application demands. In chapter VIII we studied the manual alignment between AGROVOC and the Schlagwortnormdatei, and the automatic alignment between AGROVOC and NALT. Exploring the difference allowed us to see how current automatic alignment techniques fail to emulate humans.

Goal-Oriented Approaches With respect to the quality of an ontology alignment, there are a number of measurable features that correlate with human preference. Some have to do with a specific application, like the number of relevant resources an application returns given a certain alignment, or the accuracy of predictions given an alignment. Others have to do with the alignment itself, like the percentage of sound statements in the alignment, or the average of how many human judges would suggest each mapping. Which features are useful for quality prediction depends on how well they predict satisfaction and how expensive it is to measure them.

We propose a number of evaluation methods that make use of different measurable features. In chapter III we measure Recall based on a sample set of desirable mappings that is both representative of all mappings we need for a food-safety use case, and is easy to set up. In chapter V we manually construct and assess sample sets of mappings, and use *alignment sample evaluation* to give an overview of alignment quality without a given use case. In chapter VII we use a double-annotated corpus and reverse engineer which alignments would be necessary to find documents using another vocabulary than was used for indexing. In this approach we ignore the actual retrieval strategy. In chapter VI we propose a method for *end-to-end evaluation* that could be used to measure the impact of alignments on retrieval strategies and hence to infer the quality of the alignment given a specific use case. In chapter VIII we randomly select a set of mappings and manually analyze how difficult each mapping is to find. Based on this analysis we show how well each difficulty category is covered by current alignment systems and by humans.

TUNNEL VISION When an evaluation is based on automatically generated mappings, it inherits the limitations of the method that generated the mappings. For example, if a reference alignment is based on automatic instance-based matches, it inherits the shortcomings of the classifiers that categorized the instances. If an evaluation is based on a reference alignment that was constructed on restricted sets of concepts (*e.g.* the Recall samples in chapter V listed in table V.4 on page 64), we inherit the restricted scope of the sample sets. For example, if we only align similar subhierarchies, we ignore possibly valuable matches with concepts in other subhierarchies. Such mappings are harder to find, for humans and computer alike, but can be valuable. When time constraints dictate that a reference alignment has to be constructed with a limited scope of concepts, always consider mappings that lead out of the scope, *cf.* section VI.2 and figure VI.2 on page 86.

IX.2 DISCUSSION AND FUTURE WORK

IX.2.1 REFLECTION ON THE APPLICABILITY OF ALIGNMENT TECHNIQUES

Whether automation is a viable option depends on many factors: time, money, and the desired quality of the alignment. As we concluded in the previous section, the quality of the alignment is dependent on the topic of the alignment, the type of relations, the complexity of the domain, and the kind of background knowledge that is readily available.

The effort to automate ontology alignment can outweigh the effort to create the mappings manually, especially for small alignments, *cf.* the paragraph on application of thesaurus alignment on page 78. Also, the power of communities should not be underestimated. Many hands make light work. The FAO's new AGROVOC Concept Server Workbench[3] acknowledges this. We do not trivialize the role of automatic alignment techniques, because they can do exactly what humans are not willing to do: aligning concepts that would have been obvious matches were it not that they are hidden between thousands of other concepts. This takes care of the largest part of the problem and frees resources to tackle the rest.

Automatic alignment can be a way to get a project started, after which humans take over the alignment process to find the remaining mappings. The converse is also possible. Manual alignment can provide seed mappings by which automatic techniques can be trained or by which the alignment process can be streamlined (*e.g.* based on seed mappings, large ontologies can be partitioned, *cf.* the PBM algorithm of Falcon-AO, Hu et al. (2006)).

IX.2.2 THE NATURE AND SEMANTICS OF ALIGNMENT RELATIONS

The alignment relations discussed in this thesis are outlined in figure IX.1. Listed are the SKOS alignment relations used in chapter V, along with their stronger, more specific RDF(S) or OWL counter parts, amongst which the subClassOf relation used in chapter II. Also shown is the part-whole relation used in chapter III, which does not belong to the predefined RDF(S) or OWL relations, but is clearly defined by Winston et al. (1987). Each SKOS relation shown can be considered a superproperty of the relations listed to the right of it in

[3] http://www.fao.org/aims/agrovoccs.jsp

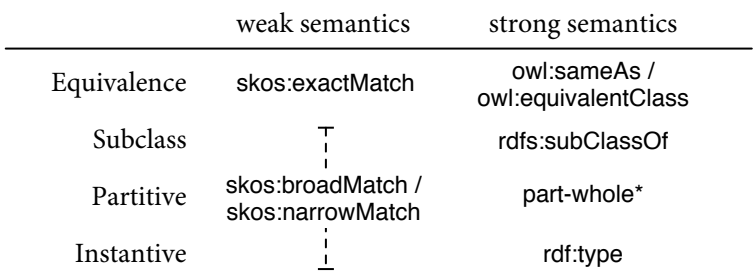


Figure IX.1: Alignment relations discussed in this thesis.

figure IX.1. For the alignment of thesauri, the SKOS alignment relations are more suitable than the RDF(S) or OWL relations, although specific applications, such as the retrieval task in chapter III can warrant these stricter relations.

