Proceedings of 1st International Workshop in Computational Linguistics for Uralic Languages (IWCLUL 2015); ‹http://dx.doi.org/10.7557/scs.2015.2›

Multilingual Semantic MediaWiki for Finno-Ugric dictionaries

Niklas Laxström
University of Helsinki
Department of Modern Languages
niklas.laxstrom@helsinki.fi

Antti Kanner
University or Helsinki
Department of Finnish, Finno-Ugrian and Scandinavian Studies
antti.kanner@helsinki.fi

December 16, 2014

Abstract

This paper introduces the concept of Multilingual Semantic MediaWiki, which can be used to build collaborative on-line projects for certain types of multilingual content. Namely, dictionaries whose users are multilingual or have different native languages. We describe two multilingual on-line dictionary projects built using the Multilingual Semantic MediaWiki framework. These projects cover Finnish, Swedish, the Sámi languages, Estonian and Ludic among others. We describe the benefits of using semi-structured data and the limitations of this particular semantic software based on the case study offered by the aforementioned projects. We evaluate these projects in terms of development and maintenance effort, number of visitors and contributors. We conclude that this is a low cost approach to increase openness and collaboration and to create more value for this kind of data.

1 Introduction

Following the trends of opening up data for the public under free licenses and of letting the public contribute, this paper describes two such projects coming from the academic sector: the Bank of Finnish Terminology in Arts and Sciences (tieteentermipankki.fi), hereafter BFT, and Sanat (sanat.csc.fi).

We used open source software to build platforms for dictionary data and imported a number of existing sources into them under a free license. We let the interested public contribute and we abolished a separate publishing step.

Building upon existing MediaWiki software, we developed the Multilingual Semantic MediaWiki (MSMW) framework to support such multilingual content and multilingual user base. A multilingual user base consists of polyglots and users with different linguistic backgrounds. By multilingual content we mean content which can be meaningfully accessed through multiple languages.

In the next section we will define our objectives and describe our projects. In the third section we give summary of previous work which we built upon. In the fourth section we describe what MSMW is. In the remaining sections we will present the results, draw conclusions and outline future work.

2 Scope of the research

Our objective is to create an on-line platform for users to collaboratively build specialized dictionaries: in this case terminologies for different fields of research or a multilingual dictionary for purposes of education in a lesser used language. In this section, we describe the general characteristics of two such dictionary projects.

The Bank of Finnish Terminology in Arts and Sciences (started in 2011) is a multidisciplinary project for gathering a permanent terminological database covering all fields of research in Finland. BFT receives funding from the Academy of Finland and University of Helsinki; it is coordinated at the Department of Finnish, Finno-Ugrian and Scandinavian Studies at the University of Helsinki. The project aims to strengthen the parallel use of languages in the academic sector by providing a reliable, easily accessible and up-to-date terminological resource. BFT is split into multiple sub projects or expert groups, each coordinating terminology work in their own field of research. Each group comprises experts of its field, while terminological consultation and guidance are offered by the BFT staff.

In BFT, each term can contain the following information, compliant with ISO standards concerning terminological work: definition, explanation, images, references, additional information, categories, expressions and related terms. There are also some

rarely used fields or topic-specific fields, like the scientific name. Expression, related terms and images are special in a sense that they have an inner structure. Expressions contain the language of the expression, the expression itself, the characterization of equivalence and whether is is recommended, to be avoided, obsolete, etc. Related terms also contain the type of relation, e.g. hyponymy or meronymy, in addition to the term itself. Images hold the name of an image file and a description.

The primary language of most fields is Finnish, but there are also a handful of fields which use Swedish, English, or Finnish Romani. Primary language means the language of term descriptions and explanations: it is the starting point for concept systems. For example, in the topic of jurisprudence, the term *tuomari* (judge) has a very different meaning if we look at it from the view point of the Finnish legal system compared with what it means in common law systems.

Sanat is an editing and publishing platform under development (since 2014) to host multiple monolingual and multilingual dictionaries from the Institute for the Languages of Finland. A prototype was developed in 2014 in collaboration with the Institute for the Languages of Finland, Lyydiläinen seura, CSC – IT Center for Science and Niklas Laxström.

Sanat is composed of more independent dictionaries and hence differs in structure from BFT. While BFT is a concept-based terminology, Sanat is a lexicographical dictionary, where the starting point is a word, not a concept as it is in the BFT. A Ludic dictionary built by the Lyydiläinen seura was chosen as a pilot dictionary to be converted and imported to the Sanat prototype. The Ludic dictionary was a word processing document which used text formatting and special symbols to denote the structure of the data.

