Draft of an internal CLARIN Standards Committee report;
please do not circulate this version.

Ver. 0.2, 10 May 2018,

Piotr Bański,
IDS Mannheim

Towards unified CLARIN recommendations for the use of standards: a pilot study on “text formats”

Contents

1. Introduction

This document is an attempt to define the stance of the CLARIN Standards Committee with respect to standards and best practices that are recommended for use by and across CLARIN centres. Following on the F2F meetings in Aix and Budapest, I sketch a proposed taxonomy and suggest a possible way to reconcile a top-down strategy that is important for CLARIN as a whole with the bottom-up freedom that is important for each individual centre. I stress that this version of the document does not reflect the position of the entire CSC and should be treated as material for discussion, rather than as normative.

In the lifetime of the CLARIN initiative, several more or less official "lists of recommended standards" have been proposed, several documents have been produced, and a few surveys have been circulated. Many of these proposed lists are publicly accessible under the CLARIN label and sometimes very different in content: in the range of norms covered, granularity of versioning, internal classification and predicted use.[1] This has resulted in a general feeling of uncertainty as to what and how exactly is recommended for CLARIN centres, while the stakes are high and rising all the time, given the energetic efforts at ensuring interoperability across CLARIN tools and centres (cf. Jan Odijk's position paper on interoperability in CLARIN), and the support that the CSC should offer to these efforts.

In the first part of this document, I look at ways of addressing the potential tension between what CLARIN centres *should* recognize as parts of a coherent interoperable network, and what each centre *wants* to do as part of its contribution to the network, on the understanding that the strength of a decentralized network lies in the expertise of its researchers, but also in an overall commonly agreed coherence of aims and ways of fulfilling them. In the latter parts, I suggest certain "vertical" and "horizontal" categorizations and apply them to the vague notion of "textual data" that was chosen (at the CAC in Budapest, 2017) as the subject of this pilot study.

1 A list of most of such recommendations has been gathered by, notably, Alex Herold and Hanna Hedeland in the CSC part of the CLARIN wiki: https://trac.clarin.eu/wiki/StandardsCommittee

2. Top-down vs. bottom-up perspectives on the implementation of standards

In the conversations on the role of the CSC in CLARIN, a top-down position is sometimes expressed stating that the Standards Committee, with the help of the Centres Committee and the NCF, should impose a uniform list of standards on the CLARIN centres. This view has some potential advantages, essentially lifting responsibility for such decisions from the individual centres by making them adhere to guidelines on standards use, formulated by the CLARIN governance. However, given the richness of formats and the diversity of goals of the individual centres, a bottom-up perspective that gathers information on the actual standards use in the particular centres and distils the overall picture out of such surveys, is also viable, definitely realistic, and probably close(r) to the expectations of many researchers. This seems all the more pragmatic given that it is at the moment not at all clear what internal devices for ensuring top-down uniformity CLARIN currently possesses.

In what I propose, I attempt to leverage the advantages of both approaches: on the one hand, the convenience of being able to formulate a general message about CLARIN capabilities and restrictions that is not only valuable to the governance but also constitutes a kind of “shield” for the individual centres, when they face the task of explaining why they do not accept or produce certain formats (and why they expect and produce others), and, on the other hand, the fact that individual centres have their own user profiles and research foci, and therefore they prioritize certain kinds of data and formats, and consequently are also willing to go an extra mile towards accepting data even if it is presented to them in exotic or obsolete formats.

In essence, I propose that the top-down perspective, fully regulated by the CSC (in concert with the other relevant committees), encompasses the fundamental standards that should be expected of CLARIN from the outside, at a high level of granularity (specified as, e.g., plain “TEI” without further qualifiers). At the same time, specifying the features for low level of granularity will remain in the hands of the particular centres, whose relationship to the list sketched here will be not so much to obey the (intentionally underspecified) directives, but rather to fill the values in, according to their specific profiles, interests, and capabilities.[2]

2 A more “authoritarian” path might need to be taken with respect to the MIME types used, for the sake of maximally enabling interoperability across centres and projects. Potential listings of potential filename suffixes (.html, .htm, etc.) should only be treated as informative (rather than normative).

