ClarinSoftwareDescription profile Documentation

Jan Odijk

1 Introduction

We present the CMDI profile ClarinSoftwareDescription. We describe the components and elements that it consists of (section 2) and its semantics (section 3). We show that the profile has been used to describe some 70 pieces of software, mostly from the Netherlands (section 4). We have validated these descriptions in various ways and set up schematronfile specific for this profile (section 5).

We describe the link between this profile and the CLARIN Language Resource Switchboard (section 6). We discuss other profiles for software currently available and show the most important differneces with the ClarinSoftwareDescription profile (section 7). We propose a specific set of facets to define a faceted search application for software and we specify how these facets can be derived from the CMDI records. As we will see, many CMDI records for software require curation, in particular the addition of some properties that are crucial for deriving the values for facets. We propose a strategy for this curation, and some specific curation proposals for CMDI records describing software (section 8). We end with concluding remarks in section 10.

2 ClarinSoftwareDescription profile

The ClarinSoftwareDescription profile enables one to describe information about software in a CMDI-compatible manner. A metadata record is a description of a resource. Metadata can be set up in many different ways, and with many different purposes. Our approach has been as follows:

- we want to describe properties that support finding the relevant resource
- we want to describe properties to document the resources
- we want to describe properties wherever this is feasible in a formal manner, i.e by strict typing and using closed or half-open vocabularies as much as possible.

Though the use of natural language cannot be avoided when describing resources, ideally natural language is avoided wherever it can be . Natural language expressions for concepts come with serious problems, in particular, they

are not unique, they are language-dependent, often ambiguous, and are over-redundant. In addition, there is the property of mnemonically usefulness, which is both a blessing and a curse.

Ideally, closed or half-open vocabularies avoid all the problems of natural language, by sticking to the following principles:

- a vocabulary in this context is a set of strings that act as labels for concepts. For each concept there is preferably only one label (uniqueness,no synonyms), and each label is associated only with one concept (no ambiguity).

- the labels are as short as possible and have no or a minimum of redundancy.

- the labels are selected on the basis of English terms (for mnemonic reasons) but are not English terms (to avoid the curse of mnemonically useful terms, i.e. that opening up their interpretations to all interpretations English assigns to these strings)

We must admit that these principles have not been fully adhered to here, and improvements in this respect can still be made.

We describe each of the components and elements of this profile in a separate subsection.

2.1 Generalinfo

This component describes general information about the resource. It is an extension of the component *cmdi-generalinfo*.[1]

The general information includes a *name* and a *title* for the resource, its *version*, *publication year*, *owner* and contact *email address*, a *url* and/or a *PID* to the resource, its *time coverage*, *release status*, any *alternative names* for the resource and a *description* of the software.

2.2 SoftwareFunction

This component describes the function of the software in terms of a *tool category, tool tasks*, relevant *research domains* and for linguistics, relevant *linguistic subdisciplines* for which it was originally developed.[2] It also describes the *language variants* that the software applies to and provides documentation about its *performance*.

[1] clarin.eu:cr1:c_1342181139620.

[2] which, of course, does not preclude its use in other research domains that were not foreseen during development.

2.3 SoftwareImplementation

Information for users on the implementation and installation of a software program. It describes how the software is *distributed*, what the *installation requirements* are, the nature of the *interface* with the user or other software, properties of the *package* that the software is delivered in (if any) and what the *input* and the *output* of the software is.

2.4 Access

This component contains information about the availability and accessibility of a resource. It is an extension of the *cmdi-access* component.[3]

It contains a reference to a *catalogue link*, information about the *license* for the resource, information about *copyright* and *copyright holder(s)* and a *contact* organisation and/or person.

2.5 ResourceDocumentation

This component describes the documentation on a resource. It offers facilities to describe the *documentation*, *publications* on the resource, a *description* of the resource and *pictures* (including *logos*) related to the resource.

2.6 SoftwareDevelopment

This component is intended for information on the history and development of the software. It offers facilities to describe the *source*(s) that the software is based on or from which it has been derived, the *project* in which the software was created, the *creator* of the software, and any planned *software updates*.

