Acknowledgments

• Work started in 2012, often interrupted...

• Many people contributed:
  – Eline Westerhout, Rogier Kraf, Erica Renckens
  – Software developers, Centre managers
  – Excellent support by Menzo Windhouwer and Twan Goosen!
  – Daan Broeder created the faceted search!
Overview

• Introduction
• Profile CLARINSoftwareDescription
• Metadata Descriptions
• Faceted Search
• Software Metadata Curation
• Concluding Remarks
• Future Work
• Recommendations
Introduction

• Discovery of Resources
  – Virtual Language Observatory (VLO)
  – Focus on description and discovery of data
  – Software not so easy to discover (SoftwareQuery)

• Here:
  – CMDI profile for software descriptions
  – Faceted Search definition
  – Strategy for curating existing software descriptions
  – + Illustration
• CLARINSoftwareDescription (CSD)
  – Properties to support discovery
  – Properties for formal documentation
• Components:
• Closed (half-open) Vocabularies wherever possible
Components

• GeneralInfo
  – Extension of cmdi-generalinfo
  – CLARIN Centre, National Project(s)

• SoftwareFunction
  – Tool category, toolTasks, research phases, research domains, linguistic subdisciplines

• SoftwareImplementation
  – (user) interface, input, output
Components

• SoftwareImplementation/Input, Output:
  – characterEncoding (UTF8), inputType (text), inputResource (lexicon), Schema (FoLiA), MimeType (text/xml), AnnotationFormat (i.a. tagset: D-COI), AnnotationType (Morphosyntax/POS), Language (nld), Description (free text)
Components

• Data ➔ Tool ➔ Modified Data
• Data + Metadata ➔ Tool ➔ Modified Data + Modified Metadata
• CSD enables description of modifications to be made to the metadata
• E.g. Alpino output
Components

• Access
  – Availability, accessibility, license, ...

• ResourceDocumentation
  – Documentation, publications, pictures, demo scenario’s, faqs, ...

• SoftwareDevelopment
  – Development project(s)

• TechnicalInfo
  – Varied technical information
Components

- **Service**
  - For the description of web services
  - Compatible with CLARIN CMDI Core model 1.0.2

- **LRS**
  - Description of the properties of a particular task for the CLARIN Language Resource SwitchBoard (CLRS)
  - Script to generate JSON for the CLRS Registry
  - Tested successfully with Frog web service
• Components, elements, possible values
  – Many have explicit semantics (in the CCR)
  – Definitions etc. created for others
  – Submitted on 2017-09-08 to the CCR Coordinators
  – None included as yet ..... 
    • One problem: the horrors of natural language
  – New ones introduced for which no definitions exist yet
Other Profiles

• 20 profiles for software in total
• Many not in use
• Major ones in use (2017-09-29/2018-10-01):
  – ToolProfile (49/32)
  – WebLichtWebService (287/410)
  – resourceInfo[resourceType=toolService] (68/94)
  – OLAC-DCMITerms (189/257)
  – LINDAT_CLARIN (??/34)
• SoftwareQuery yields 1075 results (2018-10-01)
82 CMDI descriptions
   – From NL & Flanders

Quality:
   – All validated
   – Multiple Schematron files for testing quality
   – Check on the resolution of URLs (2018-10-01):
     • 969 URLs correctly found
     • 11 URLs found but no access granted
     • 35 URLs not found
     • 4 exceptions raised
   – CLARIN Curation Module less useful
Faceted Search

• Dedicated faceted search for software
• Search Facets:
  – LifeCycleStatus, ResearchPhase, toolTask, researchDomain, linguisticsSubject, inputLanguage, applicationType, NationalProject, CLARINCentre, input modality, licence
• Display Facets:
  – name, title, version, inputMimetype, outputMimetype, outputLanguage, Country, Description, ResourceProxy, AccessContact, ProjectContact, CreatorContact, Documentation, Publication, sourcecodeURI, Project, CMDIFileLink, OriginalLocation
  – **Bold: half-open vocabulary; italics: still to do**
Faceted Search

• For the 82 NL+FL resources: [http://portal.clarin.nl/clariah-tools-fs](http://portal.clarin.nl/clariah-tools-fs)
• For NL+FL + (partially) curated versions of 286 WebLichtWebService descriptions: [http://portal.clarin.nl/clariah-tools-fs-global](http://portal.clarin.nl/clariah-tools-fs-global)
• Try it out yourself!
• I will demo it this afternoon
• Strategy of Odijk 2015
• create new metadata record
  – based on the original metadata and a curation file
  – in accordance with the required facets
  – Automatically, (each time upon harvesting)

• Curation file
  – Adds lacking information
  – Maps existing information
    • From: combinations of field values
    • To: other combinations of field values in the right vocabularies
Concluding Remarks

• Let’s make this faceted search (or a variant)
  – Part of the VLO
  – As a separate faceted search for software
• Let’s start a concerted effort to curate the existing metadata files for software
• Let’s issue clear guidelines / requirements
  – for making new metadata records for software
  – And let’s test them (automatically)!
    • Cf. https://curate.acdh.oeaw.ac.at/#!Instances
    • https://cmdi.clarin.eu/mapping/index.html
Future Work

• Publishing the profile
• Finalising the semantics
• Document the profile (in progress)
• Better facilities for parameters
• Enable paired <parameters, input, output>
• Harmonize details (e.g. naming conventions)
• Remove some redundancies
• Derive information automatically from CLAM, codemeta, WADL descriptions, ...
Recommendations

• Coordinate metadata creation nationally
• Define a minimum set of metadata elements (defined semantically)
  – For faceted search
  – For a minimal proper description
    (and test for compliance, as automated as possible)
• Use closed (‘half-open’) vocabularies whenever possible
  – But be prepared to update it regularly
  – And to upgrade it occasionally
• Avoid the horrors of natural language
Recommendations

• Enumerated lists:
  – Should not be defined with an element
  – Should be reusable with multiple elements
  – Viewing and copying the values should be possible in the Component Registry without having to edit!

• Real need for a good CMDI editor
  – Not web-based!
  – Enabling editing of multiple files at once
  – I can provide more requirements
Thanks for your attention
• Words have a (common-sense) meaning
• Words are often ambiguous / polysemous
• Words are too redundant (➔ abbreviations, acronyms)
• Words have synonyms
• Words are specific to a language

➔ use codes that are non-words instead!
(cf. the ISO language codes)
• **outputType** text  
• **characterEncoding** utf8  
• **Schema** LASSY DTD  
• **MimeType** text/xml  
• **AnnotationType**  
  — **AnnotationType** Morphosyntax/Inflection  
  — **AnnotationType** Morphosyntax/Lemma  
  — **AnnotationType** Morphosyntax/POS  
  — **AnnotationType** Morphosyntax/Word form  
  — **AnnotationType** Orthography/Token  
  — **TagSet** POSTags/DCOI Tagset  
• **AnnotationType**  
  — **AnnotationType** Syntax/Chunks  
  — **AnnotationType** Syntax/Dependency Relations  
  — **AnnotationType** Syntax/Grammatical Relations  
  — **AnnotationType** Syntax/Multiword Expressions  
  — **TagSet** Syntax/Alpino Tagset