



Metadata standards in the language sciences and how to implement them

Table of Contents

1	Introduction	2
2	CMDI	3
2.1	Introduction.....	3
2.2	Repositories that are compliant with CMDI.....	7
2.3	COMEDI – an online metadata editor for CMDI.....	8
2.4	Other CMDI tools.....	13
3	OLAC	14
4	TEI	15
4.1	Introduction.....	15
4.2	Tools	17

1 Introduction

As a researcher, it is imperative to adequately document the research data you collect and/or create, in order to enhance the interpretability and re-usability of your data. Put differently, documentation is one of the prerequisites to make your data [FAIR](#) (= Findable, Accessible, Interoperable and Reusable). In short, data documentation and metadata correspond to:

- all the information you need to easily **find the data**, cf. creator of the data, title, affiliation, subject, link with associated publication(s) etc.
- all the information you need to be able to **reproduce** the findings based on the research data, cf. for example a codebook that explains the different variable names and their levels etc.

Documentation is more effective when it is captured in the structured format of a metadata standard. Concretely, information such as the name and affiliation of the data creator is inserted into specific fields of a structured format (for example, an XML-file), so that an “information system” can efficiently read and communicate this information. Thanks to **machine-readable** metadata standards, metadata can be efficiently **exchanged between information systems**.

Metadata standards can be generic ([Dublin Core](#), [DataCite Metadata Schema](#) etc.) or discipline-specific. In this document, our focus is on **discipline-specific** metadata standards used in the language sciences. Different metadata standards and how to implement them will be discussed:

- **CMDI** (cf. Section 2)
- **OLAC** (cf. Section 3)
- **TEI** (cf. Section 4)

Other metadata standards from the social sciences can also be interesting to take into account. For example, the **DDI metadata standard** ([Data Documentation Initiative](#)) lends itself well to variable descriptions.

2 CMDI

In this section, we present the **CMDI metadata standard**, advocated by the European CLARIN infrastructure. The remainder of this section is structured as follows:

- Section 2.1 provides a general introduction of the CMDI standard.
- In Section 2.2, some examples of repositories that are compliant with CMDI are presented.
- Next, Section 2.3 discusses COMEDI, an online metadata editor that allows researchers to create CMDI metadata.
- Finally, Section 2.4 briefly comments on a few other CMDI tools.

2.1 Introduction

The CLARIN website describes the CMDI metadata standard as follows:

Metadata for language resources and tools exists in a multitude of formats. Often these descriptions contain specialized information for a specific research community (e.g. TEI headers for text, [IMDI](#) for multimedia collections).

To overcome this dispersion, CLARIN has initiated the Component Metadata Infrastructure ([CMDI](#)). It provides a framework to describe and reuse metadata blueprints. Description building blocks (“components”, which include field definitions) can be grouped into a ready-made description format (a “profile”). Both are stored and shared with other users in the [Component Registry](#) to promote reuse. Each metadata record is then expressed as an XML file, including a link to the profile on which it is based.

The CMDI approach combines architectural freedom when modelling the metadata with powerful exploration and search possibilities over a broad range of language resources.

To date, there are two supported versions of CLARIN's component metadata framework: CMDI 1.1 and CMDI 1.2. They are not interchangeable, but CMDI 1.1 metadata can easily be converted into CMDI 1.2. More information about the changes in CMDI 1.2, its place in the infrastructure and how it may affect you as a user, metadata creator, modeller, repository owner or software developer can be found on the [CMDI 1.2 page](#). The general information in this page applies to both CMDI 1.1 and CMDI 1.2.

Source: <https://www.clarin.eu/content/component-metadata>

Thanks to this common metadata standard, it becomes possible to make language resources searchable via the overarching [Virtual Language Observatory](#) ([source](#)).

Contrary to other metadata standards that are more “static”, CMDI offers some degree of flexibility. This is because researchers have a range of CMDI “metadata profiles” to choose from, for example a profile specifically designed for text corpora and another one that is used for speech corpora. If none of the existing profiles suffices to document your data, there is also a possibility to develop your own profile according to the CMDI-standard. Each of these profiles is composed of smaller building blocks called “components”. This enables re-use of certain components across various profiles, whereas other

more specific components are restricted to a particular profile. The difference between profiles and components is also explained in Figure 1:

