

CLARIN FOR LINGUISTS STORING RESOURCES IN CLARIN

Jan Odijk

LOT Summerschool

Nijmegen, 2014-06-27

OVERVIEW

- Why store resource in CLARIN?
- How to store resources in CLARIN
 - What **you** must do
 - What the **CLARIN Centre** must do

OVERVIEW

➤ Why store resource in CLARIN?

- How to store resources in CLARIN
 - What **you** must do
 - What the **CLARIN Centre** must do

WHY?

- You may benefit from it
 - Existing tools in CLARIN
 - faster production, better quality, more features
 - Search engines, analysis tools, visualisation tools
 - Can be easily combined with other data in CLARIN
- Others may benefit from it
 - Many unexpected uses of your data
 - Now or in the future

WHY?

- Openness in Science
 - Usually produced with public money
- Integrity
 - Recently many scandals with faked data
- Verifiability and Replicability of research results
 - Essential for the proper conduct of science
 - More and more journals are requiring it
- Funding Agency requires it (data management plan)
- [DANS CLIP on data sharing](#) (in Dutch)

OVERVIEW

- Why store resource in CLARIN?
- **How to store resources in CLARIN**
 - What **you** must do
 - What the **CLARIN Centre** must do

How?

- Start early
 - Preferably before you start creating data
- Contact a CLARIN **Type B** centre
 - They can help you
 - Your resource must be stored at a CLARIN centre
- Which CLARIN centre?
 - Check CLARIN-NL portal (<http://dev.clarin.nl>)
 - CLARIN-NL website [Centres](#) page
 - [Brief summary](#)

OVERVIEW

- Why store resource in CLARIN?
- How to store resources in CLARIN
 - **What you must do**
 - What the **CLARIN Centre** must do

WHAT YOU MUST DO

- Define what your data are / are going to be
- Ensure legal / ethical compliance
 - Permission from subjects to use the data for research
 - Provisions for respecting privacy matters
- Determine Metadata Contents
 - Determine what information should be included in the metadata (resource description) of your resource
 - Collect this information

WHAT YOU MUST DO

- Determine centre
- Contact them and make arrangements
- Determine CLARIN-recommended standard format(s) for your data
 - Consult with the CLARIN Centre
 - Ask help from the helpdesk helpdesk@clarin.nl
 - E.g. LMF (Lexical Markup Framework for lexicons, cf. Cornetto, DuELME, ...)
 -

WHAT YOU MUST DO

- Metadata must be in CMDI format
- CMDI provides
 - A model for metadata
 - A format for metadata
 - Tools to make metadata
- CMDI Metadata are written in XML
- It does NOT proscribe the contents of the metadata
- [Introduction to CMDI](#) cmdi@clarin.eu

WHAT YOU MUST DO

- CMDI Metadata use a metadata profile
- **Metadata profile**: a combination of
 - Metadata components
 - Metadata elements
- **Metadata Component**: a combination of
 - Metadata components (optional)
 - Metadata elements (optional)
- **Metadata Element**
 - XML element: name, value (of an explicit type), attribute-value pairs

WHAT YOU MUST DO

- This provides high flexibility
 - YOU can determine the metadata for your resource
 - By defining your own profiles, components, elements
- CMDI helps you with
 - A [profile and component editor](#) [[login required](#)]
 - A list of [commonly used profiles and components](#)
 - A metadata editor: [ARBIL](#)

WHAT YOU MUST DO

- Flexibility requires explicit semantics!
 - The CLARIN infrastructure must `know` what you mean with your metadata elements
 - Otherwise it cannot use faceted browsing in the [VLO](#) or the [Meertens Metadata Search Engine](#)

WHAT YOU MUST DO

- Explicit semantics:
 - Each element in the **data and metadata** must have a link to a CLARIN-recognized concept or data category registry
 - Most prominent data category registry in CLARIN is [ISOCAT](#)
 - [Example Data Category in ISOCAT](#)
 - [Example Link to ISOCAT in element definition](#)

WHAT YOU MUST DO

- Explicit semantics (2):
 - [RELCAT alpha version](#)
 - For relations between data categories
 - [SCHEMACAT alpha version](#)
 - For describing schemas of resources
 - [CLAVAS Vocabulary Service](#)
 - Interface to other data category registries and vocabularies
 - [ISO 639-3 Language codes](#)

