Creating specialized ontologies using Wikipedia: The Muninn Experience

Robert Warren
rwarren@math.carleton.ca
Carleton University, Ottawa, Canada

Abstract

This paper reports on the experiences of the Muninn Project in creating specialized ontologies for historical governmental and military organizations using the Wikipedia data set and its Linked Open Data companion DBpedia. The motivation for the ontologies and the extraction methods used are explained and their pitfalls reviewed. Overall Wikipedia is a very accurate knowledge base from which multi-lingual concepts can be extracted. The caveat is that while the information is almost always present, it is not always straight-forward to retrieve because of missing structures or categorization information.

1 Introduction

The Muninn Project is a cross disciplinary project that extracts information from imaged archival documents to create a complex database of events of the Great War. As part of this research a series of ontologies were designed to represent the different structures within the information being collected. In this paper we discuss the use of Wikipedia data to create two ontologies dedicated to modeling historical military and civil organizations. The intent of this paper is not to describe the process of creating the ontologies but rather act as a case study about using Wikipedia as a data source for ontologies.

As Wikipedia grows into a repository of human knowledge it has become the ideal source of general purpose data about the world and is a good base from which the skeletons of ontological repositories can be created. Similarly, DBpedia is a Linked Open Data ([2]) representation of a limited subset of Wikipedia pages making heavy use of template information boxes to generate properties between instances. We use DBpedia ([1]) in our work as a convenient interface for certain kinds of queries against Wikipedia, as well as a resource against which the ontologies terms are linked to for later data interchange and reasoning.

The paper is organized as follows: First, both ontologies are described and the main components desired from Wikipedia outlined. Secondly, the previous work in the area is briefly reviewed and the techniques used to extract information from both data-sets are described for specific pieces of information while reporting on the performance of the methodologies. Thirdly, the results of the extraction are reviewed while discussing the lessons learned from the exercise both from an extraction perspective and from the perspective of maintaining the integration between Wikipedia, DBpedia and the ontologies. We then conclude on our ongoing work in integrating archival documents to knowledge bases and automating the contribution of information back to Wikipedia.

During the course of this paper data errors, user behaviors and design decisions that can make the ontological use of Wikipedia difficult are identified. The obvious answer within the Wikipedia world has always been to “fix it yourself according to community standards”. This has not been done here in the interest of stimulating further discussion of these issues in the community.

2 Problem definition

The ontologies cover both organizations and people within civil and military settings and include the relationship that exists between them. This can mean the titles, offices, trades and roles of individuals, the relative size of certain types of organizations and their expected missions. As a historical project Muninn required support for documenting changes in name, geographic boundaries and allegiance over time to institutions which in most cases no longer existed.

Ontologies and data repositories that offer some coverage of these domains already exist. GeoNames[1] is a well known repository of geographic and place name information, the UN Food and Agriculture Organization ontology and dataset [2] provides data on many countries since 1985 while the Open eGovernment ontologies[3] were targeted to model the current United States government.

At issue with these ontologies is that they are meant to represent current truths instead of historical facts. What makes the creation of the Muninn ontologies difficult when compared with previous approaches is its dependence on temporal information. An ongoing issue with the use of Wikipedia for the specific case of historical ontologies is that in the encyclopedic perspective this is all mostly outdated information. This seems paradoxical at first, especially since Wikipedia keeps all revisions of its pages available online. However the revision history of a page does not map to the changes in human thought over human history; it only reflects the short term diversity of opinion over 11 years.

Furthermore there is (a well meaning) bias in an encyclopedic work to update the readers knowledge of a subject and this does not always map well into determining what was the historical fact. Since the Great War, a number of institutions and countries have had significant changes in their administration, borders and forms of government. France has changed through 3 republics, Canada has changed from a Dominion to an independent state while absorbing Newfoundland and extending its borders north and westward.