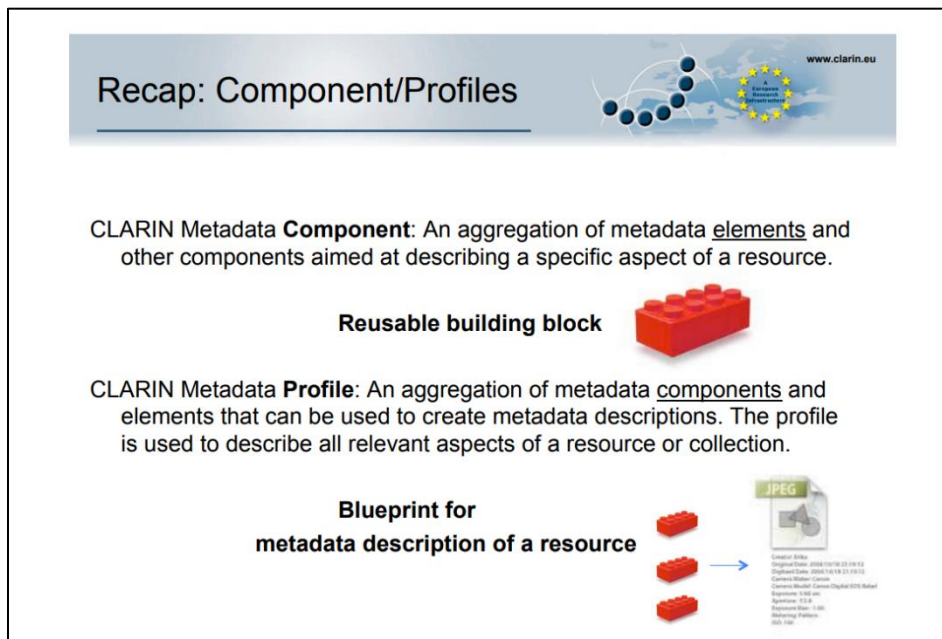


Figure 1 Components and Profiles ([source](#))

Via [this link](#), you can access the complete inventory of already registered profiles and components in the **CMDI Component Registry**. Figure 2 below contains a screenshot of the CMDI Component Registry, with additional information on the most important aspects of the registry. More information about the Component Registry can be found [here](#).

The screenshot displays the 'CMDI Component Registry' interface. At the top, there is a navigation bar with a 'Public space' dropdown, a 'Profiles' dropdown (highlighted with a red box), and buttons for '+ New', 'Edit as new', 'Delete', and 'Status'. A search bar and 'Showing 179 of 2' are also visible. Below this is a table listing profiles:

Name	Group Name	Domain Name	Creator	Description	Registration Date	Comments
CourseProfile	NaLiDa	Other	nalida	A CMDI profile for courses, for example publishing course materia...	2019-10-31	0
ExperimentProfile	NaLiDa		nalida	This CMDI profile can be used for describing psychological studie...	2019-10-31	0
ToolProfile	NaLiDa		nalida	A CMDI profile for tools.	2019-10-31	0
SpeechCorpusProfile	NaLiDa		nalida	Description for speech corpora with recordings and transcription(s)	2019-10-31	0

The 'SpeechCorpusProfile' row is highlighted in blue and has a red box around it. A red arrow points from the 'Profiles' dropdown to this row. Another red arrow points from the 'Show info' button in the right-hand menu to the details view below. The details view for 'SpeechCorpusProfile' shows the following components:

- Component: **GeneralInfo** [1 - 1]
- Component: **Project** [0 - 1]
- Component: **Publications** [0 - 1]
- Component: **Creation** [1 - 1]
- Component: **Documentations** [0 - 1]
- Component: **SpeechCorpusContext** [1 - 1]
- Component: **Access** [1 - 1]
- Component: **ResourceProxyListInfo** [1 - 1]

At the bottom left, it says 'Version 2.3.1'. At the bottom right, there is a 'CLARIN' logo and a 'Co' icon.

Possibility to obtain a list of "Profiles" or a list of "Components" (= the building blocks of a Profile).

Show info
Download XML (CMDI 1.1)
Download XML (CMDI 1.2)
Download XSD (CMDI 1.1)
Download XSD (CMDI 1.2)

Overview of the different "Components" that together constitute the "SpeechCorpusProfile".

Possibility to download the (i) XML or the (ii) XML Schema Definition of this profile, for two different versions of the CDMI standard (1.1 and 1.2).

Figure 2 CMDI Component Registry with "Profiles"

If you click on one of the components of the profile, you obtain an overview of the different “elements” that form the component. For example, Figure 3 shows the elements of the component “SpeechCorpus”.

