2016

Wikimatcher: Leveraging Wikipedia for Ontology Alignment

Helena Brooke McCurdy
Wright State University

WIKIMATCHER: LEVERAGING WIKIPEDIA FOR ONTOLOGY ALIGNMENT

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Computer Engineering

By

HELENA B. MCCURDY
B.S., Wright State University, 2014

2016
Wright State University

WRIGHT STATE UNIVERSITY
GRADUATE SCHOOL

April 27, 2016

I HEREBY RECOMMEND THAT THE THESIS PREPARED UNDER MY SUPERVISION BY Helena B McCurdy ENTITLED WikiMatcher: Leveraging Wikipedia for Ontology Alignment BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Master of Science in Computer Engineering.

Michelle Cheatham, Ph.D.
Thesis Director

Mateen M. Rizki, Ph.D.
Chair, Department of Computer Science and Engineering

Committee on
Final Examination

Michelle Cheatham, Ph.D.

Mateen M. Rizki, Ph.D.

Krishnaprasad Thirunarayan, Ph.D.

Robert E.W. Fyffe, Ph.D.
Vice President for Research and
Dean of the Graduate School

ABSTRACT

McCurdy, Helena. M.S. Department of Computer Science and Engineering, Wright State University, 2016. WikiMatcher: Leveraging Wikipedia for Ontology Alignment.

As the Semantic Web grows, so does the number of ontologies used to structure the data within it. Aligning these ontologies is critical to fully realizing the potential of the web. Previous work in ontology alignment has shown that even alignment systems utilizing basic string similarity metrics can produce useful matches. Researchers speculate that including semantic as well as syntactic information inherent in entity labels can further improve alignment results. This paper examines that hypothesis by exploring the utility of using Wikipedia as a source of semantic information. Various elements of Wikipedia are considered, including article content, page terms, and search snippets. The utility of each information source is analyzed and a composite system, WikiMatcher, is created based on this analysis. The performance of WikiMatcher is compared to that of a basic string-based alignment system on two established alignment benchmarks and two other real-world datasets. The extensive evaluation shows that although WikiMatcher performs similarly to that of the string metric overall, it is able to find many matches with no syntactic similarity between labels. This performance seems to be driven by Wikipedia's query resolution and page redirection system, rather than by the particular information from Wikipedia that is used to compare entities.

Contents

List of Figures

List of Tables

ACKNOWLEDGEMENTS

I would like to express my gratitude to my adviser and mentor, Dr. Michelle Cheatham, for giving me the inspiration and guidance in working on this project and thesis. I would also like to thank my family and friends for their enduring support during my time as an athlete and a student. Without their support I would not be where I am today. Lastly, thank you to Hermanus Botha for always being there for me throughout my two years of graduate school.

1

Introduction and Motivation

It has become increasingly difficult to effectively manage data, knowledge, and information in the World Wide Web. This is due to the vast amount of data added to the web on a daily basis. No one person, or machine for that matter, has the ability to read or process it all let alone understand it all and be able to utilize it effectively and efficiently. However, with the introduction and implementation of The Semantic Web, we will be able to do just that. The Semantic Web was first proposed by Tim Berners Lee in 2001. The primary motivation is to provide the current web with the ability to contain more structured knowledge [Berners-Lee et al. 2001]. The Semantic Web enables computational machines to handle more intelligent tasks and reason about data without explicit input or assistance from a user. Semantics in the web helps with expressing meaning and context of information which allows for novel ways of making existing data work for us. To realize the vision of the Semantic Web, ontologies are used to store the meaning, context, and the structure of data within the web. Specifically, an ontology is a representation of the concepts in a

domain and the relationships between them. Ontologies allow for the sharing and reuse of a common understanding of information structure and domain knowledge between both software and people. Ontologies also make it easy to separate domain knowledge from operational knowledge and facilitate the analysis of established domain knowledge [Noy et al. 2001] so there can be universal acceptance. It is important to have a common understanding of information structure and be able to reuse domain knowledge since many specific concepts share the same base concepts. For example, every family is different, however the underlying concept is always the same. There are parents and children and all of them are people. Using the same base ontology to describe all families ensures the ability to access, extract, and combine the same data in the same manner for every family.

1.1 Ontology Parts

Ontologies consist of classes, individuals, and properties. Figure 1.1 shows a basic example of an ontology describing athletic teams. An ontology describes a specific domain by making use of the components mentioned above. Classes within an ontology form a hierarchy which describe entities that have similar characteristics. Classes in the provided athletic team example include Thing, Person, Athlete, Coach, Team, and Sport. This ontology shows that Athlete and Coach are both subclasses of Person. A subclass is a more specific representation of its superclass. An individual within an ontology is an instance of a class. An example is Jane Smith who is an instance of type Athlete. Jane Smith has a specific name, position, and age of type string, string, and integer, respectively. Classes and individ-

uals in an ontology are very much the same as classes and instances within Object Oriented Programming. All athletes, regardless of team and sport will have the same descriptive features which are contained within the Athlete class. There are two types of properties

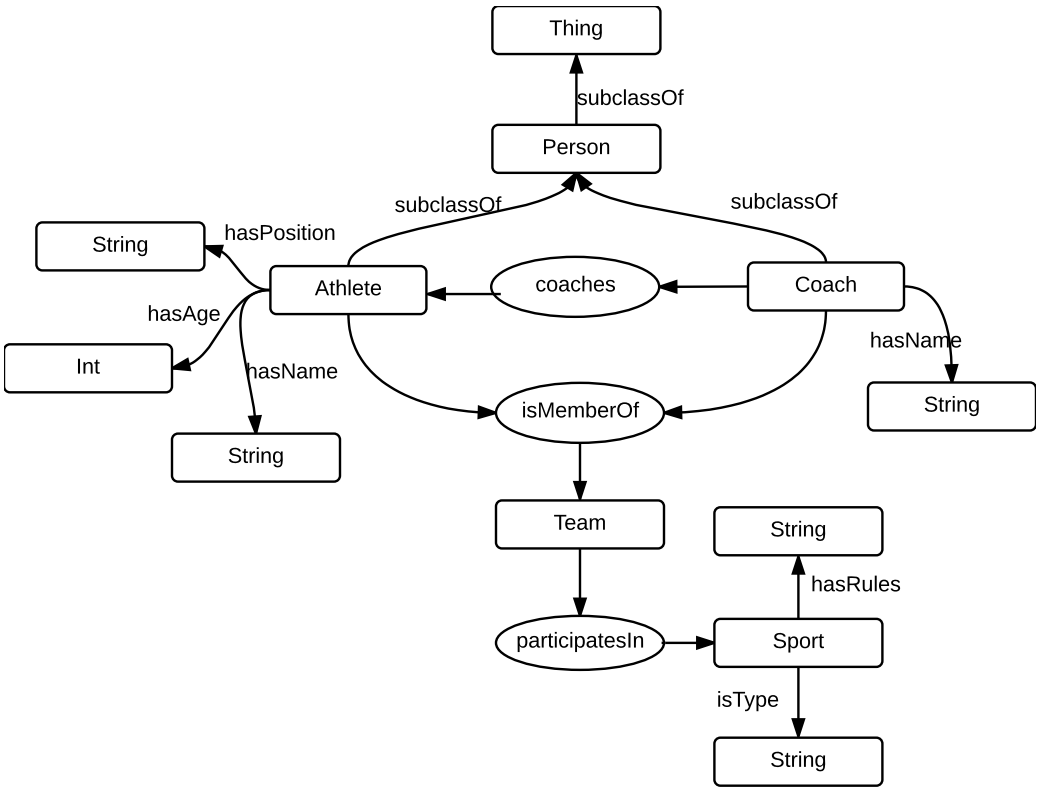

Figure 1.1: An example ontology snippet describing athletic teams

that can be found in an ontology: object properties and data properties. One individual can be related to another with object properties. For example, an individual of type Coach "coaches" an individual of type Athlete where "coaches" is the object property in this example. Other object properties from the athletic team example include "isMemberOf" and "participatesIn". Data properties relate an instance to a literal value. An example is the relationship "hasAge" between Athlete and integer. Other data properties from the athletic team ontology example include "hasName" and "hasPosition".

1.2 Ontology Alignment

There is no single correct way to design and model an ontology for a given domain. Furthermore, two ontologies describing different domains could possibly include the same objects within them. Also, many ontology design decisions must be made by designers with varied backgrounds and experiences. Their differing experiences and opinions could result in them creating different ontologies for the same domain. The end result is that even with two ontologies that represent the same domain, the overall layout and object names may not be the same. This is because engineering new ontologies is not a deterministic process. For example, Figure 1.2 shows another ontology snippet of athletic teams from the point of view of a league organizer. Both example ontology snippets indicate that athletic teams have athletes and coaches; however, each one does so differently. The goal of ontology alignment is to determine when an entity in one ontology is semantically related to an entity in another ontology [Euzenat et al. 2007].

A typical ontology alignment system will accept two different ontologies. It will then compare all or a selection of entities within each ontology and compile a set of matches, usually with a confidence score. An example of a match from the two athletic team example ontology snippets include Athlete to Player. In both ontology snippets that class represents a person who plays for a sports team. The difference between the two ontologies is that in one ontology snippet the position of the person is important whereas in the other the boolean value of fees paid is important.