There exists the parallel problem that what is true from a "common knowledge" or "crowd" perspective may not actually be true but needs to be disambiguated. For example, it is now common for people to reference both Canadian and Newfoundland soldiers in the Great War as Canadians, through at the time the Newfoundland regiment would be completely independent from Canadian command. Similarly, a monument may be labeled as dedicated to soldiers, while also including the female non-combatant medical staff. Ambiguity between other items such as rank, trade and appointment similarly exist and the difference between the driver of a car, a professional vehicle driver, the rank of driver and a driver of an artillery train need to be resolvable within the ontology.

Finally, beyond representing the temporal events, Muninn ontologies have to deal with classes, properties and instances that change over time. Group memberships, locations and nomenclature changes over time and the ontologies must be able to deal with these situations.

The Muninn ontologies are meant to support Natural Language Processing in the manner of Helmann et al. [3] that reads the plain texts extracted from archival documents. The ontologies are also meant to provide a schema with which information from other sources will be marked-up and referenced.

These sources have provided a skeleton of classes, instances and properties to be included within the ontology. These initial values provide the seeds against which Wikipedia data is to be queried and extracted from. While many involved methodologies for ontology construction exist, only simple queries against Wikipedia pages, templates and category taxonomies were performed to import additional data. This was done to concentrate on the classes and instances of the ontologies while "softer" properties such as chains of command and seniority grades could be created at a later time using statistical methodologies.

3 Previous Work

The creation of ontologies is an activity that has drawn considerable interest in both the philosophy fields and in the area of data interchange. From a philosophical perspective this is seen primarily as a type of classification problem while the data oriented community seems ontological development as a tool for data interchange.

Some of this dichotomy can be seen in contrasting some of the previous work in the field. Smith and Jackson ([7]) created the OntoRanch tool, a collaborative framework that guided a team through the creation on an ontology. The team should be guided by a philosophical researcher aided by a number of experts who curate the work of stake-holders in a specific data management need in an organization. The objective is to create authoritative ontologies that sets the standards with which the data will be reported within the organization.

A second style of research has been the work of Torniai et al. ([9]) who seeked to leverage the use of folksnomies and user note tags to enrich formal ontologies. While the addition of the terms to the ontologies is meant to require approval from a domain expert, the objective of the ontology here is to make the items that it catalogs as accessible as possible with multiple vocabularies.

The use of large corpuses of text as a data source for ontologies within restricted domains was attempted by Ruiz et al. ([6]). They made use of statistical

[1] http://www.geonames.org/

[2] http://www.fao.org/countryprofiles/geoinfo.asp

[3] http://oegov.org/

Natural Language Techniques to extract a starting set of properties between previously defined classes and instances which were tend reviewed for inclusion within the ontology.

Lastly, Ponzetto and Strube ([5]) published on the wholesale use of both article text and category data to create large scale general purpose ontologies. While their results were generally positive, they remarked that errors were present on instance data that would need manual review or external validation.

4 Experimental design

The construction of the ontologies was driven for a pressing need for recording Great War data and enabling the data interchange between different repositories and data consumers. The ontologies were designed primarily by a single designer with support from translators and cognitive scientists and input from the community.

The bootstrap of the ontology was done using the schema and data organization of publicly available datasets to identify a core skeleton of classes and instances for the ontology. Because of the complexity of the data sources and of the events involved, the construction of the ontologies is an ongoing iterative process where its completeness is judged by its ability to handle news facts from external data sources as they become available.

From the basic skeletons there several abstracts problems that we wished to solve: 1) Give a class, what are all of its known instances? 2) Given a few instances, what are other similar instances? and 3) given sets of classes or instances, what are the potential properties that relate them one to another? Lastly, the ontology is meant to support work related to the Great War, which entails that it be multilingual, multi-cultural and transcends multiple historical periods.

Most ontological design closely resembles the creation of a super-schema in that the design is primarily about classes or a taxonomy of classes into which an external database will use to classify its own instances (for a full description of these issues, see Sowa [8]). In this case, the complexity of the data over time requires a mixed model where certain classes and instances are one in the same. The typical example is that of a father, which can be both property, role instances and class of people at the same time.