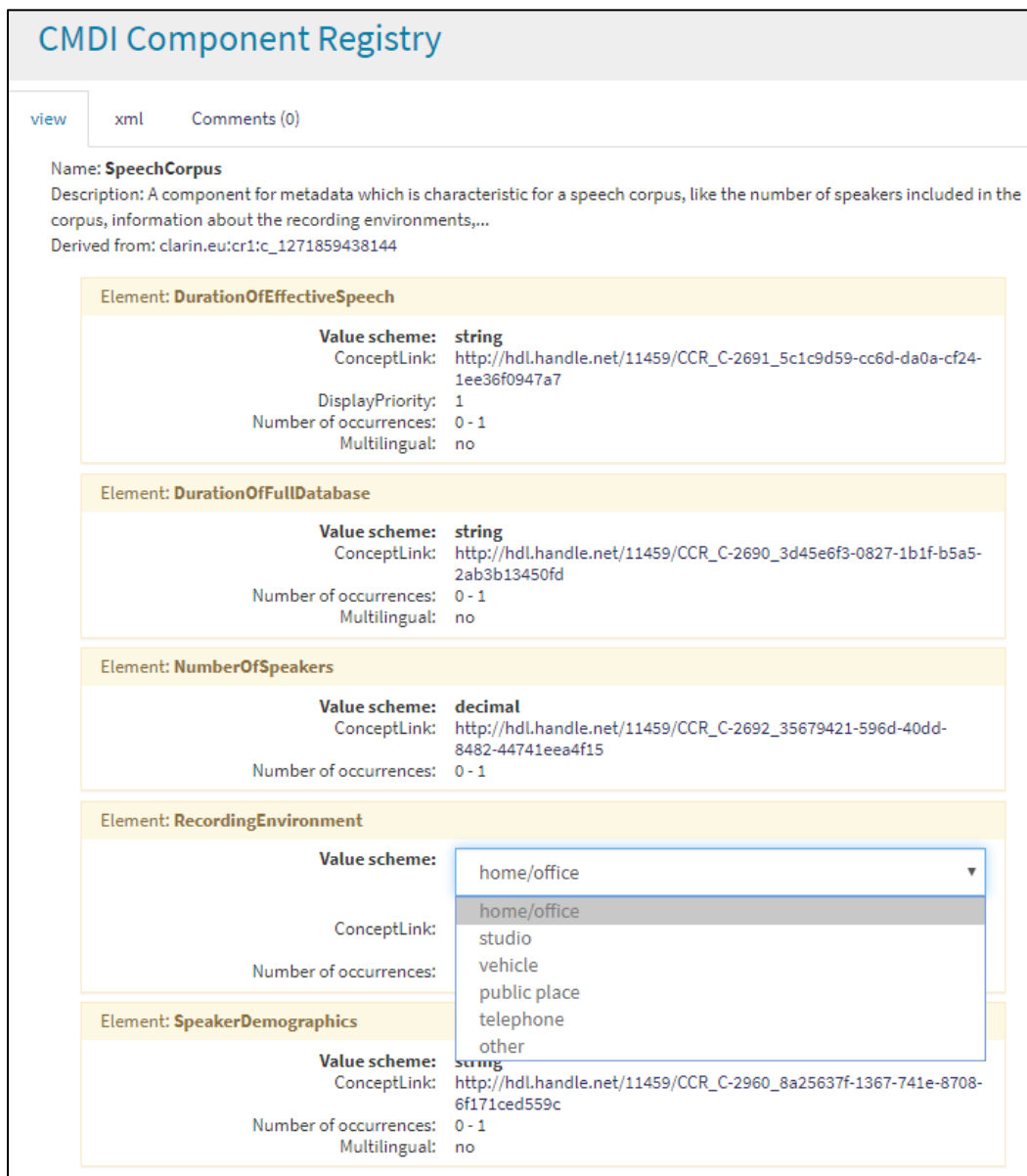


Figure 3 One of the components (“SpeechCorpus”) of the SpeechCorpusProfile, with different elements (“DurationOfEffectiveSpeech”, “RecordingEnvironment” etc.)

Some elements can be specified in a free-text field, whereas other elements require a specific value from a controlled vocabulary list (cf. “RecordingEnvironment” in Figure 3).

Finally, every element in a component has a link to a well-defined semantic concept registered in the [Concept Registry](#) (cf. “ConceptLink” in Figure 3) or should refer to other established metadata schemes, such as the Dublin Core Metadata Set. This ensures semantic [interoperability](#).

Extra details about CMDI can be found in the following resources:

- [The Component Metadata Initiative \(CMDI\) - DWDS](#)
- [Describing Research Data with CMDI - Challenges to Establish Contact with Linked Open Data \(2020\)](#)
- [Short Guide about Component Metadata - CLARIN \(2009\)](#)

2.2 Repositories that are compliant with CMDI

In order to illustrate the usefulness of the CMDI metadata standard, we present a series of repositories that actively promote the standard. These repositories are part of the larger CLARIN infrastructure and [contribute CMDI metadata to the general CLARIN metadata portal](#), the [Virtual Language Observatory](#). Via the Virtual Language Observatory, researchers can find resources and re-use them for their own research purposes.