WHAT YOU MUST DO

- Attend dedicated tutorials on CMDI and ISOCAT
 - [Regularly organized](#) in NL (each 2+ times/year)
- Usually the CLARIN Centre helps you creating the CMDI metadata
- Maximally reuse existing profiles /components
 - It will help you get better metadata
 - You do not have to reinvent the wheel

WHAT YOU MUST DO

- CLARIN
 - strongly recommends using certain components (e.g. GeneralInfo component) and
 - may require inclusion of certain properties
- Do not forget properties that are 'obvious to you', e.g.
 - Language
 - Title
 - Version

WHAT YOU MUST DO

- Live Version v. exchange/archive version
 - E.g. lexicon
 - in [Lexical Markup Framework](#) compatible [XML](#) text v.
 - database with indexes for fast search
 - Live version is ideally derived fully automatically from the exchange/archive version
- In close cooperation with the CLARIN Centre

WHAT YOU MUST DO

- Software
 - Desktop tool
 - Desktop application
 - Web service
 - [SOAP](#) / [REST](#) , use [CLAM](#) if possible
 - Web application
- There must be `metadata' for software as well!
 - Generic profile exists and is being refined
- Consult with the CLARIN Centre

OVERVIEW

- Why store resource in CLARIN?
- How to store resources in CLARIN
 - What **you** must do
 - **What the CLARIN Centre must do**

WHAT THE CLARIN CENTRE MUST DO

- Assist you with your tasks
- Assign Persistent Identifiers (PIDs) to all data and metadata
 - [Handle system](#) for assignment and resolution of PIDs
 - [example](#)
- Make Metadata harvestable
 - [OAI-PMH](#) protocol
 - Open Archives Initiative Protocol for Metadata Harvesting

WHAT THE CLARIN CENTRE MUST DO

- Store the data in the centre's repository
 - [LAMUS](#) (the Language Archive) and its documentation [online](#) or as [PDF](#)
 - [EASY](#) (DANS) and its [Help and Support Page](#)
- Make data themselves available and accessible in the CLARIN infrastructure

WHAT THE CLARIN CENTRE MUST DO

- Provisions for legal / ethical restrictions
- Long term preservation
 - [Data Seal of Approval](#)
- (Minimal) Maintenance

WHAT ABOUT EXISTING DATA?

- Adapt them to meet the CLARIN requirements
 - Data Curation
- CLARIN-NL
 - has financed many [data curation projects](#)
 - Has set up [Data Curation Service](#)
- [CLARIAH](#) (successor project, if financed)
 - Will continue these curation activities

Thanks for your attention!

DO NOT ENTER HERE

NL

A large, faint background diagram consisting of several light gray circles of varying sizes connected by thin, curved lines. The circles are arranged in a roughly circular pattern, with one circle at the top right containing the letters "NL" in a bold, white, sans-serif font.

CLARIN-NL B CENTRES

- **Meertens Institute:** resources relevant for the study of
 - cultural expressions and Language variation within the Dutch language
- **Max Planck Institute for Psycholinguistics (The Language Archive):** resources related to the study of
 - psychological, social and biological foundations of language
- **Huygens Institute:** resources related to the study of
 - history and literature of the Netherlands.
- **Institute for Dutch Lexicology (INL)**
 - relevant to the lexicological study of the Dutch language
- **Data Archiving and Networked Centres (DANS)**
 - digital research data generally



- **Koninklijke Bibliotheek (National Library):**
 - Digital books, articles, newspapers
 - Includes DBNL (Digital Library for Dutch Literature)
 - (will be available in the VLO soon)
- **Nederlands Instituut voor Beeld & Geluid (NIBG, Netherlands Institute for Sound and Vision)**
 - Audio-visual data (esp. TV and radio programmes)
 - [NIBG data via the VLO](#)
- **Utrecht University Library (UBU)**
 - Digital books, articles
 - [UBU data via the VLO](#)