The specific objects that were pursued were instances of countries (in their historical instance), conflicts, military ranks and roles (occupations) as well as the relationships that bind them. Generally, locating information within Wikipedia was strait-forward and within limits the disambiguation of terms was accurate. Most problems occurred in the extraction of the information from the mark-up and in the inferencing of relationships from the mark-up and templates. Especially in the contexts of Eras, rank and MilitaryRank properties from Templates, the users tend to abuse the labels for visual elegance instead of semantic correctness.

Lastly, Wikipedia (and DBpedia) has a bias towards the English language which at times can cause confusion with respect to the translation of the *thing* versus the translation of the *name of the thing*. This can occasionally can cause confusion and is an issue that needs to be monitored.

5 Results

The overall experience extracting information from Wikipedia was a positive one, as a matter of coverage the majority of the information on the specific instances was available but getting the properties and/or relationships between the instances proved to be a challenge. This section reports on the extraction methods used for some types of information and some of the issues that were raised during the creation of the ontology.

5.1 Countries, modern and history

Since the Muninn ontologies were designed to deal with historical events, the referencing of current Nation names and objects was not always possible or accurate. Similarly the mapping between the name given to a physical region in a specific historical era and the political entity that has effective control over the territory is more than merely complex.

Some databases, such as GeoNames, have attempted to skirt this issue by making use of the ambiguous class “Populated Place” to reference any feature. DBpedia makes use of both classes “Populated Place” and “Country” to reference nations. However the political systems that control the country change over time even through the “Populated Place” component does not. Furthermore, the autonomy and political importance of each country changes over time even through its cultural and demographic roots remains the same.

An element that was added to the Muninn ontology was the use of precedent and successor properties to track the lineage of Country instances. These are directly taken from the Country templates of Wikipedia and make the extraction on an entire demographic lineage strait-forward.

The ability to track which instances becomes the foundations or the remains of an institution allow us to track common ancestry or cultural constructs while retaining ontological consistency. As an example the German Empire page links to the Wiemar Republic which then links to the Third Reich Page and then on to the Military Occupation, East and West Germany and then an Federal Republic of Germany.

While ontologically convenient, these templates are not always followed and the use of multiple pages for a single populated place is occasionally discouraged as "too complex" by some editors that aggressively protect pages. Examples includes the page for China, which is also redirected to from the page People's Republic of China. Similarly, the Wikipedia page for Canada has currently only one page for several different eras and forms of government for the past 100 years while having the Dominion of Canada redirecting to it inappropriately.

While from an encyclopedic standpoint the topic might seem one of presentation, the use of Wikipedia data and DBpedia as an ontology becomes problematic in these cases. Ontologically, the above examples means that there is no difference between a populated place that has been active for several thousand years and a form of government that is active since 1949. In the case of Canada, the current country significantly bigger geographically and has a constitutional independence that it never did as a Dominion.

In these cases the editors should consider changing the pages to read as History of Country or use a template for a populated place as opposed to a country which is inaccurate. From the Muninn ontology perspective, we cannot reference these instances without creating a logical conflict within the ontology. Since users will undoutably attempt to use these erroneous DBpedia instances using the ontology, we deliberately declare our instances for these countries to be different from the DBpedia ones using the <owl:differentFrom> tag. This pro-actively declares the DBpedia country instance as being different from the Muninn instances and will prevent any future reasoning using this inconsistent version of the Wikipedia page.

5.2 Ranks, Roles and Appointments

Short of order of battle information, military ranks and appointments are some of the most complex elements of a military ontology in that they represent the function, social status and authority of a person within an organization. Whereas most militaries organizational units use a structure that is relatively uniform across cultures and institutions, the ranks have changed dramatically with new types of warfare and their increasing technological sophistication.

