Text bank software

The text bank coming to life

Jørg Asmussen, DSL, with input from other members of WP 2
DK-CLARIN WP 2.1+2 Technical Report. **WP 2.1 Deliverable 3 (T 12)**
Final version of June 15, 2011[1]

Document history

27-OCT-2009:

- After having successfully passed the experimental phase, eXist will be used as textbank software in WP 2.1. Thus, certain passages of this report needed be revised accordingly.
- An obsolete section on using oXygen on eXist has been deleted. Use of the textbank software will be described in a future manual.

1 Outline

[1] The most recent version can be downloaded from:
http://korpus.dsl.dk/clarin/corpus-doc/textbank-software.pdf.

This technical report gives an account of which textbank software to choose. The software should be able to store and give multi-user access to XML documents and cooperate with an editor. This report also serves as documentation for WP 1.10 and 1.21 as described in **?**.

2 Introduction

Crucial for composing corpora according to specific characteristics is a repository of texts which may be included in a corpus. This repository is intended to contain corpus-suitable texts with certain textual metadata expressed in a systematic, XML-formatted way, see **?**. A repository of this kind is called a *textbank*, see **?**.

In the following, the choice of software to realise such a text bank for (some of) the corpus workpackages under DK-CLARIN, called the *Clarin Text Bank*, CTB, is described in further detail.

3 XML database systems

The fundamental units of a corpus are texts (or text fragments) drawn from a text repository. A repository containing potential corpus texts in a common text and metadata format only is called a *textbank*. As corpus texts to be included in the text bank are XML-formatted according to a specific schema, see **?** and **?**, it seems obvious to store them in an XML database. Moreover, it is necessary to be able to access and edit these fundamental text units with a viewer/editor.

Storing the text material in a relational database would require substantial processesing of the material prior to database import and export. As such, conversion procedures introduce an extra amount of resources and complexity, they must be considered error-prone in terms of appliance and maintenance. Therefore, it appears obvious to apply an XML-based db solution. An overview of XML databases can be found on Wikipedia.

To minimise software expenditures, only open source products have been examined. Among these, projects with low or none obvious development activities were rejected, that is Xindice, myXMLDB, ozone. Sedna could be a candidate, but it is obviously not supported by the oXygen XML editor, the same seems to be the case for MonetDB and BaseX. After these exclusions the only candidate left is eXist.

4 eXist – the text bank system by choice

4.1 Advantages

- Native XML db
- Easy to install and maintain
- Built-in indexing (automatic and user-defined) which means quick searches
- XQuery is used to manipulate data
- Theoretically unlimited document size. So far, the 13,5 MB DK-PAROLE Corpus has been the largest document uploaded
- Can store up to 2^{31} documents
- Accessible from the oXygen editor
- Easy to set up as a web service. This should make it straightforward to
  - have the eXist-based text bank to fit into the envisaged DK-CLARIN infrastructure
  - develop a stand-alone text bank interface and skip the oXygen editor

4.2 Disadvantages

So far, the following disadvantage could be identified:

Non-commercial project: As for most software development projects of this kind, there is no guarantee for continuous development as well as support may be unreliable.

This disadvantage should be taken into account for future development. In particular, web services probably should not be based entirely on eXist's XQuery interface but should be encapsulated by a self-developed web service interface that accesses eXist but could be changed to access other db

4.3 Current installation

The CTB is located in the eXist XML document collection `/db/ctb` as data repository and an oXygen editor as a basic user interface which communicates with the data repository by means of a oXygen-eXist db connection. The oXygen editor will be replaced by a dedicated viewer/editor during the project period. The eXist service is installed on the host `ja-korpus.dsl.lan` until a more convenient solution is available.

Bibliography