2.7 TechnicalInfo

This component describes technical information on a resource and is mainly aimed at developers. It provides facilities to describe the *run time environment*, any *access protocols*, as well as the *programming language(s)* that have been used to implement the software.

2.8 Service

This component (CLARIN-NL Web Service description) is intended for describing properties of web services. It is compatible with the CLARIN CMDI core model for Web Service description version 1.0.2.[4]

[3] clarin.eu:cr1:c_1311927752326.

[4] This component was created by Menzo Windhouwer, and adapted to the requirements of CMDI version 1.2.

2.9 LRS

Description of the properties of a particular task for the CLARIN Language Resource SwitchBoard. Multiple LRS components can be present. See section 6.

3 Semantics

Many elements used in the profile have an explicit semantics by a reference to a concept in the CLARIN Concept Regisry (CCR). For the elements and components that do not have such explicit semantics, we provided a label, a definition and additional information in the required format and submitted it to Menzo Windhouwer (2017-09-08), the CCR gate-keeper. He submitted the proposal to the Semantic Interoperability Committee for approval, but we received no feedback as yet.

We did not include the currency codes in this list, as recommended by Menzo Windhouwer. This currency list is based on ISO4217[5] and is better dealt with by CLAVAS.[6]

In the meantime a few other elements and components have been added, and for these their semantics is yet to be defined.

4 Actual use

5 Validation and Schematron

We have tested the quality of the profile by the CLARIN curation facility (section 5.4).

5.1 Validation

5.2 Schematron file

5.3 URL and PID check

5.4 Quality

6 CLARIN Language Resource Switchboard

7 Other profiles for tools

The *resourceInfo* profile is used to describe some tools but most are data. Searching for 'tool' within this selection http://vlo.minerva.arz.oeaw.ac.at/vlo/?q=tool&fqType=_componentProfile:or&fq=_componentProfile:resourceInfo yields 68 results out of 462 metadata records in total.

[5] https://en.wikipedia.org/wiki/ISO_4217.
[6] http://portal.clarin.nl/node/4223.

Table 1: Profiles for software in the CLARIN Component Registry (2017-09-29)

Profile Name	ID	#descriptions
AnnotationTool	clarin.eu:cr1:p_1297242111880	0
BamdesTool	clarin.eu:cr1:p_1288172614017	0
resourceInfo	clarin.eu:cr1:p_1360931019836	462 (68)
Tool	clarin.eu:cr1:p_1297242111879	0
ToolComponentProfile	clarin.eu:cr1:p_1369752611626	1
ToolProfile	clarin.eu:cr1:p_1290431694581 4	49
ToolService	clarin.eu:cr1:p_1311927752306	1
ToolService	clarin.eu:cr1:p_1360230992146	0
WebApplication	clarin.eu:cr1:p_1469541567394	2
serviceDescription	clarin.eu:cr1:p_1295178776924	0
TextImagerWebService	clarin.eu:cr1:p_1469541567396	0
WeblichtService	clarin.eu:cr1:p_132065762964428	287
BASWebService	clarin.eu:cr1:p_1324638957718	1
BASWebService	clarin.eu:cr1:p_1423750293168	0
BASWebService	clarin.eu:cr1:p_1381926654686	0
CLARINWebService	clarin.eu:cr1:p_1505397653795	0
RECOGNIZER	clarin.eu:cr1:p_1430905751617	0
RECOGNIZER	clarin.eu:cr1:p_1360230992143	0
RECOGNIZER	clarin.eu:cr1:p_1393514855416	0

7.1 WebLichtWebService

The WebLichtWebService (WLWS) profile is a profile to accommodate for WebLicht web services, extending the Clarin Core Webservices Profile (version 1.0.1). It has been derived from Profile clarin.eu:cr1:p$_1$311927752335 (CLARINWebService).[7]

Some 295 descriptions based on this profile can be found in the Virtual Language Observatory.[8]

The profile contains one component, *Service* with semantics http://hdl.handle.net/11459/CCR_C-4159_ca0e6cba-cab5-b51a-f430-fdcb0756c9ac and component id clarin.eu:cr1:c_1320657629647 .