CMDI PROFILE: EXAMPLE



IPROSLA_Deliver... Data, Tools, ... http://...38110# x GrE TEL for C... Frog :: jano Lagere econ... Nieuw tabblad Nieuw tabblad Universiteit ... JoshiFest: "J... +

catalog.clarin.eu/ds/ComponentRegistry/?item=clarin.eu:cr1:c_1271859438110#app=cd01&e2ee-selectedIndex=0 Johan Bolhuis

Meest bezocht Aan de slag

Clarín Component Registry help User: anonymous login

Component Browser

Profiles Components Public space

Create new Edit as new Import filter... Showing 157 of 157

Name	Group Name	Domain Name	Creator	Description	Registration Date	Comments
GTRP_sub_location			Folkert de Vriend	Profile for locations subdivision ...	30 July 2010 10:05:03	0
imdi-corpus			Alexander Koenig	IMDI corpus profile	02 June 2010 15:16:26	0
imdi-session			CLARIN (cmdi@clarin.eu)	IMDI metadata (type: session)	21 April 2010 16:19:43	3
LexicalResourceProfile			Laura van Eerten	a profile for describing a lexical ...	26 April 2010 12:13:06	2

LexicalResourceProfile view xml Comments (2)

Name: **LexicalResource**
Group Name: clarin-nl/lexicalresource
Description: A component for describing lexical resources like different types of word lists, dictionaries and semantic networks.

Component: **cmdi-generalinfo**
Number of occurrences: 1 - 1

Component: **OriginLocation**
Number of occurrences: 0 - unbounded
Component: **cmdi-location**
Number of occurrences: 1 - 1

Component: **cmdi-project**
Number of occurrences: 0 - unbounded

Component: **cmdi-creators**
Number of occurrences: 0 - 1



Browser tabs: IPROSLA_Deliver..., Data, Tools, ... http://...38110# x GrE TEL for C... Frog :: jano Lagere econ... Nieuw tabblad Nieuw tabblad Universiteit ... JoshiFest: "J... +

Address bar: catalog.clarin.eu/ds/ComponentRegistry/?item=clarin.eu:cr1:c_1271859438110#app=cd01&e2ee-selectedIndex=1 Johan Bolhuis

Meest bezocht Aan de slag

Clarín Component Registry help

Component Browser User: anonymous login

Profiles Components Public space ▼

Create new Edit as new Import Q filter... Showing 872 of 872

Name	Group Name	Domain Name	Creator	Description	Registration Date	Comments
cmdi-language	clarin		CLARIN (cmdi@clarin.eu)	Component for describing a c...	21 April 2010 16:17:21	0
cmdi-location	clarin		CLARIN (cmdi@clarin.eu)	Component for describing a c...	21 April 2010 16:17:21	0
cmdi-mimetype	clarin		CLARIN (cmdi@clarin.eu)	List of frequently used mime t...	21 April 2010 16:17:18	3
Creation	CLARIN		nalida	Component contains informati...	31 January 2013 12:48:14	0

cmdi-location view xml Comments (0)

Name: Location

Group Name: clarin

Description: Component for describing a certain location (address, region, country, continent)

Element: Address string
 ConceptLink: <http://www.isocat.org/datcat/DC-2528>
 Number of occurrences: 0 - 1

Element: Region string
 ConceptLink: <http://www.isocat.org/datcat/DC-2533>
 Number of occurrences: 0 - 1

Component: iso-country
 Number of occurrences: 0 - 1

Component: iso-continent
 Number of occurrences: 0 - 1



IPROSLA_Deliv... Data, Tool... http://...stry/# x GrE TEL for ... Frog :: jano Lagere eco... Nieuw tabblad Nieuw tabblad Universitei... JoshiFest: "..." CLARIN-N... +

catalog.clarin.eu/ds/ComponentRegistry/#app=cd01&e2ee-selectedIndex=1 Johan Bolhuis

