

LINGUISTICS WITH CLARIN STORING RESOURCES IN CLARIN

Jan Odijk LOT Winterschool Amsterdam, 2015-01-16

1



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 (<u>http://dev.clarin.nl</u>)
 - CLARIN-NL website <u>Centres</u> 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



- 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



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



- 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



- 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



- 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 <u>commonly used profiles and components</u>
 - A metadata editor: ARBIL



- Flexibility requires explicit semantics!
 - The CLARIN infrastructure must `know' what you mean with your metadata elements
 - Otherwise it cannot use faceted browsing in the <u>VLO</u> or the <u>Meertens Metadata Search Engine</u>



- 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 was <u>ISOCAT</u>
 - Example Data Category in ISOCAT
 - Example Link to ISOCAT in element definition



- Explicit semantics (2):
 - RELCAT alpha version
 - For relations between data categories
 - SCHEMACAT alpha version
 - For describing schemas of resources
 - Recently changed to CLARIN Concept Registry
 - <u>CCR Browser</u>
 - an OpenSKOS instance hosted by Meertens Institute
 - <u>CLAVAS</u> <u>Vocabulary Service</u>
 - Interface to other data category registries and vocabularies
 - ISO 639-3 Language codes



- Attend dedicated tutorials on CMDI and ISOCAT/CCR
 - <u>Regularly organized</u> 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



• 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
 - Period of the language
 - Standard language or dialects
 - Title
 - Version



- Add properties that are important from a linguistic perspective:
 - Which linguistic annotations does it contain
 - For which subdisciplines of linguistics is it mostly relevant
 - Does it involve text, audio, video, or databases?
- For tools
 - What are the linguistic properties of its input and output (which annotations, which annotation schemata, pos-tags, formats, etc)



- Live Version v. exchange/archive version – E.g. lexicon
 - in <u>Lexical Markup Framework</u> compatible <u>XML</u> 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



Software

- Desktop tool
- Desktop application
- Web service
 - <u>SOAP</u> / <u>REST</u> , use <u>CLAM</u> 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
 - <u>example</u>
- Make Metadata harvestable
 - <u>OAI-PMH</u> protocol
 - Open Archives Initiative Protocol for Metadata Harvesting



WHAT THE CLARIN CENTRE MUST DO

- Store the data in the centre's repository
 - <u>LAMUS</u> (the Language Archive) and its documentation <u>online</u> or as <u>PDF</u>
 - 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
- <u>CLARIAH</u> (successor project)
 - Will continue these curation activities



Thanks for your attention!





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



CLARIN-NL D CENTRES



• 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)
 - <u>NIBG data via the VLO</u>
- Utrecht University Library (UBU)
 - Digital books, articles
 - UBU data via the VLO

				Prc)FI	LE	E: E	ΞΧΑ	MPI	-)
IPROSLA_Deliv	ver C Data, To	ols, http://38110# 🗙 💽	GrETEL for C Frog :: jano	Lagere econ	Nieuw tabbla	ad Niei	euw tabblad	🚺 Universiteit	JoshiFest: "'J	+	- F	1 🕅
🗲 🗰 🖉 ca	atalog. clarin.eu /ds/Co	omponentRegistry/?item=clarin.eu:	cr1:c_1271859438110#app=cd01&e2ee	-selectedIndex=0			⊽ C ⁱ 8 + J	ohan Bolhuis	٩	☆自	↓ 🏠	≡
A Meest bezocht	t 📋 Aan de slag											
Clarin Compone	ent Registry											help
Compon	ent Browse	er								User: and	onymous	login
Profiles (Components		Public space	e V								
Create new	Edit as new	Import						Q filter.		Showing	157 of 157	
Name		Group Name	Domain Name	Creator		Description	n	Registration	Date		Commen	Its
GTRP_sub_loca	ition			Folkert de Vriend		Profile for lor	cations subdivis	ion 30 July 2010	10:05:03		0	•
imdi-corpus				Alexander Koenig	1	IMDI corpus	profile	02 June 2010	15:16:26		0	
imdi-session				CLARIN (cmdi@clarin.e	eu)	IMDI metada	ata (type: sessi	on) 21 April 2010	16:19:43		3	
LexicalResource	eProfile			Laura van Eerten		a profile for (describing a lex	ical 26 April 2010	12:13:06		2	
			LexicalResourceProfile	view	xml	Cor	omments (2)					
	Name: LexicalRe	source										-
	Group Name: cla	rin-nl/lexicalresource										
	Description: A co dicti	mponent for describing lexical res onaries and semantic networks.	ources like different types of word lis	sts,								
	Component: cm	di-generalinfo										_
	Number of occurre	nces: 1-1										Ξ
	Component: Orig	ginLocation										
	Number of occurre	nces: 0 - unbounded										F
	Compo	nent: cmdi-location										
	Numbe	r of occurrences: 1 - 1										
	Number of occurrent	ar-project										
	Component:	di-creators										
	Number of occurre	nces: 0 - 1										
	Componenti	r)										•
												4 * 1 1