The WebLichtWebService is limited to the description of web services. It, surprisingly, provides no way to formally encode in which language(s) data can be that the tool applies to, or which languages it yields as output. It also offers no formal way of encoding the tasks that the tools can perform.

7.1.1 Service

The (Nalida) component Service clarin.eu:cr1:c_1320657629647 is used in the profile WebLichtWebService.

[7] See http://lux17.mpi.nl/isocat/clarin/ws/cmd-core/index.html.

[8] by using the query encoded in this URL: https://vlo.clarin.eu/search?q=_componentProfile:WebLichtWebService.

CSD uses the component Service `clarin.eu:cr1:c_1505397653788`, which is identical to the CLARIN-NL component Service (`clarin.eu:cr1:c_1320657629631`) except for the fact that the ServiceDescriptionLocation is not an empty component anymore (to make it compatible with CMDVersion 1.2). There is one more apart from the ones mentioned already, but it is not relevant in this context: the component Service for BAS Web services `clarin.eu:cr1:c_1423750293167`.

All Service components are derived from the CLARIN Core Web Service `clarin.eu:cr1:c_1505397653787`.

We list the elements and components of the WebLichtWebService Service component, and specify the relation with CSD elements and components

Attribute: CoreVersion (string, closed vocabulary). Identical in CSD

Element: Name (string). Identical in CSD, and can be mapped to *General-Info/Name*

Element: Description (string). Maximally 2.

Attribute: type (optional) only one possible value: 'short'. The Element *Description* without the attribute can be mapped to the CSD element *GeneralInfo/Title*, while this element with the attribute can be mapped to the CSD element *GeneralInfo/Name*.

Element: TypeOfWebservice no equivalent in CSD

Element: url (anyURI).

Element: LifeCycleStatus CSD element GeneralInfo/LifeCycleStatus

Element: PublicationDate (string).

Element: LastUpdate)string).

Component: ServiceDescriptionLocation (1 - 1)

Component: Contact (1 - 1)

Component: Creation [1 - 1]

Element: Topic not used

Component: Creators (0 - 1)

Component: CreationToolInfo (0 - unbounded) not used

Component: Annotation (0 - 1) not used

Component: Source (0 - unbounded) not used

Component: Descriptions (0 - 1)

The Parameter component differs from the CSD Parameter component in that it has a few additional elements:

Component: Operations

- WebServiceArgValue (type: String)
- AllowManualSelectionFallback (type: Boolean)

7.2 ToolProfile

Description: A CMDI profile for tools.

7.2.1 GeneralInfo

`clarin.eu:cr1:c_1290431694495`

Element: ResourceName

Element: ResourceTitle

Element: ResourceClass only *Tool* occurs here in the actual records

Element: Version

Element: LifeCycleStatus

Element: StartYear

Element: CompletionYear

Element: PublicationDate

Element: LastUpdate

Element: TimeCoverage

Element: LegalOwner

Element: Genre Occurs with values such as other, unspecified but also with (informal) indications of the software task, e.g. *OCR*, *Visualizer*, *Data visualizing*, and *Tagger*

7.2.2 Project

7.2.3 Publications

7.2.4 Documentations

7.2.5 Creators

7.2.6 Access

7.2.7 Copyright

7.2.8 ResourceContext

7.2.9 ToolContext

ToolType allows any string as value, which offers flexibility but also unnecessary variability, such as variants in different languages (English and German), spelling variants (upper case and lower case), alternative formulations (synonyms), etc. ToolType values and mapping to CSD ToolTask (after whitespace trimming):

Analysewerkzeug (1)

analysis (1)

analysis tool (1)

annotation search (1)

Annotation Tool (3)

annotation tool (3)

annotation visualization (1)

Annotationseditor (1)

