1 ALIM into ILC4CLARIN

This document describes the procedure used to extract information from ALIM texts to populate the ILC4CLARIN repository with literary sources. Data, scripts, and additional software to create the item are contained in a dedicated GitHub repository, see Section 1.1.
The exported items are archives which contain the following metadata files:

metadata_local.xml, dublin_core.xml, and metadata_metashare.xml

All of them are populated with data extracted from elements of the <teiHeader>. The complete mapping for Literary Sources is described in section 2.1, while the one for Documentary Sources in Section 3.1.

1.1 GitHub repository

The main GitHub repository is available at https://github.com/cnr-ilc/alim2clarin-dspace. Its structure is the following:

A "corpus" folder which contains the <teiHeader>s of Literary Sources and the <teiCorpus> of Documentary Sources;

A "template" folder which contains the prototypical metadata_metashare.xml file. This file is the same file for each item in the repository.

Saxon-HE-9.9.1-3.jar. A XSLT and XQuery Processor available through Maven, https://mvnrepository.com/artifact/net.sf.saxon/Saxon-HE;

tei2dublin_core.xls. A XSLT file to create the dublin_core.xml file;

tei2metadata_local.xls A XSLT file to create the metadata_local.xml file;

tei2clarin.sh The executable shell script.

2 Literary Sources

2.1 Metadata Extraction from teiHeader

Both the dublin_core.xml and metadata_local.xml files are filled from the tei2clarin.sh script. The script loops over the "corpus" folder and creates all XML files using specific XSLT files. Before calling the XSLT files using Saxon-HE, the script extracts the identifiers from the original files. Identifiers are used to link the items in the ILC4CLARIN repository to the uris where the sources are, such as http://it.alim.unisi.it/dl/resource/194.

The main instruction of the shell script is:

```
java -jar Saxon-HE-9.9.1-3.jar -o:<OUT> -xsl:<XSLT>
-s:<FILE> id=<ID>
```

Where <OUT> is either dublin_core.xml or metadata_local.xml; <XSLT> is either tei2dublin_core.xls or tei2metadata_local.xls; <FILE> is the <teiHeader> parsed file and <ID> its identifier.

2.1.1 dublin_core.xml

The elements of the dublin_core.xml files are filled by three sources of information:

- (i) Parameters used in the scripts;
- (ii) Constants encoded in the stylesheet;
- (iii) Textual content and attribute values in the child nodes of the <teiHeader>[1].

[1]The <teiHeader> is described in details in the second Chapter of the TEI P5 Guidelines, available online: https://tei-c.org/release/doc/tei-p5-doc/en/html/HD.html

The principal parameter in (i) is the identifier (id) of the document. It is a numeric value qualified as uri, passed to the XSLT stylesheet as a parameter, because it is not contained in the TEI-XML document itself but it is contained as a prefix in the parsed file.[2] Examples of (ii) are the contributors to the *TEI-fication* of the Latin files. The tei2dublin_core.xls scripts uses a <lookup:table> to map the initials of the contributors to their full name: JD $->$ John Doe. About (iii), all the required data are extracted from different tags of the <teiHeader>. The type of all the documents is *corpus* and the language (qualified by *ISO*, namely ISO 639-3) is *lat*. The contributor, qualified as author, is extracted from the <title> in the <titleStmt> (title statement) of the <fileDesc> (file description) section in the textHeader, whereas the title, always in Latin, is extracted from the title element of the same section. The description, always in Italian, is the addition of the <resp> (responsibility) and <name> elements contained in the <respStmt> (statement of responsibility) of the <titleStmt> section. The date, qualified as issued, is extracted from the attribute value @when of the date element in the publicationStmt (publication statement). The publisher, in Italian, is *ALIM Archivio della Latinità Italiana del Medioevo*. The source, qualified as uri is composed by a constant part, suffixed by the numeric identifier of the document. The subjects in Italian are extracted from the <term>s of the <keywords> contained in the <profileDesc> (profile Description) section. Two fixed subjects in English are added: *Latin* and *Middle Ages*, in order to improve the research through the VLO. Finally, rights is setted to *Creative Commons - Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND4.0)* with the qualifiers uri *http://creativecommons.org/licenses/by-nc-nd/4.0/* and label *PUB*.

2.1.2 metadata_local.xml

The elements of the metadata_local.xml files are filled by two sources of information:

- (i) Parameters used in the scripts;
- (ii) Constants encoded in the stylesheet;
- (iii) Textual content and attribute values in the child nodes of the <teiHeader>

As for the dublin_core.xml files, the parameter in (i) is the identifier, which is used to link the item in the repository to its corresponding file: $123 ->$ http://it.alim.unisi.it/dl/resource/**123**.
Constants in (ii) are about some technical values such as, for example, the fact that the item has no physical file: <dcvalue element="has" qualifier="files" language="*">no</dcvalue> as well as a constant about the address of the help desk. Information about the size of the files, in words and characters with and without spaces, extracted from the <measure> element in the <extent> section of the <fileDesc>, (iii).

3 Documentary Sources

As claimed above, the documentary sources are grouped in collections (corresponding to printed volumes) with a general <teiHeader>, but also the singular headers of each document contained into the collection. We export information only from the general header with the same criteria applied to the literary sources, but the contributor is declared *Auctores Varii* (multiple authors).

3.1 Metadata Extraction from teiHeader

3.1.1 dublin_core.xml

TBA

3.1.2 metadata_local.xml

TBA

[2] Please note that DSPACE expects a valid URI. If this is not provided a new handle is created. The entry <dcvalue element="identifier" qualifier="uri" language="*">123</dcvalue> is to make the file valid.