The ability to accurately align ontologies without substantial manual involvement is

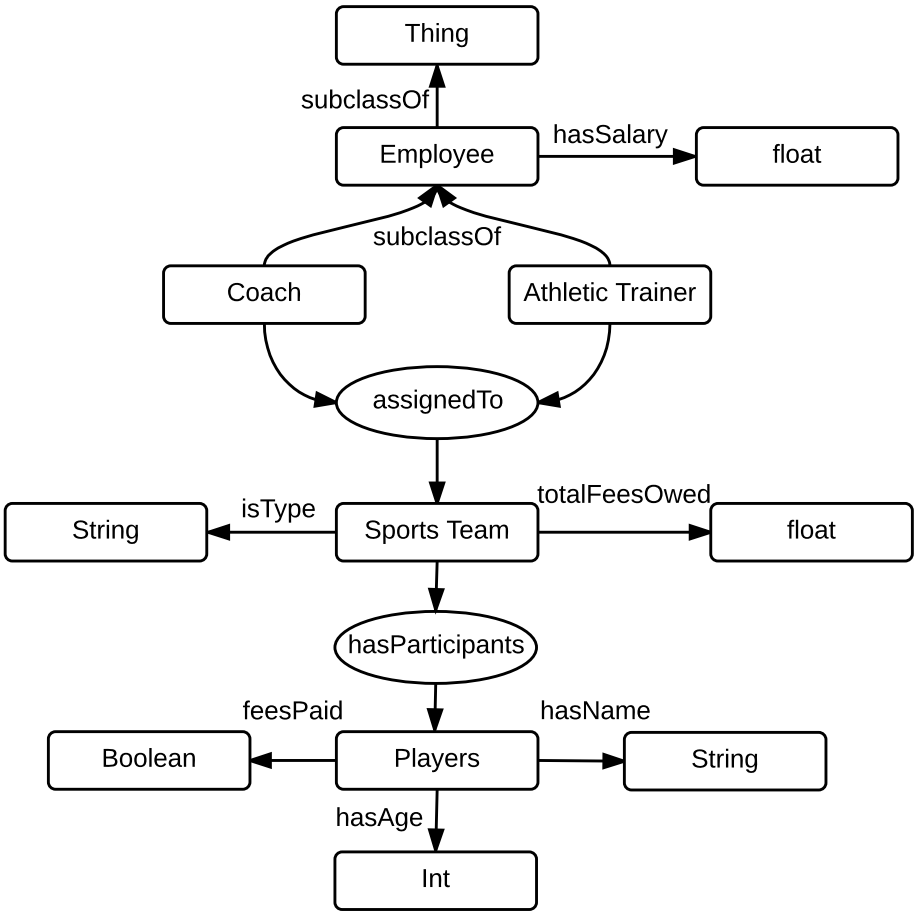

Figure 1.2: An example ontology snippet describing athletic teams

essential to the development of applications that leverage the potential of the Semantic Web. There are various common methods used for ontology alignment, each of which utilizes a different aspect of the information that can be gathered from the ontologies being aligned [Euzenat et al. 2004]. One method, terminological, uses the syntactic relatedness between entity names or other stings gathered from each entity. Previous work showed that much of the semantic meaning inherent in current ontologies is encapsulated in the names (i.e. labels) that ontology designers assign to entities. For this reason, ontology matching algorithms based solely on lexical comparison of entity labels perform surprisingly well

[Cheatham and Hitzler 2013] [Cheatham and Hitzler 2014]. An example of a terminological match between the two ontology snippets is the match of 'Coach' from example 1 to 'Coach' from example 2. Another method for ontology alignment is to compare the structural or design similarity between two ontologies. This can be the internal (the value range of attributes) or external (the relationship of one entity to other entities) structure of the ontologies. An example of a structural match from the two example ontologies is 'Athlete' from example 1 to 'Player' from example 2. This is because both 'Athlete' and 'Player' are very similar in what defines them and both 'hasAge' and 'hasName' can be easily matched first. A third method for ontology alignment is to consider the semantic meaning or interpretation of entity labels. Ontology alignment researchers frequently speculate that leveraging external knowledge sources such as thesauri, dictionaries, encyclopedias, or search engine results could improve the performance of current alignment algorithms [Cheatham and Hitzler 2013]. An example of an alignment that could be generated from gathering external knowledge is 'Team' from ontology 1 to 'Sports Team' from ontology 2. For example, if Wikipedia is used the article of team discusses the article of sports team which could lead to a match.

In this work we explore the utility of using Wikipedia as an external knowledge source to enable the semantic comparison of entities by an ontology alignment algorithm. In particular, we seek to answer the following questions:

- What, if any, information from Wikipedia is useful for ontology alignment?
- How can this information be used in an automated alignment system?

- Is Wikipedia useful for aligning some types of ontologies but not others? How can we distinguish between the two cases?

1.3 Motivation

Wikipedia was chosen because it provides a vast amount of data written by both specialists in every field and by average people. Wikipedia is the largest known encyclopedia in the world. At the time this document was written, there are 5,109,197 unique articles in the English version of Wikipedia and over 35 million articles in 291 different languages across all versions of Wikipedia. The site has 27,800,340 unique users and there are currently 133,529 active users who create, edit, and validate pages in the encyclopedia. On average, 20,000 new articles are added to Wikipedia every month, covering any and every topic. Up to date statistics can be found on Wikipedia:Statistics[1]. In addition to the web interface, Wikipedia provides access to this vast amount of information through an easily accessible API.

While this is not the first time Wikipedia has been incorporated into an alignment algorithm, this work improves upon previous work by systematically comparing different approaches to leveraging the data available on Wikipedia (e.g. simple text parsing, article snippet analysis, page term overlap). This work analyzes performance on a wider variety of datasets, including both general and domain-specific data.

[1] `https://en.wikipedia.org/wiki/Wikipedia:Statistics`

2

Background

The importance of entity labels for ontology alignment is widely known. Many current ontology alignment tools use string similarity metrics based on labels to determine relatedness, e.g. [Cheatham and Hitzler 2013], [Stoilos et al. 2005], and all of the systems mentioned in [Shvaiko and Euzenat 2005]. Many researchers have also suggested that adding more knowledge by using the labels to consult an external knowledge source can improve alignment quality. Approaches along this line include domain-aware ontology matching [Slabbekoorn et al. 2012a] and matching using the WordNet thesaurus [Miller 1995].

An ontology alignment system using a domain-aware method first attempts to identify the domain of a source ontology so that it can restrict matches to only entities that exist within that domain. This type of matcher typically consults an external data source relevant to the domain of interest to gather more information about each entity, which then informs the matching process. This method shows improvement over domain-unaware methods in many situations, [Slabbekoorn et al. 2012b], [Ritze and Paulheim 2011], [Shamdasani

et al. 2011]. It obviously relies on the existence of a relevant machine-accessible domain-specific knowledge source. Many such sources are available for biomedical topics because that community has found semantic web technologies to be a valuable tool [Pesquita et al. 2009]. Examples of biomedical ontologies include the 514 ontologies in BioPortal[1].

Using synonyms from WordNet to expand the possible equivalent terms for an entity label is one of the most common thoughts by alignment researchers regarding leveraging external knowledge sources [Lin and Sandkuhl 2008]. However, we have shown in previous work [Cheatham and Hitzler 2013] that the results of this approach are generally poor. The problem is two-fold: for common words the large number and topic variety of synonyms add significant noise to the alignment process, which increases false positives. On the other hand, the coverage of WordNet is very limited, to the extent that it doesn't contain enough technical/scientific words or jargon to be of any use when aligning domain-specific ontologies.

The difficulties inherent in automatically determining the subject domain of an ontology and finding an appropriate domain-specific knowledge source, along with the limitations of WordNet for use in ontology matching led us to consider Wikipedia as a possible "intermediate" knowledge source that contains both domain-specific and general-purpose information.

[1] `http://bioportal.bioontology.org/`

2.1 Wikipedia Based Matchers

Several other researchers have considered Wikipedia for evaluating the semantic similarity between two concepts, including the developers of WikiRelate [Strube and Ponzetto 2006], BLOOMS [Jain et al. 2010], Wikipedia Linked-Based Measure (WLM) [Witten and Milne 2008], and WikiMatch [Hertling and Paulheim 2012]. These alignment systems differ in either how Wikipedia was utilized or how the data gathered from Wikipedia was analyzed.

The WikiRelate approach uses Wikipedia's page category hierarchy to compute relatedness between pages corresponding to entities. The approach computes similarity values using a path-length algorithm that determines the shortest and most informative path between pages. Though the category approach outperformed approaches using WordNet, it was found that Wikipedia combined with both Google and WordNet performed best on benchmark datasets. Similarly, BLOOMS also utilizes Wikipedia's category hierarchy. A forest (group of trees) is created from the categories returned by Wikipedia for a particular search term and is compared to the forest for a second term using a standard set similarity metric. BLOOMS performs well in respect to other alignments systems on both linked open data as well as on benchmark datasets because it is able to utilize the noisy data from the Wikipedia categories effectively. WLM differs from the previous two approaches by using internal Wikipedia links as opposed to using categories. It uses both the links pointing to a Wikipedia page and links pointing outward from the page. These sets of links are then compared to determine relatedness. WLM differs from many other Wikipedia-based approaches in its response to failed queries. Generally if a query does not return a specific

article, a system will give up and use an alternative similarity metric that does not rely on information from Wikipedia. Instead, WLM uses the Wikipedia general search function to attempt to return related data in this situation. Finally WikiMatch utilizes Wikipedia's articles and language links. It uses no structural information from within the ontology itself. WikiMatch computes a similarity score based on the titles of the articles pulled. It does not beat state of the art systems but does perform better than the benchmark of the string metric approach on OAEI datasets.