The properties of ranks and appointments differ based on the institution and the historical period in which it is used. The well known rank of Sargeant is widely used in most militaries but its responsibilities vary. The use of the archaic English spelling of Sarjeant indicates someone of Sargeant rank whom is either part of the small set of modern British units that still use the rank or a common Sargeant whom was in a military before the late 1930's. A Havildar is the equivalent rank of Sargeant in the modern Indian and Pakistan armies as well as in the British Indian Army. However, a rank of Havildar in the British Indian Army would still be consider junior to the actual rank of Sargeant in the British Army.

Tracking these equivalences and relative comparisons is what makes this section of the ontology complex and useful for data analysis. It allows end users to compare the relative authority and responsibility of different individuals within organizations and make queries that, for example, compare the relative salaries of equivalent ranks across armies.

The classification of previously known trades, ratings, appointments and ranks was done in a series of queries on Wikipedia page titles while concurrently checking for a disambiguation page. Most sections of the Muninn ontologies were simply built through the extraction and the linking of different classes of instances. A stronger focus on properties was needed with these instances due to the differences in ranks between countries, branch of the armed service and historical area.

A number of instances of ranks and appointments as well as their relationships were already known from parsing other data sources. Given that these instances were already known to exist, we only wished to recover any additional properties from Wikipedia. The initial "bootstrap" set consisted of about 748 different ranks and appointments within the Commonwealth armies of the Great War without references to gender, rank or civilian or military use.

These were queried against Wikipedia pages and 283 concepts were found to already be within Wikipedia with 116 concepts having disambiguation pages. Manual review indicated that 19 of the non-disambiguated concepts were improperly allocated and needed to be removed. A second smaller experiment which removed any compound words or modifiers to the ranks (e.g.: Staff Sargeant becomes simply Sargeant) yielded a second list of 274 concepts, 242 of which existed with 166 of them having disambiguation pages. The same 19 non-disambiguated entities with improper (in context) definitions were observed in both sets.

The disambiguation pages were not always foolproof as the actual item to be disambiguated could be a source of disagreement. A typical example of this is the Wikipedia page for Corps which is meant for the military context. The disambiguation page only deals with the upper most level of context disambiguation, leaving the actual Corps page to represent two different concepts within a military organization (a formation versus an administrative unit).

This situation is a concern for ontologies and linking Wikipedia to ontologies: the page represents two distinct concepts within the same knowledge domain. If the domains were different, it would be possible to ignore the problem since both senses would never be used concurrently. However, in the Muninn context the equivalent DBpedia instance will always reference the two concepts concurrently, making both the DBpedia and Muninn data-sets logically inconsistent.

It has been the practice to link Muninn instances to the DBpedia instances using the <owl:sameAs> for greater interoperability. In the cases where the DBpedia instance is ambiguous we use instead the <owl:differentFrom> construct to prevent anyone from mistakingly using an improper information.

Redirection pages are especially effective at locating different gender forms, making the resolution of the actual occupation simple. For example, both waiter and waitress both redirect to the Waiting staff Wikipedia. This allows for the identification of different forms at the cost of needing to resolve the difference as gender-based manually. By locating other pages which redirect back to this page we can identify previously unknown forms, such as "server" which was not contained as an occupation within the original data-set.

Recovering the temporal aspects can be a problem: an example the Wikipedia page for determining current Czech Republic ranks in a NATO structure now reflects changes in the Czech Army as of 2011. While enough information remains within the page history to map the previous rank names, the media and insignia information has since been removed for licensing reasons. It is also not possible to recover the actual date of the change since Wikipedia edits naturally lag the official decision.

Wikipedia pages on comparative ranks in different wars and armies are available that classify ranks of multiple countries according to the NATO rank classification scheme. These pages are extremely useful to generate relationships of seniority between different ranks across different organizations. While all of these pages use the same style of template, the problem is that the mark-up of the contents is again generated for the visual representation of the information instead of its logical representation. This makes the extraction of these comparative properties overly difficult compared with the manual entry of the properties at a later date. On occasion this can work to our advantage as with the Template:Military_ranks page where seniority is rendered by the listed order of the specified ranks. Extraction is thus simply done through scripting through updates may not be possible in the future.