Each term in the Ludic dictionary can contain the following linguistic information: basic form; word class; variants and their inflection in different dialects; example sentences translated in different languages, usually Finnish and Russian. There are more than 1 500 Ludic terms in Sanat.

3 Past work: Semantic MediaWiki

We chose Semantic MediaWiki (or SMW; semantic-mediawiki.org) as our base. SMW is explained in this section. Section 4 will explain our extension of it, MSMW.

A wiki approach has multiple properties which have proven to be useful for our projects. All changes are immediately visible. Wikis are cheap to host and it is easy to keep the software up to date; occasionally spam fighting might become a problem. Wikis enable collaboration, as many users are already familiar with them through Wikipedia and it is possible to build an intuitive and predictable interface. People

work on content areas according to their interests and expertise, so some very specialized areas can thrive, while there are naturally gaps in the data.

Semantic MediaWiki is used by over 1600 websites as of 2014, for very different purposes [1]. Semantic MediaWikis include:

- the Finnish Järviwiki (jarviwiki.fi), which contains information about all the lakes in Finland and where users contribute their observations;
- translatewiki.net, created by Niklas Laxström, which is the second biggest SMW site in number of pages (4 millions as of October 2014) [2] and uses SMW for auxiliary functions;
- WikiApiary.com, which is the second biggest SMW site in number of semantic property values (30 millions as of October 2014) [1].

SMW has already been described extensively, in particular by Krötzsch and Vrandečić [3]; here we will only summarise why we chose it as a platform for our projects. SMW satisfied the following requirements:

- store and preserve all the information contained in the pre-existing thesaurus structure;
- offer a user-friendly interface to add and edit data, namely semantic forms which most users can use without knowing any markup[1];
- expose the data in meaningful and attractive ways, for instance on the main page, topic-based portals and other query interfaces to the data; allow researchers to search expressions with a certain form.

Moreover, with SMW the structure is not defined in the programming code, but instead the semantic relations and forms are defined in the wiki, with wiki markup. It is also easy to modify this structure on the fly, unless manual updates of the existing data are implied, and even then modifications are retrospectively possible to some extent. This lowers the barrier to build the dictionary structure. Practically all available resources (in terms of funding and time) can be used to customise the wiki platform for the intended purpose.

4 Multilingual Semantic MediaWiki

For both dictionaries, given their goals, we defined a set of requirements for multilingual support. We call a wiki which satisfies these requirements a Multilingual Seman-

[1] In fact, previous research has stressed how SMW is suitable for inexperienced users as well [4], so that the usability of web 2.0 and richness of the semantic web are not in opposition [5].

tic MediaWiki (MSMW). In this section we explain how we satisfied the requirements by using existing solutions or developing one ourselves.

Semantic wikis have been attempted in the past, often based on Controlled Natural Language [6, 7] and with a multilingual approach [8], showing users reach a high level of consensus for content [9]. However, we find that all previous approaches failed to fully internationalize the user experience as we aim; and produced wikis which rarely are still on-line and in use. Therefore, the need of a more systematic and robust approach arises: the properties which we define next.

We need a semantic wiki provided with

1. automatic guesses of the user's preferred language,
2. manual language switching,
3. input methods,
4. web fonts for language support,
5. translatable documentation,
6. translatable forms,
7. translatable content interface,
8. structured multilingual content.

Firstly, we consider *multilingual* something that strives to equally support all languages: hundreds rather than few. MediaWiki aspires to be "internationalized, with equal support for all languages," [10] and is being localised in over 350 languages, including right-to-left languages, with full support for any language specificity [11]. MediaWiki is the only existing platform satisfying our multilingualism requirement; no existing CMS, semantic platform or wiki engine can satisfy all the 8 properties, other than MediaWiki. This made us choose SMW as our base platform and call MSMW a system which satisfies all the 8 properties.

The MSMW framework is meant to leverage this extensive language support, by making sure that it extends to all semantic features of the wiki without degradations. As for features and interfaces shared between wikis, any defect or lack of translations should be fixed in the upstream MediaWiki code and in translatewiki.net respectively, to benefit all installs. As for content and interfaces specific to one wiki, they should be translatable on the wiki itself.

Parts 1.–5. are readily available to any MediaWiki instance by installing the MediaWiki Language Extension Bundle (MLEB)[2]. Parts 1.–4. constitute *basic language support* and are provided by the Universal Language Selector extension (included in MLEB), which uses information given by the user's browser, geo-location of the user's

[2] https://www.mediawiki.org/wiki/MLEB

IP address and Unicode Common Locale Data Repository (CLDR) to choose and suggest most likely languages[3]. The user can easily choose language manually when the Universal Language Selector fails to infer correctly.