IPROSLA_Deliver	Data, Tools,	http://38110# ×	GrETEL for C	Frog :: jano	D Lagere ec	on Nieu	w tabblad	Nieuw tabblad	🚺 Universiteit	JoshiFest: "J	+		e X
🗲 🗰 🖉 catalog.clarin	eu/ds/Compone	ntRegistry/?item=clari	n.eu:cr1:c_12718594381:	LO#app=cd01&	e2ee-selectedIndex=1			⊽ C" [🖁	🔻 Johan Bolhuis	م	☆ 自	+ 1	r E
Meest bezocht Aan d	le slag												
Clarin Component Registr	v												help
Component Br	rowser										User: a	nonvmou	us login
component bi				Public	snare V							,	
Profiles Component	ts			Public .	space v								
Create new Edit a	as new Imp	ort					-		Q filter		Showing	9 872 of	872 🔊
Name	Group N	lame	Domain Name		Creator		Descriptio	n	Registration Date			Comn	nents
cmdi-language	clarin				CLARIN (cmdi@clari	n.eu)	Component	for describing a c	21 April 2010 16:17:2	1		0	_
cmdi-location	clarin				CLARIN (cmdi@clari	n.eu)	Component	for describing a c	21 April 2010 16:17:2	1		0	
cmdi-mimetype	clarin				CLARIN (cmdi@clari	n.eu)	List of frequ	ently used mime t	. 21 April 2010 16:17:1	8		3	
Creation	CLARIN				nalida		Component	contains informati	. 31 January 2013 12:4	8:14		0	•
			cmdi-	location	view	xml	0	Comments (0)					
Name: Location													
Group Name: clarin													
Description: Componer	nt for describing a	a certain location (ad	dress, region, country,	continent)									
Element: Address str	ring												
ConceptLink:	http://www.iso	cat.org/datcat/DC-25	28										
Number of or	courrences: 0 -	1											
ConceptLink:	http://www.iso	cat.org/datcat/DC-25	33										
Number of or	Number of occurrences: 0 - 1												
Component: iso-count	ry												
Number of occurrences:	0-1												
Component: iso-continent													
Number of occurrences:	0 - 1												

CLARIN 💩

CLARIN C					PLOD
IPROSLA_Deliv C Data, Tool http:/stry/# × 🛐 GrE	TEL for Frog :: jano	Lagere eco Nieuw tabblad	Nieuw tabblad	🚺 Universitei 💽 JoshiFest: " 💽 CLARIN	I-N + 🗖 🖻 🖻
Certailog.clarin.eu/ds/ComponentRegistry/#app=cd01&e	2ee-selectedIndex=1		V	7 C 🛛 🖁 🔻 Johan Bolhuis	오☆ 🖻 🖡 🏦 🚍
Meest bezocht Aan de slag					
Clarin Component Registry					help
Component Browser				User: (Ddijk101@soliscom.uu.nl settings
Parfiler Commente	Pu	blic space 🛛 🔻			
Create new Edit as new Import				Q, filter	Showing 872 of 872
Name Group Name	Domain Name	Creator	Description	Registration Date	Comments
WrittenResource CLARIN-D Version 1.2	phonetics	71A3DD8C90E89334@imu.de	Group of recordings o	informati 05 February 2013 14:12:47	0
author CLARIN-D: DTA-Basisformat	phonetics	Axel Herold	the author of a textua	al resource 16 September 2013 15:59:33	°
	fileDass		Comments (0)		
Group Name: CLABIN D. Diversisionmat Description: bibliographical information on autho Element: title string ConceptLink: http://www.isocat.org/dz Documentation: the title of a work act DisplayPriority: 1 Number of occurrences: 1 - unbounded Multilingual: false AttributeList: type main v level a v decimal Component: author Number of occurrences: 0 - unbounded Component: editor Number of occurrences: 0 - unbounded	fileDesc r/editor and title of a text atcat/DC-2545 cording to the classification in ed	view xml	Comments (0)		