3. Vertical feature matrix

I would like to propose grouping standards along three major parameters listed below together with their possible range of values:

1. Direction[3]
   a. bidirectional (ingest and export; this is very much a shorthand parameter)[4]
   b. ingest
   c. export
2. Urgency
   a. recommended
   b. optional (for e.g. “local best practices”)
   c. discouraged (in most cases with 3c)[5]
3. Status
   a. current
   b. deprecated
   c. outdated
   d. – (unmarked) when multiple variants of the standard exist that vary in the “urgency” parameter

The overall goal for the above matrix was to make it simple to manage (and to adjust the values within minimally 1-year cycles), but at the same time reasonably expressive.

The first parameter describes (or regulates) the direction of the data flow: whether it is (broadly speaking) ingested into the centre or exported from it (the value “bidirectional” serves as shorthand).

The “Urgency” setting describes CLARIN’s recommendation concerning the given standard (or best practice, or local practice). This is where the setting may be normative in the top-down fashion (“recommended”) or leave the decision to the centre (“optional”) or express a top-down frown that is at the same time meant to shield a centre from demands to support a format that is obsolete or in some way suboptimal. “Urgency” is a global

3 **[** please note: I may be having (or imagining) a problem with parameter (1): I would like to treat it both as informative about the direction of the data flow, but also as normative from the perspective of the CSC: in cases where the CSC fully specifies some format for “ingest” or “export”, the individual centres should have no (or very little) wiggle room left **]**

4 .. and I’m thinking of eliminating it altogether, because it’s theoretically somewhat redundant (although definitely useful for practical purposes). I’ve been pondering using a value of “internal” here, as well, for tool formats that should not be used for data exchange outside of the tools that use them, such as TCF, CWB, SkE, FoLiA, ANNIS (not exactly a text format), TEI-TXM, (LIFT for lexical data), etc. TCF is an example of a borderline status: it is a tool format that has visibly been developed for internal purposes but given the interface nature of the pipeline that it is used for, some centres have offered data for download in this format, to make it easier for users to subject the data for further processing with WebLicht.

5 The intended meaning of (2c) is “less than optional”. This is a top-down value supposed to shield the individual centres – while a centre may choose to support it, no pressure should be used on the centre because of the “optional” setting.

setting, so the value “optional” is a way to handle local (best) practices. The value “discouraged” is paired with the “outdated” value of the last parameter.

The “Urgency” parameter expresses the perceived status of the standard within the fields of interest to CLARIN, with “outdated” reserved for formats such as FeldPartitur or COCOA for data and e.g. Word Perfect for submissions of texts for ingestion into corpora. While the extremes are handled with the values “current” and “outdated”, there is a large grey area of practices potentially coming up in status or standards apparently already on the downward slope, and I have not been able to find a neutral and satisfactory label for those (values I have pondered were, among others, “deprecated” (not quite fitting emerging best practices), “peripheral” (with “current” visualized at the core), “partial” (neutral but too vague) or “marginal”. In this version of the document, I will continue to use “deprecated”, while noting that a better label should be sought for.

4. Horizontal divisions among types of textual data (“text formats”)

The initial idea for this particular task was to be a quick pilot study or demo on what the unification of CLARIN’s recognized pool of standards might look like in the domain of “textual data” or “text formats”. However, the phrase “textual data” can be understood in many ways, and at least several of them are relevant to CLARIN and should be clearly distinguished, also because they involve different perspectives on the role of standards and consequently different classifications: for example, while “PDF” (in nearly all variants) is in many cases a good format for submitting *documentation* in, even when it is not tagged (that is, without structural divisions), it is far from perfect as a carrier of textual data for analysis.

“Text”, on one interpretation, is what we expect in the documentation that accompanies the data submitted to (or generated by) CLARIN centres. When addressing this type of data, I will from now on use the term “documentation”.

“Text” is also the primary object of interest of most CLARIN centres: it is the (more or less) raw data that we annotate and research. I will use the term “text content” for this meaning.

Text is also the format that document metadata for a variety of textual and non-textual formats comes in, and is harvested from. The “Metadata” category is intentionally left unfilled in this very document, for the sake of its scope restrictions. See Appendix A for

a skeletal table that should be filled in by (or at least in connection with) colleagues from the metadata task force.