Input methods and web fonts complement the support provided by browsers and operating systems, using JavaScript and web standards for font delivery. Lack of fonts is a common problem for many Indian languages. Lack of input methods is also common for many Indian languages, as well as small languages not included in computer standards and people who live abroad or are traveling.

Part 5, translatable documentation, can be achieved with the Translate extension (included in MLEB). When pages are prepared for translation according to the documentation of the Translate extension, they can easily be translated by translators using a dedicated translation interface inside the wiki. The interface provides common translation tools like translation memory, machine translation service integration, translation notes and most importantly change tracking. *Change tracking* ensures that translated versions are never out of date by integrating missing and outdated translations with the source language.

Content interface means that some elements of the interface are defined on the wiki, but follow the user's interface language. This feature was used, for example, with headings and labels. The Translate extension also provides a way to tag those elements so that they can be translated. In Translate's documentation, this method is called unstructured element translation[4]. Parts 6. and 7. are an application of unstructured element translation to SMW, which to our knowledge has not been done before.

Structured multilingual content means the wiki has data input forms which can accept multilingual content and is able to store and display such multilingual content. For example, in BFT, the list of expressions in different languages for each term are multilingual content. To design structured multilingual content, one has to understand what parts of the data can be multilingual. Multilingual content does not necessarily follow the user's interface language as content interface does.

In practice this means providing, for fields which accept content in different languages, an additional field where to set the language of linguistic content. Correct language tagging in HTML output is important for search engines and application of web fonts. Language annotations of the data are used in semantic queries and by third party users of your data.

[3] https://www.mediawiki.org/wiki/Universal_Language_Selector/FAQ#How_does_Universal_Language_Selector_determine_which_languages_I_may_understand

[4] https://www.mediawiki.org/wiki/Help:Extension:Translate/Unstructured_element_translation

5 Other semantic structure characteristics

In MediaWiki, all content is split across pages. A page is like a web page in that it has no fixed length, unlike printed pages. A page has content on a specific topic, usually defined by the page title, which is also used in the unique address of the page. Pages are sometimes also known as articles if it fits the type of content, like in Wikipedia. Furthermore, namespaces are used to separate different types of content, e.g. help pages are in a separate namespace. Namespace appears in page title as prefix separated by colon, for example `Help:Editing`.

For BFT we created multiple additional namespaces. Each sub group, which we call a terminology, has its own namespace. Thanks to this `Kielitiede:kieli` (language in linguistics) and `Eläintiede:kieli` (tongue in zoology) are two separate terms. When links are created to other pages, the namespace is not usually visible in the link text. We gave each terminology a separate color, which is shown in page titles and links across the interface. This allows users to know in which terminology they are.

In addition to each terminology having its own namespace, we also created a namespace for all the expressions. The pages in the expressions namespace contain information which relates to the surface forms like word class and language. The pages link back to all terms in any terminology which contain that expression. In the case that multiple expression have same form, they will share the same page with information for both.

What caused most headaches for us were the limitations of the semantic structure. The basic tool we have is a subject-property-value triplet, where the subject is always implicitly a page in the wiki. Properties we can define freely, but the values cannot have inner structure; in other terms, SMW can only store 2D data, not 3D or N-dimensional data.

For example, we have a word *talo* in Finnish and we want to give multilingual examples of sentences where it is used. We were unable to say that the subject `talo` had a property `example-sentence` with value `{fi: talo paloi; en: house burned}`. We tried to store the data without semantic relations, but that did not work either, because semantic forms have limitations with so-called multiple instance templates[5]. In the end this problem was solved in different ways in BFT and Sanat. In BFT we do not have such complex embedded structure. In Sanat we moved that kind of data into separate pages (subpages in MediaWiki), which forced us to provide editing controls directly on the page itself, hence mixing up *view mode* and *edit mode*.

[5] `https://www.mediawiki.org/wiki/Extension:Semantic_Forms/Defining_forms#Multiple-instance_templates`

From Sanat's characteristics follows the main structural difference from BFT: there is no shared global namespace for expressions. Information related to expressions is included in the term pages. Multilingual examples of terms are stored in separate pages due to reasons described above. The section "Related terms" is replaced by a generic section "See also" on the same page. Examples are given as sentences with translations.