CLARIN TO		IS	50		TC	C)
Communication Matching Matching	Lagere ec Nieuw tabblad	Nieuw tabblad	🚺 Universit,	🖸 JoshiFes	t: 🧊 ISOca >	Arbil Th	ISOcat >	+ •	- đ	
A https://catalog.clarin.eu/isocat/interface/index.html					⊽ C' 🖁 -	Johan Bolhuis	٩	☆ 自	•	≡
Aan de slag										
J isocat							🚷 Wel	come Jan Oo	lijk 🔹 🔞 H	ielp *
2545 🛍 🖏 😒	🕞 🖨 🗉 🔯 🧟 🕼	i i		search	result for '2545'				₩ 🕄	
🖻 🏦 My Workspace	# Name	Version	Administration	n st Registration	stati Check Type	Owned by	Scope			-
B Private	2545 resource title	1:0	private	private	🖌 open	Athens Core	public			
JOTest										
🗄 📌 Shared										
🗄 🧑 Public										
1										-
	+ ⊜ ∪ ₪ ∕	* 🕉 * 🐨 🖬	•	reso	urce title - 1:0					
				2. Descri	ption Section					
	Profile	Metadata								
	[-] 2.1 English Language Sec	tion								
	Language	English (en)								Ξ
	2.1.1 Name Section									
	Name Name Clature	resource title								- []
	2 1.2 Definition Section	admitted name								
	Definition	The title is the c	omplete title of	the resource with	out any abbreviation	ns.				
	Source	CLARIN								
	[+] 2.2 Catalan Language Se	ction								
	rui 2.2 Creation Language Fr	sction		_						

CLAR			LIN	1K	то	ISC	C		DC		0)
IPROSLA_Deliv	C Data, To	ool http:/stry/# 🗙 🋐 Gr	ETEL for Frog :: jano	Lagere eco	Nieuw tabblad	Nieuw tabblad	🚺 Univers	itei 💽 JoshiFest: ".	CLARIN-N	+		
(¢) 🖬 🖉 cata	alog. clarin.eu /ds/	/ComponentRegistry/#app=cd01&	e2ee-selectedIndex=1				⊽ C ⁱ 8 •	r Johan Bolhuis	م	☆ 自	•	Ξ
Meest bezocht	🗌 Aan de slag											
Clarin Componen	t Registry											help
Compone	ent Brows	ser							User: Odijk:	101@solisco	m.uu.nl se	ttings
Profiles Co	manante		Pul	blic space 🔻								
Create new	Edit as new	Import						Q filter	1	Showing	g 872 of 872	2
Name		Group Name	Domain Name	Creator		Description		Registration Date			Commen	its
media-session		CLARIN-D Version 1.2	phonetics	71A3DD8C90E	89334@lmu.de	Group of recordings	s of at leas	05 February 2013 14:1	12:47		0	-
WrittenResource		CLARIN-D Version 1.2	phonetics	71A3DD8C90E	89334@lmu.de	Component contair	ns informati	05 February 2013 14:1	13:45		0	=1
author		CLARIN-D: DTA-Basisformat		Axel Herold		the author of a tex	tual resource	16 September 2013 1	5:59:33		0	•
			fileDesc	view	 xml	Comments ((0)					
(Group Name: (CLARIN-D: DTA-Basisformat										•
1	Description: bi	ibliographical information on auth	or/editor and title of a text									
	Element: title		dataat/DC-2545									
	Docu	<u>imentation</u> : the title of a work a	according to the classification in	etype								_
	Displ	layPriority: 1										Ē
	Num	ber of occurrences: 1 - unbound	ded									
	Multi Attributel ist:	ilingual: false										
	type	main										
	level											
	n	decimal										
	Component: a	author										
	Number of occur	rrences: 0 - unbounded										
	Component: e	editor										
1	Number of occur	rrences: 0 - unbounded										Ţ
•		.			11111							ŀ



