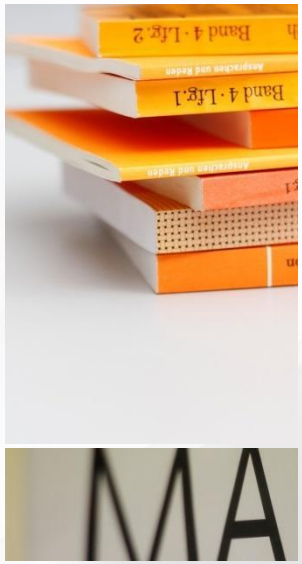

Piotr Bański
IDS Mannheim

Roadmap towards the unification CLARIN-endorsed standards [DRAFT VERSION]

Aix-en-Provence, 26 October 2016

Outline

- About this presentation
- Motivation
- Inspiration
- Main features of the proposed architecture
- A quasi-algorithmic approach to filling out the matrix
- Interlude: TEI flavours
- "Hands-on”: let's bite on this
- Roadmap

About this presentation

- What this presentation is not: yet another list of standard names, by a fairly anonymous guy, to share the fate of those before it
- What it is: a proposal for a dynamic and updateable system to tackle the standardization issues in CLARIN as they occur (and/or pass away)
- S/BP = Standards, *de-facto* standards, best practices

The rationale

- several confusing sources for standards recommendations in CLARIN
  - the sources are far from unified
  - they have different roles to fulfill
  - they have come into existence at different points in time

The rationale (cont.)

- Express need to consolidate the standards recommendation
  - in view of the continuous expansion of CLARIN ERIC and the need for up-to-date guidelines
  - with a view to sustainability
  - to ensure a firm basis for interoperability considerations (cf. Jan Odijk's 2016 position paper)
  - to strengthen and stabilize the links with DARIAH over the PARTHENOS project (its WP4)

Goals

- uniform information on the standards landscape in CLARIN
- binding and common across the centres, old and new
- friendly towards the users
- responsive towards the emerging and receding standards and best practices (S/BP)
- distinguishing among the levels of endorsement (with consequences for interoperability considerations)

Inspirations (selected)

- “common knowledge”
- EAGLES recommendations (three-level CES conformance)
- our own work on ISO CQLF
- previous standards proposals ("Interoperability and standards”, Standards DB, “Standards for LRT”, CLARIN wiki, FAQ, ...)
- Jan Odijk's 2016 proposals regarding interoperability

Grouping of standards (basic)

- Level 1 (baseline level, expected of all centres)
- Level 2 (wider scope, centres that can/have to afford that)
- Level 3 (“special treat”)
- Level D (deprecated!); this doesn't mean that centres mustn't support such standards – they are simply not expected to be supported

Potential mechanisms for grouping

- weighted standards (not all are equal and not all influence the resulting score equally)
- dependencies (e.g., if one then obligatorily another, e.g. IETF BCP 47 → ISO 639)
- **simple, tiered** (chosen here; the dependencies can be expressed and categorised elsewhere)

The graphic that isn't

- This is where a figure should appear, but I will add it in the archival version of these slides. And now, imagine...
  - a tree, 4 siblings: Lev1, Lev2, Lev3, DEPRECATED
  - each of these divided into “general”, “protocols”, “terminology / ontology”, “metadata”, “media”, “general text formats”, “LRT text formats”, “character encoding”
- Alternative presentation, fully equivalent: a set of four tables, with the above subdivisions
- Yet another: fancy visualisation as in the Standards DB

Crucial assumptions

- These are not levels "awarded" to CLARIN centres,
  - but rather levels characterising the CLARIN offer in general, from which centres select
- The membership in each grouping is **fluid**
- It gets **re-evaluated** by this Committee, each year
- Each year, we produce a time-stamped version of the grouping and announce it, here and in correspondence with each centre
- The history of the given S/BP entry is recorded in the DB

Added value?