Finally, when talking about “textual formats”, we often mean annotated data or annotations (“analytical metadata”), even where the annotated object itself is not textual (e.g. in the case of audio/video streams or facsimiles).[6] This document only looks at a subset of such data, with somewhat fuzzily defined borders. I return to this point in section 4.4.

One might wonder about why this document excludes e.g. lexical data or transcription of spoken language. The reasons are purely practical and have to do with the intended finiteness of this very document and the limits of expertise of its author. These data formats, together with binary formats, are definitely intended to appear in the final CSC deliverable.

Please note that the tables below are supplied as illustration – for discussion within the CSC, not as yet another set of lists of recommended standards. This document in the current version, or its fragments, are not to be quoted as normative for CLARIN.

The service in which this information is planned to be eventually stored, maintained and published is located at https://standards.clarin.eu/sis/ (the sources are provided at https://github.com/clarin-eric/standards). Before that happens, the CSC may produce temporary lists, preferably with TTL (time-to-live) or “expiration date” specified.

4.1. Documentation formats

Please recall that this is not meant as exhaustive. Please feel encouraged to add to this list (the tables will be put online for the CSC members to edit).

documentation format type	parameters		
Postscript	ingest	optional	–
PDF/A	bidirectional	optional	current
PDF/A-1	ingest	recommended	current
PDF/A-2	ingest	recommended	current
PDF/A-3	ingest	optional	current
PDF/E	ingest	optional	current

6 I am told that the Metadata Curation Taskforce called this aspect of text “structuredDataset”.

documentation format type	parameters		
PDF/VT	bidirectional	optional	current[7]
PDF XFA forms	bidirectional	discouraged	outdated
PDF/X	ingest	recommended	current
PDF/X	bidirectional	optional	current
PDF (other)	bidirectional	optional	–
plain text (UTF-8)[8]	bidirectional	recommended	current
plain text (ASCII)	bidirectional	recommended	current
plain text (EBCDIC)	bidirectional	discouraged	outdated
plain text (Latin-1)	bidirectional	optional	current
plain text (other encodings)[9]	bidirectional	optional	–
TEI P5 (always with ODD), any flavour[10]	ingest	recommended	current
HTML 3.2	ingest	recommended	deprecated
HTML 4.x	bidirectional	recommended	current
XHTML	bidirectional	recommended	current
HTML 5	ingest	recommended	current
HTML 5	bidirectional	optional	current
DocBook	ingest	recommended	current
DITA	bidirectional	recommended	current
markdown	ingest	recommended	current
markdown	bidirectional	optional[11]	current
LaTeX (package references satisfied)	ingest	recommended[12]	current
Open Document Format (odt)	bidirectional	recommended	current

7 Built on top of PDF/X-4, it has to inherit the values for PDF/X
8 Preferably, without the BOM character.
9 Conversion to UTF-8 is encouraged, unless not feasible.
10 TEI documents *must* be accompanied by the corresponding ODD and *should* be accompanied by at least one schema document derived from the ODD.
11 Markdown is a family of plain-text, human-readable formats. The optionality concerns the choice of the flavour for export.
12 For the purpose of documentation, it is encouraged to submit the compiled PDF files alongside LaTeX sources.

documentation format type	parameters		
Office Open XML, (docx)	bidirectional	recommended	current
OpenOffice.org XML (sxw)	bidirectional	discouraged	outdated[13]
Microsoft Word (doc)	bidirectional	optional	deprecated[14]
Rich Text Format (.rtf)	ingest	recommended	deprecated[15]
Rich Text Format (.rtf)	export	optional	deprecated

On PDF/A, see https://www.pdfa.org/publication/pdfa-in-a-nutshell-2-0/

4.2. Metadata formats

See Appendix A for a skeletal table.