Lastly, we attempted to extract additional ranks from Wikipedia using both DBpedia classes and the Wikipedia category systems. The use of Wikipedia templates, and its DBpedia MilitaryRank property to identify ranks proved to be inefficient as opposed to category information. Searching the property yielded only 203 unique military ranks after extensive human reviews and even then only 158 of these pages were actual Wikipedia pages instead of strings.

Searching the Category:Military_ranks_by_country category tree enabled the location of 1,200 different ranks including country of allegiance but not the armed service branch. Interestingly, a number of Canadian army ranks are not reachable through this query since they are (appropriately) classified as appointments. Since this is only known category of appointments, it also segregates this knowledge from the main Military ranks categories.

These pockets of knowledge occur constantly within Wikipedia and are consistent with problems in other knowledge bases in that "chunking" occurs. Linking certain information is easier than others and thus natural, but undesirable, clustering of the knowledge occurs. Thus we find that U.S. Navy rating (ranks) pages are separated in their own categories, German Military ranks are also in their section but direct equivalences are not made.

Lastly, one of the most appreciated benefits of Wikipedia is its ability to provide context-sensitive translation for instance labels in several languages. DBpedia is especially good at making multi-lingual synopsis of pages available in Linked Open Data Format. Using Wikipedia, we were able to make sure that all concepts within the Muninn ontologies were translated in French, English and German.

This also points out an ongoing problem for multilingual knowledge bases that professional translators are only too aware of: *the translation of the term is not the same as the translation of the concept.* This is particularly evident both in the Muninn ontology and in Wikipedia in that the knowledge naturally clusters culturally. Take for example the Wikipedia pages for Sergeant and Oberfeldwebel which are equivalent military ranks in the English and German speaking armies. While both pages are translated into the other language, two different pages exist for the same rank.

Category pages also have different content based on the language they are in even if on the same topic. For examples the German language category Kategorie:Dienstgrad_(Bundeswehr) contains 75 pages about ranks within the German army while the English page for the same category Category:Military_ranks_of_Germany contains only 55. Also, the German rank pages have convenient templates that make the extraction of juniorRank and SeniorRank properties convenient for ontological construction which is lacking in Canadian and British Army pages.

Hence the creation of Rank and Appointments within the Muninn ontologies from Wikipedia data was an iterative process as new pockets of information was located on an opportunistic basis. In a process that is similar to “dipping queries”, an initial set of instances and properties were extracted by a query of the data. These results pointed to a number of others pages and categories to be exploited using a different style of queries, which then imported a different set of instances and properties. This repetition is necessary because the creation of the content on Wikipedia is really a random walk by the crowd over time over several topics and linking content in a coherent scheme requires more long term effort than adding a single item.

6 Discussion

Wikipedia primary goal has always been to be a end-user editable encyclopedia. Over time extensions have been added, such as templates, which were initially meant to ensure visual consistency and organization. Efforts such as DBpedia have made this data available in a linked open data format that is suitable for linking to and generating knowledge bases.

The coverage in breath of topics means that Wikipedia is rapidly becoming one of the main repositories of knowledge on the web. In keeping with its original mandate to be end user editable, the tools and recording mark-up used by wikipedia is geared towards the representation or the rendering of the text on the screen versus the recording of the relationships between the instances.

Templates were created to ensure a consistent look and feel to the wikipedia pages and while these can be a proxy for logical relationships, this is not always true. Tables that could be used to infer a relationship between two instances are sometimes only used for typesetting purposes while the semantics of templates are sometimes abused to provide a good looking visual effect without any underlying meaning. The extraction of relevant relationships behind this mark-up is sometimes too demanding when compared with its re-creation from scratch.

It is to be noted that the use of template fields tends to follow the specificity of the template; widely used templates tend to have field filled with any information while specifically targeted templates enjoy high quality data quality.

A typical example is the Era field which contains a large amounts of references to Star Wars books with the rest being English, Chinese and Japanese historical periods. Similarly, the rank property contains mostly numerical rank information along with some Military rank information and civil appointments.

