Towards a metadata infrastructure for online dictionaries

Michal Boleslav Měchura

Natural Language Processing Centre, Masaryk University, Brno, Czech Republic
Fiontar and School of Irish, Dublin City University, Ireland
New English–Irish Dictionary Project, Foras na Gaeilge, Dublin, Ireland
`valselob@gmail.com`

1 Introduction

When dictionaries started migrating from paper to screens in the early 2000s, it was a huge innovation: one could now search a dictionary quickly and accurately in one or two seconds instead of having to leaf through the pages of a book. Even today some people are still awe-struck by this. But for most of us that evolutionary step is now in the past: online dictionaries have become the new normal. Which begs the question, what is the next step going to be in the evolution of human–dictionary interaction? I think it is going to be **aggreggation**: people will increasingly be wanting to access multiple dictionaries simultaneously, from one place. A similar trend already exists in other branches of the reference industry including libraries (in the form of country-specific or city-specific library portals[1]) and in scientific publishing (in the form of 'discovery' portals, often operated by universities for their staff and students[2]).

In their current incarnations, our online dictionaries do not support aggregation very well. There are hundreds if not thousands of online dictionaries in the world, but finding them and finding things inside many of them at the same time is a challenge for web users. General search engines like Google do not meet this challenge very well because (1) they do not distinguish between searches for factual content and searches for linguistic information (*tell me about cats* versus *tell me about the word 'cat'*) and (2) they fail to promote trustworthy dictionaries over 'quick and dirty' predatory click-bait. Clearly, some kind of dictionary-specific search engine or portal is needed.

ENeL's Work Group 1 has given itself the mission of building a European Dictionary Portal, `dictionaryportal.eu`. We have done reasonably good work, given the circumstances. But the circumstances given were not favourable. Out main outcome is the realization that one does not simply build a dictionary portal: one must build up the necessary infrastructure first.

2 The missing infrastructure

Europe's dictionary infrastructure can be best described as non-existent. Unlike our railways, our mobile phone networks and other infrastructural industries which have successfully standardized themselves for the mutual benefit of all, Europe's dictionaries are medieval city-states who eye each other with suspicion, especially those of their neighbours who speak the same language. The

[1] Like `cistbrno.cz` which meta-searches all libraries in the city of Brno or `knihovny.cz` which includes all (major) libraries in the Czech Republic.

[2] Like Masaryk University's `discovery.muni.cz`.

infrastructure which we need in order to evolve out of that state can be divided into two areas: **dictionary discovery** and **dictionary access**.

In **dictionary discovery**, what is missing is a usable scheme for describing, categorizing and ranking dictionaries. We tried to create one in Work Group 1, at least implicitly, while working on our dictionary inventory [7], [8] and later while implementing a radically simplified version of it in the portal. The scheme we have ended up with is not entirely satisfactory. What we need is a thoroughly thought-through metadata standard for describing dictionaries.

In **dictionary access**, we live in a world where each online dictionary is its own island with no documented, machine-executable protocol for performing searches or locating individual entries. Building an aggregator, such as portal website, is difficult in these circumstances: one needs to reverse-engineer each individual website before one can know how to link to it and whether one can link to it at all. Some websites make this so difficult it is practically impossible.

In the rest of this paper I will go through these two areas in more detail, outlining the work that needs to be done to build the infrastructure we do not currently have. Then I will come back again to the idea of dictionary portals and other applications that could one day be built on top of such infrastructure.

3 Towards a metadata standard for dictionaries

By metadata I mean 'data about data'. In other words, not the contents of a dictionary but information about it as a whole: what language or languages it has, what kind of audience it is intended for, which headwords it has, whether or not it has pronunciation and so on.

Thinking about metadata takes us away from lexicography and closer to library science. Throughout its evolution – especially in its latest, digital stage – library science has created a number of machine-readable standards for recording metadata about intellectual works such as books and articles, including online publications. These standards range from expressive and complicated ones like MARC [9] to simple ones like Dublin Core [1], including half-way compromises like MODS [10]. Disappointingly, neither of these is perfectly suited to capturing metadata about dictionaries. For example, distinguishing between a dictionary's object language and metalanguage is not something existing library standards manage easily. We need a dedicated metadata standard for dictionaries which would allow us to capture relevant information about dictionaries such as:

- Which languages the dictionary contains and what roles they have there: which is the object language (the language the dictionary describes) and which, if different, is the metalanguage (the language in which the descriptions are made); which is the source language (the language of the headwords) and which is the target language (the language of the translations). The scheme needs to handle unusual cases gracefully, such as when a dictionary has several metalanguages.
- Which subset of the object language the dictionary covers: general vocabulary (LGP[3]), specialized terminology (LSP[4]) and if so, in which discipline or disciplines, phraseology, idioms, a particular dialect and so on.

[3] Language for General Purposes

[4] Language for Specific Purposes

- Whether the dictionary is intended for encoding or for decoding or for some other lexicographic function, but bearing in mind that the boundaries can be fuzzy and that dictionary publishers can be deliberately vague about this.
- What kind of user the dictionary is intended for: native speakers, second-language learners, children, adults and so on; again bearing in mind that these categories are fuzzy and cannot always be determined.
- What kind of information the dictionary provides: is it mainly an orthographic dictionary, a morphological dictionary, an etymological dictionary? Again, bearing in mind that while some dictionaries fit one of these descriptions perfectly, many others do not fit a single category or the publisher may be deliberately vague about it.

This is only an early and rough catalogue of requirements for a dictionary metadata standard. A closer scrutiny would certainly reveal more detail.