4.3. Text content formats

Please note that “text content” as understood here does not preclude basic structural (and consequently basic semantic) divisions, for example into sections (and also headers), paragraphs, or sentences, and finally into individual tokens. For example, column-based formats can alternate between relatively “pure” text (even if tokenized, with an ID in the first column) and annotated (with the rest of the columns filled in). The same is true of many XML-based (or near-XML) formats, when the annotations are left out. On the other hand, text content formats also include those where single textual divisions are stored as mammoth-sized single lines, terminated with a newline.

The values in the table below owe a lot to the IANUS directives (https://www.ianus-fdz.de/), already shared by some CLARIN centres.

In the table below, the “recommended” value has several assumptions attached that can be gathered under the heading of “common sense” or a version of the pragmatic principle of relevance, namely that the submitter will do their best to make the process of up-conversion of the text as painless as possible. Naturally, we as the CSC/CLARIN can only provide general labels in such cases, and “recommended” is exactly such a general

13 The CSC strongly encourages using the Open Document Format families (.odt, .ods, etc.) over the obsolete OpenOffice.org format families (.sxw, .sxc, etc.).

14 For Microsoft-based word-processing formats, the CSC encourages the use of Office Open XML (docx) instead of the legacy formats saved with the .doc extension.

15 RTF is a documented but proprietary format, and centres are encouraged to use standardized non-proprietary formats instead.

label. Other related principles involved here are cost-effectiveness and feasibility. If the CSC recommends a standard/BP as “recommended”, the recommendation refers to clean data that is cost-effective to process (as opposed to e.g. obfuscated or non-uniform data that is nevertheless syntactically valid according to the given standard’s definition).

text content format	parameters		
PDF/A	bidirectional	optional	current
PDF (preferably tagged)	bidirectional	optional	–
plain text (UTF-8)[16]	bidirectional	recommended[17]	current
plain text (ASCII)	bidirectional	recommended	current
plain text (EBCDIC)	bidirectional	discouraged	outdated
plain text (Latin-1)	bidirectional	optional	current
plain text (other encodings)	bidirectional	optional	–
TEI P5 (always with ODD+schema), any flavour[18]	bidirectional	recommended	current
TEI P4	ingest	optional	outdated
TEI P4	export	discouraged	outdated
DocBook	ingest	recommended	current
DocBook	export	discouraged	current
DITA	ingest	recommended	current
DITA	export	discouraged	current
XML (with a schema)[19]	bidirectional	recommended	current
SGML	bidirectional	discouraged	outdated
HTML (various versions)	bidirectional	recommended	–
XHTML	bidirectional	recommended	current
EPUB 3.x	bidirectional	recommended	current
EPUB 2.x	bidirectional	optional	deprecated

16 Preferably, without the BOM character.
17 Export as UTF-8 is recommended if feasible.
18 For export of text content in the TEI, the flavour (essentially, the ODD description) is determined by the given centre.
19 An assumption here is that the schema documents form a black-box layer from the perspective of CLARIN: if a given metaformat is accepted, at least one corresponding schema format must be accepted; additionally, by the TEI definition of conformance, a document claiming to be TEI-conformant *must* be accompanied by the ODD definitions and documentation.

text content format	parameters		
HTML 5	ingest	recommended	current
HTML 5	export	optional	current
markdown	ingest	recommended	current
markdown	export	optional	current
Mediawiki plain text markup	ingest	recommended	current[20]
Mediawiki plain text markup	export	optional	current
LaTeX (package references satisfied)	ingest	optional	current
Open Document Format (odt)	bidirectional	recommended	current
Office Open XML (docx)	bidirectional	recommended	current
OpenOffice.org XML (sxw)	bidirectional	discouraged	outdated[21]
Microsoft Word (doc)	bidirectional	optional	deprecated[22]
Rich Text Format (.rtf)	ingest	recommended	deprecated[23]
Rich Text Format (.rtf)	export	optional	deprecated
TIPSTER	ingest	optional	outdated
TIPSTER	export	discouraged	outdated

4.4. Annotated data in text corpora: gross divisions

Recall that this pilot study targets a subset of textual annotated data, restricted to the types most commonly seen in this type of resource, and is meant to start a discussion rather than deliver a ready product. Consequently, the following types of formats (or content) are *excluded from the present sample*, for practical reasons:

- terminology and lexical standards,
- metadata standards/vocabularies,
- knowledge/concept/ontology description (RDF family: RDF, RDFS, SKOS, OWL-*),

20 See e.g. https://github.com/IDS-Mannheim/WikiI5Converter for a converter.
21 The CSC strongly encourages using the Open Document Format families (.odt, .ods, etc.) over the obsolete OpenOffice.org format families (.sxw, .sxc, etc.).
22 For Microsoft-based word-processing formats, the CSC encourages the use of Office Open XML (docx) instead of the legacy formats saved with the .doc extension.
23 RTF is a documented but proprietary format, and centres are encouraged to use standardized non-proprietary formats instead.