The weak semantics of thesauri leave many things underspecified that can cause reasoning conflicts when the alignment relations are very strict. For example, one thesaurus can use the term 'snails' to indicate all snails, while the other uses the term to refer to the species. A book can be about snails in general, but not about certain individual snails, or vice versa.[4] When these two different senses of a term are aligned using owl:equivalentClass the meaning of both terms is compromised. A skos:exactMatch does not imply logical equivalence between the concepts. It does not even imply that the preferred and alternative labels (*i.e.* descriptor and non-descriptor terms, the concept name and its synonyms) of the two concepts are transferred to the other thesaurus. The mapping only gives access to them. Whether this access is exploited can be decided later by rules or by a user. Currently, there is discussion about dividing the current exactMatch relation into a relation indicating real synonymy and a more relaxed relation, closeMatch. This would allow exactMatch to become more meaningful. Which would allow exactMatch to become transitive, while remaining weaker than owl:equivalentClass or owl:sameAs.

OWL reasoners, even given global extensions to OWL like C-OWL (Bouquet et al., 2003), can not deal well with contradictions at the moment. Strict mappings or bridging rules can only work if there is a high level of consensus about the atomic concepts of a field.[5] The current state-of-the-art is far away from being able to automatically induce a sound bridge rule between the topographical and administrative 'Ireland' in the example on page 74. A human would have to do modeling to formalize the properties that connect and divide the different senses of 'Ireland', see figure V.11 on page 74. Given any level of formalization, there will always be cases where additional modeling is required to connect different views. People just disagree, and while they do, the world also changes.[6] Sometimes it is even unclear how concepts differ, while it is clear that they do. One promising solution is to relax the reasoning mechanism, *cf.* Huang et al. (2005); Huang and van Harmelen

*OWL does not have a predefined partitive relation. In this thesis we defined our own general partitive relation, which is a superproperty of the six types of partitive relations described by Winston et al. (1987).

[4]Then there is also the subject as such, which is not the same as the species or specimens.

[5]Disregarding for a moment whether there is such a thing as one single truth.

[6]Another field altogether, see Klein (2004).

(2008). In a sense, this approach attempts to model human intuition about which implication are more likely to be true than others. This approach has not been widely adopted yet. Another solution is, as was hinted at before, to use alignments with weak semantics. Such light-weight alignments can promote the forming of consensus by connecting conflicting, but similar, concepts by putting the judgement in the capable hands of the user. To simply know that two concepts are very similar, although their definitions conflict, can be valuable in applications where users can select useful results by "cherry picking" from a set of candidates, *cf. e.g.* the result list of search engines or galleries, where a few bad items are permissive.

An implicit assumption in ontology alignment is that between two ontologies there is *one* correct alignment. If we consider the two different concepts named 'Ireland' in the previous paragraph, for example, then the assumption is that there is one sound way to connect these two concepts. Possibly, in this case, this one way is: One 'Ireland' is the state which has sovereignty over a physical region that is part of a larger physical region that is the island represented by the other 'Ireland'. Alignments formulated in a restricted set of general alignment relations, such as the SKOS Mapping Vocabulary, generalize over this complex mapping. For example, 'Ireland' broadMatch 'Ireland' is a generalization of: 'Ireland' has sovereignty over φ, φ part of 'Ireland'. The weaker the semantics of the alignment relations are, the greater the loss of accuracy is between the "one true alignment" and the actual alignment. Weak relations give up some completeness for soundness and simplicity. This is sometimes seen as an unnecessary sacrifice, because given enough time, the parties involved in the alignment could engineer a more complete, more specific, closer approximation of the correct alignment. However, there are three things that stand in the way of this approach. First, the parties themselves are sometimes unable to decide the exact semantics of their concepts, especially for weak semantic structures like thesauri, and this is a prerequisite for a complex alignment. Second, complex alignments necessarily consist of many different relation types, due to the complexity of the world they represent. There are domain-independent relations, such as subclass and the other relation types considered in this thesis, but each domain and each perspective on that domain has its own additional relations, like has sovereignty over. Like in regular ontology engineering, one has to decide on a relatively small set of reusable relation types, or the ontology will become unwieldy. Third, there is no time. Automatic alignment techniques would have to be able to derive complex alignments with many relation types. At this moment they are not. Regardless of whether the implicit assumption of one correct alignment is true, we are stuck with a limited number of alignment relations for the moment. If we want to extend this set of relation types we should try to find those that can be reused in many applications and that we can teach automatic alignment systems to find.