The following non-exhaustive list of repositories is compatible with the CMDI metadata standard:

Repository	Information about data deposition and required metadata
Dutch Language Institute (INT)	https://portal.clarin.inl.nl/doc/information about deposition INT.pdf
The Language Archive	https://archive.mpi.nl/tla/deposit-manual-tla
A Resource Centre for Humanities Related Research in Austria	https://arce.acdh.oeaw.ac.at/browser/formats-filenames-and-metadata
CLARIN Centre at the University of Copenhagen	https://repository.clarin.dk/repository/xmlui/page/metadata
CLARINO Bergen Centre	https://repo.clarino.uib.no/xmlui/page/metadata
LINDAT	https://lindat.mff.cuni.cz/repository/xmlui/page/metadata
Hamburger Zentrum für Sprachkorpora	https://corpora.uni-hamburg.de/hzsk/en/corpus-hosting

Table 1 Examples of repositories that are compatible with the CMDI metadata standard

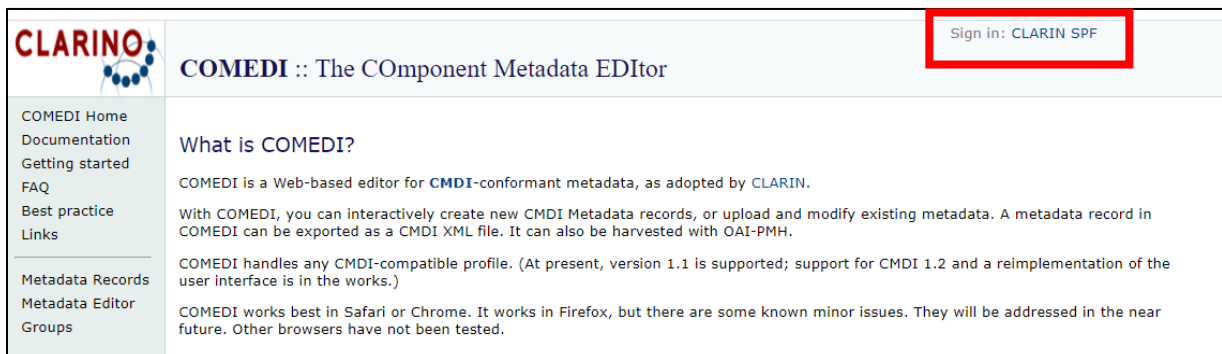
CLARIN also offers the option to register CMDI metadata about data that are available “elsewhere” via the [Language Resource Inventory](#). This option can be used if you have, for example, a project website with data that you want to make findable through the [Virtual Language Observatory](#). In this way, it is not necessary to upload the data themselves to a CLARIN repository. However, be aware that it is in general better to upload the data to a trustworthy repository, because this offers more guarantees for

sustainable long-term preservation. Self-maintained archiving solutions are often more subject to deprecation over time. Of course, this does not mean that a custom-made database cannot be a valid option if the data you are working with require such a custom-made solution. But always check already existing archiving options before resorting to such a solution!

2.3 COMEDI – an online metadata editor for CMDI

If the data repository where you archive your data does not offer any tools to generate CMDI-compliant metadata, you can use [COMEDI](#) – an online metadata editor for CMDI – to (1) select an appropriate “CMDI-profile” for your data, (2) fill out the template of the selected profile in the web-based editor and (3) save the metadata in xml-format. See [this link](#) for an example of such an xml-file. More information on how to get started with COMEDI can be found [here](#).

In order to gain access to the COMEDI metadata editor, you have to login by clicking on the “CLARIN SPF”-label in the top right corner of [this webpage](#), as highlighted in Figure 4.



The screenshot shows the COMEDI website interface. At the top left is the CLARINO logo. The main header reads "COMEDI :: The COmponent Metadata EDItor". In the top right corner, there is a button labeled "Sign in: CLARIN SPF" which is highlighted with a red rectangular box. A left-hand navigation menu contains links for "COMEDI Home", "Documentation", "Getting started", "FAQ", "Best practice Links", "Metadata Records", "Metadata Editor", and "Groups". The main content area is titled "What is COMEDI?" and contains introductory text about the tool's purpose and supported browsers.

Figure 4 Login via CLARIN SPF

Next, you search for your research organization via the “Home organisation list” and click on it, see Figure 5 below.