- semantic annotation,
- machine translation / translation memory formats,
- lexical formats extending from dictionaries to Framenet/Verbnet/Nombank/Wordnet resources,
- database formats expressible as text (tables) or flavours of JSON.

This is expressly not to say that such formats are outside the scope of CLARIN – quite on the contrary, the present document is expected to mesh with similar studies done by, or informed by, experts in each specific area. Here, I concentrate on a subset of formats used for the purpose of annotating text collections with relatively basic grammatical information.

It is at this level of specificity that I expect the top-down strategy to give way to bottom-up reporting, and the CSC to catalogue the variation and ideally restrict itself to marking specific formats as outdated and discouraged, and others as "current" and "optional". In other words, I believe that the experience of the past years and "past standards lists" teaches us that we can hope to ensure uniformity[24] only down to a certain level of granularity, below which it is only reasonable to expect informative (rather than normative) statements.

What follows is a rough, intentionally shallow division of formats for annotated data usable for corpus-level text resources up to text-based treebank description. It is not a theoretical division – its *practical* purpose is to cut the entire range of annotated data formats up into smaller, more manageable groups. I believe that in a list containing the formats below, the values of the "Urgency" parameter will in most cases be "optional", in order to guarantee the centres and researchers the maximum reasonable level of freedom. This version of this document does not provide a table; the proposed values are mentioned in the text.

24 "Ensuring uniformity" cannot naturally be assumed to be a permanent state; it crucially assumes an active CSC that responds to the changes and progress on the "annotation market", taking action at least once per year.

The tentative groupings of formats are as follows:

4.4.1. structured plain text

In this, group, one can minimally distinguish between:

4.4.1.1. column-based formats...

… extending from CSV and TSV into, on one side “TSV on steroids” such as the CoNNL-* formats (the most recent of them being CoNNL-u used in Universal Dependencies, with CoNNL-X as a somewhat obsolete ancestor), into relational tables on the other side (recall that this document skips relational formats, but it’s worth to mark the transition/meeting point).

Other obsolete examples include NeGra (variant 3 or 4). A current example is PropBank, which is a format dependent on Penn Treebank (see below).

TreeTagger and MMAX2 also qualify here, both as tool formats.

4.4.1.2. bracketed formats...

...of which the outstanding example is Penn Treebank (ver. I and II). Given that conversion tools from this format into more restrictively structured formats exist, it feels obvious that the “Urgency” parameter can only be set to “optional” in this case, while “Status” could be “deprecated” in favour of other representations.

A more exotic example of this category, from the point of view of a language technologist, would be LMNL.

4.4.2. near-XML formats

This is a very unequal bunch of formats, consisting among others of XML’s legitimate ancestor, SGML, in its many possible variants (some of which are well-known under the name “HTML”), and consequently TEI P3 and CES, the last of which naturally feels like a borderline category, because of its relatively smooth historical transition into XCES.

A more exotic example in this category would be COCOA, definitely obsolete and discouraged.

Current formats in this category include SkE/NoSke, which is basically a column-based format clad in an XML-like outer garment for sentence-boundary marking, as well as CWB, which is a span-based, attribute-less pseudo-XML. While these are thriving formats, it is obvious that they should only function as “tool formats” rather than exchange or

storage formats. CLARIN should definitely discourage the usage of these formats for both storage and exchange. (Please note, that this should not be misconstrued as a suggestion not to use SkE or CWB – the intention is merely to state/remind that tool formats are not suited for environments outside the given tool).

