Towards Interoperability in CLARIN
Version 2.0 2017-11-20

Jan Odijk
Utrecht University, the Netherlands
j.odijk@uu.nl

Abstract

This paper proposes (1) a general scheme for specifying interoperability; (2) *very tentative* specific instantiations for two data types (text and audio) of the general interoperability scheme, as illustration of the general scheme; (3) considerations on a policy for CLARIN ERIC to ensure continued efforts on interoperability.

1 Introduction

This is a position paper in which I make some specific proposals related to interoperability in CLARIN. In section 2 we discuss some general considerations with regard to interoperability. I make a specific (but tentative) technical proposal for interoperability in CLARIN in section 3. I elaborate this proposal very tentatively for one specific subcase (text) in section 4.1, and for a second one (audio) in section 4.2. This elaboration is mainly intended to try out the scheme developed and can at best form a basis for discussions that hopefully will lead to a concrete broadly supported proposal. Section 5 considers some policy options for CLARIN to set up and maintain activities to define interoperability and for incentives for national consortia and individual researchers to work towards interoperability. I finish with concluding remarks and describe future work in section 6.

2 Interoperability

Interoperability in general, but also within CLARIN, is still a major problem. Standards and best practices[1] are of course essential for interoperability. Though inventories of standards and best practices have been made (See here and here), and some recommendations with regards to standards have been made (Kemps-Snijders et al., 2009), there is no clear guideline to member consortia and individual researchers on how to work towards interoperability nor a clear incentive for them to do so. Interoperability, however, is a crucial factor for the success of CLARIN, because the strongest motivation for researchers to invest in adhering to CLARIN requirements is if they see that they can benefit from it (for example by using CLARIN tools that apply to their data, if they are interoperable).

It is not realistic, in my view, to expect that just working on standards and best practices will bring a solution in the short term, and perhaps it will never bring a full solution. Unified standards are not necessarily the best option (because they accommodate too much, or because they accommodate too little). In addition, there are many data in legacy formats and many tools that work with legacy formats and we cannot afford to ignore them. A standard will only be successful if it is accompanied by many tools that are better or in other ways more attractive than existing tools, and that will not be easy to achieve.

Whatever may be the case, we have to be pragmatic and offer clear guidelines for interoperability in the CLARIN context and incentives to work towards it. I make a specific proposal for exactly this in this paper. I focus on interoperability between software (tools, applications, web services, etc) and data. I propose a general scheme for specifying interoperability within CLARIN, distinguishing two dimensions (the *data* and the *metadata* dimensions) with 4 levels at the data dimension, and elaborate this in detail

[1] I will use the terms *standards* and *best practices* interchangebly, and in *standards* I include *de facto* commonly used formats etc. even if they are no officially recognized standards.

for two data types. There will be several different interoperability schemes, depending on the data type, and the aspects of the datatype being considered.

3 Interoperability Proposal

Here I concentrate on interoperability between software (tools, applications, web services, etc) and data. CLARIN has to specify in detail, for a given data type, which formats it supports. We say that a tool that applies to this data type is interoperable if it works on data that are in such a supported format.

I make a very specific, but tentative, proposal for this, which is mainly intended to make as clear as possible what I have in mind, and which is limited by my restricted knowledge and expertise in the matter. But I hope it can form a good starting point for further discussions.

I distinguish formal (syntactic) interoperability from semantic interoperability and define separate requirements for them. I define different requirements for different data types (e.g. natural language text, audio, video, pictures, structured data, and possibly specific subtypes if that would be required). I identify, for each data type, a number of formats that CLARIN wants to support. I distinguish a number of priority levels (4 levels in the current proposal along the *data* dimension), where each of the levels 2-4 includes the lower levels and adds new requirements. In addition there is an orthogonal *metadata* dimension that can be reached independently of the other levels.[2] The requirements may be different for input and output.

The general scheme for interoperability of software that applies to data type T is as follows:

General requirements

Multiple files If the software takes a file as input, it must be able to deal with multiple files, of multiple supported formats, in hierarchical folders and in compressed files such as zip, 7z, gz, tar.gz

Basic Encoding The software must be able to deal with the CLARIN-supported encodings of the basic units of data type T

Wrapping Software that enriches data in a format F must be able to include these enrichments in a copy of the input and output the input combined with the enrichments in format F (assuming that the enrichments can be expressed in F)

This is our proposal for levels that should be distinguished along the *data* dimension

Level 1 General requirements + internationally recognized or de facto standards / best practices

Level 2 level 1 + highly prioritized regionally / domain-specific recognized or de facto standards / best practices

Level 3 level 2 + common formats in use in everyday life

Level 4 level 3 and less prioritized recognized / de facto standards / best practices

For the *metadata* dimension only one level exists ('accepts CMDI metadata in combination with the data') which can be met or not (binary distinction).

