Brede Wiki: Neuroscience data structured in a wiki

Finn Årup Nielsen

Center for Integrated Molecular Brain Imaging, Copenhagen, Denmark;**
DTU Informatics, Technical University of Denmark, Lyngby, Denmark;
Neurobiology Research Unit, Copenhagen, Denmark
fn@imm.dtu.dk, http://www.imm.dtu.dk/~fn/

Abstract. Setup in January 2009 the *Brede Wiki* contains data from neuroscience, particularly from published neuroimaging peer-reviewed papers. Data is stored in simple MediaWiki templates and it can automatically be extracted and represented in an SQL format. Off-wiki Web-scripts can use the SQL database so items in the wiki can be queried efficiently, e.g., to find close brain activations to a given coordinate. Templates are maintained simple so data extraction is more or less complete.

1 Background

The complexity of neuroscience data has resulted in multiple neuroinformatics databases. One particular kind of database records brain activations foci from published peer-reviewed neuroimaging papers. The foci are reported with respect to a so-called stereotaxic space, such as the *Talairach space* [1], so coordinates are reasonably comparable across studies. One of the first neuroinformatics databases, *BrainMap*, records this kind of data and had its Web-presence in the early Web [2]. It continues in a modified version, and a few other coordinate databases have emerged: AMAT, SumsDB and our own Brede Database [3, 4]. Many of the neuroinformatics databases handle data entry manually. This is time consuming and may be the reason why databases are not complete. Our Brede Database uses Matlab for data entry and XML for storage. The scheme does not encourage collaborative and incremental data entry. We have constructed a plug-in that searches the database from the popular *SPM* program [5]. Extensions to the plug-in could potentially upload coordinates to the database, but that would need a submission interface on the server side.

Bioinformatics has embraced wiki technology, with e.g., the wikis SNPedia, WikiGenes, WikiProtein, WikiPathways.[1] Although some neuroinformatics wikis exist they are mostly text-oriented and used for documentation rather than for recording data or describing its content more semantically.

** Thanks to the Lundbeck Foundation for funding.

[1] The English Wikipedia page *Portal:Gene Wiki/Other Wikis* presently lists other bioinformatics wikis.

We have gained some experience in processing the templates of Wikipedia (more specifically the *journal* field of the *cite journal* template for scientific citations) and for submitting the processed templates to statistical analyses [6, 7]. The DBpedia effort [8] has shown the possibility for large-scale databasing with data extracted from Wikipedia content.

Here I describe a system, the *Brede Wiki* (http://neuro.imm.dtu.dk/wiki/), that use MediaWiki and its template functionality to record neuroinformatics data. Associated scripts can extract the template content from the entire XML dump of the wiki and automatically construct SQL representation of the data. The SQL database enables a more advanced query than is possible with the present standard MediaWiki.

2 Methods

To avoid making data extraction unnecessarily complex and to match with the SQL language the application of MediaWiki templates has been kept simple: Present templates do not nest, have lower-case characters (except for first letter) and wiki markup is avoided within the template field values. The following is an example of an application of the `paper` template:

```
{{Paper
 | author1 = Daniela Balslev | author2 = Finn Nielsen
 | author3 = Olaf B. Paulson | author4 = Ian Law
 | title = Right Temporoparietal Cortex Activation during
          Visuo-proprioceptive Conflict
 | journal = Cerebral Cortex
 | volume = 15 | issue = 2 | pages = 166-169 | year = 2004
 | pmid = 15238438 | doi = 10.1093/cercor/bhh119 | wobib = 128
}}
```

In the present wiki one page may contain multiple templates, e.g., a page for a specific neuroimaging paper can contain the `paper` template as well as multiple templates for brain coordinates: the `Talairach coordinate` template. Most template definitions format the template content by construction of an infobox as known from Wikipedia. What is somewhat different from Wikipedia is the extensive use of wiki links within the template definitions. For example, for the `paper` template wiki links to the authors and journal is constructed, with MediaWiki template definitions like `[[{{{author1}}}]]`. External links are created from the external database identifiers, such as DOI and the PubMed Identifier (PMID). For the templates which usually come in sets, such as `Talairach coordinate`, each application of a template defines a row in a table.