Meest bezocht Aan de slag

Clarín Component Registry help

Component Browser User: Odijk101@soliscom.uu.nl settings

Profiles Components **Public space** ▼

Create new Edit as new Import filter... Showing 872 of 872

Name	Group Name	Domain Name	Creator	Description	Registration Date	Comments
media-session	CLARIN-D Version 1.2	phonetics	71A3DD8C90E89334@lmu.de	Group of recordings of at leas...	05 February 2013 14:12:47	0
WrittenResource	CLARIN-D Version 1.2	phonetics	71A3DD8C90E89334@lmu.de	Component contains informati...	05 February 2013 14:13:45	0
author	CLARIN-D: DTA-Basisformat		Axel Herold	the author of a textual resource	16 September 2013 15:59:33	0

fileDesc **view** xml Comments (0)

Group Name: CLARIN-D: DTA-Basisformat
 Description: bibliographical information on author/editor and title of a text

Element: title string
 ConceptLink: http://www.isocat.org/datcat/DC-2545
 Documentation: the title of a work according to the classification in @type
 DisplayPriority: 1
 Number of occurrences: 1 - unbounded
 Multilingual: false

AttributeList:
 type: main ▼
 level: a ▼
 n: decimal

Component: author
 Number of occurrences: 0 - unbounded

Component: editor
 Number of occurrences: 0 - unbounded



Browser tabs: Too... http://...istry/# GrE TEL fo... Frog :: jano Lagere ec... Nieuw tabblad Nieuw tabblad Universit... JoshiFest... ISOca... x Arbil | Th... ISOcat - ...

Address bar: https://catalog.clarin.eu/isocat/interface/index.html

Search: Johan Bolhuis

Meest bezocht Aan de slag

ISOcat Welcome Jan Odijk Help

2545

search result for '2545'

#	Name	Version	Administration st	Registration stati	Check	Type	Owned by	Scope
2545	resource title	1:0	private	private	✓	open	Athens Core	public

resource title - 1:0

2. Description Section

Profile	Metadata
[-] 2.1 English Language Section	
Language	English (en)
2.1.1 Name Section	
Name	resource title
Name Status	admitted name
2.1.2 Definition Section	
Definition	The title is the complete title of the resource without any abbreviations.
Source	CLARIN
[+] 2.2 Catalan Language Section	
[-] 2.3 Croatian Language Section	



IPROSLA_Deliv... Data, Tool... http://...stry/# x GrETEL for ... Frog :: jano Lagere eco... Nieuw tabblad Nieuw tabblad Universitei... JoshiFest: "..." CLARIN-N... +

catalog.clarin.eu/ds/ComponentRegistry/#app=cd01&e2ee-selectedIndex=1 Johan Bolhuis

Meest bezocht Aan de slag

Clarin Component Registry help

Component Browser User: Odijk101@soliscom.uu.nl settings

Profiles Components Public space ▾

Create new Edit as new Import filter... Showing 872 of 872

Name	Group Name	Domain Name	Creator	Description	Registration Date	Comments
media-session	CLARIN-D Version 1.2	phonetics	71A3DD8C90E89334@lmu.de	Group of recordings of at leas...	05 February 2013 14:12:47	0
WrittenResource	CLARIN-D Version 1.2	phonetics	71A3DD8C90E89334@lmu.de	Component contains informati...	05 February 2013 14:13:45	0
author	CLARIN-D: DTA-Basisformat		Axel Herold	the author of a textual resource	16 September 2013 15:59:33	0

fileDesc view xml Comments (0)

Group Name: CLARIN-D: DTA-Basisformat

Description: bibliographical information on author/editor and title of a text

Element: title string

ConceptLink: <http://www.isocat.org/datcat/DC-2545>

Documentation: the title of a work according to the classification in @type

DisplayPriority: 1

Number of occurrences: 1 - unbounded

Multilingual: false

AttributeList:

type main ▾

level a ▾

n decimal

Component: author

Number of occurrences: 0 - unbounded

Component: editor

Number of occurrences: 0 - unbounded

LOGIN

- CLARIN attempts to maximize open and free access to resources
 - with as little restrictions as possible
 - no login unless it cannot be avoided
- Sometimes, a login (Authentication and Authorisation, AAI) is required, e.g.
 - Because there are legal and/or ethical restrictions on the data
 - To identify you and assign you your own workspace / data
 - To enter with your own personal settings

LOGIN






- CLARIN is a distributed infrastructure
 - How can we avoid that you have to login again and again?
 - How can we avoid that you have to remember many user names and passwords?
 - How can we avoid that CLARIN has to securely store user names, passwords and possibly other privacy-sensitive information?