IX.2.3 A User's Perspective

It is hard to give recommendations on how to proceed with ontology alignment for an end user of the alignment without knowing the use case that warrants the alignment. We attempt to do so anyway in figure IX.2, if only to clarify our view on interactive ontology alignment. Perhaps this outline can serve as a suggestion to parties that want to undertake

ontology alignment.

Currently, alignment systems focus on step 8 of this process. However, most of the analysis steps under 5 could be automated. Also, the partitioning of the ontologies in step 6 can be partially automated. Perhaps the inherently manual steps, like the politics of the first two steps, and the formulation of scenario's, can even benefit from computerized assistance.

Useful things to remember when planning an alignment project are that verifying mappings takes considerably less time than constructing mappings by searching for matching concepts in the ontologies, see section V.4 on page 55, and that verifying suggested mappings requires considerably less domain expertese than construction from scratch, see the paragraph on inter-judge agreement in section V.4.1 on page 60.

IX.2.4 A Computer Scientist's Perspective

Only part of the alignment problem can be solved by current techniques. Provided that the labels of the ontologies use the same language, most of the simple mappings can be found and some of the harder. There are very few experimental results about truly multilingual alignment. To reach beyond the simple mappings, completely new alignment techniques will have to be investigated. A concrete example is that background knowledge should be incorporated into the alignment process, *cf.* Aleksovski (2008). Not only third-party ontologies should be used as sources, *cf.* Sabou et al. (2007), but also semi-structured sources of knowledge, like Wikipedia, free text, and other concrete-domain data. Also, social interaction should be investigated as a source of knowledge. In general, people underestimate how much of their world views are inherited from discourse and other forms of social interaction, as opposed to direct empirical evidence. Instance-based alignment is a very dependable and objective method, *cf.* Isaac et al. (2007). The reason why it is not used more than it is today, is lack of instance data. Social applications, like flickr[7], or games like the ESP game, *cf.* von Ahn (2006), could be a valuable source of instance data.

Apart from the actual alignment task, there is another enduring problem for computer scientists: How to determine which alignment relations should be investigated. In essence, this is the same problem as deciding which set of primitives to use when building an ontology. Although it is possible to reason top down which relations could be further specified, we think the best way to approach this problem is from the bottom up, by generalizing over the specific (sometimes ad hoc) requirements of applications.

Ontologies are aligned to make it possible for people to work together. If we are to avoid the pitfalls of ontology unification with ontology alignment, we should pay attention that alignment formalisms do not get in the way. Mappings should not force people to restrict or change their views. They should enable parties to tie their ontologies together step by step and gradually form consensus.

[7] http://www.flickr.com

1. Decide who are stakeholders and hear them.
2. Decide how future maintenance of the alignment will be done and who will be responsible for the alignment.
3. Analyze requirements of the alignment for the use case.
4. Formulate prototypical trial scenario's, *i.e.* topics for *end-to-end* or *relevance-based evaluation* that represent the future use of the alignment in the context of the use case.
5. Analyze the similarity of the ontologies. This encompasses, amongst other things, the following tasks (listed in arbitrary order):
    - Find shared collections of instances, *e.g.* books indexed with subjects from both ontologies.
    - Analyze how the schema's overlap, *i.e.* the meta-ontology, like OWL or SKOS, and which relations they share.
    - Analyze how the ontologies overlap in topic.
    - Analyze which features are shared, *e.g.* which percentage of the labels, or other datatype properties, overlap.
    - Analyze naming or structural convensions, *e.g.* common identifiers, like CAS numbers or country codes.
    - Analyze perspectives, *e.g.* whether anatomical concepts are organized by function or by structure.
6. Divide the alignment into hard, easy, important and less important parts, based on which relations, topics, features, naming convensions, and perspectives can be used to match concepts.
7. Consult evaluation studies, like those of the OAEI, to find out which automatic systems can handle each part. If there are no relevant studies available, consider performing a small sample evaluation.
8. Perform automatic alignment for the easy and less important parts.
9. Depending on the availability of resources, perform manual alignment for hard and important parts.
10. Possibly manually verify automatically aligned parts.
11. Perform trial scenario's and fix bugs in the alignment, and iterate this step.

Figure ix.2: Suggested outline for new alignment projects.