An example of a page within the Brede Wiki is displayed in Figure 1 with the `paper` template as the upper right infobox and with three formatted applications of the `Talairach coordinate` template at the bottom. The `x`, `y` and `z` fields of the template can be combined in a query to external specialized coordinate search engines. These links are display in the right-most column.

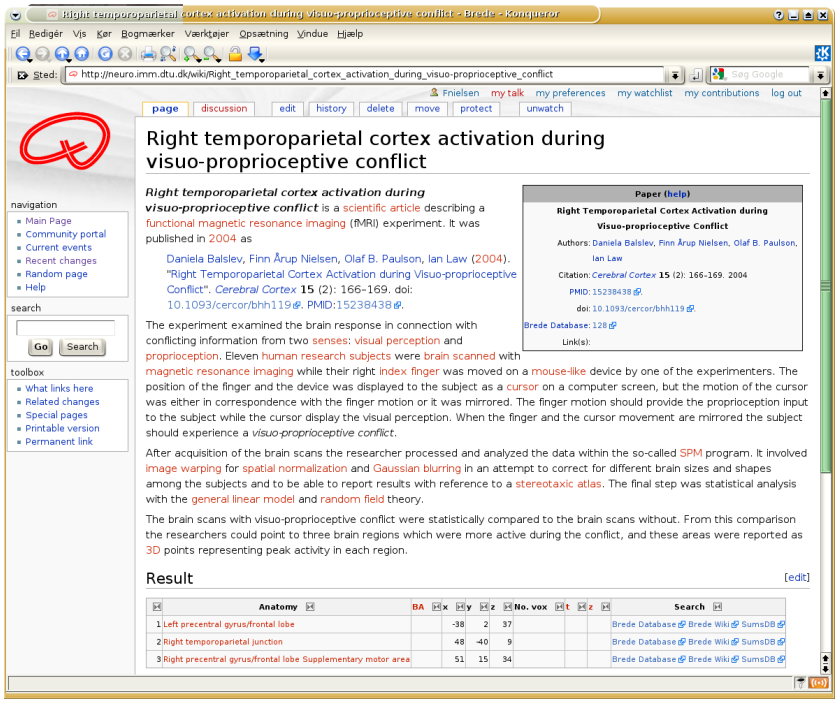

Fig. 1. Screenshot of Brede Wiki with a page about a scientific article.

As the templates are not nested relatively simple Perl regular expressions retrieve them with `m/{{(.*?)}}/sg` and subsequently extract the template name and its content with

```
m/([a-z][a-z0-9]*(?:[ _][a-z0-9]+)*)\s*(\|.*)?/si
```

Spaces are later substituted with underscores. Another regular expression in the same style extracts name-value pairs from each field. The extracted content is written to SQL tables, where the master table is presently defined for SQLite as

```
CREATE TABLE brede(id INTEGER, pid INTEGER, title, tid INTEGER, template,
                   field, fid INTEGER, value);
```

`id` is the row identifier, `pid` an identifier for the wiki page which title is also (redundantly) represented in the `title` column. `tid` and `template` are identifier and name of the template, while `field` is the field name, `fid` the field number and `value` the value, i.e., the actual content of the field. An insert with the `author2` field from the above displayed `paper` template example may be issued as:

```
INSERT INTO brede VALUES(387, 31, 'Right temporoparietal cortex
                 activation during visuo-proprioceptive conflict',
                 136, 'paper', 'author', 2, 'Finn Nielsen');
```

Apart from the master table several other tables are built: One for each template. For the `paper` template the SQL definition for the corresponding table may look like the following:

```
CREATE TABLE brede_paper(__tid, __pid, __title,
             _author1, _author2, _author3, _author4,
             _title, _journal, _volume, _issue, _pages, _year,
             _pmid, _doi, _wobib);
```

The final number of columns will depend on the number of different field names discovered during reading of the XML dump. The table names are prefixed with `brede_` and the column names with an underscore to avoid clashes between Brede Wiki names and SQL reserved words. A specialized search engine use the constructed SQL database when searching for nearby Talairach coordinates to a query coordinate [9].

At the time of writing 43 papers were represented in the Brede Wiki, 31 which potentially contain Talairach coordinates. In comparison the Brede and BrainMap databases have presently 186 and 1711 papers, respectively. Apart from templates for papers and Talairach coordinates, the Brede Wiki has also templates for, e.g., brain regions, researchers, subject groups and brain volume results.