To further explain these existing Wikipedia based matchers, a simple example using the label *player (game)* and the label *athlete* will be walked through. For both WikiRelate and BLOOMS the categories of both labels will be retrieved from Wikipedia. The categories for *Athlete* include 'Sports Terminology' and 'Sports Competitors'. The category for *Player (game)* is 'Game Terminology'. The BLOOMS method then pulls the parent categories to a depth of four. A parent category for 'Sports Terminology' is 'Game Terminology', meaning many of the parent categories will be overlapping for these two labels. This will result in a higher confidence score for the match. In the case of WikiRelate, the shortest category path between the two pages is found, namely from *Player (game)* to 'Game Terminology' to 'Sports Terminology' to *Athlete*. From this path a similary score is calculated. WLM gathers the internal links for both pages rather than the categories. Some of the links gathered for *Athlete* include: 'sport', 'professional', some particularly well known athletes, and different types of sports that athletes participate in. Some of the links gathered for *Player (game)* include: 'game', 'gamer', 'player of the match', and some well known

games. A match confidence score is then computed by measuring the overlap of these internal links. Lastly, WikiMatch pulls the page title and article for each of the labels, fragments, and comments found within the ontologies being aligned. In the case of this simple example the articles and page titles 'Athlete' and 'Player (Game)' would be found. Then the language links to translated articles are pulled for each page title, examples for *Athlete* include 'Atleta', 'Atleto', 'Sportler', among others. No language links exist for *Player (game)*. This information is then used to produce a confidence value for the match between the entities.

The work presented here differs from these previous efforts in the type of data available on Wikipedia that is used for similarity computation. Rather than focusing on categories or links, this work evaluates the utility of article text, snippet text, and page terms for ontology alignment. This work also explores the utility of secondary information sources available on Wikipedia when an exact page can not be found for a particular entity label.

2.2 Evaluation of Ontology Alignment Systems

The Ontology Alignment Evaluation Initiative[2] (OAEI) is a worldwide initiative to evalute the inpact and usefullness of ontology alignment systems. The OAEI hosts a yearly workshop at which ontology alignment systems are tested on various dataset tracks and their strengths and weaknesses are evaluated. Each alignment system and its results on said tracks are then published for further evaluation. The OAEI provides a platform for

[2] `http://oaei.ontologymatching.org/`

researchers to share what they have learned and to improve upon the overall processes and techniques of ontology alignment.

As mentioned, the OAEI consists of eight different tracks, which include:

- Benchmark - this synthetic dataset varies from year to year. It is designed to have wide coverage and to expose the strength of weaknesses of alignment systems. It is designed by systematically modifying established ontologies.

- Anatomy - a real world dataset consisting of two biomedical ontologies. One describes the anatomy of a human, and the other describes the anatomy of a mouse. Both ontologies are contain around 3000 entities and have a high number of overlapping terms.

- Conference - consists of 16 small real world datasets focused on the domain of conference organization. These datasets were created by gathering information about conference organization from a wide variety of sources.

- Multifarm - consists of a subset of the datasets from the conference track that have been translated into nine different languages. This is useful for evaluating an alignment system's multilingual capabilities.

- Large BioMedical Ontologies (largebio) - consists of three large biomedial datasets: the Foundational Model of Anatomy (FMA), SNOMED CT, and the National Cancer Institute Thesaurus (NCI). This track tests the scalability of alignment systems, as each dataset contains tens of thousands of semantically rich classes.

- Interactive matching evaluation (interactive) - user interaction is incorporated into ontology alignment systems in an attempt to further improve overall alignment results. The datasets used are from the anatomy, conference, and largebio tracks. This track was introduced when researchers found alignment tools are slowly hitting the upper bound of performance for fully automated systems.[3]

- Instance Matching - consists of synthetically generate datasets that focus on the matching of instances.

- Ontology Alignment for Query Answering (oa4qa) - this track simulates an ontology-based data access (ODBA) use case for ontology alignment, in which a database is organized according to one ontology and a user is querying based on a different ontologies. The ontologies used are from the conference track.

For the purpose of this work, both the conference and anatomy datasets will be used to evaluate the performance of Wikimatcher. These two data were chosen because they provide a dataset with frequent general terms and a dataset with very specific terms. This work is applicable to any of the other tracks that involve entities that might have Wikipedia pages. If the foreign language versions of Wikipedia were also used, then the multifarm track would also be relevant.

Reference alignments for each track are provided to verify each pairwise match generated by alignment systems. When evaluating the effectiveness of an ontology alignment system on each of the OAEI tracks, the metrics used are precision, recall and f-measure.

[3] `http://oaei.ontologymatching.org/2015/interactive/`

Precision is computed using Equation 2.1, where TP is true positives and FP is false positives. Precision measures the fraction of matches made that are correct. It is important when an alignment has many potentially viable matches that need to be narrowed down.

$$precision = \frac{TP}{TP + FP} \tag{2.1}$$

Recall is computed using Equation 2.2, where TP is true positives and FN is false negatives. Recall measures the fraction of correct matches made in comparison to all possible correct matches established by the reference alignment. Recall is important when an alignment system is trying to generate match possibilities to be filtered down in later stages of the alignment system.

$$recall = \frac{TP}{TP + FN} \tag{2.2}$$

F-measure is the overall accuracy of an alignment when both precision and recall is equally taken into account. That is, it is the harmonic mean of precision and recall. Equation 2.3 is used to compute f-measure for a generated alignment.

$$f - measure = 2 * \frac{precision * recall}{precision + recall} = \frac{2 * TP}{2 * TP + FP + FNl} \tag{2.3}$$

Many of the tracks in OAEI contain matches that can be found by an exact label match. To remove the exact matching as a factor in the results, OAEI uses a metric called recall+. Recall+ is computed the same way as recall however, all exact matches are removed before the number of true positives and false negatives are found. This metric determines an

alignment system's effectiveness in finding non-trivial matches. Each of these scores results in a value between 0 and 1, where a higher value means a better performing system. The combination of these metrics provide a thorough insight into how each alignment system performs on each of the tracks.

2.3 Top Performers

There were 22 total participants in the OAEI 2015 workshop, 14 participated in the conference track and 15 in the anatomy track [Cheatham et al. 2015]. Of the 14 participants 11 performed better overall than the basic string metric alignment. The top three performers and the basic string metric on the conference track are as seen in Table 2.1. Of the 15 participants on the anatomy track, nine preformed better than the basic string metric alignment. The top three performers and the basic string metric on the anatomy track are as seen in Table 2.2. The top performing system for both the conference and anatomy track was AgreementMakerLite (AML).

AgreementMakerLite by [Faria et al. 2015] is based on the AgreementMarker ontology system [Cruz et al. 2009]. It utilizes both external knowledge and internal knowledge gathered form the ontologies being aligned. AML has nine different modules that can be used. The collection of modules that are employed depends on the input ontologies. The modules include: translation of foreign languages, string matching metrics, word matching algorithms, structural matching, property matching, and matching utilizing external knowledge sources. The external knowledge sources include three different bio-medical

Table 2.1: Results of top alignment systems and basic string equivalency at OAEI 2015 on conference track

	F-Measure	Precision	Recall
AML	0.74	0.84	0.66
Mamba	0.72	0.83	0.64
XMAP	0.68	0.85	0.56
StringEquiv	0.56	0.8	0.43

Table 2.2: Results of top alignment systems and basic string equivalency at OAEI 2015 on anatomy track

	F-Measure	Precision	Recall
AML	0.944	0.956	0.931
XMAP	0.896	0.928	0.865
LogMapBio	0.891	0.882	0.901
StringEquiv	0.766	0.997	0.622

ontologies and WordNet. Whereas other systems employ one or two of the matching techniques, AML brings them all together, producing top alignment results across all OAEI tracks.

Another top performer in both the conference and anatomy tracks was XMap (eXtensible matching) [Djeddi et al.]. XMap utilizes three different stages of alignment: terminological, structural, and alignment. The terminological layer combines finding string similarity of entity names with semantic data gathered from within the ontologies being aligned. The structural layer combines computing similarity between the external structural hierarchy of datasets, along with computing the similarity of internal structure of concepts. The alignment stage combines the information and matches gathered in the previous two stages to produce a final alignment. XMap performs well in both the anatomy

and conference tracks. XMap produces results with a higher precision in the conference track than any other system.

The Mamba alignment system also performed very well on the conference dataset [Meilicke]. Mamba utilizes the labels and logical entities to make hypotheses and assumptions about matches. It then uses Markov Logic to identify constraints and formulate a final alignment. An in-depth description of this approach can be found in [Meilicke and Stuckenschmidt 2015].

On the anatomy dataset LogMapBio also performed well, with a 0.891 f-measure. LogMapBio is a variation of the LogMap system [Cuenca Grau and Jimenez-Ruiz 2011]. The LogMap alignment system uses a combination of lexical indexation, structural indexation, computation of initial 'anchor mappings', and mapping repair and discovery. LogMapBio extends LogMap by including the use of the external knowledge source BioPortal. BioPortal provides LogMapBio with the top bio-medical ontologies to pull knowledge from. This configuration allows LogMap to perform well on biomedical datasets.

The work in this thesis is largely orthogonal to these full-featured alignment systems. The work presented here could be combined with existing complete alignment systems by providing additional semantic knowledge about entities being aligned. For instance, a Wikipedia-based matcher could be easily added as a tenth module in AML, an additional stage in XMap (which would allow false positives generated from Wikipedia to be culled in later stages), or an additional external knowledge source in LogMapBio. Achieving strong results via a Wikipedia-based approach to ontology alignment could therefore potentially

be used to improve the performance of the top-performing alignment systems when aligning ontologies pairs with entities likely to be covered on Wikipedia.