Chunker (2)

concordancer (1)

Converter (1)

Converter Framework (1)

corpus conversion (1)

Corpus index and search tool (1)

corpus manager (2)

corpus search (1)

editor (1)

format exporter (2)

format importer (2)

Konkordanzwerkzeug (1)

Korpusmanager (1)

Lemmatizer (2)

lemmatizer (1)

Morphological Analyser (1)

morphological analyzer (1)

morpho-syntactic tagger (1)

OCR scanned text (1)

parser (2)

Parser (6)

POS Tagger (2)

query tool (1)

search engine (26)

sentence boundary detector (1)

Suchwerkzeug (1)

Tagger (1)

tokenizer (1)

Transcription / Annotation tool (1)

transcription tool (1)

Traskriptionseditor (1)

visualisation (1)

visualization tool (1)

Visualizer (2)

XForms library (1)

8 Facets for software

The current facets of the VLO have been selected to search for data, not for software. They are derived from CMDI elements by mapping rules as specified in http://lux17.mpi.nl/isocat/clarin/vlo/mapping/index.html.

Several facets have no or hardly any meaning for software, e.g. *subject* and *genre*, perhaps also *collection*.

The facet discovery system does not discover

- country
- organisation
- license
- nationalProject

The *modality* facet could be derived from input mimetypes, but are not discovered. There is no element for *keywords* but these could be derived from the description text.

ToolProfile: Resourceclass (http://hdl.handle.net/11459/CCR_C-3806_e55e9ed6-b099-c21d-a634-3c7f4d22a215) with values *Tool* (http://hdl.handle.

net/11459/CCR_C-4362_bbceebba-02fc-2417-5433-3402a33e350a) or *Toolchain* (http://hdl.handle.net/11459/CCR_C-4368_324c1065-aa9d-8321-51e0-f4d04022b51d)

A separate faceted search view should be made available for software with a set of facets specific for software. The first problem with thus is that the VLO must be able to distinguish metadata records for software from metadata for data. To my knowledge, that distinction is currently not made systematically. Surely, most metadata records only specify this not at all or only implicitly, and to my knowledge one cannot mark CMDI profiles in thus way (as intended for describing data v. software).

For the ToolProfile profile, it s ResourceClass (http://hdl.handle.net/11459/CCR_C-3806_e55e9ed6-b099-c21d-a634-3c7f4d22a215) with values *Tool* (http://hdl.handle.net/11459/CCR_C-4362_bbceebba-02fc-2417-5433-3402a33e350a) or *Toolchain* (http://hdl.handle.net/11459/CCR_C-4368_324c1065-aa9d-8321-51e0-f4d04022b51c
All valid records currently in the VLO have the value *Tool*.[9]

in particular:

- Priority 1

  tooltask derive it for:

  General from http://hdl.handle.net/11459/CCR_C-2500_a16d939a-58e3-121d-aaa3-0

  ClarinSoftwareDescription from http://hdl.handle.net/11459/CCR_C-2500_a16d939a-58e3-121d-aaa3-05237ec2d206

  WebLichtWebService from the mapping table provided

  ToolProfile values from the element with semantics http://hdl.handle.net/11459/CCR_C-3810_7d3b783a-d2c6-51fc-01ab-6a8a658d94c8 in accordance with the table provided

  research domain derive it for:

  General from http://hdl.handle.net/11459/CCR_C-2467_f4e7331f-b930-fc42-eeea-0

  ClarinSoftwareDescription from http://hdl.handle.net/11459/CCR_C-2467_f4e7331f-b930-fc42-eeea-05e383cfaa78

  WebLichtWebService from the mapping table provided

  ToolProfile http://hdl.handle.net/11459/CCR_C-3796_e89bb008-3e2e-1f70-afa5-e5

  linguistics subject derive it for:

  General from http://hdl.handle.net/11459/CCR_C-2527_77dad635-e029-acca-2fa8-f

  ClarinSoftwareDescription from http://hdl.handle.net/11459/CCR_C-2527_77dad635-e029-acca-2fa8-f2fa5751d3af

  WebLichtWebService from the mapping table provided

  ToolProfile http://hdl.handle.net/11459/CCR_C-3796_e89bb008-3e2e-1f70-afa5-e5

  input mimetype derive it for:

  General from http://hdl.handle.net/11459/CCR$_C$-2571$_2$be2e583-e5af-34c2-3673-93359ec1f7df inside Input component

[9] There are four invalid records consisting of an HTML document instead of an XML CMDI document.