3 Discussion

There are both advantages and disadvantages with the Brede Wiki compared to our previous system. Some of the advantages are online versioning, immediate access to entered data, incremental addition of data, the possibility for free format text descriptions as well as discussion pages. Furthermore the data structure is extensible, i.e., it is relatively easy for editors to add new 'columns' (fields).

Among the disadvantages are the issue of vandalism, quality control and database consistency. A paper may have multiple sets of coordinates that arise from different brain scanners, multiple subject groups and multiple experiments. If these components are described on a single wiki page the connection between them will need to be indicated with keys of some kind, e.g., to say that the fourth `Talairach coordinate` resulted from analysis of brain scans from the second `Subject group`.

The wiki does not solve the data entry problem *per se*. The data in the Brede Wiki has so far for the most part been entered manually in the raw wiki text. A small Matlab script can convert results from SPM so they appear in the Brede Wiki template format. Our small fielded wiki with form entry for personality genetics association studies [10] can also output its data for inclusion in the Brede Wiki. Entry with forms within the wiki would be a natural next step as well as possibly the adding of Semantic MediaWiki functionality [11]. Yet another option for data entry is scripts that automatically setup wiki pages from information in other databases, — an approach taking for *Gene Wiki* in Wikipedia [12]. Using this scheme it should be possible to augment the Brede Wiki with information from the Brede Database.

In our small fielded wiki of personality genetics we can perform on-the-fly meta-analysis and data plotting. The vision is that the Brede Wiki can constitute the basis for large-scale Web-based meta-analyses similar to that of the (non-wiki) AlzGene database [13].

4 Conclusion

The Brede Wiki is one of the first steps in neuroinformatics with Web 2.0 and with a high degree of structured content. It shows the possibility to output structured MediaWiki content to an SQL database.

References

1. Talairach, J., Tournoux, P.: Co-planar Stereotaxic Atlas of the Human Brain. Thieme Medical Publisher Inc, New York (January 1988)
2. Fox, P.T., Lancaster, J.L.: Neuroscience on the net. Science **266**(5187) (November 1994) 994–996
3. Derrfuss, J., Mar, R.A.: Lost in localization: The need for a universal coordinate database. NeuroImage, doi:10.1016/j.neuroimage.2009.01.053 (2009)
4. Nielsen, F.Å.: The Brede database: a small database for functional neuroimaging. NeuroImage **19**(2) (June 2003) Presented at the 9th International Conference on Functional Mapping of the Human Brain, June 19–22, 2003, New York, NY. Available on CD-Rom.
5. Wilkowski, B., Szewczyk, M., Rasmussen, P.M., Hansen, L.K., Nielsen, F.Å.: Coordinate-based meta-analytic search for the SPM neuroimaging pipeline. In: International Conference on Health Informatics (HEALTHINF 2009). (2009)
6. Nielsen, F.Å.: Scientific citations in *Wikipedia*. First Monday **12**(8) (August 2007)
7. Nielsen, F.Å.: Clustering of scientific citations in Wikipedia. In: Wikimania. (2008)
8. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: DBpedia: A nucleus for a web of open data. In: The Semantic Web. Volume 4825 of Lecture Notes in Computer Science., Heidelberg/Berlin, Springer (2008) 722–735
9. Szewczyk, M.M.: Databases for neuroscience. Master's thesis, Technical University of Denmark, Kongens Lyngby, Denmark (2008) IMM-MSC-2008-92.
10. Nielsen, F.Å.: A small wiki for personality genetics. 37th Annual Meeting on Biochemistry and Molecular Biology: Frontiers in Genomics (October 2008)
11. Krötzsch, M., Vrandečić, D., Völkel, M.: Semantic MediaWiki. In Cruz, I., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., Aroyo, L., eds.: The Semantic Web - ISWC 2006. Volume 4273 of Lecture Notes in Computer Science., Berlin/Heidelberg, Springer (2006) 935–942
12. Huss, III, J.W., Orozco, C., Goodale, J., Chunlei, Batalov, S., Vickers, T.J., Valafar, F., Su, A.I.: A gene wiki for community annotation of gene function. PLoS Biology **6**(7) (July 2008) e175
13. Bertram, L., McQueen, M.B., Mullin, K., Blacker, D., Tanzi, R.E.: Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. Nature Genetics **39**(1) (January 2007) 17–23