The disambiguation of terms with multiple contexts was satisfactory through the problem of multiple languages is still an issue of concern; especially when certain pages such as Category pages are not fully translated across languages. It should be easy to do so automatically due to the short length of the labels, which would ensure against pockets of knowledge being isolated due to their lack of translation to other languages.

The next logical evolution of Wikipedia is towards the creation of secondary data sets that will serve new purposes besides that of an online encyclopedia (WikiTravel, WikiNews, etc...). This will necessitate some additional efforts by the end user and data quality checking since mistakes in the initial entry will perpetuate themselves to other datasets. These types of situations already occur in OpenStreetMaps where a small editing error can lead to entire countries becoming submerged under the sea overnight.

A current concern is how to best annotate the ontology in order to record the specific version of the Wikipedia page that created the ontology instance imported and/or the version of the DBpedia instance that was linked to. Early work by Hepp and Bachlechner ([4]) suggested this approach and it would be useful in preventing user changes from having unintended consequences on downstream knowledge bases.

A partial solution that could be implemented within the DBpedia dump of Wikipedia is the addition of Wikipedia versioning tags to the terms. This would allow external knowledge systems to ensure the validity of links they are providing and by automatically alerting their domain experts to changes in the upstream Wikipedia articles.

7 Conclusion

Crowd sourced knowledge bases such as Wikipedia and Open Street Maps are primarily oriented towards the aggregation of crowd sourced knowledge. As the amount of knowledge increases it becomes desirable to make use of this information to create machine readable ontologies to support increasingly sophisticated end user applications.

In this paper we reported on an attempt to create ontologies using the Wikipedia data-set. Its use proved to be a great asset in classifying existing knowledge and in expanding the instances of known classes and their properties. Its use for the generation of property information required methods that are less robust that simple category or linkage extraction and some human review proved necessary to capture this information.

In the future work, the ontology will be augmented with annotation properties tracking the Wikipedia page versions used to create the ontology properties and terms. When possible, it would be desirable to follow the lead of Ruiz et al. ([6]) and seek ways to export information back into Wikipedia using this lineage information.

References

[1] Sören Auer and Jens Lehmann. What have innsbruck and leipzig in common? extracting semantics from wiki content, 2007.

[2] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data - the story so far. In T. Heath, M. Hepp, and C. Bizer, editors, *Special International Journal on Semantic Web and Information Systems*, 2009.

[3] Sebastian Hellmann, Claus Stadler, and Jens Lehmann. The german dbpedia: A sense repository for linking entities. In Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann, editors, *Linked Data in Linguistics*, pages 181–190. Springer Berlin Heidelberg, 2012.

[4] M. Hepp, D. Bachlechner, and K. Siorpaes. Harvesting Wiki Consensus-Using Wikipedia Entries as Ontology Elements. *First Workshop on Semantic Wikis.*

[5] Simone Paolo Ponzetto and Michael Strube. Deriving a large-scale taxonomy from wikipedia. In *AAAI*, pages 1440–1445. AAAI Press, 2007.

[6] Maria Ruiz-Casado, Enrique Alfonseca, and Pablo Castells. From wikipedia to semantic relationships: a semi-automated annotation approach. In Max Völkel and Sebastian Schaffert, editors, *SemWiki*, volume 206 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2006.

[7] Steve Smith and Thomas W. Jackson. Harvesting information from the internet to construct ontologies. *Journal of Emerging Trends in Computing and Information Sciences*, 3(2):211–224, February 2012.

[8] John F. Sowa. The role of logic and ontology in language and reasoning. In R. Poli and J. Seibt, editors, *Theory and Applications of Ontology: Philosophical Perspectives*, volume 1787, pages 231–263. Springer, 2010.

[9] Carlo Torniai, Jelena Jovanovic, Scott Bateman, Dragan Gasevic, and Marek Hatala. Leveraging folksonomies for ontology evolution in e-learning environments. In *ICSC*, pages 206–213. IEEE Computer Society, 2008.