Bibliography

Aleksovski, Z. (2008). *Using Background Knowledge in Ontology Matching*. PhD thesis, Vrije Universiteit Amsterdam.

Aleksovski, Z., Klein, M., ten Kate, W., and van Harmelen, F. (2006a). Matching unstructured vocabularies using a background ontology. In *Proceedings of Knowledge Engineering and Knowledge Management (EKAW)*.

Aleksovski, Z., ten Kate, W., and van Harmelen, F. (2006b). Exploiting the structure of background knowledge used in ontology matching. In *Proceedings of the International Workshop on Ontology Matching (OM-2006)*.

Aleksovski, Z., van Hage, W. R., and Isaac, A. (2007). A survey and categorization of ontology-matching use cases. In *Proceedings of the International Workshop on Ontology Matching (OM-2007)*.

ANSI/NISO (1974-2005). Z39.19-2005 guidelines for the construction, format, and management of monolingual controlled vocabularies.

Antoniou, G. and van Harmelen, F. (2004). *A Semantic Web Primer*. MIT Press. ISBN 0-262-01210-3.

Artale, A., Franconi, E., Guarino, N., and Pazzi, L. (1996). Part-whole relations in object-centered systems: an overview. *Data & Knowledge Engineering*, 20(3):347–383.

Avesani, P., Giunchiglia, F., and Yatskevich, M. (2005). A large scale taxonomy mapping evaluation. In *Proceedings of the Int. Semantic Web Conf. (ISWC)*.

Berland, M. and Charniak, E. (1999). Finding parts in very large corpora. In *Proceedings of ACL 1999*.

Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The semantic web. *Scientific American Magazine*.

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic acids research*, pages 267–270.

Bouquet, P., Giunchiglia, F., van Harmelen, F., Serafini, L., and Stuckenschmidt, H. (2003). C-owl: Contextualizing ontologies. In *Proceedings of the International Semantic Web Conference (ISWC 2003)*.

Brickley, D. and Guha, R. (2000). *Resource description framework (RDF) schema specification 1.0*. W3C.

Brin, S. (1998). Extracting patterns and relations from the world wide web. In *WebDB Workshop at 6th International Conference on Extending Database Technology, EDBT'98*.

Buscaldi, D., Rosso, P., and Arnal, E. S. (2005). A wordnet-based query expansion method for geographical information retrieval. In *Working Notes for the CLEF 2005 Workshop*.

Castano, S., Ferrara, A., and Messa, G. (2006). Results of the hmatch ontology matchmaker in oaei 2006. In *Proceedings of the International Workshop on Ontology Matching (OM-2006)*.

Chklovski, T. and Pantel, P. (2004). Verbocean: Mining the web for fine-grained semantic verb relations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*.

Cimiano, P. and Staab, S. (2004). Learning by googling. *SIGKDD Explor. Newsl.*, 6(2):24–33.

Clarke, S. G. D. (1996). Integrating thesauri in the agricultural sciences. In *Compatibility and Integration of Order Systems. Research Seminar Proceedings of the TIP/ISKO Meeting*.

Cochran, W. G. (1977). *Sampling Techniques*. John Wiley & Sons Ltd, 3 edition. ISBN 0-471-16240-X.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Curino, C. A., Orsi, G., and Tanca, L. (2007). X-som results for oaei 2007. In *Proceedings of the International Workshop on Ontology Matching (OM-2007)*.

d'Aquin, M., Baldassarre, C., Gridinoc, L., Angeletou, S., Sabou, M., and Motta, E. (2007). Characterizing knowledge on the semantic web with watson. In *Proceedings of the 5th International Evaluation of Ontologies and Ontology-based Tools Workshop (EON 2007)*.

Daumé, III, H. (2004). Svmsequel tutorial manual . http://www.cs.utah.edu/~hal/SVMsequel/svmsequel.pdf.

Doerr, M. (2001). Semantic problems of thesaurus mapping. *Journal of Digital Information*, 1(8).

Ehrig, M. and Euzenat, J. (2005). Relaxed precision and recall for ontology matching. In *Proceedings of K-CAP 2005 workshop on integrating ontologies*, pages 25–32.

Etzioni, O., Cafarella, M., Downey, D., Shaked, A.-M. P. T., Soderland, S., Weld, D. S., and Yates, A. (2004). Methods for domain-independent information extraction from the web: An experimental comparison. In *Proceedings of the AAAI Conference*.

Euzenat, J. (2004). An api for ontology alignment. In *Proceedings of the 3rd International Semantic Web Conference (ISWC 2004)*.

Euzenat, J. (2007). Semantic precision and recall for ontology alignment evaluation. In *Proceedings of IJCAI 2007*, pages 348–353.