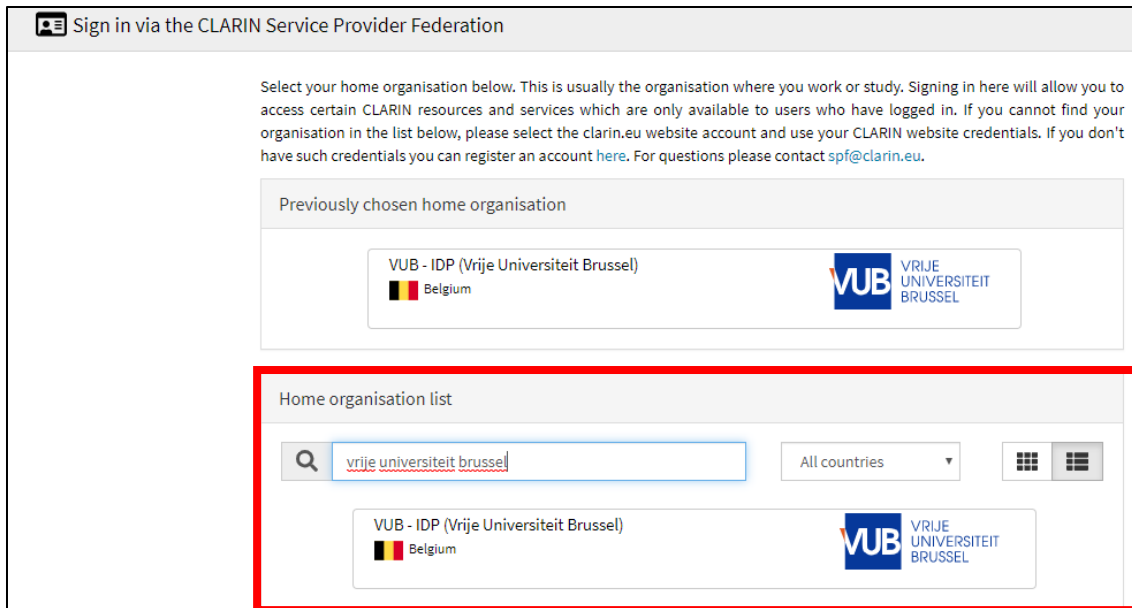


Figure 5 Sign in via the CLARIN Service Provider Federation

After you have authenticated yourself via your credentials, you are redirected back to the COMEDI website and it is possible to use the COMEDI metadata editor. Click on “Metadata Records” in the left column and you should obtain a screen as visualized in Figure 6.

CLARINO Sign out (NIEK VAN WETTERE)

COMEDI :: Metadata Records

COMEDI Home
Documentation
Getting started
FAQ
Best practice
Links
Metadata Records
Metadata Editor
Groups

Create a metadata record

Add a resource metadata record, starting from an empty profile

Choose profile:

... or provide a valid CMDI Profile ID:

Please provide an identifying name for the resource. This name will be used internally. Please keep the identifier reasonably short. Avoid spaces, tabs etc.

Identifier:

Upload a metadata record

Upload an existing syntactically valid CMDI 1.1 metadata record. The file name (without extension) will be used as identifier.

Choose file: No file chosen

... or batch upload metadata files using the REST API. See the documentation for details.
(session-index=_19e79b48b83ee0fcd14fe8a5993d4af5)

Clone a metadata record

Clone an existing metadata record, giving it a new identifier. The Self link URL will be cleared.

Choose an existing record:

Provide an identifying name for the cloned record.

Identifier:

Search in metadata records

Freetext search in existing metadata records. All matching records will be listed.

Figure 6

COMEDI – Metadata Records

In most cases, you will want to create a new metadata record based on an existing CMDI-profile. To do this, select an appropriate CMDI profile in the box “Create a metadata record”. It is also necessary to provide a self-chosen “Identifier”, i.e. an identifying name for the resource. Finally, click on “Go”, which leads you to a page where you can fill out the metadata template for the chosen CMDI profile, as shown in Figure 7.

Creation Date: 2020-05-04
Last Change Date: 2020-05-04T07:27:54Z

Status: in-progress ▼ | The metadata record is not valid. (Show details)

Access: Read/Write (owner: NIEK VAN WETTERE)
Groups: (Members of these groups have write access to the record)

Save a snapshot | Select a snapshot to revert to: 2020-05-04T07:27:54Z (current) ▼

Resources | **Resource common info** | TROLLing part

Component #1 Previous Next | View Edit

Resource common info [1/1] | [Remove content](#) | [Hide content](#)

Groups together general information, typically relevant for resources of any kind. Consists of (partially modified versions of) a subset of the common components in the 4 resourceInfo profiles by META-SHARE v3.0.

Resource type [1/1] | [CCR]

Indicates the type of the resource

corpus ▼

Identification info [1/1] | [Select existing Identification info](#) | [Hide content](#)

Groups together information needed to identify the resource. Derived from clarin.eu:cr1:c_1349361150743 (identificationInfo by META-SHARE 3.0)

Resource name [1/1] + | [CCR]

The full name by which the resource is known; the element can be repeated for the different language versions using the "lang" attribute to specify the language.

lang:

Description [1/1] + | [CCR]

Provides the description of the resource in prose

lang:

Figure 7 Example metadata template

Importantly, you have to switch from “View” to “Edit” (cf. the blue box in Figure 7) in order to be able to add metadata information in the dedicated fields of the web form. As long as there are required metadata fields that lack information, the status of the metadata record will indicate that “the metadata record is not valid” (cf. the green box in Figure 7). If you click on “Show details” next to the status, you will get an overview of the missing metadata information. Make sure to address all the different “components” that constitute the CMDI profile you selected.