Using semantic queries, we also automatically created two reverse dictionaries: Finnish to Ludic and Russian to Ludic. We expect that in Sanat the fields will be more customised for each dictionary, as opposed to BFT where all sectors of research contain the same fields to a great extent.

6 Outcomes

We found out that we could come up with working prototypes in just few hours, including the time to set up MediaWiki with many extensions. After the initial launch of the BFT, we were also able to quickly satisfy user feedback thanks to the flexibility of the platform.

We made an extension to MediaWiki, MixedNamespaceSearchSuggestions[6], to show more suggestions when the user types something in the search box and to often eliminate the need for a full text search. First, suggestions are shown from all namespaces (dictionaries) at once, with no need for the keyword to match the namespace name. Second, the namespace is shown next to each suggested title. The extension is released with an open source license.

While developing BFT we found out that some form elements, like certain types of buttons, did not allow translation with the approach used in MSMW. We submitted patches to fix some of these, but not all cases have been fixed yet.

In 2014, BFT has reached about 30 000 concept article pages, corresponding to about 70 000 terms in 35 languages. The contents of the concept articles vary from terminology to terminology, as different work groups present different choices in work flow and differing stages of progress. For example terminology of Jurisprudence contains nearly 2 000 articles with extensive contents, while Epidemiology contains not much more than Finnish-English term lists. Even still, according to user survey conducted in the spring of 2014, 85 percent of the respondents said that they had found information they were looking for either completely or partly.

BFT's constant activity of 20 to 40 monthly active editors has ensured a constant growth of the dictionary[7]. Editors are mostly academics from across Finland. For

[6] https://www.mediawiki.org/wiki/Extension:MixedNamespaceSearchSuggestions

[7] https://wikiapiary.com/wiki/Tieteentermipankki.fi

comparison, the Finnish Wiktionary has around 30 editors active in a given month[8]. Some active contributors helped the wiki's development beyond their edits, for example by negotiating a licence for an existing terminology with hundreds of terms to be imported into the wiki. Expert participation on the platform is not thus limited to only writing new terminological records, but encompasses also selecting, revising and updating existing terminological records for import. The added value of bringing resources to BFT is the possibility to integrate separate terminologies in to one easily accessible resource. According to BFT's user survey of 2014, its resources are widely used by undergraduate students in those fields, where the contents have reached sufficient extension.

Since Sanat is still not publicly launched, we cannot qualify its success in terms of users and contributors. We can say it took less than 40 hours to develop it, including conversion of an existing dictionary in a format suitable for import in the wiki.

7 Conclusion

BFT is a successful project which has benefited from the support for multilinguality. It does not compete with general purpose dictionary projects like Wiktionaries or OmegaWiki due to its specific scope and customisations to support terminology work.

MSMW enhances a regular SMW instance by making it suitable for multilingual users and multilingual content. This is different from for example Wikipedia and Wiktionary, where each language version is a separate instance with a separate community. Wikipedia and Wiktionary also do not use SMW to structure their data. The novel part of MSMW is the idea of combining MediaWiki, MLEB and SMW and the ways how to best use them together to provide additional value. We have contributed to all of the components to make them work better together.

By using MSMW on a projects like BFT we know that MSMW works in practice. The main issue is the manual work needed to set up unstructured element translation. Also the issues with untranslatable elements in forms, until fixed, make MSMW a less compelling approach. Even with these issues, MSMW is already useful because of the benefits in rapid prototyping and the superior language support.

Our code changes have been integrated in MediaWiki and extensions or released as new open-source extensions to MediaWiki where applicable.

[8] https://stats.wikimedia.org/wiktionary/EN/TablesWikipediaFI.htm#editor_activity_levels

8 Future work

MSMW extensions can be further developed to reduce the manual work needed for a new MSMW wiki to be translatable using the unstructured element translation feature. Currently, the Universal Language Selector is not employed for language selection in forms, but this should be relatively easy to add.

The idea of separate *view* and *edit modes* may be considered outdated by the supporters of in-place editing [12]. This might become an issue in the future if people start seeing our platform as outdated and hard to use. Replacing the form paradigm with in-place editing would be a huge undertaking in SMW.

Integration of our data with external data sources is still to be solved. SMW provides machine-readable application programming interfaces (API). They are not, however, tied to any general vocabulary, which means that a developer would have to map the properties and values manually, and each wiki can have different structure. A simple standardized format should be developed for our dictionaries to be used as data providers for tools such as Content translation[9], currently in development by the Wikimedia Foundation.

