CLARIN i Danmark
Common Language Resources and Technology Infrastructure

En dansk forsknings-infrastruktur

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

CLARIN i Danmark
Common Language Resources and Technology Infrastructure

En dansk forsknings-infrastruktur

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

CLARIN i Danmark
Common Language Resources and Technology Infrastructure

En dansk forsknings-infrastruktur

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

CLARIN i Danmark
Common Language Resources and Technology Infrastructure

En dansk forsknings-infrastruktur

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

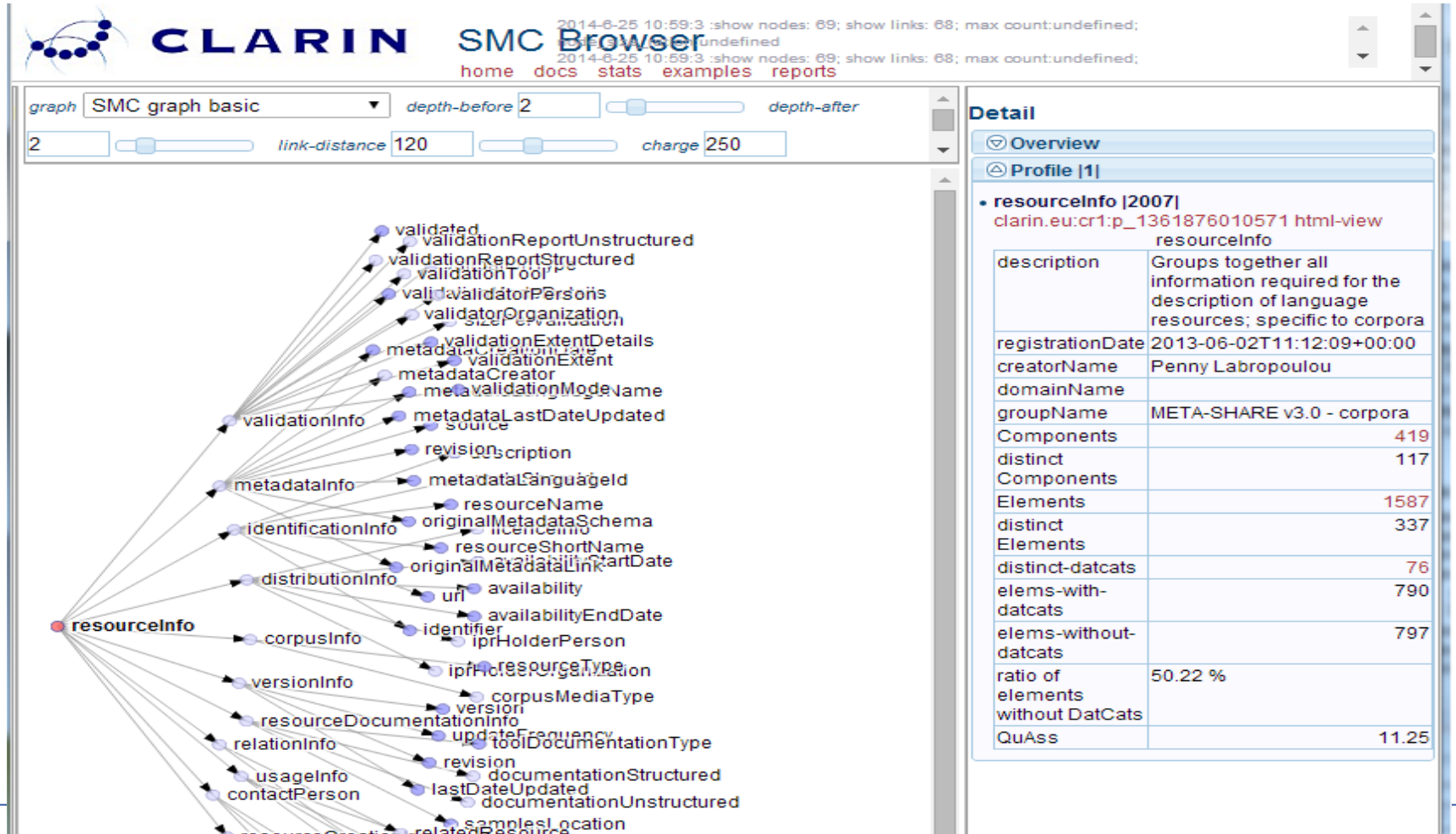
CLARIN i Danmark
Common Language Resources and Technology Infrastructure

