Diple, modular methodology and tools for heterogeneous TEI corpora

Glorieux, Frédéric
frederic.glorieux@enc.sorbonne.fr
École nationale des chartes

Canteaut, Olivier
olivier.canteaut@enc.sorbonne.fr
École nationale des chartes

Jolivet, Vincent
vincent.jolivet@enc.sorbonne.fr
École nationale des chartes

The *École nationale des chartes* publishes a variety of electronic corpora,[1] focused on historical sources (medieval, but also, modern and contemporary). A dictionary,[2] a collection of acts,[3] or a manuscript[4] are very different types of documents, each requiring different structures and interfaces. A narrative manuscript needs a table of contents, a dictionary, fast access to headwords, and acts, the ability to sort by dates. Each editorial project should allow customization, but efficient development requires that the tools and corpora are as normalized as possible. New needs emerge, such as natural language processing research, requiring large corpora with normalized metadata sets and word tagging. For several months we have been working on a platform to address these needs: *Diple* is a collection of tools to organize modular production, publication, and searching of electronic corpora.

1. Modular schemas

The TEI guidelines (500 XML elements, 1500 pages of documentation) allow endless variations in encoding, even for identical objects. For example, *italic* in our corpora has been encoded with different combinations of <(hi|emph) rend="(italique|italic|ital.|i|itlaic|...)">. After several years of development, with different encoders, each electronic edition becomes an independent software, with its own encoding, mistakes, workarounds, with also different technologies for publication or fulltext searching.

Diple starts with housekeeping. First, for all our tagged texts, we wrote a precise *document type definition* (in Relax-NG syntax) in order to define three main and shared schemas :

- file metadatas (<teiHeader>)
- general text (blocks and inlines)
- structure for a specific type of text (ex: acts and charters, dictionaries)

The normalization of the corpora is a more sustainable investment than new software. These shared schemas are extremely helpful for normalizing and validating XML instances, and therefore allow us to take advantage of earlier TEI editions. Of course, the *Diple* TEI schema is modular, allowing customization for each editorial project.[5] An editor can then focus on the specificities of each edition. Are named entities sufficiently tagged to generate automatic indexes? Are the sentences chunked, the words lemmatized?

Moreover, this work of normalization of our XML corpora is a small price to pay to factorize our code, for instance to create a standard XSLT engine: the screen transformation of a new corpus conforming to those schemas is done by this engine, increasing our publication productivity. In the end, the XSLT of a specific edition is short, focusing on the very specific aspects of the corpus (its custom schema) related to a research project, the main part of the publication job being done by the *Diple* XSLT engine. The same logic applied for presentation CSS.

2. Shared interface components, documents driven

A publication system usually allows templating and plugins. A good software architecture should be conceived in this way, but scholarly editions don't function like a CMS. Templating systems are usually designed to effect easy change of colours, to deliver the same feature under different designs. In a scholarly collection, books could share a cover, but follow very different structures. Rather than constrain all corpora to a single template, the *Diple* system provides different components, allowing

an electronic corpus editor to compose the interface best suited to his text. Headers or footers are easy to share, but beyond that, one project might require a fulltext search box, another a database query, another a sidebar table of contents. Design snippets or plugins are conceived of as portal bricks, easy to compose in a server page (PHP), and are kept as simple and light as possible. If a local variable, function or object could have a general interest, it should be shared.

3. Text engines for research, retrieving and concordances

Navigation, tables and indexes, should answer most of the user's needs; but a search box is also an important navigation tool. *Diple* ensures a canonical electronic publication, with persistent addresses, so that different text engines can be plugged around the edition. Corpora may require different approaches. A collection of items, like a dictionary or cartularies, needs at first a retrieving engine to get an item conveniently, by a keyword in full-text, but also dates, headwords and other metadatas. There are also texts for which no divisions are relevant; a concordance report is much more informative, displaying all occurrences in context. Different tools offer different views, documented XML allows us to generate what an engine likes. We have successfully used MySQL full-text indexes[6] for navigation interfaces, PhiloLogic[7] for concordances, Lucene[8] is very efficient to retrieve items, and we learned to use IMS Corpus Workbench (CWB)[9] for future lemmatized corpora. But sometimes we also simply use mixed scripts (shell, XSLT, SAX...) to run a specific experiment on a word or a semantic field.

4. Conclusion

Diple grows and adapts with each new corpus, rapidly incorporating other corpora, an idea worth generalizing. All our code will soon be released under a free software license. Anyone can download, read, and try *Diple*. We don't claim it will work for all your TEI documents, but if they conform to the schemas, you will quickly get nice results on the screen.

References

Bourgain, Pascale, Vieillard, Françoise (coord.) (2002). *Conseils pour l'édition des textes médiévaux.* Fascicule III. Textes littéraires. Paris: Éd. du CTHS, École des chartes.

Guyotjeannin, Olivier, Vieillard, Françoise (coord.) (2001). *Conseils pour l'édition des textes médiévaux.* 'Conseils généraux'. Paris: Éd. du CTHS, École des chartes.

Olivier Guyotjeannin (coord.) (2001). *Conseils pour l'édition des textes médiévaux.* 'Actes et documents d'archives'. Paris: Éd. du CTHS, École des chartes.

McCandless, Michael, Hatcher, Erik, Gospodnetić, Otis (2010). *Lucene in action.* Manning Publications Co., 2e ed.. `http://www.manning.com/hatcher3/`.

TEI Consortium (ed.). *TEI P5: Guidelines for Electronic Text Encoding and Interchange.* `http://www.tei-c.org/release/doc/tei-p5-doc/fr/html/REF-ELEMENTS.html`.

Wooldridge, Russon (1997). *Les Débuts de la lexicographie française.* Toronto: EDICTA2e éd.. `http://www.chass.utoronto.ca/~wulfric/edicta/wooldridge/`.

Notes

1. `http://elec.enc.sorbonne.fr/`
2. `http://ducange.enc.sorbonne.fr/`
3. `http://elec.enc.sorbonne.fr/cartulaires/`
4. `http://elec.enc.sorbonne.fr/sanctoral/`
5. For example, `http://elec.enc.sorbonne.fr/cartulairesIdF/src/schema.htm`
6. `http://dev.mysql.com/doc/refman/5.0/en/fulltext-search.html`
7. `http://philologic.uchicago.edu/`
8. `http://lucene.apache.org/java/`
9. `http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/`