Euzenat, J., Mochol, M., Shvaiko, P., Stuckenschmidt, H., Šváb, O., Svátek, V., van Hage, W. R., and Yatskevich, M. (2006). Results of the ontology alignment evaluation initiative.

Euzenat, J., Mochol, M., Shvaiko, P., Stuckenschmidt, H., Šváb, O., Svátek, V., van Hage, W. R., and Yatskevich, M. (2007). Results of the ontology alignment evaluation initiative.

Euzenat, J. and Shvaiko, P. (2007). *Ontology matching*. Springer-Verlag, Heidelberg (DE). ISBN 978-3-540-49611-3.

Fellbaum, C., editor (1998). *WordNet: An Electronic Lexical Database*. MIT Press. ISBN 978-0-262-06197-1.

Finkelstein-Landau, M. and Morin, E. (1999). Extracting semantic relationships between terms: Supervised vs. unsupervised methods. In *International Workshop on Ontological Engineering on the Global Information Infrastructure*, pages 71–80.

Foster, I. and Kesselman, C. (1999). *The Grid: Blueprint for a New Computing Infrastructure*. Morgan Kaufmann Publishers. ISBN 1-55860-475-8.

Friis, T., Goodier, J., Stage, E., and König, E. (1993). Unified Agricultural Thesaurus: Feasibility study on the use of an universal thesaurus in agricultural science databases. Report.

Geleijnse, G. and Korst, J. (2007a). Creating a Dead Poets Society: Extracting a social network of historical persons from the web. In *Proceedings of the International Semantic Web Conference (ISWC 2007)*.

Geleijnse, G. and Korst, J. (2007b). Improving the accessibility of a thesaurus-based catalog by web content mining. In *Proceedings of the First International workshop on Cultural Heritage on the Semantic Web*.

Geleijnse, G., Korst, J., and Pronk, V. (2006). Google-based information extraction. In *Proceedings of the sixth Dutch-Belgian Information Retrieval workshop (DIR 2006)*.

Girju, R., Badulescu, A., and Moldovan, D. (2003). Learning semantic constraints for the automatic discovery of part-whole relations. In *Proc. of the HLT-NAACL*.

Girju, R., Moldovan, D., Tatu, M., and Antohe, D. (2005). On the semantics of noun compounds. *Computer Speech and Language*, 19:479–496.

Girju, R., Nakov, P., Nastase, V., Szpakowicz, S., Turney, P., and Yuret, D. (2007). Semeval-2007 task 04: Classification of semantic relations between nominals. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220. http://tomgruber.org/writing/ontolingua-kaj-1993.htm.

Guarino, N. and Welty, C. (2004). An overview of ontoclean. In Staab, S. and Studer, R., editors, *Handbook on Ontologies*, pages 151–171. Springer Verlag.

Hearst, M. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational Linguistics*.

Hellweg, H., Krause, J., Mandl, T., Marx, J., Müller, M. N. O., Mutschke, P., and Strötgen, R. (2001). Treatment of semantic heterogeneity in information retrieval. Technical report, GESIS-IZ.

Herzberger, L. O. B. (2006). *e-Science and the VL-e Approach*, pages 58–67. Springer Verlag. ISBN 978-3-540-33245-9.

Hodge, G. (2000). *Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files*. Council of Library and Information Resources, report 91 edition. http://www.clir.org/pubs/abstract/pub91abst.html ISBN 1-887334-76-9.

Hollink, L. (2006). *Semantic Annotation for Retrieval of Visual Resources*. PhD thesis, Vrije Universiteit Amsterdam. http://www.cs.vu.nl/~laurah/thesis/thesis.pdf.

Hollink, L., van Assem, M., Wang, S., Isaac, A., and Schreiber, G. (2008). Two variations on ontology alignment evaluation: Methodological issues. In *Proceedings of 5th European Semantic Web Conference 2008 (ESWC 2008)*.

Hood, M. W. and Ebermann, C. (1990). Reconciling the CAB thesaurus and AGROVOC. *IAALD Quarterly Bulletin XXXV*, 3:181–185.

Hu, W., Cheng, G., Zheng, D., Zhong, X., and Qu, Y. (2006). The results of falcon-ao in the oaei 2006 campaign. In *Proceedings of the International Workshop on Ontology Matching (OM-2006)*.

Hu, W., Zhao, Y., Li, D., Cheng, G., Wu, H., and Qu, Y. (2007). Falcon-ao: results for oaei 2007. In *roceedings of the International Workshop on Ontology Matching (OM-2007)*.

Huang, Z. and van Harmelen, F. (2008). Using semantic distances for reasoning with inconsistent ontolgies. In *Proceedings of the International Semantic Web Conference (ISWC 2008)*.