En dansk forsknings-infrastruktur

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

CLARIN i Danmark
Common Language Resources and Technology Infrastructure

En dansk forsknings-infrastruktur

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

CLARIN i Danmark
Common Language Resources and Technology Infrastructure

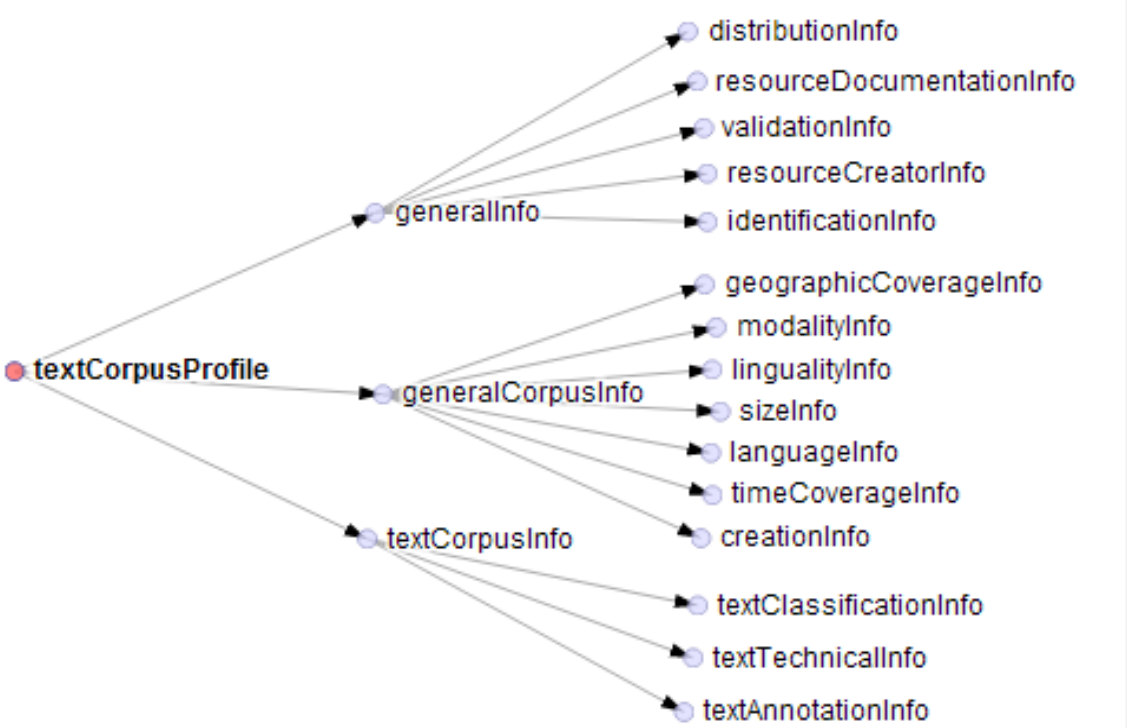
En dansk forsknings-infrastruktur

CLARIN SMC Browser

2014-6-25 10:47:54 :show nodes: 19; show links: 18; max count:undefined; node_size_ration:undefined
2014-6-25 10:47:54 :show nodes: 19; show links: 18; max count:undefined; node_size_ration:undefined
2014-6-25 10:47:54 :count max: 9609; node_size_ratio: 1.2253188360585991