3

Approach

WikiMatcher is a proof-of-concept alignment system that leverages knowledge from Wikipedia in order to generate an alignment between two ontologies. In keeping with the exploratory nature of this matcher, scalability is not currently a main design driver – the algorithm compares every element in one ontology to every element in the other ontology. For each pair of entities, the labels are used to query Wikipedia in order to gather additional information about the concepts. This information is then used as input to a similarity metric. If the similarity of the two entities is above a threshold, they are considered equivalent and added to the overall alignment.

Wikimatcher's alignment process can be thought of in terms of three distinct stages: preprocessing, knowledge gathering, and similarity computation. In the preprocessing phase, the entity labels are extracted and normalized. These normalized labels are then used to query Wikipedia during the knowledge gathering phase. Finally, the information gathered is used to compare entities and generate the overall alignment. The remainder of

this section describes each of these operations in more detail.

3.1 Preprocessing

The primary goals of the preprocessing phase are to mitigate differences between ontologies that arise due to different naming convensions (e.g. camelCase versus underscores_as_dividers) and to produce a version of the entity label that is most likely to generate a relevant search result on Wikipedia. The OWLAPI[1] is first used to extract all of the classes, data properties, object properties, and individuals in an ontology. Custom code then determines the appropriate label for each entity. Labels are extracted based either on the entity's URI (i.e. using the substring after the last #) or, if the URI does not contain a meaningful label, the rdf-schema label attribute. Labels are then tokenized and put into lowercase (e.g. workEmailAddress becomes "work email address").

3.2 Knowledge Gathering

Once an appropriate label for each entity is available, it is used to query Wikipedia in order to collect more information about the meaning, context, and use of that label. Wikipedia provides a useful API called the MediaWiki Action API, which we use extensively in WikiMatcher. The MediaWiki Action API is a web service which provides access to the contents of any Wikipedia page via an HTTP request.[2]

[1] `http://owlapi.sourceforge.net/`
[2] `https://www.mediawiki.org/wiki/API:Main_page`

The uniqueness of WikiMatcher lies in the variety of information it uses from Wikipedia in order to improve alignment accuracy. In particular, the contents of the relevant Wikipedia article, a snippet of the article's content, and a list of other page terms that are relevant to the article are collected for each label. Specifically, an article is the complete plain text of a desired Wikipedia page. An article snippet is the first three sentences of an article. Page terms include the page title, all redirects (alternative titles that point to the same Wikipedia article), and aliases (terms associated with a page). If any of that content is not available then the results of Wikipedia's search query provides alternative data.

The query made to the MediaWiki API is an HTTP request which includes the entity label to query for, as well as parameters that specify the desired data to be gathered from the Wikipedia page. In order to facilitate the replication of the results presented here, or the utilization of this approach by other ontology alignment researchers, the exact queries made to the MediaWiki API in this work are given below:

```
ARTICLE_URL = "https://en.wikipedia.org/w/api.php?action=                1
    query&prop=pageprops|extracts&format=xml&explaintext=&
    exsectionformat=plain&ppprop=disambiguation&rawcontinue=&
    titles=" + LABEL + "&redirects=&maxlag=5"

ARTICLE_SNIPPET_URL = "https://en.wikipedia.org/w/api.php?               2
    action=query&prop=pageprops|extracts&format=xml&
    exsentences=2&exintro=&explaintext=&exsectionformat=plain
    &ppprop=disambiguation&rawcontinue=&titles=" + LABEL + "&
```

```
    redirects=&maxlag=5"

PAGE_TERMS_URL = "https://en.wikipedia.org/w/api.php?action=   3
    query&prop=pageprops|pageterms|redirects&format=xml&
    ppprop=disambiguation&wbptterms=alias&rdprop=title&
    rdnamespace=0&rdlimit=max&rawcontinue=&titles=" + LABEL +
     "&redirects=&maxlag=5"

WIKI_SEARCH_URL = "https://en.wikipedia.org/w/api.php?action   4
    =query&list=search&format=xml&srsearch=" + LABEL + "&
    srnamespace=0&srinfo=suggestion&srprop=snippet%7
    Cisfilematch&srlimit=10&rawcontinue=&redirects=&maxlag=5"
```

The first query (1), collects the complete article data. This is encyclopedic information about the entity label, which provides a broad overview of the concept and its relation to other concepts. Specifically, each Wikipedia article is a comprehensive summary which includes references and titles of related topics.[3] The response from the MediaWiki API is very similar to the response for the snippet query demonstrated in Figure 3.1, however the "extract" field would contain the entire article.

The second query (2) collects the snippet text for the article (the first three sentences of the introductory paragraph), which is meant to provide enough context for users of Wikipedia to get a sense of the article's overall content. An alignment system based on snippet data would reduce the amount of data that it would be necessary to gather for

[3] https://en.wikipedia.org/wiki/Wikipedia:What_is_an_article

each entity and “normalize” the information available for each entity, which might result in more directly comparable data related to each entity. An example of the reply sent from Wikipedia in response to a request for the snippet data about the topic “femur” can be found in Figure 3.1.


```
{
    "query": {
        "pages": {
            "188497": {
                "pageid": 188497,
                "ns": 0,
                "title": "Femur",
                "extract": "The femur (/\u02c8fi\u02d0m\u0259r/) (pl.
femurs or femora (/\u02c8f\u025bm\u0275r\u0259/)), or thigh bone, is the
most proximal (closest to the center of the body) bone of the leg in
tetrapod vertebrates capable of walking or jumping, such as most land
mammals, birds, many reptiles such as lizards, and amphibians such as
frogs. In vertebrates with four legs such as dogs and horses, the femur
is found only in the rear legs."
            }
        }
    }
}
```


Figure 3.1: An example of the response from the Wikipedia API after a snippet query (2) is made.

The next query (3) retrieves the page title, redirects and aliases of the entity label. This can be thought of as a set of synonyms for the label. Using this set of terms rather than article or snippet content further reduces the amount of text involved and potentially increases the information content of each word, which may improve matching precision. An example of the Wikipedia response to this type of query can be found in Figure 3.2. Each type of page term is separated into its own section of the response and needs to be parsed and combined. These terms contain alternate spellings, common incorrect spellings,

common alternative names, and synonyms for a page title.

```
<?xml version="1.0"?>
<api>
  <query>
    <pages>
      <page _idx="188497" pageid="188497" ns="0" title="Femur">
        <terms>
          <alias>
            <term>os femoris</term>
            <term>os longissimum</term>
            <term>thigh bone</term>
          </alias>
        </terms>
        <redirects>
          <rd ns="0" title="Thigh bone" />
          <rd ns="0" title="Femora" />
          <rd ns="0" title="Thighbone" />
          <rd ns="0" title="Largest bone" />
          <rd ns="0" title="Femu" />
          <rd ns="0" title="Femir" />
          <rd ns="0" title="Fuemur" />
          <rd ns="0" title="Fumuer" />
          <rd ns="0" title="Fumeur" />
          <rd ns="0" title="Feumur" />
          <rd ns="0" title="Fumur" />
          <rd ns="0" title="Femur bone" />
          <rd ns="0" title="Upper leg bone" />
          <rd ns="0" title="Femoral bone" />
          <rd ns="0" title="Os femoris" />
          <rd ns="0" title="Os longissimum" />
          <rd ns="0" title="Facies patellaris femoris" />
          <rd ns="0" title="Patella surface of femur" />
          <rd ns="0" title="Shenton&amp;#039;s Line" />
          <rd ns="0" title="Femurs" />
          <rd ns="0" title="Thigh bones" />
          <rd ns="0" title="Thigh-bone" />
          <rd ns="0" title="Femoral bones" />
          <rd ns="0" title="Femur bones" />
          <rd ns="0" title="Largest bones" />
        </redirects>
      </page>
    </pages>
  </query>
  <limits redirects="500" />
</api>
```

Figure 3.2: An example of the response from the Wikipedia API after the pages terms query (3) is made.

For each query (of the first three types) made to the MediaWiki API, there are four possible responses: found, redirected, ambiguous and missing. We now describe each of these:

- **Found:** A page is found with an exact or near-exact match (Wikipedia will perform simple normalization steps[4]), thus the desired data is returned.

- **Redirected:** The given label redirects to an alternate, yet related, page title. In this case the desired data is still returned.

- **Ambiguous:** The given label is ambiguous and therefore has multiple possibly-related pages. In this case, no data is returned for the article and snippet queries (1, 2), and WikiMatcher launches a general search query (4) in an attempt to gather *some* information about the entity label. On the other hand, the page terms query (3) still returns the desired data in this situation since the terms associated with the page are still relevant synonyms.

- **Missing:** No page is found for the given label, and therefore no data is returned. In this case a general search query (4) is always attempted.

There are only two possible results for the general search query (4) from the MediaWiki API: the response will either contain data or it won't. If the response does contain data, it consists of a list of relevant page names along with a short extract of the part of each page where the query term appears. An example general search query response that contains

[4] `https://www.mediawiki.org/wiki/API:Query\#Title_normalization`

data is illustrated in Figure 3.3. In this example, the response contains three page results for the label “Gut epithelium”. Each of these consists of a page name and a short extract of the part of the page where the term “Gut epithelium” appears.

```
<?xml version="1.0"?>
<api>
  <query-continue>
    <search sroffset="3" />
  </query-continue>
  <query>
    <search>
      <p ns="0" title="Delta endotoxin" snippet="cleaved in some members. Once
activated, the endotoxin binds to the &lt;span
class=&quot;searchmatch&quot;&gt;gut&lt;/span&gt; &lt;span
class=&quot;searchmatch&quot;&gt;epithelium&lt;/span&gt; and causes cell lysis
by the formation of cation-selective channels" />
      <p ns="0" title="Verocytotoxin" snippet="Haemolytic uraemic syndrome. The
toxin has two parts. The A part damages &lt;span
class=&quot;searchmatch&quot;&gt;gut&lt;/span&gt; &lt;span
class=&quot;searchmatch&quot;&gt;epithelium&lt;/span&gt; through inhibiting its
protein synthesis, facilitating entry to the" />
      <p ns="0" title="Myxobolus cerebralis" snippet="the worm, the spores
extrude their polar capsules and attach to the &lt;span
class=&quot;searchmatch&quot;&gt;gut&lt;/span&gt; &lt;span
class=&quot;searchmatch&quot;&gt;epithelium&lt;/span&gt; by polar filaments. The
shell valves then open along the suture line" />
    </search>
  </query>
</api>
```

Figure 3.3: An example of the response from the Wikipedia API after the search query (4) is made.