Huang, Z., van Harmelen, F., and ten Teije, A. (2005). Reasoning with inconsistent ontologies. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence (IJCAI'05)*.

Huang, Z., van Harmelen, F., and ten Teije, A. (2006). *Reasoning with Inconsistent Ontologies: Framework, Prototype and Experiment*, chapter 5. John Wiley & Sons Ltd.

Hull, D. (2000). Evaluating evaluation measure stability. In *Proceedings of SIGIR 2000*.

Ikeda, K., Nagaoka, S., Winkler, S., Kotani, K., Yagi, H., Nakanishi, K., Miyajima, S., Kobayashi, J., and Mori, H. (2001). Molecular characterization of bombyx mori cytoplasmic polyhedrosis virus genome segment 4. *Journal of Virology*, 75:988–995.

Isaac, A., van der Meij, L., Schlobach, S., and Wang, S. (2007). An empirical study of instance-based ontology matching. In *Proceedings of the International Semantic Web Conference (ISWC 2007)*.

Kalfoglou, Y. and Schorlemmer, M. (2003). Ontology mapping: the state of the art. *The knowledge engineering review*, 18(1):1–31.

Kamps, J. (2004). Improving retrieval effectiveness by reranking documents based on controlled vocabulary. In *Advances in Information Retrieval: 26th European Conference on IR Research (ECIR)*.

Katrenko, S. and Adriaans, P. (2007). Learning relations from biomedical corpora using dependency trees. In *KDECB, LNBI, vol. 4366*.

Kazai, G., Lalmas, M., and de Vries, A. P. (2004). The overlap problem in content-oriented xml retrieval evaluation. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004)*.

Kekäläinen, J. (2005). Binary and graded relevance in ir evaluations–comparison of the effects on ranking of ir systems. *Information Processing and Management*, 41(5):1019–1033.

Klein, M. (2004). *Change Management for Distributed Ontologies*. PhD thesis, Vrije Universiteit Amsterdam. http://www.cs.vu.nl/~mcaklein/thesis.

Lapata, M. (2002). The disambiguation of nominalizations. *Computational Linguistics*, 28(3):357–388.

Lauser, B., Johannsen, G., Caracciolo, C., Keizer, J., van Hage, W. R., and Mayr, P. (2008). Comparing human and automatic thesaurus mapping approaches in the agricultural domain. In *Proceedings of the International Conference on Dublin Core and Metadata Applications*.

Li, Y., Li, J., Zhang, D., and Tang, J. (2006). Result of ontology alignment with rimom at oaei06. In *Proceedings of the International Workshop on Ontology Matching (OM-2006)*.

Li, Y., Zhong, Q., Li, J., and Tang, J. (2007). Result of ontology alignment with rimom at oaei'07. In *Proceedings of the International Workshop on Ontology Matching (OM-2007)*.

Liang, A., Sini, M., Chang, C., Li, S., Lu, W., He, C., and Keizer, J. (2005). The mapping schema from chinese agricultural thesaurus to AGROVOC. In *Proceedings of the fifth Conference of the European Federation for Information Technology in Agriculture, Food and Environment and the third World Congress on Computers in Agriculture and Natural Resources (EFITA/WCCA 2005)*.

Lin, D. (1998). Dependency-based evaluation of minipar. In *Proceedings of the Workshop on the Evaluation of Parsing Systems, First International Conference on Language Resources and Evaluation*, Granada, Spain.

Lindberg, D. A. B., Humphreys, B. L., and McCray, A. T. (1993). The Unified Medical Language System. *Methods Information in Medicine*, 31(4):281–291.

Maedche, A., Motik, B., Silva, N., and Volz, R. (2002). Mafra – a mapping framework for distributed ontologies. In *Proceedings of the 13th European Conference on Knowledge Engineering and Knowledge Management (EKAW)*.

Mao, M. and Peng, Y. (2006). Prior system: Results for oaei 2006. In *Proceedings of the International Workshop on Ontology Matching (OM-2006)*.

Massmann, S., Engmann, D., and Rahm, E. (2006). Coma++: Results for the ontology alignment contest oaei 2006. In *Proceedings of the International Workshop on Ontology Matching (OM-2006)*.

Mayr, P., Mutschke, P., and Petras, V. (2008). Reducing semantic complexity in distributed digital libraries: treatment of term vagueness and document re-ranking. *Library Review*, 57(3):213–224.

Mayr, P. and Petras, V. (2008a). Building a terminology network for search: the komohe project. In *Proceedings of the International Conference on Dublin Core and Metadata Applications*.

