Reusing CMDI components for a text corpus profile - towards a generic text corpus profile

Lene Offersgaard, Dorte Haltrup Hansen

University of Copenhagen
leneo@hum.ku.dk, dorte@hum.ku.dk

CAC2014
Overview of the talk:

1. Background
2. Reuse of CMDI profiles and components
3. A generic text corpus profile
4. A minimal set of obligatory metadata
5. A minimal text corpus profile
6. Wrapping up
Background

- Preparatory phase in CLARIN-DK: deposited text resources in the repository
- Text resources have metadata in CMDI-TEI
- Now organizing the TEI resources in collections: text corpora
- First step is to find a metadata scheme to use

- We would like to re-use a CMDI schema or components if possible:
  - Others already spend time and effort on creating profiles
  - It is tedious to do it from scratch
  - Collaboration and re-use of data easier when using same metadata description
Reuse of CMDI profiles and components I

- We have inspected and used existing profiles in the Component Registry
- Currently 162 public profiles
- Difficult to get an overview of existing CMDI corpus profiles: E.g. from CLARIN-D, CLARIN-NL, META-SHARE
- Naming of profiles are free
- Search function, but difficult to guess the names, e.g. resourceInfo, textCorpusProfile,
- Difficult to inspect and compare profiles -> use the SMC Browser
Reuse of CMDI profiles and components II

http://clarin.oeaw.ac.at/exist/apps/smc-browser/index.html

E.g. resourceInfo profile: 419 components
Reuse of CMDI profiles and components III

- Difficult to compare profiles in the Component Registry
- Even with the SMC Browser: only compare on component-id’s
- A component that is changed in any (small) way gets a new id

- Nice if in future possible to:
  - Link between components that are related
  - Versioning of components
A generic text corpus profile I

- Granularity: how fine grained?
- We promote having a large generic well-structured profile with loose bounds:

  - **Large**: options to specify a wealth of information for a text corpus
    - Different kinds of text corpora
    - Different characteristics

  - **Loose bounds**: 
    - Different interest in metadata creation

  - **Well-structured**: three top level nodes
A generic text corpus profile II

- A restructure of the META-SHARE resourceInfo: wide and deep!
- Very different on the surface but with 255 of 274 overlapping elements with the same ISOcat DC definitions
- NaLiDa component chosen for Documentation
A generic text corpus profile III

- **Structure**: easy to choose where to add new information
- **Large**: easy to take a copy of profile and delete what is not wanted
- **Flexible**: only a few obligatory elements
**DK-CLARIN Fagsprogligt Korpus - Byggeri og Anlæg**

**Title:** DK-CLARIN Fagsprogligt Korpus - Byggeri og Anlæg

**Title:** DK-CLARIN LSP Corpus - Construction domain

**Language:** da

**Resource Type:** text corpus

**Resource Identifier:** hdl:11221/3410-8400-0001-D

**Linguality Type:** monolingual

**Creator:** University of Copenhagen

**Creator:** The Danish Language Council

**Publication Date:** 2011

**Description:**


**Description:**

"DK-CLARIN LSP Corpus - Construction domain" is a part of the Danish DK-CLARIN LSP corpus consisting of seven sub-corpora from following subject domains: Agriculture, Construction, Economics, Environment, Health, IT and Nanotechnology. Texts in the Construction Domain come from Statens Byggeforskningsinstitut, Erhvervs- og byggestyrelsen and Murefagets Oplysningsråd and have been collected in the DK-CLARIN project, WP2.2, 2008 - 2011. All texts are in XML TEI/P5 format (TEIP5DKCLARIN-format), with tokenisation, sentence and paragraph segmentation, pos-tagging, lemmatisation and termhood annotation placed in separate text external spangroups. The corpus consists of 577,392 words in 35 files. Communicative setting/Number of files: expert->expert (18) expert->advanced (6) expert->basic (11)

**Format:** text/xml

**License:** CLARIN_ACA-NC

**Availability:** academic

**Sponsor:** DK-CLARIN

**Time Coverage:** 2000-2010

**Geographic Coverage:** Denmark

**Text Type:** Technical/professional reports

---

An example:

CAC2014 - Soesterberg, NL - 25 October 2014
A minimal set of obligatory metadata for a profile I