ClarinSoftwareDescription from http://hdl.handle.net/11459/CCR$_C$-2571$_2$be2e583-e5af-34c2-3673-93359ec1f7df inside Input component

WebLichtWebService from http://hdl.handle.net/11459/CCR_C-2571_2be2e583-e5af-34c2-3673-93359ec1f7df inside Input component

ToolProfile from http://hdl.handle.net/11459/CCR_C-2571_2be2e583-e5af-34c2-3673-93359ec1f TextTechnical/MimeTypes/MimeType in Input

input language derive it for:

General from http://hdl.handle.net/11459/CCR_C-2484_669684e7-cb9e-ea96-59cb-a25fe89b9b9c inside Input component

ClarinSoftwareDescription from http://hdl.handle.net/11459/CCR_C-2484_669684e7-cb9e-ea96-59cb-a25fe89b9b9d inside Input component

WebLichtWebService from the mapping table provided

ToolProfile http://hdl.handle.net/11459/CCR_C-2484_669684e7-cb9e-ea96-59cb-a25fe89b9b9d InputLanguage/Language/LanguageName

input language ISO code derive it for:

General

ClarinSoftwareDescription http://hdl.handle.net/11459/CCR_C-2482_08eded24-4086-7e3f-88e5-e0807fb01e17 in Input

WebLichtWebService from the mapping table provided

ToolProfile http://hdl.handle.net/11459/CCR_C-2482_08eded24-4086-7e3f-88e5-e0807fb01e17 in InputLanguage

output mimetype General from http://hdl.handle.net/11459/CCR$_C$-2571$_2$be2e583-e5af-34c2-3673-93359ec1f7df inside Output component

ClarinSoftwareDescription from http://hdl.handle.net/11459/CCR$_C$-2571$_2$be2e583-e5af-34c2-3673-93359ec1f7df inside Output component

WebLichtWebService from http://hdl.handle.net/11459/CCR_C-2571_2be2e583-e5af-34c2-3673-93359ec1f7df inside Output component

ToolProfile from http://hdl.handle.net/11459/CCR_C-2571_2be2e583-e5af-34c2-3673-93359ec1f iside Output

output language derive it for:

General from http://hdl.handle.net/11459/CCR_C-2484_669684e7-cb9e-ea96-59cb-a25fe89b9b9c inside Output component

ClarinSoftwareDescription from http://hdl.handle.net/11459/CCR_C-2484_669684e7-cb9e-ea96-59cb-a25fe89b9b9d inside Output component; if not found, take them from inside the Input Component

WebLichtWebService from the mapping table provided

ToolProfile http://hdl.handle.net/11459/CCR_C-2484_669684e7-cb9e-ea96-59cb-a25 inside Output component;

Application type web/desktop app, command line interface http://hdl.handle.net/11459/CCR_C-3786_21c37142-994f-63a8-5a5b-a9fce07681a7

User Interface derive it for:

General

ClarinSoftwareDescription

WebLichtWebService

ToolProfile

Modalities derive it for:

General

ClarinSoftwareDescription toolcategory? mimetype?

WebLichtWebService mimetype?