- 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



- 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: <u>Shibboleth</u>
 - 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 🗙	Detail info Academia 🛛 🗙	C Data, Tools, Demonstrat 🗙	http://catalogentRegistry/# 🗙	http://catalogentRegistry/# ×	W Shibboleth (Internet2) 🗴 🕂		P X
Catalog. Harin.eu/ds/Com	Dana ponentRegistry/#			⊽ C 🛛 🗧 Google	2 ☆ 自	🖡 🏠	≡
Meest bezocht 🗍 Aan de slag	lage						
Clarin Component Registry							help
Component Browser					User: a	anonymous	login
Profiles Components		Public space	• •				
Create new Edit as new	Import				Q filter Showin	ig 157 of 157	7 🔊
Name	Group Name	Domain Name	Creator	Description	Registration Date	Commer	nts
AnnotationTool			Eric Sanders	Description of a tooladapted fro	09 February 2011 14:13:59	0	A
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	- 11
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	- 11
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	- 11
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	•





Universiteit Utrecht	
Solis-id	
Odijk101	
Wachtwoord	
•••••	
	Login
Wachtwoord vergeten?	
Meer informatie over Solis-id: medewerkers	/ <u>studenten</u>
Kijk ook op <u>www.uu.nl/ict</u>	🚟 English

CLARIN			LC	DGIN			2	
C Data, Tools, Demonstrators	s × http://catalog.cistr	y/index.jsp# × +					- đ	Σ
Catalog.clarin.eu/d	s/ComponentRegistry/index.j:	sp#		▼ C 8 + Google	م	/ ☆ 自		Ξ
Meest bezocht Aan de slag	3							
Component Brow	/SET	Public s	pace V	Description	Q filter	<101@soliscom	n.uu.nl sett	ings
AppotationTool	Group Name		Eric Sanders	Description	09 February 2011 14:12:59		0	2
ArthurianEiction		Other	Rik Hoekstra	Profile for Arthurian Firtion data	04 September 2012 14:50:05		0	
BamdesLexicalResource		computational linguistics	Dieter Van Uvtvanck	Lexical Resource as used by BA	27 October 2010 15:47:42		0	1
BamdesMultimodalCorpus		computational linguistics	Dieter Van Uytvanck	Oral Corpus as used by BAMDES	27 October 2010 16:00:41		0	ъ
BamdesOralCorpus		computational_linguistics	Dieter Van Uytvanck	Oral Corpus as used by BAMDES	27 October 2010 16:00:05		0	1
BamdesTool		computational_linguistics	Dieter Van Uytvanck	Tool as used by BAMDES (for th	27 October 2010 15:48:54		0	ъ
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	Т
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	Т

Matej Durco

collection

•

0

minimal collection profile (start... 15 October 2010 20:58:37

CLARIN (LOGIN	
🔲 Detail info Academia	CLARIN VLO - "DE ACHTTIEN http://catalog.cistry/index.jsp# Adelheid ::	
Catalog.clarin.eu/adelhe	id/main/ 🖉 🤁 🛛 🐨 Google	오☆ 🖨 🖶 🎓 🚍
🔊 Meest bezocht 🗍 Aan de sla	9	
	Adelheid	
	This webservice annotates historical Dutch texts with word class tags and lemmas.	
	Start a new Project Project ID: Start project	
	Projects	
	Show 10 entries Se	earch:
	Project ID Last changed No data available in table	÷
	Showing 0 to 0 of 0 entries	First Previous Next Last
	Fowered by CLAM v0.7.6 - Computational Linguistics Application Mediator by Maarten van Gompel <u>Induction of Linguistic Knowledge Research Group</u> , <u>Tilburg University</u> CLAM is funded under <u>CLARIN-NL</u> projects TICCLops (09-011), coordinated by Martin Reynaert, and TTNWW, WP1 and WP2, respectively coordina Antal van den Bosch.	ated by Martin Reynaert and





• Return Page