Mayr, P. and Petras, V. (2008b). Cross-concordances: terminology mapping and its effectiveness for information retrieval. In *Proceedings of the World Library and Information Congress: 74th IFLA General Conference and Council*.

McGuinness, D. L. (2003). Ontologies come of age. In *The Semantic Web: Why, What, and How*. MIT Press.

Meier, P. (1953). Variance of a weighted mean. *Biometrics*, 9(1):59–73.

Meilicke, C. and Stuckenschmidt, H. (2007). Applying logical constraints to ontology matching. In *KI 2007: Advances in Artificial Intelligence*, pages 99–113.

Miles, A. and Bechhofer, S. (2004). Skos simple knowledge organization system reference. http://www.w3.org/TR/skos-reference/.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM (CACM)*, 38(11):39–41.

Mochol, M., Jentzsch, A., and Euzenat, J. (2006). Applying an analytic method for matching approach selection. In *Proceedings of the 1st International Workshop on Ontology Matching (OM-2006)*, pages 37–48.

Moldovan, D., Badulescu, A., Tatu, M., Antohe, D., and Girju, R. (2004). Models for the semantic classification of noun phrases. In *Proceedings of the Computational Lexical Semantics Workshop at HLT-NAACL 2004*.

Nagy, M., Vargas-Vera, M., and Motta, E. (2007). Dssim - managing uncertainty on the semantic web. In *Proceedings of the International Workshop on Ontology Matching (OM-2007)*.

Nastase, V., Sayyad-Shirabad, J., Sokolova, M., and Szpakowicz, S. (2006). Learning noun-modifier semantic relations with corpus-based and wordnet-based features. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI 2006)*, pages 781–787.

Nastase, V. and Szpakowicz, S. (2003). Exploring noun-modifier semantic relations. In *Proceedings of the 5th International Workshop on Computational Semantics (IWCS 2005)*.

Nixon, L. and Mochol, M. (2004). Prototypical business use cases. Knowledge Web Project deliverable D1.1.2.

Obrst, L. (2006). The ontology spectrum and semantic models. presentation. http://ontolog.cim3.net/file/resource/presentation/LeoObrst_20060112/OntologySpectrumSemanticModels–LeoObrst_20060112.ppt.

Patel, M., Koch, T., Doerr, M., and Tsinaraki, C. (2005). Semantic interoperability in digital library systems. DELOS Project deliverable D5.3.1.

Pedersen, T., Patwardhan, S., and Michelizzi, J. (2004). Wordnet::similarity - measuring the relatedness of concepts. In *Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI-04)*.

Porzel, R. and Malaka, R. (2004). A task-based approach for ontology evaluation. In *Proceedings of the ECAI Workshop on Ontology Learning and Population: Towards Evaluation of Text-based Methods in the Semantic Web and Knowledge Discovery Life Cycle*.

Rahm, E. and Bernstein, P. A. (2001). A survey of approaches to automatic schema matching. *VLDB Journal*, 10(4).

Raymond, E. S. (1999). *The Cathedral and the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary*. O'Reilly Media. ISBN 1-56592-724-9.

Rosario, B. and Hearst, M. (2001). Classifying the semantic relations in noun-compounds via domain-specific lexical hierarchy. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing (EMNLP 2001)*.

Rosario, B., Hearst, M., and Fillmore, C. (2002). The descent of hierarchy, and a selection in relational semantics. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., and Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, 8(3):382–439.

Ruenes, D. S. (2007). *Domain Ontology Learning from the Web*. PhD thesis, Universitat Pulitècnica de Catalunya.

Sabou, M., Garcia, J., Angeletou, S., d'Aquin, M., and Motta, E. (2007). Evaluating the semantic web: A task-based approach. In *Proceedings of the International Semantic Web Conference (ISWC 2007)*.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proc. of International Conference on New Methods in Language Processing*.

Schreiber, G., Akkermans, H., Anjewierden, A., Dehoog, R., Shadbolt, N., Vandevelde, W., and Wielinga, B. (1999). *Knowledge Engineering and Management: The CommonKADS Methodology*. MIT Press. ISBN 0-262-19300-0.

Schreiber, G., Wielinga, B., de Hoog, R., Akkermans, H., and de Velde, W. V. (1994). Commonkads: A comprehensive methodology for kbs development. *IEEE Expert: Intelligent Systems and Their Applications*, 9(6):28–37.

Schulz, S. and Hahn, U. (2005). Part-whole representation and reasoning in formal biomedical ontologies. *Artificial Intelligence in Medicine*, 34(3):179–200.

Shvaiko, P. and Euzenat, J. (2005). A survey of schema-based matching approaches. *Journal on Data Semantics*, 3730:146–171.

