Deposition of data at the Dutch Language Institute

The Repository “CLARIN INT Centre” gives access to language resources and tools from the Dutch Language Institute (INT) and other organisations.

Collection Policy

An important part of the mission of the Dutch Language Institute (Instituut voor de Nederlandse Taal – INT) is providing access to Dutch source material in the form of historical and contemporary corpora, dictionaries, lexical digital databases, grammars, including the required technical tools. Apart from the data we produce ourselves we also accept resources from other organisations.

Scope of the collection

We are interested in digesting all kind of resources that pertain to the Dutch language. However, we use a number of criteria for selection:

- Size: Is the size of the resource of sufficient interest for scientific research and/or commercial exploitation.
- Non-Redundancy: Is the resource sufficiently unique and cannot be found elsewhere.
- Quality: Is the resource well documented and the data clearly organized.
- Sustainability: Is the data provided in a sustainable format (See ‘Supported Data formats’).

Deposition Process

If you would like to deposit your data at het INT then send a message to ‘servicedesk@ivdnt.org’. We will then check with you whether the data fits in our collection. We might ask you for some additional information, and if necessary, will ask you to sign an agreement (See ‘Deposition Agreement’). The data will be safely archived at the INT and we will create a product page for your resource and a PID that can be used for referencing.

Guidelines for deposition

Data presented for deposition need to be supplied with all information that is essential for sustainable data management and future use.

Moreover, the data should be provided in standard formats. See ‘supported data formats’.
For archiving purposes, a minimum set of metadata in valid CMDI format either needs to be provided by the data producer or is extracted by INT from data and documentation.

Data producers are encouraged to supply additional documentation of the data or links to publications (using persistent identifiers) about the data. The publications are stored in the repository (given that permission is granted) while the documentation is archived in the OAI/PMH accessible data repository.

Responsibilities

The data producer will always remain the proprietor of the data. CLARIN INT Centre receives a copy of which it must take good care, according to the terms of the license contract and the terms and conditions for use.

The CLARIN INT Centre also makes copies, for example for the benefit of backup and looks after them well.

In case of an emergency we are able to build up an entire new database composed of all files we backed up and stored safely at another location. Quarterly and yearly tapes are stored at http://backupned.nl.

Preservation

To ensure the integrity of the data sets, for every deposited file a checksum (md5 type) is made which allows us to check for defects of the data over the years.

Once deposited, files in data sets are never changed and only minor changes to the metadata are allowed. For example: correction of spelling, minor changes in documentation, additional documentation added. Changes to the data themselves will be issued as a new version of the dataset, which will obtain a new persistent identifier. These changes are only made in close collaboration with the producer of the dataset.

Authenticity

Data producers hand over the materials to us. We do not change the data, except by adding metadata if required.

If applicable, we create collection-level objects which provide a context for the embedded data sets. The repository maintains links to other relevant materials (e.g. articles, theses, documentation, related data) and to software and tools that have been used in production of the data, if applicable. The identity of a depositor is ensured by the required login using CLARIN SpF for identification.

Deposition Agreement

If the product is not released with an open license, a signed deposition agreement is required. That can be our own standard agreement, of an agreement from the CLARIN Licensing Framework (https://www.clarin.eu/content/clarin-licensing-framework) or any tailor-made agreement.

Supported Data Formats

Type	**Preferred**		**Acceptable**
Text Documents	PDF/A (.pdf)		Open Office XML (.docx)
	ODT (.odt)		Rich Text File (.rtf)
			PDF other than PDF/A (.pdf)
Plain text	Unicode text (.txt)		ASCII (.txt)
Markup languages	XML (.xml)		SGML (.sgml)
	HTML (.html)		Wikitext
	XHTML (.html)		
	Related files like .css, .xslt, .js, etc.		
Spreadsheets	ODS (.ods)		Office Open XML Work-book (.xlsx)
	CSV (.csv)		
Databases	SQL (.sql)		
	Open Document Database (.odb)		
Raster Images	JPEG (.jpg, .jpeg)		
	TIFF (.tiff)		
	PNG (.png)		
	JPEG 2000 (.jp2)		
Vector Images	SVG (.svg)		EPS (.eps)
			PostScript (.ps)
Audio	Broadcast Wave Format (.bwf)		WAVE (.wav)
	Material Exchange Format (.mxf)		MP3 (.mp3)
	Matroska Multimedia Contai-ner (.mka)		AAC (.aac, .m4a)
	FLAC (.flac)		AIFF (.aif, .aiff)
	OPUS (.opus)		OGG (.ogg)
Video	Material Exchange Format (.mxf)		MPEG-4 (.mp4)
	Matroska Multimedia Contai-ner (.mka)		MPEG-2 (.mpg)
			AVI
			QuickTime (.mov, .qt)
Geographical Informations Sys-tems (GIS)	GML (.gml)		Mapinfo (.tab, related files)
	Mapinfo Interchange Format MIF/MID (.mif, .mid)		
RDF	RDF/XML (.rdf)		
	Trig (.trig)		
	Turtle (.ttl)		
	Ntriples (.nt)		
	JSON-LD		