Finally, to download the xml-file containing the metadata that were added by the researcher, you can click on the link of the self-chosen identifier, as indicated in Figure 8.



Figure 8 Identifier for metadata record

The xml content will be displayed in your web browser. In order to download it on your local laptop/pc, follow the instructions depicted in Figure 9 (for Google Chrome).

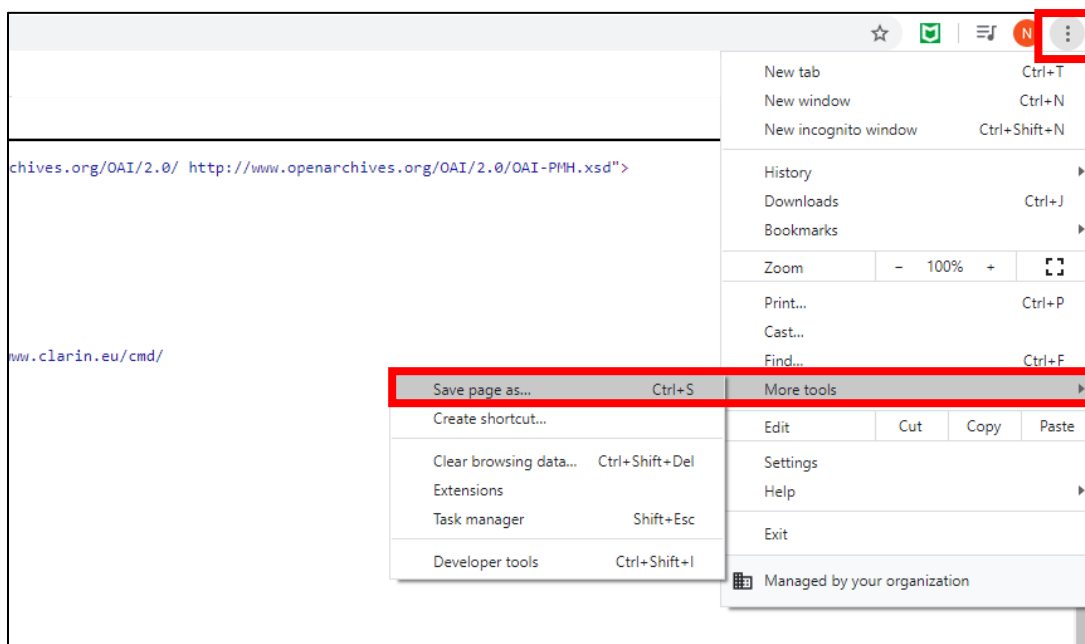
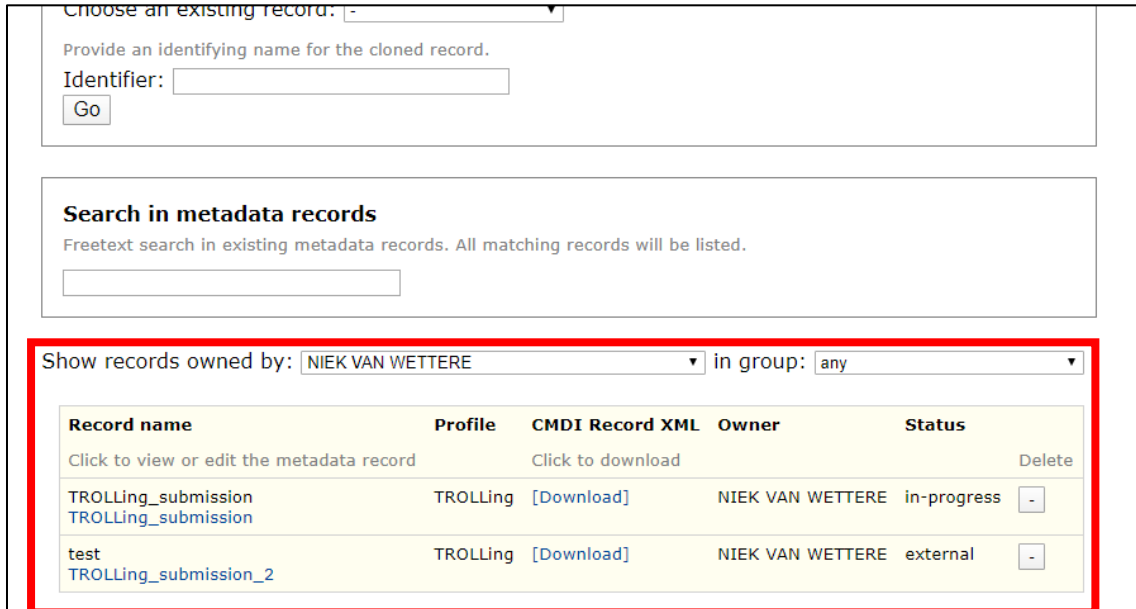


Figure 9 How to save the webpage in Google Chrome

If you want to check whether your downloaded xml file contains valid CMDI metadata, you can upload the xml file to this online [CMDI validator](#).