ToolProfile http://hdl.handle.net/11459/CCR_C-2490_44bc38a3-1799-4149-c791-40a

- Priority 2

Operating System derive it for:

General http://hdl.handle.net/11459/CCR_C-2572_0a6b62f1-52e0-f0a2-fd72-d75e64

ClarinSoftwareDescription http://hdl.handle.net/11459/CCR_C-2572_0a6b62f1-52e0-f0a2-fd72-d75e6465a32c

WebLichtWebService none

ToolProfile http://hdl.handle.net/11459/CCR_C-2572_0a6b62f1-52e0-f0a2-fd72-d75

license derive it for:

General

ClarinSoftwareDescription

WebLichtWebService

ToolProfile

- Priority 3

organisation derive it for:

General http://hdl.handle.net/11459/CCR_C-2979_8030473e-bbcb-6b87-3fd2-905544

ClarinSoftwareDescription http://hdl.handle.net/11459/CCR_C-2979_8030473e-bbcb-6b87-3fd2-90554429ec50

WebLichtWebService http://hdl.handle.net/11459/CCR_C-2979_8030473e-bbcb-6b87-3fd2-90554429ec50

ToolProfile http://hdl.handle.net/11459/CCR_C-2979_8030473e-bbcb-6b87-3fd2-905

country derive it for:

General

ClarinSoftwareDescription none- to be added; same semantics but only occuurs in LanguageVariety, which should be excludd

WebLichtWebService not

ToolProfile Country/CountryName http://hdl.handle.net/11459/CCR_C-3792_68c770a4-d58c-46dd-d429-5609ce5f81c3

project derive it for:

General

ClarinSoftwareDescription Project/name http://hdl.handle.net/11459/CCR_C-2536_13fc5f10-c14a-1f64-a669-32736f6d3ef5

WebLichtWebService none

ToolProfile Project/name http://hdl.handle.net/11459/CCR_C-2536_13fc5f10-c14a-1f64-a669-32736f6d3ef5

national project : WLS, CSD, TP: none

Further needed:

Name derive it for

General from http://hdl.handle.net/11459/CCR_C-2544_3626545e-a21d-058c-ebfd-241c0464e7e5 or http://hdl.handle.net/11459/CCR_C-4160_192be757-0d8f-f4fe-b10b-d3d50de92482

ClarinSoftwareDescription from http://hdl.handle.net/11459/CCR_C-2544_3626545e-a21d-058c-ebfd-241c0464e7e5; GeneralInfo/name

WebLichtWebService from http://hdl.handle.net/11459/CCR_C-4160_192be757-0d8f-f4fe-b10b-d3d50de92482; Service/Name

ToolProfile http://hdl.handle.net/11459/CCR_C-2544_3626545e-a21d-058c-ebfd-241c0464e7e5; GeneralInfo/ResourceName

Title derive it for

General from http://hdl.handle.net/11459/CCR_C-2545_d873f2ab-2a2f-29d6-a9ab-260cde57f227

ClarinSoftwareDescription from http://hdl.handle.net/11459/CCR_C-2545_d873f2ab-2a2f-29d6-a9ab-260cde57f227; GeneralInfo/title

WebLichtWebService from Service/Description[not(@type)], i.e with the *type* attribute absent;

ToolProfile http://hdl.handle.net/11459/CCR_C-2545_d873f2ab-2a2f-29d6-a9ab-260cde57f227; GeneralInfo/ResourceTitle

PiD derive it for

General from CMD/MDSelfLink

ClarinSoftwareDescription from CMD/MDSelfLink

WebLichtWebService from CMD/MDSelfLink

ToolProfile from CMD/MDSelfLink

9 DiRT

http://dirtdirectory.org/

Analyze data
Interpret data Annotate
Model data Archive data
Analyze networks between my data Capture information
Organize data Clean up data
Preserve data Collaborate
Program Comment
Publish Communicate
Record audio/video Analyze the content of my data
Analyze relationships between pieces of data Contextualize data
Share Convert files
Analyze the geographical aspect of my data Create
Store data Crowdsource data enrichment/analysis
Analyze the structure of my data Design
Analyze the stylistics of my data Find information
Theorize Disseminate data
Transcribe audio, video or manuscripts Add markup to an object
Translate Enrich metadata about an object
Visualize data Collect information
Build a website Add identifiers to data
Write

10 Concluding Remarks