A potential annotation format that should probably be mentioned is HTML+RDFa, which is by definition borderline between XML and near-XML, serving in HTML, XHTML, and HTML 5 contexts. Centres might wish to import such data, while their export should probably discouraged (in a top-down fashion) in favour of more expressive (and more standardized) formats.

4.4.3. XML-based formats

These naturally form the largest and the noisiest group, with many subdivisions, which I will mark only very crudely, because of the pragmatic angle of the present document.

4.4.3.1. Treebanking formats

- Tiger XML (outdated)
- Salsa (an extension of TigerXML for Framenet description, hence also outdated)
- Tiger2 (local format, enhancing Tiger XML)
- PML (Prague Markup Language) is a very robust format that is nevertheless rather localized and accompanied by powerful but not fully maintained tools; I believe that “optional” is the best label here, for many reasons
- GrAF, as a graph-based standoff pivot format for the exchange of ISO LAF-compatible graph-based data, can also serialize treebanks
- PAULA would be classified here as well; it is not clear to me what is the “urgency” status of this format
- Among these, one should also mention TEI-NKJP (the TEI customization of the Polish National Corpus, a layered stand-off format, capable of describing treebanks and used for that purpose). I am not aware of any data exchange being done in this definitely “current” format, hence its Urgency should at best be “optional”

Note that it is already obvious that in many cases, explicit classification by the CSC with attributes other than “optional” may be taken as a subjective or even political judgement, and that is one more reason why I believe that only a minimal degree of top-down

uniformity should be imposed by our committee in a live and potentially sensitive environment.

4.4.3.2. TEI-based formats

In many contexts, the TEI Guidelines are mentioned as a standard, but one has to bear in mind that that is in a way a quantitative rather than qualitative statement: numerous projects use various variants of the TEI for the purpose of a plethora of kinds of text annotation, but there exists no single TEI (and users are explicitly discouraged from using TEI-all in any production systems). TEI Guidelines offer a toolkit for preparing and documenting a great number of serializations of various data models, and some of these serializations may eventually evolve into a best practice and then become formally standardized (or they may be explicitly defined in order to serialize a standardized data model, as it has happened in the practice of at least ISO TC 37 SC 4).

Below, I enumerate “flavours” of the TEI that I believe should be guaranteed a place among *de facto* standards or at least “emerging best practices” for large parts of the CLARIN community:

- TEI-DTA (BBAW),
- TEI-I5 (local to IDS Mannheim)
- TEI-CMC
- TEI/ISO Transcription of spoken language – mentioned here for the sake of completeness (it serializes models that are not in the focus of this document)
- potentially TEI-NCP (publicized format of the Polish National Corpus), probably as a local format used by the members of the NCP consortium (that includes at least three CLARIN centres)
- another current TEI flavour is TEI-TXM, a tool format used by the Textometrie suite, mentioned here for the sake of completeness only – no data exchange via this format is expected.

It is worth mentioning that CLARIN developers have now for a while used the format-variant parameter to MIME types used in Web services (for example, TEI-DTA is identified by the following string: “application/tei+xml;format-variant=tei-dta”). This is not standardized and in fact runs against certain current IETF RFCs, but, on the other hand, it demonstrates the healthiest kind of input to potential standardization, being an emerging

bottom-up community practice. I return briefly to the issue of MIME types in the Summary section.

It is also worth mentioning that the most recent version of the TEI Guidelines (published in January 2018) contains a set of attributes for basic grammatical description at the token level, which might result in the emergence of more TEI-based formats for text corpora. (See http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-att.linguistic.html).

4.4.3.3. Other XML-based formats

This is another non-uniform group of formats, some of which have been mentioned in the "treebanking" section as well. Again, probably most or all of them should have Urgency set to "optional".

- PML
- PAULA
- XCES (variant specified; provide DTD/schema)
- GRaF understood strictly as a pivot format for the exchange of LAF-compatible graph-based data
- FoLiA (local to several Dutch/DLU centres)
- KAF (the KYOTO project)
- TCF (a tool format; capable of describing higher-level grammatical relationships though listing it as a treebanking format would probably not be proper)

There also exist isolated formats, such as Telugu Treebank or Sinica Treebank (cf. D5c-3), which do not imply the existence of any standards and should be clearly "discouraged" and probably "outdated".