- The answer: [Shibboleth](#)
 - [When you log in](#), you are directed to [a login with your own institute](#)
 - You then [log in with you institute's user name and password](#)
 - The institute server then confirms that you are a trusted person, and [you can enter this part of the CLARIN infrastructure](#)
 - it does **not** pass on any sensitive information such as your user name or password
 - If you now go [to another part of the CLARIN infrastructure](#) that requires login, it 'knows' that you are already logged in, so you do not have to do this again (Single Sign On, SSO)



Standards for LRT - Standard... x Detail info | Academia x C Data, Tools, Demonstrat... x http://catalog...entRegistry/# x http://catalog...entRegistry/# x Shibboleth (Internet2) - ... x +

catalog.clarin.eu/ds/ComponentRegistry/#  Google    

Meest bezocht Aan de slag

Clarin Component Registry help [login](#)

Component Browser User: anonymous

Profiles Components **Public space** Showing 157 of 157

Name	Group Name	Domain Name	Creator	Description	Registration Date	Comments
AnnotationTool			Eric Sanders	Description of a tooladapted fro...	09 February 2011 14:13:59	0
ArthurianFiction		Other	Rik Hoekstra	Profile for Arthurian Fiction data...	04 September 2012 14:50:05	0
BamdesLexicalResource		computational_linguistics	Dieter Van Uytvanck	Lexical Resource as used by BA...	27 October 2010 15:47:42	0
BamdesMultimodalCorpus		computational_linguistics	Dieter Van Uytvanck	Oral Corpus as used by BAMDES...	27 October 2010 16:00:41	0
BamdesOralCorpus		computational_linguistics	Dieter Van Uytvanck	Oral Corpus as used by BAMDES...	27 October 2010 16:00:05	0
BamdesTool		computational_linguistics	Dieter Van Uytvanck	Tool as used by BAMDES (for th...	27 October 2010 15:48:54	0
BamdesWrittenCorpus		computational_linguistics	Dieter Van Uytvanck	Written Corpus as used by BAM...	27 October 2010 15:59:28	0
Bedevaartbank			Folkert de Vriend	Profile for Bedevaartbank	30 July 2010 9:33:53	0
Boedelbank			Folkert de Vriend	Profile for Boedelbank	30 July 2010 9:36:25	0
cmdi-virtual-collection			CLARIN (cmdi@clarin.eu)	A component for describing per...	21 April 2010 16:18:30	0
collection			Matej Durco	minimal collection profile (start...	15 October 2010 20:58:37	0



Standards for LRT - Standard... x Detail info | Academia x Data, Tools, Demonstrat... x http://catalog...entRegistry/# x Clarin EU Service Provider x Shibboleth (Internet2) - ... x

catalog.clarin.eu/mw1/sds/discojuice?entityID=https%3A%2F%2Fsp.catalog.clarin.eu&return=https%3A%2F%2Fcatalog.clarin.eu%2FShibboleth.sso%2FL Google

Meest bezocht Aan de slag

- Return Page

Sign in to **Clarin EU Service Provider**
Select your Identity Provider

If you cannot find your institution in the list above please select the "Clarin.eu website account" and use your credentials of the CLARIN website. For questions please contact webmaster@clarin.eu.


- Clarin.eu website account
European Union
- Universiteit Utrecht**
Netherlands
- RUG
Netherlands
- Universiteit Leiden
Netherlands
- Graafschap College
Netherlands
- TERENA Secretariat
Netherlands
- Universiteit van Amsterdam
Netherlands
- Tilburg University
Netherlands
- Vrije Universiteit
Netherlands
- NIOO (KNAW)
Netherlands



Data, Tools, Demonstrators... x Universiteit Utrecht: Login x +

https://login.services.uu.nl/nidp/saml2/sso?id=42&sid=0&option=credential&sid=0

Meest bezocht Aan de slag




Universiteit Utrecht

Solis-id

Wachtwoord

[Wachtwoord vergeten?](#)

Meer informatie over Solis-id: [medewerkers](#) / [studenten](#)

Kijk ook op www.uu.nl/ict  [English](#)