Shvaiko, P., Euzenat, J., Stuckenschmidt, H., Mochol, M., Giunchiglia, F., Yatskevich, M., Avesani, P., van Hage, W. R., Šváb, O., and Svátek, V. (2007). Description of alignment evaluation and benchmarking results. Knowledge Web Project deliverable D2.2.9.

Stephens, M., Palakal, M., Mukhopadhyay, S., and Raje, R. (2001). Detecting gene relations from medline abstracts. In *Proceedings of the Sixth Annual Pacific Symposium on Biocomputing*, pages 483–496.

Stuckenschmidt, H., Scerri, T., Bhogal, R., van Buel, J., Crowlesmith, I., Fluit, C., Kampman, A., Broekstra, J., and van Mulligen, E. (2004). Exploring large document repositories with rdf technology: The dope project. *IEEE Intelligent Systems*, 19(3):34–40.

Šváb, O., Svátek, V., and Stuckenschmidt, H. (2007). A study in empirical and casuistic analysis of ontology mapping results. In *Proceedings of the European Semantic Web Conf. (ESWC)*.

Tatu, M. and Moldovan, D. (2005). A semantic approach to recognizing textual entailment. In *Proceedings of the Human Language Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*.

Turney, P. (2004). The multitext project home page, university of waterloo, school of computer science. http://www.multitext.uwaterloo.ca.

Turney, P. (2005). Measuring semantic similarity by latent relational analysis. *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI-05)*, pages 1136–1141.

Turney, P. (2006). Expressing implicit semantic relations without supervision. In *Proceedings of the 21st International Committee on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL)*.

Turney, P. and Littman, M. (2005). Corpus-based learning of analogies and semantic relations. *Machine Learning*, 60(1-3):251–278.

van Hage, W. R., Isaac, A., and Aleksovski, Z. (2007). Sample evaluation of ontology-matching systems. In *Proceedings of the 5th International Evaluation of Ontologies and Ontology-based Tools Workshop (EON 2007)*.

van Hage, W. R. and Katrenko, S. (2007). Uvavu: Wordnet similarity and lexical patterns for semantic relation classification. In *Proceedings of SemEval 2007*.

van Hage, W. R., Katrenko, S., and Schreiber, G. (2005). A method to combine linguistic ontology-mapping techniques. In *Proceedings of the International Semantic Web Conference (ISWC 2005)*, pages 732–744.

van Hage, W. R., Kolb, H., and Schreiber, G. (2006). A method for learning part-whole relations. In *Proceedings of the International Semantic Web Conference (ISWC 2006)*, pages 723–735.

van Hage, W. R., Kolb, H., and Schreiber, G. (2008a). Relevance-based evaluation of alignment approaches. In *Submitted for publication*.

van Hage, W. R., Sini, M., Finch, L., Kolb, H., and Schreiber, G. (2008b). The OAEI 2006 food task: An analysis of a thesaurus mapping task. *Submitted for publication*.

van Rijsbergen, C. J. (1979). *Information Retrieval*. Butterworths. ISBN 978-0-408-70929-3.

Vizine-Goetz, D., Hickey, C., Houghton, A., and Thompson, R. (2004). Vocabulary mapping for terminology services. *Journal of Digital Information*, 4(4).

von Ahn, L. (2006). Games with a purpose. *IEEE Computer Magazine*.

Voorhees, E. (1998). Variations in relevance judgments and the measurement of retrieval effectiveness. In *Research and Development in Information Retrieval*, pages 315–323.

Voorhees, E. and Tice, D. (2000). Building a question answering test collection. In *Proceedings of SIGIR*.

Wang, S., Isaac, A., van der Meij, L., and Schlobach, S. (2007). Multi-concept alignment and evaluation. In *Proceedings of the Int. Workshop on Ontology Matching*.

Wang, T. D., Parsia, B., and Hendler, J. (2006). A survey of the web ontology landscape. In *Proceedings of the 5th International Semantic Web Conference (ISWC)*.

Wiederhold, G. (1991). Intelligent integration of diverse information. In *Proceedings of the First Workshop on Information Technologies and Systems (WITS'91)*.

Wiederhold, G. (1994). Interoperation, mediation, and ontologies. In *Proceedings of the International Symposium on Fifth Generation Computer Systems; Workshop on Heterogeneous Cooperative Knowledge-Bases*.

Winston, M. E., Chaffin, R., and Herrmann, D. (1987). A taxonomy of part-whole relations. *Cognitive Science*, 11:417–444.

Zeng, M. L. and Chan, L. M. (2004). Trends and issues in establishing interoperability among knowledge organization systems. *Journal of the American Society for Information Science and Technology*, 55(3):377–395.