I define labels (which could be associated to graphical 'laundry tags' as in Creative Commons), as follows

(1) CLARIN-(FRM|SEM)-(INP|OUT)-(TXT|ATX|AUD|AV|STR) (NA|*{1,4})

with the meanings as indicated in Table 1.

So for example, the label CLARIN-FRM-OUT-AUD defines syntactic (FRM) interoperability for software yielding audio (AUD) output (OUT) in CLARIN (CLARIN).

[2] Called level 5 in previous versions of this document.

Laundry Tag	Meaning
CLARIN	It is a CLARIN interoperability specification
FRM	Concerning formal interoperability
SEM	Concerning semantic interoperability
INP	For input of data
OUT	For output of data
TXT	For natural language text
ATX	For annotations on natural language text
AUD	For audio
AV	For audio-visual data
STR	For structured data
NA	Star system not applicable
*	# Stars indicate the level: 1 .. 4 for the *data* dimension, with 4 the highest level of interoperability
✓	Presence of a check mark indicates interoperability for the *metadata* dimension.

Table 1: Laundry tags and their interpretation

4 Elaboration

I *very tentatively* present an elaboration of the general proposal for natural language text (section 4.1) and for audio (section 4.2). I emphasize that these are provided just as an illustration of how the scheme *could* work, and to test this properly one must go into some detail. But a definitive proposal requires input from users, experts and national consortia. However, I hope that the current proposal can form good basis for further such discussions.

4.1 Natural language text

I elaborate the generic specification in detail for natural language text. I mean here

- running natural language text (so no structured numerical data in a textual format such as csv or xml)
- encoded in a textual format (so no text that is contained in a picture)

Here I formulate interoperability requirements on the natural language text only. Note that I will mention below many formats that can encode annotations on text, but this specification deals with the representation of the natural language text in such formats only, not on the annotations (separate requirements are needed for that). I formulate requirements both for input (section 4.1.1) and for output (section 4.1.2).

Though Level 1 interoperablity for textual input and output is relatively easy to achieve[3], I submit that not a single tool or application in CLARIN already reaches Level 1 (CLARIN-FRM-INP-TXT * and CLARIN-FRM-OUT-TXT *), but I hope that I will be proven wrong.

4.1.1 Input

I first concentrate on *input* for software operating on text. The requirements for this are labeled in accordance with what was said above as CLARIN-FRM-INP-TXT.
General Requirements:

- the *Multiple files* and *Wrapping* requirements hold
- **Basic Encoding** = Character encoding: Unicode is obligatory, with a preference for UTF8 encoding; a Byte Order Mark should be accepted in UTF-16 encodings, and, when present, both *BigEndian* and *LittleEndian* encoding should be accepted. ISO-* and ASCII encodings are desirable, but not required.

[3] at least in comparison to interoperability of annotations on textual resources.

- Plain text must be accepted in multiple varieties, to be specified by parameters:
  - Unanalysed running natural language text (the default)
  - Unit-split text, in which the units are separated from each other by a unique non-ambiguous character sequence. The parameters to specify this are a label for the unit and the character sequence separator. A number of units are recognized by the tool: token, sentence, paragraph, section, chapter, part; and if possible/relevant acted upon.
  - Labeled unit-split text: each unit distinguished can be preceded by a label separated from the unit by a unique non-ambiguous character sequence, and the label is processed as metadata on the unit.
  - Text in one or more columns in a CSV[4] file, as specified by a sequence of column counters.

Level 1 General requirements + internationally Recognized or de facto Standards / best practices

TEI, XCES, CHAT, EAF, plain text.

This probably requires a more detailed specification: which variant of TEI, etc. to be decided upon after consultation of the experts and input from the NCF about actual usage in their countries

Level 2 level 1 + Regionally / domain-specific recognized or de facto standards / best practices

- FoliA (NL), TCF (DE, Weblicht), LASSY XML (NL), NAF, Prague Markup Language (PML) (CZ) , TIGER-XML,

Level 3 level 2 + common formats in use in everyday life

- RTF, MSword 7, MSWord 10, OpenOffice ODT, LibreOffice ODT, PDF , CSV, ePub, HTML

It is enough here to have a mapping from these formats into plain text, with loss of lay-out features, font information, highlighting etc.

Level 4 level 3 and less prioritized recognized / de facto standards / best practices

- CES, TMX, ALTO, Shoebox / Toolbox, Tipster, RDF, CoNLL-U, CoNLL-X, Penn Treebank format, Susanne format, NEGRA format, PAULA XML (Potsdamer AUstauschformat Linguistischer Annotationen), LIF, PRAAT TextGrid.

For the *metadata* dimension the criterion is whether the software accepts CMDI metadata in combination with the data. The software reads a CMDI file and processes the actual data in part based on the information found in the CMDI file.

Of course, I hope that people will make and share convertors and wrappers so that no unnecessary duplication of work is done: CLARIN-NL offers web services for converting Alto, text, Word, HTML, en ePub into TEI or FOLIA. (OpenConvert) and there are open source convertors such as those of pandoc that can perhaps be used.