If the general search response contains data (as it does in the example for ”gut epithelium”), it is parsed differently depending on whether the original goal was query (1), (2), or (3). When the original goal was to collect the article (1) or snippet (2) data, the snippet data for all of the pages within the general search results are combined and returned as the desired data. If the original goal was to retrieve the page terms (3), the titles of all of the pages within the general search results (up to a limit of ten) are combined and returned. This is because the the combined snippets are the best approximation available to the arti-

cle extract and snippet extract, as they will allow a similar type of comparison to be made during the alignment process. Similarly, the page titles correlate to the list of terms gathered from the page term query in the sense that they can be used to compute values for the same type of similarity metrics.

3.3 Alignment Generation

Similarity between two entities is calculated differently depending on what knowledge from Wikipedia is being used in the alignment process. The various possibilities are discussed in detail in the appropriate subsection within Chapter 4. Regardless of which similarity metric is used, the resulting value for each entity pair is in the range [0,1], with 0 representing no similarity and 1 representing perfect similarity according to that metric. As mentioned previously, all entities from the first ontology are compared to all entities from the second ontology. As this process unfolds, the best possible match (i.e. the one with the highest similarity) for every entity from both ontologies is tracked.

After all pair-wise comparisons have been made, this list of best matches for each entity is filtered to produce the alignment. For every entity, if the confidence value associated with its best match is greater than or equal to a predetermined threshold then the match is added to the alignment. In the event of a "tie" (i.e. there are multiple matches with the same confidence value) for an entity, none of those matches are included in the alignment. The rationale for this is that any choice made between two matches with equal similarity would be arbitrary, and precision would likely suffer.

4

Evaluation and Analysis

The evaluation procedure for this work was designed to explore the performance of using Wikipedia for ontology alignment, as implemented in WikiMatcher, compared to that of a non-semantic label comparison method. Crucially, this approach is not compared against a full-featured alignment system, because it is meant to result in a component that could be used by *any* existing system, in particular to develop a set of potential matches that can then be pruned down by more computationally intensive procedures.

Rather than simply judging the performance of the best available configuration of the WikiMatcher system, the goal of this section is to evaluate the contribution of each component of the system so that the end result will be an effective ontology matcher (or matcher component) that contains *no extraneous elements*. In support of this, three different tests were performed in order to evaluate the utility of article data, snippet data, and page terms (i.e. title, redirects, and aliases) to the alignment process. Each of these approaches reduces the volume of information available about each entity while also (theoretically) reducing

the amount of noise introduced to the matching process.

4.0.1 Datasets

Four different test cases were used to evaluate the effectiveness of WikiMatcher, henceforth known as anatomy, conference, hydro, and geo. Two of the test sets used here are from the Ontology Alignment Evaluation Initiative (OAEI). The first test case, anatomy, is from the OAEI anatomy track, which consists of two bio-medical ontologies: one describing mouse anatomy and the other describing human anatomy. The second test set, conference, is from the OAEI conference track, which consists of seven ontologies describing the organization of a conference. The other two test sets, geology realms and hydro, are real-world ontology matching problems. The geology case contains a pair of ontologies. The first is the Environment Ontology[1] (EnvO) which contains entities describing biomes, environmental features, and environmental material. The second is a compilation of ontology snippets from NASA's Semantic Web for Earth and Environmental Ontology[2] (SWEET) realm ontology collection. SWEET was originally a taxonomy of terms from the Global Change Master Directory that has evolved into a shallow ontology. These ontology snippets describe ocean, land surface, terrestrial hydrosphere, atmosphere, and geosphere, among other geology related topics. For this work they have been merged into a single ontology. The hydro dataset consists of six different ontologies that contain terms from the surface hydrographic domain. The Surface Water Ontology is based upon an analy-

[1] https://bioportal.bioontology.org/ontologies/ENVO/?p=summary
[2] https://sweet.jpl.nasa.gov/graph?domain=Realm

sis of the National Hydrography Dataset (NHD) of the US Geological Survey [Varanka and Usery 2015]. A subset of the EnvO ontology which relates to water-based environments was considered, i.e. those entities related to the EnvO class ”hydrographic feature” (ENVO_00000012) [Buttigieg et al. 2013]. HY_Feature attempts to model hydrographic systems governed by disparate global geographic entities and authorities in a uniform way. For this work, the portion of the model related to the HY_SurfaceHydroFeature was used [Dornblut and Atkinson 2014]. The Surface Water Network Ontology was developed by a group of geographers and ontologists during a modeling session in 2013. The model contains surface water features and corresponding containing features within the terrain [Sinha et al. 2014]. The HydroGazetteer was developed to support semantic gazetteer functions involving topology. It includes hydrographic surface water entities and spatial relationships between them, with a focus on topological links [Vijayasankaran 2015]. The final ontology in this dataset is the realmHydro module of NASA's SWEET ontology.

4.0.2 Evaluation Techniques

The performance of WikiMatcher on each of the three tests for each of the four test sets has been compared to that of an alignment system based solely on the Levenstein string similarity metric. Levenstein computes the edit distance between two strings, which is the number of insertions, deletions, and substitutions required to transform one string into another. The baseline alignment system compares all entities from one ontology to all entities from the other and keeps the best match for each entity. Any entity for which there

are multiple “best” matches is not included in the alignment. The results of the Levenstein alignments are shown in Table 4.1. The metrics used to evaluate WikiMatcher include precision, recall and f-measure. The results tables included in this chapter will not include f-measure and recall scores for both the geo and hydro datasets. This is because there is no verified ’gold-standard’ reference alignment for these test cases that can be used to verify the results generated by WikiMatcher. This means there is no false negative value, so recall and f-measure can not be computed. To verify the results of hydro and geo in this work, an individual with no connection to this work was consulted and used to confirm the matches generated. This person was given the entity labels matched along with the relations that could be extracted from within the ontologies to determine if the match was correct or incorrect. An example of the information provided to this individual is presented in Figure 4.1.

```
MATCH: http://cegis.usgs.gov/SWO/Estuary|http://sweet.jpl.nasa.gov/2.3/realmHydroBody.owl#Estuary
        Question: Estuary = Estuary?
                Relations for Estuary:
                        Every Estuary is a Waterbody
                        No Estuary is a Rapids
                        No Coastline is a Estuary
                        No Estuary is a IceMass
                        No Estuary is a SeaOrOcean
                        No Estuary is a SubmergedStream
                        No Estuary is a Waterfall
                        No Estuary is a SinkOrRise
                        No Estuary is a LakeOrPond
                        No Estuary is a StreamOrRiver
                        No Estuary is a SpringOrSeep
                        No Estuary is a Watercourse
                        No Estuary is a Impoundment
                        No Estuary is a BayOrInlet
                Relations for Estuary:
                        Every Estuary hasSubstance that is a Sediment
                        Every Estuary is a BodyOfWater
                        Every Estuary hasSubstance that is a BrackishWater
```

Figure 4.1: An example of the information given to determine if a match was correct or incorrect.

Table 4.1: Results of Levenstein alignment

	Anatomy	Conference	Hydro	Geo
True Positives	982	144	83	73
False Positives	20	45	0	0
False Negatives	534	161	NA	NA
F-Measure	0.78	0.58	NA	NA
Precision	0.98	0.76	1.0	1.0
Recall	0.65	0.47	NA	NA

4.1 Articles

There are three levels of possible similarity when entire Wikipedia articles are used for comparing two entities. Obviously, if the labels of both entities are identical, they will return the same article if they are used to query Wikipedia. Because this would be a waste of time and bandwidth, entities with syntactically equal labels are declared equivalent with a confidence value of 1.0 without doing any knowledge gathering. If there is only one such equivalent match[3], then this relationship will be included in the final alignment during the alignment generation phase.

Even if two labels are not syntactically equal, they may still return the same article when used to query Wikipedia due to the normalization and page redirects the query system employs automatically. In this case there is again no need to compare the article text, since it is obviously identical. For these situations, WikiMatcher assigns a confidence value of 0.9. The exact confidence value chosen is not critical – the important point is to represent

[3]Intuitively, it seems that exact lexical matches would always be unique since a single ontology will not use the same label for two entities. This is indeed the case when WikiMatcher is employed; however, some alignment systems remove stopwords when preprocessing entity labels, in which case multiple distinct labels may be conflated (e.g. hasEmail and Email might both be converted to email).

that this type of match involves slightly more uncertainty than one in which the labels of both entities are identical. Because WikiMatcher only keeps the match with the highest confidence value for each entity, this has the effect of preferring string similarity above background knowledge. In practice we have found that this avoids cases in which the system removes a correct match in favor of an incorrect one or does not generate any match for an entity (because multiple potential matches have the same confidence value) even when there is an obvious best answer.