home docs stats examples reports

graph SMC graph basic depth-before 2 depth-after 2 link-distance 120 charge 250



Detail

Overview

Profile |1|

- **textCorpusProfile |378|**
  clarin.eu:cr1:p_1386164908461 html-view

textCorpusProfile

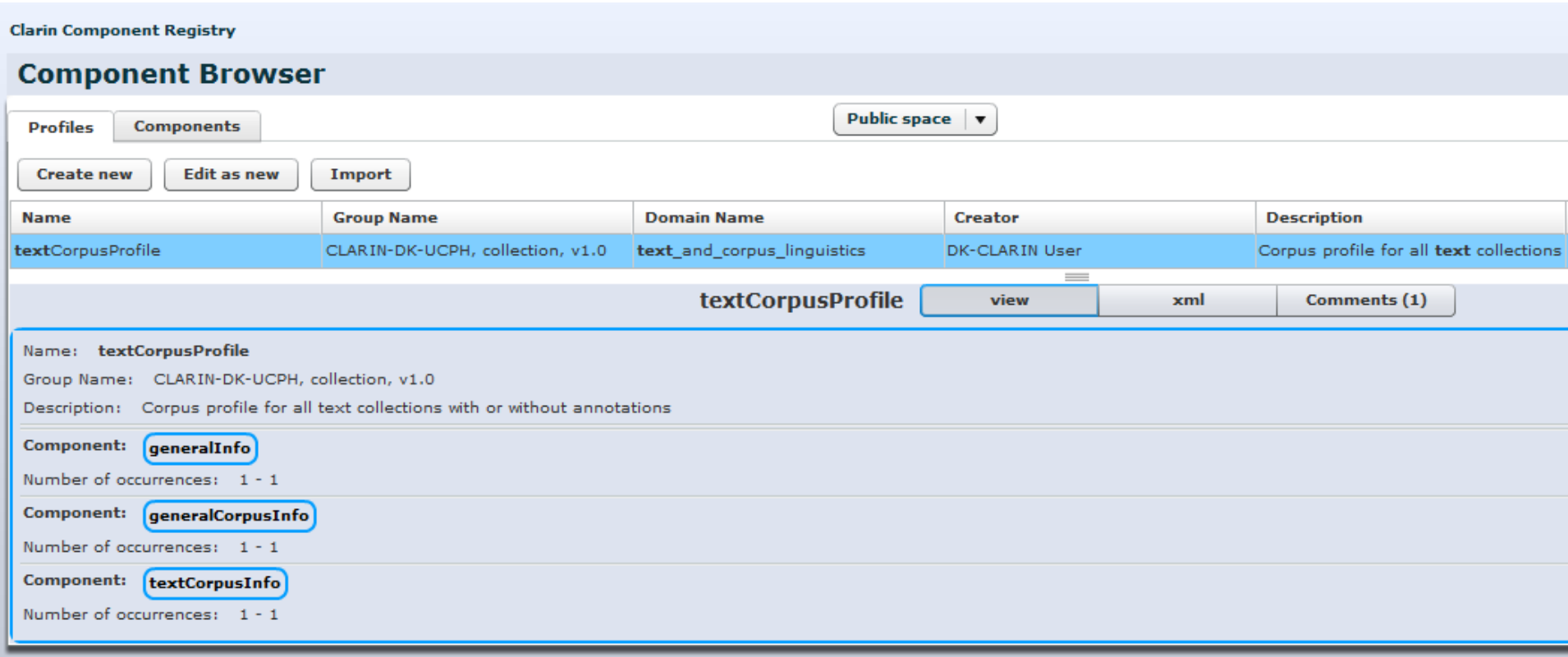
description	
registrationDate	
creatorName	
domainName	
groupName	unknown-profile
Components	103
distinct Components	59
Elements	274
distinct Elements	119
distinct-datcats	53
elems-with-datcats	194
elems-without-datcats	80
ratio of elements without DatCats	29.20 %
QuAss	3.99

CLARIN i Danmark
Common Language Resources and Technology Infrastructure

En dansk forsknings-infrastruktur

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



CLARIN - DK
Common Language Resources and Technology Infrastructure

En forsknings-infrastruktur for humanister

Ressource ID

DK-CLARIN Fagsprogligt Korpus - Byggeri og Anlæg

Detaljer | CMDI metadata | Elementer i samlingen | Download | Tilføj til kurv

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:
"Fagsprogligt Korpus - byggeri og anlæg" er en del af "DK-CLARIN Fagsprogligt Korpus" som dækker 7 fagområder: Byggeri og Anlæg, IT, Landbrug, Klima og Miljø, Nanoteknologi, Sundhed og Medicin samt Økonomi. Teksterne i domænet for byggeri og anlæg er indsamlet fra Statens Byggeforskningsinstitut, Erhvervs- og Byggestyrelsen og Murerfagets Oplysningsråd i projektet DK-CLARIN arbejdspakke WP2.2, 2008-2011. Teksterne er i XML TEIP5-format (TEIP5DKCLARIN-format), suppleret med selvstændige annoteringer der dækker tokenisering, sætnings- og paragrafsegmentering, pos-tagging, lemmatisering og termstatus. Korpusset indeholder 577.392 ord fordelt på 35 filer. Kommunikationstype (antal filer): expert->expert (18) expert->advanced (6) expert->basic (11)
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 Murerfagets Oplysningsråd and have been collected in the DK-CLARIN project, WP2.2, 2008 - 2011. All texts are in XML TEIP5 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

CLARIN i Danmark
Common Language Resources and Technology Infrastructure

En dansk forsknings-infrastruktur

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

CLARIN i Danmark
Common Language Resources and Technology Infrastructure

En dansk forsknings-infrastruktur

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

CLARIN i Danmark
Common Language Resources and Technology Infrastructure

Obligatory elements

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

CLARIN i Danmark
Common Language Resources and Technology Infrastructure

En dansk forsknings-infrastruktur

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