- confusion → less confusion
- several sources → single source
- clear basis for further steps towards strengthening the interoperability in CLARIN
- recognition and incorporation of the fact that S/BP and the associated recommendations grow, mature, and get obsolete – and we should react to them dynamically

A way to deal with it

- through thorough discussion in the Standards Committee and among the interested parties
- starting here and now, as long as the overall proposal (or at least its core grouping) is accepted; I definitely don't expect to leave with the final version
- need support from the Centres Committee
- finding a way to explain that grading standards is not a territorial issue but rather one of common sense coupled with technological abilities and the users' expectations

In the long run

- Each Centre is expected to declare which ones out of the pool of standards it supports
  - This information can be used for assessing the interoperability potential
  - (If interoperability fails, something is wrong, investigate the causes – but this is a separate story)

- Scratch all the existing online standards lists, as long as they may be taken as normative rather than informative, and replace them with a single resource (the One Ring)

- keep the Standards DB as the informational front-end (it's there, it's fancy, it's flexible)

In the long run..

- The central resource (the One Ring) could be fully dynamic (wiki), but there are known cons (it would get messy quick)
- It could be semi-dynamic (with e.g. me or whoever is appointed, 2 people would be better)
- This solution bears the danger of subjectivity (and may invite non-constructive criticism), but this is why the Committee has the ultimate say in this

In terms of manpower

I can take the role of the “fil[lt]er” from a simple resource towards the Standards DB

What I need:

- feedback, here and later (= soon)
- support in the form of a stamp of approval
  - from this Committee
  - from the Centres Committee
  - eventually from the BoD (?)
- a friendly nudge, from time to time

Assumptions and entailments of the proposed grouping

- Level 1 carries the entailment of being a full endorsement for outside use
- The suggestions on following slides are to be seen as an expression of violable constraints when converting from the existing sources
- Each step of the conversion should be verified (because the base proposals are dated)

Assumptions and entailments..

"state" in the "Standards for LRT" doc (no automation, verification):

- "proven" → maybe Lev1
- "ready" → maybe Lev2
- "in progress" → maybe Lev3

Assumptions and entailments...

“advise” in the “Standards for LRT” doc (no automation, verification):

- “obligatory” → likely Lev1
- “recommended” → likely Lev2
- “neutral” → likely Lev3

Assumptions and entailments....

"recommendation" from the wiki table (again, no automation):

- "obligatory" → maybe Lev1
- "recommended" → maybe Lev1 or Lev2
- "acceptable" → Lev2 or Lev3
- "neutral" → maybe Lev3
- "not recommended" → DEPRECATED

Assumptions and entailments.....

a stamp from a standardizing body → Level1 unless the standard is dated (SGML) or underapplied (ISO 1957)

or the opposite:

- lack of a stamp of a standardization organization: careful examination before Level 1 is assigned

Deprecated

Examples for the DEPRECATED category:

- an exaggerated example for character encoding: EBCDIC
- SGML (TEI P3, CES), ISO 639-2, ISO 1957 (for lexicon markup)
- XML 1.1?
- ...

Shifting group membership

- previous versions of standards:
  - TEI P4 still in Level3 but we should set a date for deprecation, maybe
  - note that several standards for the encoding of transcribed speech may be shifted due to the forthcoming ISO-TEI standard (ISO 24624:2016) that encompasses them all
- standards that don't appear successful
- standards that appear to be on the uptake

Interlude: TEI (vs. XML) flavours

- Recall the recent discussion on media types on the Standards list – there is no single “TEI P5”, and services have to deal with this, somehow, e.g. by using `application/tei+xml;format-variant=tei-iso-spoken`
- I have created a ticket #1483 requesting green light to move towards IETF for an update of xml+tei with the proposed parametres, and to survey the community for others; I recently received a mandate to act on this

Interlude: TEI (vs. XML) flavours..

- This IS big, also because of the following:
  - any parameters postulated for application/xml (as opposed to tei+xml) are in fact a counter-standardization step
  - they simply have no chance of ever getting approved
- The only way to use parameters with the other CLARIN XML proposals (e.g. TCF) is to "go TEI", i.e., to make them TEI applications (with their own ODDs) and then to use the parameters defined for the TEI
- Back to the issue at hand: there is little point in having “TEI P5” on our lists, we have to adopt finer granularity, and keep an eye on similar cases

Finally

If there is time, we can do a hands-on, applying some of the above suggestions to the 2009 "Standards for LRT" document, reexamining its proposals and combining them with others from my dropbox and the Net, however...

... this is not what I think of as the goal of this meeting.

My goal will be achieved if we agree on a roadmap towards a system (close to the one) presented here: a cyclically re-evaluated dynamic resource rather than a one-off list, shakily "valid" for a few years.

Thank you!

banski @ids-mannheim.de