5. Summary and remaining issues

This document is an internal report to the CLARIN Standards Committee, not meant to be circulated outside of it, for fear of being falsely taken to be yet another list of (lists of) recognized standards. Tables from it *may* be made available to the CSC (and other interested parties) separately as e.g. Google documents, solely for the purpose of restructuring the content in a collective fashion.

A very important issue only briefly mentioned above is the need for an agreed repertoire of MIME types (often with additional parameters) to identify the content of

documents by Web services. MIME types have no simple correspondence to the rows of the tables above: sometimes, many rows would be served by a single MIME type, but sometimes a single row has several MIME types corresponding to it “in the wild”, and that is probably an even more dramatic problem, from the point of view of interoperability. I did not want this set of issues to obscure the system outlined here, but I do stress that I consider it fundamental that the CSC regulates this aspect of format use across CLARIN as well, and I expect the final form of the information provided by us to also include a repertoire of advocated, permitted, as well as discouraged MIME types for each relevant format. Much work towards this goal has already been done by Hanna Hedeland.

An interesting aspect emerging from the presentation above is that a single list where a format has a single set of parameters, is probably not feasible: the comparison of the lists for documentation formats and text content formats shows that, depending on the context of use, different values should sometimes be proposed.

It has been mentioned at the Aix meeting that one of the description items for each standard should be the user group that the standard addresses. Given how general some standards are and that sometimes they are embedded deep into tools used by various communities (sometimes for various reasons), this information item can be very tricky to formulate. It obviously depends on the adopted taxonomy of research communities, and that appears to be opening a separate set of issues, so I propose to set this proposal aside while we concentrate on discussing the present, narrower, proposal.

Lastly, it should be mentioned that the system assumes a clear date stamp, with Time-to-Live (or “expiration date”) specified, and, last but not least, active involvement of the CSC, either as a whole (given the importance of the task), or in the form of a standing task force, with updates released in minimally yearly intervals.

Acknowledgements

While it might be deemed pompous to include acknowledgments in this type of draft deliverable, I would feel uneasy and close to dishonest without thanking the following colleagues. I owe the realization of the importance of the “clash” between the top-down view and the bottom-up approach to conversations with Thomas Schmidt and to the work done by Hanna Hedeland. Susanne Haaf has kindly read the first version of this pilot and provided comments that made me change or add several details (naturally, the blame for potential misinterpretation and final result is mine to carry). I treated the silence on the part of the other colleagues with whom the first version was shared as a gentle suggestion for me to put some more work into it, which I followed.

References

[unordered and unformatted for now]

[D5c-3] Hinrichs, Erhardt and Vogel, Iris (eds). 2010. “Interoperability and Standards”. CLARIN deliverable. URL

[Kemps-Snijders et al., 2009] “Standards for LRT, v. 6”, a 2009 recommendation by Marc Kemps-Snijders, Núria Bel, Peter Wittenburg, Daan Broeder, Dieter van Uytvanck, Laurent Romary, Erhard Hinrichs and Gerhard Budin. URL.

Hedeland, Hanna (ed.). 2015?. “CE-2014-0421-relevant-formats”. Survey conducted by Hanna Hedeland (more info?). URL

CLARIN Standards Committee wiki page – other recommendations listed there.

Odijk, Jan. 2017. Towards interoperability in CLARIN. Version 2.0. Manuscript circulated among the CSC, 2017-11-20.

Appendix A: Metadata formats

The left-hand values here were mostly copied from Hanna Hedeland’s survey results as well as from Kemps-Snijders et al., 2009. The repertoire and values should be established in concert with, minimally, the Metadata Curation Taskforce.3

metadata type	**parameters**		
CMDI 1.2	bidirectional	recommended	current
COMA			
IMDI			
XML			
CSV			
XSLX			
RTF			
DOCX			
DOC			
RTF			

metadata type	parameters		
ODT			
TEI Header			
Dublin Core (this is content, not format)			
OLAC			
METS			
MPEG21 DID			
MPEG7 “elements of text annotation”			
EAD			
MARC			
ORE			