In the final case, queries to Wikipedia for the entity labels return two different articles. In this situation we employ a basic word presence based method to compare the articles [Pang et al. 2002]. If the article returned when entity A's label contains the label for entity B within its text and vice versa, a confidence value of 0.8 is assigned to that entity pair. If the text inclusion is only present in one direction (i.e. the article returned for entity A contains the label for entity B within its text, but the article returned for entity B does not mention entity A), then the confidence value is set to 0.7. While our results show that this basic technique performs quite well, we do plan to evaluate the effectiveness of more advanced text comparison methods in the future.

The results of the article alignment are shown in Table 4.2. Of particular interest in this table are the results on the anatomy test set. Levenstein has a higher precision, while WikiMatcher based on articles has a slightly higher recall. This is a recurring theme: leveraging information available on Wikipedia in the ontology alignment process can uncover matches that have no string similarity. In the case of anatomy, recall+ was computed and found to

Table 4.2: Results of WikiMatcher article alignment

	Anatomy	Conference	Hydro	Geo
True Positives	1016	141	91	76
False Positives	122	84	0	12
False Negatives	500	164	NA	NA
F-Measure	0.77	0.53	NA	NA
Precision	0.89	0.63	1.0	0.86
Recall	0.67	0.46	NA	NA

be 0.15 in comparison to Levenstein's recall+ score of 0.09. Some notable matches made by WikiMatcher when using articles include that of *adenohypophysis* to *anterior lobe of the pituitary gland*, *midbrain* to *mesencephalon*, and *brachiocephalic trunk* to *innominate artery*.

A non-inclusive selection of other notable matches made by WikiMatcher using articles that do not have strong syntactic similarity is included in Table 4.3. In general, WikiMatcher is capable of using articles to uncover more non-obvious true positives than a string-based approach, but with a corresponding increase in false positives that keeps F-measure essentially the same between the two methods. This is a very useful result, as later phases in the alignment process can draw on established techniques, such as inconsistency checking and repair [Meilicke et al. 2007] to reduce the number of false positives, but it is difficult to find viable methods to identify more true positives. Of the true positive matches made by the article alignment for the anatomy dataset, 933 came from the entity labels being syntactically equal. Of the non-trivial matches 83 came from entity labels that had an exact page or was redirected to an exact page. No other information returned from Wikipedia resulted in a match being made.

Another important point that can be gleaned from Table 4.2 is the rather dismal performance of WikiMatcher on the conference test set. WikiMatcher found roughly the same number of correct matches as Levenstein while identifying more than twice as many false positives. Furthermore, WikiMatcher made only one unique match (*email to e-mail*), whereas Levenstein made four (*organization* to *organisation* (three times, for different ontology pairs) and *sponsorship* to *sponzorship*). This is empirical support for something other researchers in this area have mentioned previously: there is likely no such thing as a single approach to ontology alignment that performs well for all types of alignment tasks [Cheatham and Hitzler 2013; Eckert et al. 2009]. The one unique match made by WikiMatcher came from a finding the exact page in Wikipedia and having the other label redirected to the same page. The general problem on the conference test set for WikiMatcher is the generality of the labels involved – words like "paper" have a much wider variety of meanings than terms like "gut epithelium".

In the cases of hydro and geo, WikiMatcher had more true positives than Levenstein. However, Levenstein had fewer false positives. All of the matches made by Levenstein for both hydro and geo were exact syntactic matches. Though WikiMatcher allowed for more false positives it was able to find interesting non-trivial matches including 'floodbank' to 'levee' for hydro and 'red clay' to 'uitsol' for geo.

Table 4.3: Sample matches of WikiMatcher article alignment.

profunda femoris artery = deep femoral artery
lienal vein = splenic vein
kidney cortex = renal cortex
obliquus externus abdominis = external oblique muscle
white fat = white adipose tissue
forebrain = prosencephalon
midbrain = mesencephalon
synovial joint = diarthrosis
triquetral = triangular bone
xiphisternum = xiphoid process
podzol = spodosol
floodbank = levee

4.2 Snippets

As with the article case, when snippets are used to augment the information available about an entity, multiple levels of similarity are possible. As was done with articles, if the labels for two entities are identical, they are matched with a confidence of 1.0 without bothering to query Wikipedia. If the labels are different but they return the same snippet text, they are matched with a confidence of 0.99 (in order to avoid confounding exact syntactic matches). The remaining possibility is that two entity labels return different snippet text. In the case of complete articles, the text was compared based on the presence or absence of the potentially-matching entity's label in the article. This is not suitable for snippet comparison – the shorter length of the text resulted in an inordinate number of false negatives. Instead, snippets are compared based on a standard bag-of-words string similarity metric. A term vector is generated for each snippet (after removing stopwords[4]) and the Jaccard

[4] http://www.lextek.com/manuals/onix/stopwords1.html

Table 4.4: Results of WikiMatcher snippet alignment

	Anatomy	Conference	Hydro	Geo
True Positives	1035	142	92	76
False Positives	140	84	0	13
False Negatives	481	163	NA	NA
F-Measure	0.77	0.53	NA	NA
Precision	0.88	0.63	1.0	0.85
Recall	0.68	0.47	NA	NA

similarity coefficient is employed to compute the similarity between the two vectors. This value is used as the confidence value of a match between the two entities.

Similar to the article configuration, the snippet-based alignment has a better recall than Levenstein for the anatomy dataset. Table 4.5 shows notable non-obvious matches made based on snippets. These matches were all overlooked by the Levenstein based matcher. The recall+ score for anatomy improved to 0.18 from the article alignment which is twice the recall+ score that Levenstein achieved. The snippet-based approach performs worse than Levenstein for the conference dataset in terms of both precision and recall. WikiMatcher also performs worse than Levenstein in precision for both hydro and geo. Though, as with the article alignment, for hydro and geo the snippet method has more true positives than Levenstein. Most of these extra matches are non-trivial.

Using snippets produces more true positives than using the full text of the Wikipedia articles, but snippets also result in more false positives on all datasets besides hydro. For the anatomy dataset the snippet method allowed for 19 true positive matches, with only 9 false positive matches, to be found using the results from the search-search results. This is an improvement over the article method for the search results. For the geo dataset, more

Table 4.5: Sample matches of WikiMatcher snippet alignment.

cranium = skull
auditory tube = eustachian tube
glomerular capillary endothelium = endothelium of the glomerular capillary
liver sinusoid = hepatic sinusoid
profunda brachii artery = superior profunda artery
sublingual gland = sublingual salivary gland
intercostales = intercostal muscle
occipital cortex = occipital lobe
ventricular septum = interventricular septum
atrial septum = interatrial septum
downstream = is downstream to
organic material = organic matter

false positives than true positives resulted from using the search query. This points to the search query having some potential however, more work is needed to better analyze the data gathered.

4.3 Page Terms

The page terms test is similar to the snippet test, however the term vectors are generated from the page terms data returned from Wikipedia rather than from the snippet data. Jaccard similarity coefficient is again used to compute the similarity between the two term vectors, which is then used as the confidence value of the potential match.

As with the previous two alignments, using page terms from Wikipedia to augment entity labels results in an alignment with better recall than a Levenstein based approach on the anatomy dataset. Some of the notable non-syntactically similar matches identified through the use of terms are shown in Table 4.7. Once again, the results of WikiMatcher

Table 4.6: Results of WikiMatcher terms alignment

	Anatomy	Conference	Hydro	Geo
True Positives	1036	141	91	76
False Positives	124	86	0	12
False Negatives	480	164	NA	NA
F-Measure	0.77	0.53	NA	NA
Precision	0.89	0.62	1.0	0.86
Recall	0.68	0.46	NA	NA

in this configuration perform worse than the Levenstein baseline on the conference dataset with respect to both precision and recall.

The results of this test on the anatomy and hydro datasets indicate that basing entity similarity on page terms may be a more accurate approach than using either snippets or full articles in many cases. The primary reason for this seems to be that page terms remain useful even when a particular entity label is not associated with a specific Wikipedia page and a general search must be conducted (Case (4) in the description from Chapter 3). More than 400 pairwise entity comparisons from the anatomy test set devolved to this general search approach. The results from this type of query to Wikipedia contain limited textual data but all possible page titles, which is a major boon for the term-based approach. Matches produced by the term-based approach in these situations were correct ten times as often as they were incorrect for the anatomy test set. This approach also reduced the number of false positives for both the anatomy dataset and the geo dataset.

Table 4.7: Sample matches of WikiMatcher terms alignment.

cerebellar vermis = vermis
forebrain = prosencephalon
spinal ganglion = dorsal root ganglion
spiral organ = organ of corti
brachiocephalic trunk = innominate artery
bony labyrinth = osseous labyrinth
lateral geniculate nucleus = external geniculate body
cervical vertebra 2 = c2 vertebra
stomach mucosa = gastric mucosa
profunda femoris artery = deep femoral artery
ocean trench = deep sea trench
waterbody = body of water

4.4 When A Wikipedia-based Approach Is Viable

Looking at all three tests, the results suggest that WikiMatcher performs significantly better when more pages related to the entities in the ontologies to be aligned are found in Wikipedia (as opposed to a query resulting in a list of potentially matching pages or devolving to a general text search). The utility of any external knowledge source for ontology alignment is constrained by its coverage of the ontologies' subject areas. Figure 4.2 shows the degree of coverage within Wikipedia for the anatomy, conference, hydro, and geo datasets.