- Could CMDI / CLARIN ERIC require a core set of metadata?
- NOT implemented as a mandatory component
  - Allowable to use already defined standards, not just OLAC
- NOT implemented as ISOcat references
  - ISOcat references cover e.g. a number of titles
- => list of Xpaths (now partly required by VLO)
- Could they be stored in the Component Registry?
  - for investigation of profiles
  - for other aggregating search facilities
- Could OLAC be used? Not all elements optimal as mandatory
A minimal set of obligatory metadata II

- We suggest a core set of metadata consisting of 9 elements:
  1. Title
  2. Description
  3. Identifier
  4. Type
  5. Creator
  6. Date
  7. Rights
  8. Language
  9. Format

- From the core OLAC metadata set the following are left out: subject, coverage, publisher, relation, source, and contributor

- We suggest correspondence with VLO facets

- Consider to have semi-closed pick lists, that can be versioned
A minimal text corpus profile

- The implemented textCorpusProfile contains the suggested core elements

- In addition another 4 metadata elements is mandatory:
  1. Size
  2. ResourceShortName
  3. Linguality
  4. ProjectName (OLAC contributor)

- Resulting in a generic text corpus profile with total of 274 elements and a minimum of 13 elements
## Obligatory elements

<table>
<thead>
<tr>
<th>textCorpusProfile</th>
<th>OLAC</th>
</tr>
</thead>
<tbody>
<tr>
<td>textCorpusProfile</td>
<td></td>
</tr>
<tr>
<td>1 generalInfo</td>
<td></td>
</tr>
<tr>
<td>1.1 identificationInfo</td>
<td></td>
</tr>
<tr>
<td>resourceName</td>
<td>Title</td>
</tr>
<tr>
<td>resourceShortName</td>
<td>Dcterms:alternative</td>
</tr>
<tr>
<td>resourceDescription</td>
<td>Description</td>
</tr>
<tr>
<td>resourceType</td>
<td>Type</td>
</tr>
<tr>
<td>resourceIdentifier</td>
<td>Identifier</td>
</tr>
<tr>
<td>1.2 resourceCreatorInfo</td>
<td></td>
</tr>
<tr>
<td>creationEndDate</td>
<td>Date</td>
</tr>
<tr>
<td>resourceCreator</td>
<td>Creator</td>
</tr>
<tr>
<td>organizationInfo</td>
<td>Contributor</td>
</tr>
<tr>
<td>organizationName</td>
<td></td>
</tr>
<tr>
<td>fundingProject</td>
<td>Rights</td>
</tr>
<tr>
<td>projectInfo</td>
<td>Rights</td>
</tr>
<tr>
<td>projectName</td>
<td></td>
</tr>
<tr>
<td>1.3 distributionInfo</td>
<td></td>
</tr>
<tr>
<td>availability</td>
<td>Language</td>
</tr>
<tr>
<td>licenceInfo</td>
<td>Format, dcterms:extent</td>
</tr>
<tr>
<td>licence</td>
<td></td>
</tr>
<tr>
<td>2 generalCorpusInfo</td>
<td></td>
</tr>
<tr>
<td>2.1 lingualityInfo</td>
<td></td>
</tr>
<tr>
<td>lingualityType</td>
<td></td>
</tr>
<tr>
<td>2.2 languageInfo</td>
<td></td>
</tr>
<tr>
<td>languageId</td>
<td></td>
</tr>
<tr>
<td>2.3 sizeInfo</td>
<td></td>
</tr>
<tr>
<td>Size, sizeUnit</td>
<td></td>
</tr>
<tr>
<td>3 textCorpusInfo</td>
<td></td>
</tr>
<tr>
<td>3.2 textTechnicalInfo</td>
<td></td>
</tr>
<tr>
<td>textFormatInfo</td>
<td>Format</td>
</tr>
<tr>
<td>mimeType</td>
<td></td>
</tr>
</tbody>
</table>
Wrapping up

- **Issues:**
  - Finding the right profile to use in the Component Registry
  - Versioning in Component Registry important

- **Suggestions:**
  - Having a core set of mandatory metadata elements
  - Use part of the core OLAC metadata set
  - Store these as XPaths
  - Expand the Components Registry to enable storing of the Xpath mappings