4 Towards a standardized protocol for dictionary access

By dictionary access I mean things like sending search requests to a dictionary, obtaining search results in a machine-readable format and sending users to individual entries. Ideally, all dictionary websites should do these things in roughly the same way, but we have a long way to get to that level of standardization. To get there, we can once again take inspiration from standards that already exist outside lexicography.

One such standard is OpenSearch [4]. Most web users know it in the form of a search box somewhere in the corner of their web browser. You can type one or more keywords into this box and it will redirect you to a page on Google or some other searchable website with results for those keywords. As you browse the web you can add more websites to the search box. This is possible because many websites provide an OpenSearch-compliant plugin which, among other things, tells your browser how to construct a search URL for that website. The OpenSearch standard was created in the early 2000s by A9 (a subsidiary of Amazon.com) and remained little used for a long time, but in the last decade or so OpenSearch plugins have become quite a common sight on the web, mainly thanks to the fact that browsers now usually have an OpenSearch search box. The nice thing about OpenSearch is that an OpenSearch plugin can be written by anybody, even by people not associated with the website the plugin is for. A popular repository of OpenSearch plugins written by third parties is the Mycroft Project [3] (formerly run by the Mozilla Foundation, now hosted at Oregon State University). Many dictionary websites already offer OpenSearch plugins but many still do not. Moreover, OpenSearch is not a completely perfect match for dictionary websites because it does not easily allow for things like language selection: if a dictionary can be searched in more than one language direction, it needs to provide two separate OpenSearch plugins. But, a version of OpenSearch adapted especially for dictionaries would be an attractive proposition for lexicography.

Another useful web standard is Sitemaps [5], originally created by Google and now an open-source specification ‘sponsored’ jointly by Google, Yahoo and Microsoft. A sitemap is a machine-readable file which a website publisher can put on their website. It tells search engines which pages exist on the website, what their URLs are, what languages they are in, how often they are likely to be updated and so on. This helps search engines index a website more reliably, without having to guess where everything is or reverse-engineer anything. Again, the Sitemaps standard is so generic that it fails to capture many things that are important on dictionary websites, but a dictionary-specific version of it is possible and desirable.

Taking all these sources of inspiration together, we need a standardized scheme in which dictionary websites could communicate the following facts about themselves to the world:

- How the dictionary can be accessed online: whether a subscription is required and if so, how to find out whether the user already has one, whether the user will be asked to agree to terms and conditions, and so on.
- How to compose a search URL. If the dictionary supports several languages, then how to compose a search URL for each allowed combination of them. If the dictionary is multilingual but can be searched without giving a language, then again, how to compose a URL for such a search.
- How to find out whether the dictionary has any results for a given search query. An aggregator application (such as a portal website) could use this to know which dictionaries can be offered to a user and which not, depending on whether they do or do not have the content the user requires.

Again, this is an early rough draft of requirements, not a deep analysis. The point is, however, that the standard proposed here would not require dictionary publishers to give up their data. Early thinking on the European Dictionary Portal included suggestions that dictionary publishers should be forced to make the contents of their dictionaries freely available in machine-readable formats (LMF [2], TEI [6] etc.) so that a portal like ours could then 'mash' them up and present them to the user on a single screen. I do not think this vision is achievable. Dictionary publishers tend to have an aversion to such proposals. Dictionary publishers mainly want to bring people to their own websites – even non-commercial ones because public awareness and traffic numbers are what supports them in their struggle for relevance and funding. The infrastructure we hope to build must respect this, otherwise it will not be widely adopted. The scheme must work on the basis of metadata rather than data, and must normally culminate in the redirection of a human user to a page inside the dictionary's own website.

5 Conclusion

The metadata infrastructure envisaged here is not just meant to support the European Dictionary Portal. In a way, our own portal website is unimportant. What is important is that anyone and everyone would be able to build aggregator applications that use machine-readable metadata about dictionaries, including:

- Portal websites, either strongly curated like our own European Dictionary Portal or permissive 'catch-all' ones, topic-specific portals, language-specific portals, country-specific portals and so on.
- Plugins for web browsers or e-book readers, or completely stand-alone applications which users could have on their computers and which would allow them to build up their personal library of dictionaries and search them all from one place.
- Links between dictionaries: imagine an online dictionary which, when it is unable to return any search results, gives you recommendations for other dictionaries that do have matches for your query.

This paper is, then, a suggestion to the e-lexicography community to create a metadata standard for dictionaries. What motivates it is the realization that simply publishing dictionaries as stand-alone websites is not good enough any more. This is no longer innovative. What people want now is to be able to search many dictionaries simultaneously, without having to visit them one by one. This will be the next step in the evolution of human-dictionary interaction. But to deliver it we need to build up the necessary infrastructure first.

References

1. Dublin Core Metadata Initiative. http://dublincore.org/.
2. ISO 24613:2008 Language Resource Management – Lexical Markup Framework (LMF). http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=37327.
3. Mycroft Project. http://mycroftproject.com/.
4. OpenSearch. http://www.opensearch.org/.
5. Sitemaps. https://www.sitemaps.org/.
6. Text Ecoding Initiative Guidelines: 9. Dictionaries. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html.
7. Gerbrich de Jong. How to interpret the Excel file with the list of dictionaries. Report from ENeL short-term scientific mission, 8 2014. http://www.elexicography.eu/wp-content/uploads/2015/04/Explicative-article-STSM-Gerbrich-de-Jong.docx.
8. Gerbrich de Jong. Report on a survey of European dictionaries. ENeL workshop, Bled, Slovenia, 9 2014. http://www.elexicography.eu/wp-content/uploads/2014/11/Bled-2014-enel_de-Jong.ppsx.
9. Library of Congress. MARC (Machine Readable Cataloguing). http://www.loc.gov/marc/.
10. Library of Congress. MODS (Metadata Object Description Schema). http://www.loc.gov/standards/mods/.