As mentioned previously, when a query is made to Wikipedia for a particular term, there are several possible results. In the best case, an article exists that exactly matches the query term, and that article is returned. The next-best outcome occurs when the query term does not have an article associated with it directly, but the term does exist in Wikipedia as a "redirect" to another page. In this case that redirect article is returned. Sometimes

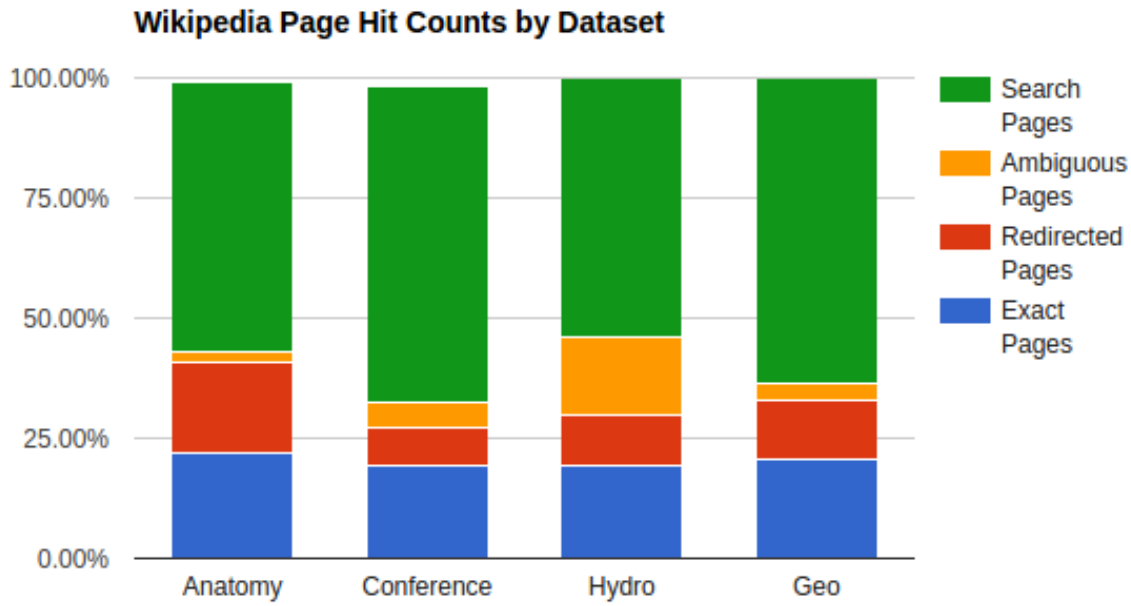

Figure 4.2: Wikipedia page hit counts for the anatomy, conference, hydro, and geo datasets

Wikipedia does not have a specific article indexed for a query term. In these ambiguous situations, rather than returning a single article, a list of possibly relevant articles is returned. In the worst case, Wikipedia does not have the query term indexed to any articles. In this case, a text search is performed on all articles, similar to a Google search, and the top articles that mention the query term are returned in a list.

It is important to realize that each of these levels generally adds more noise to the query results, which is likely to decrease the accuracy of an alignment system that relies on the queries. It is evident from Figure 4.2 that the query results are much more specific for the anatomy test set than for the other test sets. Nearly 40 percent of queries for anatomy labels return a single article (either exactly or after a redirection), while this is the case for only about 25 percent of conference terms, 27 percent of hydro terms, and 29 percent of geo

terms. This is likely to be a part of the reason for the large difference in performance of WikiMatcher on the different test sets.

When looking at the type of Wikipedia response to queries for entities in potential matches, it is noticeable that results vary considerably based on the type of query response from Wikipedia. In the anatomy test set, many "interesting" matching came from cases in which either one of the entities resulted in an exact page hit and the other a redirected page hit or both resulted in a redirected page hit. This is because pages in Wikipedia use redirects to apply many names (synonyms) to an article. This allows syntactically unrelated terms to be matched in a highly accurate way. An example is 'brown fat', which redirects to the page found for 'brown adipose tissue'. This pattern holds for the hydro and geo datasets as well: "Interesting" matches tended to come from a found-redirected or redirected-redirected case. Examples from geo include 'podzol' to 'spodosol' and 'red clay' to 'ultisol'. An example from hydro includes 'floodbank' to 'levee'.

While the type of response to Wikipedia queries is a good predictor of accuracy, the particular information used from Wikipedia is not. Neither precision nor recall varies much regardless of whether the articles, snippets, or page terms from Wikipedia are used for the similarity calculation. This is evident in Figure 4.3 and Figure 4.4. This again suggests that the mere *existence* of Wikipedia pages for the entity labels within the ontologies being aligned (and the redirection system that sends requests for similar concepts to the same page) is the biggest impact on overall performance, rather than the particular information used from those pages.

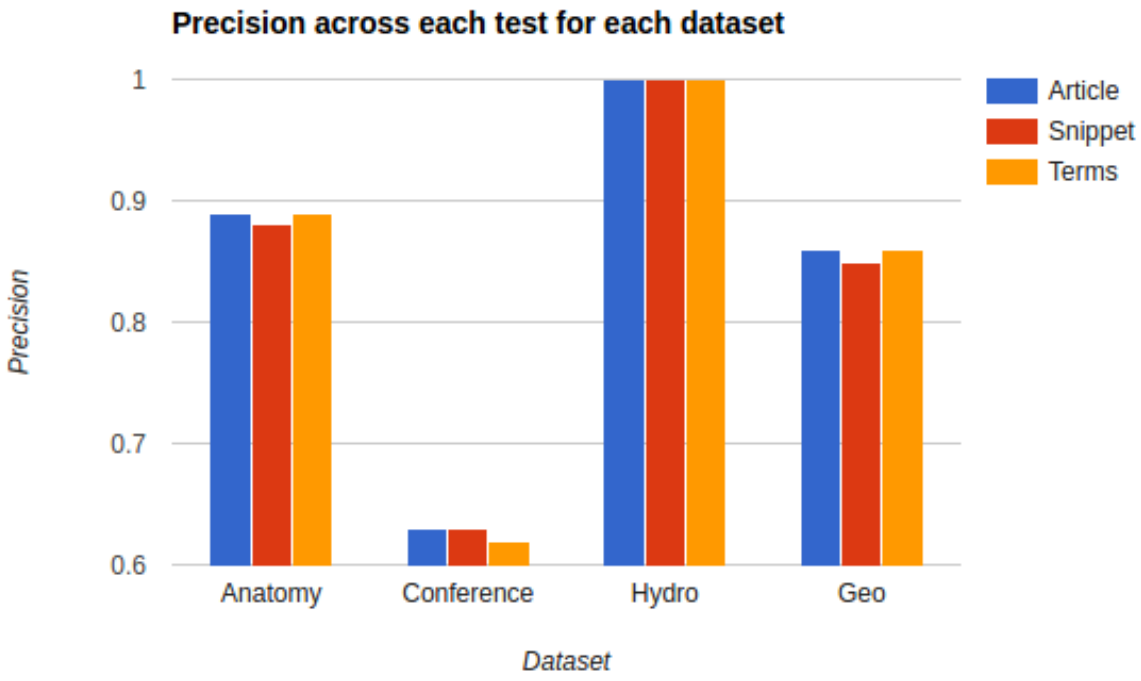

Figure 4.3: Precision scores for each dataset across each of the three tests.

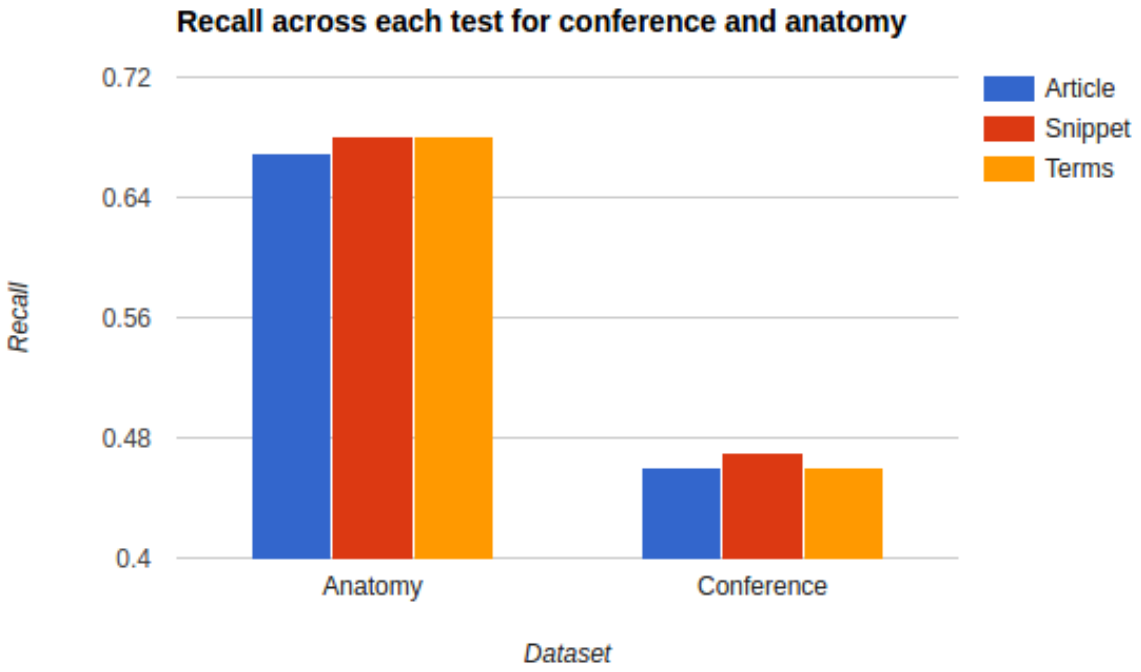

Figure 4.4: Recall scores for each dataset across each of the three tests.