A different method to visualize the metadata is via html (= more readable than xml). In order to obtain this view, you can use the URL <http://clarino.uib.no/comedi/metadata-record?identifier=<identifier>>, where the last part <identifier> should be replaced by the name of the identifier for the metadata record in question (cf. red box in Figure 8).

If you want to come back to your metadata record at a later stage and edit it, this is possible. Upon login, you will find at the end of the webpage referred to in Figure 8 a list of the already created metadata records. This is exemplified in Figure 10.



Choose an existing record: [-]

Provide an identifying name for the cloned record.

Identifier:

Search in metadata records

Freertext search in existing metadata records. All matching records will be listed.

Show records owned by: in group:

Record name	Profile	CMDI Record XML	Owner	Status
Click to view or edit the metadata record		Click to download		Delete
TROLLing_submission TROLLing_submission	TROLLing	[Download]	NIEK VAN WETTERE	in-progress -
test TROLLing_submission_2	TROLLing	[Download]	NIEK VAN WETTERE	external -

Figure 10 Overview of already created metadata records

The CMDI Record XML download option directs you to a webpage with a view on the metadata.

2.4 Other CMDI tools

Besides COMEDI, there are a couple of other options to generate CMDI metadata, for example by means of a general XML editor (only for advanced users) or other tools specifically designed for CMDI (e.g. Arbil, no longer actively supported, or [CMDI Maker](#)). Some of these options are presented [here](#). New [tools](#) are being developed, such as [CMDIExplorer](#).

3 OLAC

An alternative to the CMDI metadata standard is the [OLAC metadata standard](#) (Open Language Archives Community). This standard is less specific than CMDI, but is also fairly wide-spread. Essentially, the OLAC metadata standard is largely based on the generic metadata standard Dublin Core, complemented with some extra descriptors for language resources:

The OLAC metadata standard [\[OLAC-Metaddata\]](#) follows the generic resource description standard known as "Qualified Dublin Core". In order to meet the specific needs of the language resources community, the OLAC metadata standard incorporates an extension mechanism that makes it possible to describe language resources with greater precision.

Source: <http://www.language-archives.org/REC/olac-extensions.html>

These extensions concern:

- code for discourse types: the genre of a language resource
- code for identifying languages
- code for linguistic field
- code for linguistic data types
- code for participant roles: the role of a Contributor in the development of a language resource is determined more precisely

Repositories that use OLAC include [CoCoON](#) and [Ortolang](#).

Finally, there is also an [OLAC Language Resource Catalog](#) that enables you to search through the OLAC metadata records.

4 TEI

Finally, we discuss the TEI metadata standard. A brief introduction is given in Section 4.1. Next, some useful tools are presented in Section 4.2. Given that TEI is a vast subject, our concise presentation will only be able to address some essential points.

4.1 Introduction

[TEI](#) (Text Encoding Initiative) is an encoding standard for the representation of texts in digital form, based on XML, a so-called mark-up language.¹ In general, it pertains to the following two aspects:

1. Metadata associated with the digital textual document(s) in its/their entirety. These metadata are contained in the [TEI Header](#). This part of the TEI metadata standard is comparable to the aforementioned metadata standards CMDI and OLAC, in the sense that it gives information about the data as a whole. The “classic” TEI Header can be enriched with elements that characterize [language corpora](#), [transcriptions of speech](#) etc.
2. Metadata inserted within the textual document (= poem, letter, interview transcript or other textual object), i.e. “inline annotation”, in order to structure the text and highlight/document certain aspects.

Both aspects are exemplified in Figure 11:

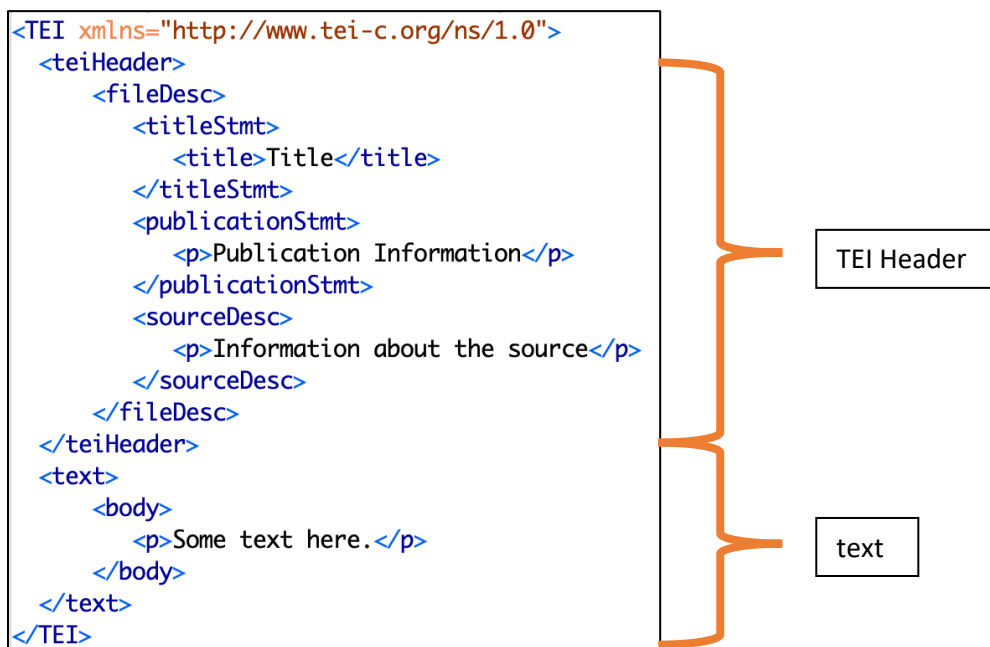


Figure 11 TEI Header and text ([source](#))

¹ Note that there also exist xml-based standards specifically for corpus annotation, such as the [TCF format](#).

Figure 12 illustrates how a text can be marked up to highlight certain elements, such as person names (cf. the tag “persName”), place names (cf. the tag “placeName”) and dates (cf. the tag “date”).

```
<text>
  <body>
    <p>The American poet, critic and publisher <persName>T S Eliot</persName> was born into a comfortable and historically distinguished family in <placeName>St. Louis</placeName>, <placeName>Missouri</placeName> in <date>1888</date>. He studied at Smith Academy and then Harvard, where he undertook an eclectic range of courses before settling on a BA in what would now be called Comparative Literature and an MA in English Literature.</p>
    <p>He spent a year studying at the Sorbonne in Paris, and returned to Harvard to work on the philosophy of consciousness. This can be seen as influential in his earlier poetry, much of which is concerned with fractured perceptions and mental illness.</p>
  </body>
</text>
```

Figure 12 TEI text ([source](#))

TEI also lends itself to more complex tasks, such as the encoding of parts of speech (POS-tagging), i.e. assigning morphosyntactic categories (adjective, noun etc.) to words. For example, in Figure 13, each word of the English sentence ‘We’re going on vacation to Brazil for a month!’ is POS-tagged (cf. the tag “pos=”).

The English sentence ‘We’re going to Brazil.’ tagged with the [CLAWS-5](#) tagset, arranged inline (with significant whitespace).

```
<p><w pos="PNP">We</w>
<w pos="VBB">'re</w> <w pos="VVG">going</w> <w pos="PRP">to</w> <w pos="NP0">Brazil</w><pc pos="PUN">.</pc></p>
```

The English sentence ‘We’re going on vacation to Brazil for a month!’ tagged with the [CLAWS-7](#) tagset and arranged sequentially.

```
<p>
<w pos="PPIS2">We</w>
<w pos="VBR">'re</w>
<w pos="VVG">going</w>
<w pos="II">on</w>
<w pos="NN1">vacation</w>
<w pos="II">to</w>
<w pos="NP1">Brazil</w>
<w pos="IF">for</w>
<w pos="AT1">a</w>
<w pos="NNT1">month</w>
<pc pos="!">!</pc>
</p>
```

Figure 13 POS-tagged sentence ([source](#))

In sum, thanks to the TEI encoding, extra information can be added to the raw text. Moreover, since the metadata was added in a structured, machine-readable way, the text can be **computationally parsed** in order to extract relevant strings and perform other data analysis operations.

The following online resources offer extra (basic) information on TEI and XML:

- [Text encoding and the Text Encoding Initiative - dariahTeach](#)
- [Why do we encode? - dariahTeach](#)
- Examples of possible use-cases, provided by the [Standardization Survival Kit](#).

Examples of data repositories that adhere to the TEI metadata standard are the [Oxford Text Archive](#), [LAUDATIO](#) and the [TextGrid Repository](#). Testimonials about the use of TEI in research can be found [here](#) and [here](#).

4.2 Tools

We present some tools that can help you to apply TEI in practice:

- A generic xml editor, for example [oXygen](#).
- The TEI standard is open to customization by its users. Some established customizations are provided [here](#). A tool that can be used to customize the standard is [Roma](#). More information on the tool can be found on this [webpage](#) (and [here](#)), and in this [video tutorial](#).
- The aforementioned CMDI Component Registry contains a couple of profiles that implement the TEI Header.
- The (German) Deutschen Textarchiv provides a [web form](#) that facilitates the creation of [DTABf](#) conformant TEI Headers.
- An [online TEI metadata editor](#) provided by [Ortolang](#) (French).