Other researchers working with SMW stress the importance of alternative modes of accessing lexicographical data such as maps, timelines and charts [13]. We have not yet explored how to apply MSMW in such a context.

Finally, MSMW is defined strictly by the actual software we have used. A more general, software independent approach can be developed for creating multilingual collaborative content management systems by starting from the important concepts such as the separation between the software interface, the user created content interface and the user created content.

Acknowledgements

Niklas Laxström, co-author of this paper, is a MediaWiki consultant and has received funding from the University of Helsinki to develop BFT and from Lyydiläinen seura to develop Sanat. He also works with Wikimedia Foundation which is a contributor to MediaWiki software [14].

We thank Federico Leva for detailed comments on the manuscript and helpful discussions regarding this work and previous research, aided by his extensive knowledge of wikis.

[9] `https://www.mediawiki.org/wiki/Content_translation`

References

[1] Jamie Thingelstad. Semantic statistics. `https://wikiapiary.com/wiki/Semantic_statistics`, October 2014. WikiApiary, as of this writing, monitors 24,997 MediaWiki websites, of which over 21,000 were provided by WikiTeam and Federico Leva. Some well known Semantic MediaWikis are not included in the statistics because they don't publicly expose the required data.

[2] Daniel Zahn's wikistats. List of largest wikis. `http://wikistats.wmflabs.org/largest_html.php?s=total_desc&th=0&lines=20`, October 2014.

[3] Markus Krötzsch and Denny Vrandečić. Semantic mediawiki. In Dieter Fensel, editor, *Foundations for the Web of Information and Services*, pages 311–326. Springer Berlin Heidelberg, 2011.

[4] François Bry, Sebastian Schaffert, Denny Vrandečić, and Klara Weiand. Semantic wikis: Approaches, applications, and perspectives. In Thomas Eiter and Thomas Krennwallner, editors, *Reasoning Web. Semantic Technologies for Advanced Query Answering*, volume 7487 of *Lecture Notes in Computer Science*, pages 329–369. Springer Berlin Heidelberg, 2012.

[5] Anupriya Ankolekar, Markus Krötzsch, Thanh Tran, and Denny Vrandecic. The two cultures: Mashing up web 2.0 and the semantic web. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 825–834, New York, NY, USA, 2007. ACM.

[6] Jie Bao, Paul R Smart, Nigel Shadbolt, Dave Braines, and Gareth Jones. A controlled natural language interface for semantic media wiki. In *3rd Annual Conference of the International Technology Alliance (ACITA'09)*, September 2009. Event Dates: 23rd - 24th September 2009.

[7] Pradeep Dantuluri, Brian Davis, Pierre Ludwick, and Siegfried Handschuh. Engineering a controlled natural language into semantic mediawiki. In Michael Rosner and NorbertE. Fuchs, editors, *Controlled Natural Language*, volume 7175 of *Lecture Notes in Computer Science*, pages 53–72. Springer Berlin Heidelberg, 2012.

[8] Kaarel Kaljurand and Tobias Kuhn. A multilingual semantic wiki based on attempto controlled english and grammatical framework. In Philipp Cimiano, Oscar Corcho, Valentina Presutti, Laura Hollink, and Sebastian Rudolph, editors, *The Semantic Web: Semantics and Big Data*, volume 7882 of *Lecture Notes in Computer Science*, pages 427–441. Springer Berlin Heidelberg, 2013.

[9] Kaljurand Kaarel, Kuhn Tobias, and Canedo Laura. Collaborative multilingual knowledge management based on controlled natural language. *Semantic Web*.

[10] MediaWiki.org. Principles. `https://www.mediawiki.org/w/index.php?title=Principles&oldid=1185610`, October 2014.

[11] MediaWiki.org. Languages in MediaWiki architecture. `https://www.mediawiki.org/w/index.php?title=Manual:MediaWiki_architecture&oldid=1092154#Languages`, October 2014.

[12] Alan Cooper, Robert Reimann, David Cronin, and Christopher Noessel. *About Face: The Essentials of Interaction Design*. John Wiley & Sons, 2014. See several passages on edit in place and «Design principle: allow input wherever you have output».

[13] Bruno Bon and Krzysztof Nowak. Wiki lexicographica. linking medieval latin dictionaries with semantic mediawiki. In *Proceedings of eLex 2013*, 2013.

[14] MediaWiki.org. Differences between Wikipedia, Wikimedia, MediaWiki, and wiki. `https://www.mediawiki.org/w/index.php?title=Differences_between_Wikipedia,_Wikimedia,_MediaWiki,_and_wiki&oldid=1240490`, October 2014.