The text bank

A platform for managing text collections
DRAFT VERSION

DK-CLARIN WP 2.1 Technical Report
Jørg Asmussen, DSL, with input from other WP 2 mebers
Final version of March 5, 2014[1]

Deliverables concerned

D1 Text registry DSL as well as DSN collect Infomedia text material, parts of which are likely to be included in the WP 2.1 corpus. Therefore, a way of registering texts needs to be established. A registry allows tracing and eliminating possible duplicate texts. The text registry functionality is part of the CMRS. **Outcome:** Report.

D3 Decision on text bank system A text bank system is necessary for project-internal text administration. Investigations of different approaches to such a system will be carried out. Two general options seem viable – either one based on a relational db or on an XML db. The text bank system is the core component of the CMRS. **Outcome:** Report.

D4 Text supplier registry A registry of active and potential text suppliers needs to be designed as an integrated component of the CMRS. **Outcome:** Report.

D5 Implementation of text bank system The chosen text bank approach (see D3) implemented (possibly with a GUI) as component of the CMRS. **Outcome:** Report and a project-internal service.

[1] A more recent version may be available at:
http://korpus.dsl.dk/clarin/corpus-doc/textbank.pdf.

Outline of this document

This technical report gives an account of the text bank, that is, of its intended functions as text repository and administrative registry as well as its implementation. The db software should be able to store and give multi-user access to XML text documents, manage text supplier data, and facilitate the detection of duplicate text material. This report also serves as documentation for WP 1.10 and 1.21 as described in Asmussen (2008).

1 Introduction

Crucial for composing corpora is a repository of texts from which texts according to a specific profile can be selected. The corpus itself is considered a separate resource. This repository is intended to contain corpus-suitable texts with certain textual metadata expressed in a systematic way that can be expressed as XML, see Asmussen et al. (2013). A repository of this kind is called a *text bank*.[2] A text bank may provide some administrative functions as well like handling text supplier information.

In the following, the implementation of the text bank is described in further detail. This is followed by a description of the text registry and the text supplier registry functionality, see Figure **??**.

2 Implementation

2.1 XML vs. relational db systems

The fundamental units of a corpus are *text items* drawn from a text repository. A repository containing potential corpus texts in a standardized text and metadata format only is called a *text bank*. As corpus texts to be included in the text bank

[2] See also Asmussen (2013a).

are XML-formatted according to a specific schema, see Asmussen et al. (2013) and Asmussen (2013b), 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 processing 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 minimize 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.

2.2 eXist – the text bank system by choice

2.2.1 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

2.2.2 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

2.2.3 Current implementation and set-up

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.

2.2.4 User Interfaces

3 Features

3.1 Text repository

3.2 Text registry

3.3 Text supplier registry

4 Alternative approaches

Bibliography

Asmussen, J. (2008). DOT's Sprogteknologiske Drejebog. Udviklingsopgaver i forbindelse med *ordnet*-projektet. Technical report, Det Danske Sprog- og Litteraturselskab, ja-korpus.dsl.lan/doc/drejebogen.pdf.

Asmussen, J. (2013a). Aim and concepts. Technical report, DK-CLARIN, korpus.dsl.dk/clarin/corpus-doc/concepts.pdf.

Asmussen, J. (2013b). Text formatting. Technical report, DK-CLARIN, korpus.dsl.dk/clarin/corpus-doc/text-format.pdf.

Asmussen, J. et al. (2013). Text metadata. Technical report, DK-CLARIN, korpus.dsl.dk/clarin/corpus-doc/text-header.pdf.