[Data, Tools, Demonstrators...](#) x <http://catalog.c...istry/index.jsp#> x +

[catalog.clarin.eu/ds/ComponentRegistry/index.jsp#](#)
Google
☆
↓
↑
☰

Meest bezocht Aan de slag

Clarin Component Registry help

Component Browser User: Odijk101@soliscom.uu.nl [settings](#)

Profiles Components Public space

Create new Edit as new Import
Q filter... Showing 157 of 157

Name	Group Name	Domain Name	Creator	Description	Registration Date	Comments
AnnotationTool			Eric Sanders	Description of a tooladapted fro...	09 February 2011 14:13:59	0
ArthurianFiction		Other	Rik Hoekstra	Profile for Arthurian Fiction data...	04 September 2012 14:50:05	0
BamdesLexicalResource		computational_linguistics	Dieter Van Uytvanck	Lexical Resource as used by BA...	27 October 2010 15:47:42	0
BamdesMultimodalCorpus		computational_linguistics	Dieter Van Uytvanck	Oral Corpus as used by BAMDES...	27 October 2010 16:00:41	0
BamdesOralCorpus		computational_linguistics	Dieter Van Uytvanck	Oral Corpus as used by BAMDES...	27 October 2010 16:00:05	0
BamdesTool		computational_linguistics	Dieter Van Uytvanck	Tool as used by BAMDES (for th...	27 October 2010 15:48:54	0
BamdesWrittenCorpus		computational_linguistics	Dieter Van Uytvanck	Written Corpus as used by BAM...	27 October 2010 15:59:28	0
Bedevaartbank			Folkert de Vriend	Profile for Bedevaartbank	30 July 2010 9:33:53	0
Boedelbank			Folkert de Vriend	Profile for Boedelbank	30 July 2010 9:36:25	0
cmdi-virtual-collection			CLARIN (cmdi@clarin.eu)	A component for describing per...	21 April 2010 16:18:30	0
collection			Matej Durco	minimal collection profile (start...	15 October 2010 20:58:37	0



Detail info | Academia x CLARIN VLO - "DE ACHTTIEN ... x http://catalog.c...istry/index.jsp# x Adelheid :: x +

catalog.clarin.eu/adelheid/main/ Google

Meest bezocht Aan de slag

Adelheid

This webservice annotates historical Dutch texts with word class tags and lemmas.

[Start a new Project](#)

Project ID:

Projects

Show entries Search:

Project ID	▲ Last changed
No data available in table	

Showing 0 to 0 of 0 entries

Powered by CLAM v0.7.6 - Computational Linguistics Application Mediator
by Maarten van Gompel
[Induction of Linguistic Knowledge Research Group](#), [Tilburg University](#)

CLAM is funded under [CLARIN-NL](#) projects [TICCLops](#) (09-011), coordinated by Martin Reynaert, and TTNW, WP1 and WP2, respectively coordinated by Martin Reynaert and Antal van den Bosch.



Detail info | Academia x CLARIN VLO - WFT x Historische Woordenbo... x http://catalog....try/index.jsp# x Adelheid :: x handle persistent identif... x

catalog.clarin.eu/vlo/record?5&docId=hdl_58_10032_47_d5fd1beb2eb754afe3ff3ea550649e69&fq=nationalProject:CLARIN-NL&fq=collection:INL+Ta

Meest bezocht Aan de slag

- Return Page

RESOURCE S

- Hide all metadata fields

Reference to resource	Resource description	Resource MIME type	Resource Proxy ID
http://hdl.handle.net/10032/dae95348ec916dbe27bea418e0cd8545	Resource	text/html	resource
http://hdl.handle.net/10032/05ce0c0cb68b4770619040f962ca8bfd	SearchPage	text/html	SearchPage
http://hdl.handle.net/10032/12824827a77b9602cc66840a62aedf43	LandingPage	text/html	LandingPage

METADATA CONTENT

- LexicalResourceProfile
 - LexicalResource
 - GeneralInfo
 - Name: WFT
 - Name: Wurdboek fan de Fryske taal online in de GTB
 - Version: 1.0
 - Owner: Fryske Akademy



NL

- [Return Page](#)