5

Conclusion and Future Work

Ontologies are an important feature to realize the power of a semantically connected web. If data can be given context, meaning, and relationships we can open the door to a wide range of future possibilities and advancements. Developing accurate automated ontology alignment techniques is an important part of the future of the semantic web. By being able to effectively and efficiently align ontologies we are able to reuse and connect data across the web more easily. Aligning ontologies can prove very difficult though due to differences in people creating such ontologies. Many different methods for ontology alignment have been proposed. This work focused on using semantic data gathered from the external knowledge source Wikipedia. Wikipedia provides an abundance of ever-changing and ever-growing data that can be leveraged to increase the information about each of the ontology entities that is available to an alignment system.

Chapter 2 discussed other work that has utilized Wikipedia for ontology matching. This work differs in that it uses different aspects of Wikipedia as-well-as forming fully

designed alignment systems whereas we are evaluating the use of Wikipedia so it can be integrated with other systems. The rest of this work presented the WikiMatcher approach for leveraging Wikipedia for ontology alignment. Three different knowledge sources from Wikipedia were utilized, namely the article, snippet, and page terms. From trial and error we learned that the most difficult part of gathering information from Wikipedia is when a page did not exist for a particular entity. In these cases we utilized Wikipedia's search function to gather possible pages along with the text surrounding the search term within those pages. This information was then used to form matches between entities in two or more ontologies. Results were presented which demonstrate that although WikiMatcher produces alignments that are only on par with a basic Levenstein-based matching system in terms of overall F-measure, it is capable of identifying a larger percentage of "interesting" non-trivial matches that have little or no overall string similarity. This important aspect of the approach, along with the fact that more and more data becomes available on Wikipedia each day, indicates that this approach may have significant merit when combined with other ontology alignment techniques. Whereas the Levenstein approach can only form matches between syntactically comparable labels, WikiMatcher can find matches between terms with no syntactic resemblance. In the final section of the Analysis and Evaluation Chapter, we discussed when Wikipedia, and by extension WikiMatcher, should be use for ontology alignment. It was found that the coverage of entity labels within Wikipedia plays a large role in whether or not WikiMatcher will provide meaningful non-trivial matches.

While the results presented here need to be further verified on a wider variety of in-

formation sources from within Wikipedia and ways for comparing them, we are led to suspect that the performance of existing alignment systems that use Wikipedia may be largely driven not by the particular information from Wikipedia that they use, but rather by Wikipedia's manually curated synonym sets, as encoded in the site's query resolution and page redirection system. Further, we propose that such systems should explore how to handle non-specific query results in order to improve their overall performance, in particular their recall of non-syntactically similar matches.

Presently only simple comparison techniques were used on the gathered data. In order to test the hypothesis mentioned above, an exploration of the utility of more advanced text analysis and data mining techniques for comparing the snippet, article, and term data should be conducted. Such techniques include TF-IDF comparison on snippet and article data or pattern recognition using machine learning. It would also be instructive to attempt to reproduce the results of other Wikipedia-based alignment systems and then explore how much of their performance is due to the inherent synonym sets.

Finally, we recognize that there is likely not a "one size fits all" single best approach to ontology alignment. As a result, it would be interesting to combine WikiMatcher with other alignment tools to determine if the combination of different semantic alignment systems or a combination of a semantic system with a syntactic or structural system would benefit the final alignment created.

References

BERNERS-LEE, T., HENDLER, J., LASSILA, O., ET AL. 2001. The semantic web. *Scientific american 284,* 5, 28–37.

BUTTIGIEG, P. L., MORRISON, N., SMITH, B., MUNGALL, C. J., LEWIS, S. E., CONSORTIUM, E., ET AL. 2013. The environment ontology: contextualising biological and biomedical entities. *J. Biomedical Semantics 4*, 43.

CHEATHAM, M., DRAGISIC, Z., EUZENAT, J., FARIA, D., FERRARA, A., FLOURIS, G., FUNDULAKI, I., GRANADA, R., IVANOVA, V., JIMÉNEZ-RUIZ, E., ET AL. 2015. Results of the ontology alignment evaluation initiative 2015. In *10th ISWC workshop on ontology matching (OM)*. No commercial editor., 60–115.

CHEATHAM, M. AND HITZLER, P. 2013. String similarity metrics for ontology alignment. In *The Semantic Web–ISWC 2013*. Springer, 294–309.

CHEATHAM, M. AND HITZLER, P. 2014. The properties of property alignment. *Ontology Matching*, 13.

CRUZ, I. F., ANTONELLI, F. P., AND STROE, C. 2009. Agreementmaker: efficient matching

for large real-world schemas and ontologies. *Proceedings of the VLDB Endowment 2,* 2, 1586–1589.

CUENCA GRAU, B. AND JIMENEZ-RUIZ, E. 2011. Logmap: Logic- based and scalable ontology matching.

DJEDDI, W. E., KHADIR, M. T., AND YAHIA, S. B. Xmap: Results for oaei 2015.

DORNBLUT, I. AND ATKINSON, R. 2014. Ogc hy_features: a common hydrologic feature model. *Open Geospatial Consortium Technical Report OGC 11-039r3, 55pp*.

ECKERT, K., MEILICKE, C., AND STUCKENSCHMIDT, H. 2009. Improving ontology matching using meta-level learning. In *The Semantic Web: Research and Applications*. Springer, 158–172.

EUZENAT, J., SHVAIKO, P., ET AL. 2007. *Ontology matching*. Vol. 333. Springer.

EUZENAT, J., VALTCHEV, P., ET AL. 2004. Similarity-based ontology alignment in owl-lite. In *ECAI*. Vol. 16. 333.

FARIA, D., MARTINS, C., NANAVATY, A., OLIVEIRA, D., BALASUBRAMANI, B. S., TAHERI, A., PESQUITA, C., COUTO, F. M., AND CRUZ, I. F. 2015. Aml results for oaei 2015. In *ISWC International Workshop on Ontology Matching (OM), CEUR Workshop Proceedings*.

HERTLING, S. AND PAULHEIM, H. 2012. Wikimatch–using wikipedia for ontology match-

ing. In *Proceedings of the 7th International Workshop on Ontology Matching*. Citeseer, 37–48.

JAIN, P., HITZLER, P., SHETH, A. P., VERMA, K., AND YEH, P. Z. 2010. Ontology alignment for linked open data. In *The Semantic Web–ISWC 2010*. Springer, 402–417.

LIN, F. AND SANDKUHL, K. 2008. A survey of exploiting wordnet in ontology matching. In *Artificial Intelligence in Theory and Practice II*. Springer, 341–350.

MEILICKE, C. Mamba-results for the oaei 2015.

MEILICKE, C. AND STUCKENSCHMIDT, H. 2015. A new paradigm for alignment extraction. In *Proceedings of the Tenth International Workshop on Ontology Matching (OM 2015)*.

MEILICKE, C., STUCKENSCHMIDT, H., AND TAMILIN, A. 2007. Repairing ontology mappings. In *AAAI*. Vol. 3. 6.

MILLER, G. A. 1995. Wordnet: a lexical database for english. *Communications of the ACM 38,* 11, 39–41.

NOY, N. F., MCGUINNESS, D. L., ET AL. 2001. Ontology development 101: A guide to creating your first ontology.

PANG, B., LEE, L., AND VAITHYANATHAN, S. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. Association for Computational Linguistics, 79–86.

PESQUITA, C., FARIA, D., FALCAO, A. O., LORD, P., AND COUTO, F. M. 2009. Semantic similarity in biomedical ontologies. *PLoS Comput Biol 5,* 7, e1000443.

RITZE, D. AND PAULHEIM, H. 2011. Towards an automatic parameterization of ontology matching tools based on example mappings. In *Proc. 6th ISWC ontology matching workshop (OM), Bonn (DE)*. 37–48.

SHAMDASANI, J., BLOODSWORTH, P., MUNIR, K., RAHMOUNI, H. B., AND MCCLATCHEY, R. 2011. Medmatch–towards domain specific semantic matching. In *Conceptual Structures for Discovering Knowledge*. Springer, 375–382.

SHVAIKO, P. AND EUZENAT, J. 2005. A survey of schema-based matching approaches. In *Journal on Data Semantics IV*. Springer, 146–171.

SINHA, G., MARK, D., KOLAS, D., VARANKA, D., ROMERO, B. E., FENG, C.-C., USERY, E. L., LIEBERMANN, J., AND SOROKINE, A. 2014. An ontology design pattern for surface water features. In *Geographic Information Science*. Springer, 187–203.

SLABBEKOORN, K., HOLLINK, L., AND HOUBEN, G.-J. 2012a. Domain-aware ontology matching. In *The Semantic Web–ISWC 2012*. Springer, 542–558.

SLABBEKOORN, K., HOLLINK, L., AND HOUBEN, G.-J. 2012b. Domain-aware ontology matching. In *The Semantic Web–ISWC 2012*. Springer, 542–558.

STOILOS, G., STAMOU, G., AND KOLLIAS, S. 2005. A string metric for ontology alignment. In *The Semantic Web–ISWC 2005*. Springer, 624–637.

STRUBE, M. AND PONZETTO, S. P. 2006. Wikirelate! computing semantic relatedness using wikipedia. In *AAAI*. Vol. 6. 1419–1424.

VARANKA, D. E. AND USERY, E. L. 2015. An applied ontology for semantics associated with surface water features. *Land Use and Land Cover Semantics: Principles, Best Practices, and Prospects*, 145.

VIJAYASANKARAN, N. 2015. Enhanced place name search using semantic gazetteers.

WITTEN, I. AND MILNE, D. 2008. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy, AAAI Press, Chicago, USA*. 25–30.