4.1.2 Output

The requirements for *output* for software operating on text are labeled in accordance with what was said above as CLARIN-FRM-OUT-TXT. They apply to software that yields natural language text in textual format as output.

The requirements are essentially the same as for input, with the following proviso's:

- There is no requirement to be able to output text in formats that are proprietary. So level 3 is easily reached once level 2 has been reached

- For the metadata dimension the software must produce new data as output together with an updated CMDI file that formally describes what the software did to the original input data, and which configuration was used.

[4] or similar files with a different column separator.

4.2 Audio

I define, very tentatively, the requirements for audio. The requirements for input and ouput are identical.

General requirements: The *Multiple files* and *Wrapping* requirements hold, but the *Basic Encoding* requirement is not defined for audio.

Level 1 General requirements plus internationally recognized or de facto standards / best practices

- single channel raw header-less PCM, uncompressed audio files in RIFF WAV, AIFF, or AU format. These formats are containers usually containing LPCM-encoded audio and a header that specifies properties of the LPCM encoding such as sample rate (#samples per second), bit depth (# bits per sample), endianness (order of bytes in a memory word) and number of channels. Sampling rates: 48kHz and 16kHz.
- MPEG-1, MPEG-2 and MPEG-4
- EAF: A file in ELAN Annotation Format (EAF) is a container encoded in XML that can contain references to media files through the MEDIA_URL and MIME_TYPE attributes of its MEDIA_DESCRIPTOR element. A tool must be able to deal with an EAF file if the file contains references to media files in CLARIN-supported audio formats, and then it should process these media files.

Level 2 Level 1 + regionally / domain-specific recognized or de facto standards / best practices

- NIST format

Level 3 Level 2 + level 2 + common formats in use in everyday life

- lossy formats such as MP3 and AAC

Level 4 Level 3 + less prioritized recognized / de facto standards / best practices

- 8kHz sampling rate, A-law and mu-law companding,
- multiple channels

For the *metadata* dimension the criterion is whether the software accepts CMDI metadata in combination with the data. The software reads a CMDI file and processes the actual data in part based on the information found in the CMDI file.

5 Policy and Incentives

In this section I mention some options for a concrete policy for CLARIN to set up and maintain activities to define interoperability and for incentives for national consortia and individual researchers to work towards interoperability.

5.1 Policy

CLARIN ERIC has to organize itself in such a way that (1) concrete proposals concerning interoperability, in particular on a general schema for interoperability and specific instantiations of the general schema for particular data types are created; (2) consults users, the standards committee, the national coordinators and any relevant specialists in the matter, (3) the CLARIN BoD can adopt a particular proposal for CLARIN. It is possible (and actually very likely) that no consensus can be achieved among all CLARIN member with regard to certain matters. In such cases a proposal should made anyway, but it should explicitly mention any objections from national consortia or individual researchers that cannot be accommodated, and it justifies why this cannot be accommodated. The CLARIN BoD then takes a decision on making the proposal a CLARIN requirement, or to request a modified proposal.

How to best organise the creation of proposals is open for discussion. Many different forms can be thought of (e.g. special committee, one of the existing committees, research teams working on specific types of resources, etc.

5.2 Incentives

Incentives can be implemented in many ways but I concentrate here on an implementation in terms of clearly stated targets, competitiveness, prestige and money.

The existence of clear targets and guidelines for interoperability will aid national consortia and individual developers: they know exactly what should be done, can plan for it, and can justify a national plan for doing so against reviewers by referring to the CLARIN targets.

With the star system I hope to initiate a constructive competition between individual researchers and between national consortia: everybody wants to be the best and therefore to have most stars and check marks.

I propose that CLARIN awards the individual who contributed most to interoperability in the preceding year with an Interoperability Prize. CLARIN also awards, once a year, the national consortium that contributed most to interoperability in the preceding year (relative to national budget) with a prize, hence prestige. Details of the awards are to be elaborated but could involve money or funding of activities.

It is our hope and expectation that such incentives can be made very attractive and will stimulate work on interoperability.

6 Concluding remarks

This paper proposed (1) a general scheme for specifying interoperability; (2) very tentative specific instantiations for two data types (text and audio) of the general interoperability scheme, as illustration of the general scheme; (3) some considerations for a policy for CLARIN ERIC to ensure continued efforts on interoperability.

Acknowledgements

This work crucially uses data and/or tools made available through the CLARIN infrastructure. The work was financed by CLARIAH-NL, an NWO project in the Netherlands.

References

[Kemps-Snijders et al.2009] Marc Kemps-Snijders, Núria Bel, Peter Wittenburg, Daan Broeder, Dieter Van Uytvanck, Laurent Romary, Erhard Hinrichs, and Gerhard Budin. 2009. Standards for LRT. https://www.clarin.eu/sites/default/files/Standards%20for%20LRT-v6.pdf, January.