Towards a consolidated LOD vocabulary for linguistic annotations

Agenda / minutes document

# 1. Main points

Apr, 23 2020, 10:00-11:30 CET

* Goal: (community) standard for an RDF representation of linguistic annotations on the web
* GitHub repo <https://github.com/ld4lt/linguistic-annotation>
* Along with RDF-based vocabularies (NIF, WA, etc.) also consider ISO standards (LAF and domain-specific standards)
  + **TODO@Thierry+others**: find a way to make these documents or related documentations (e.g., publications) accessible to all participants, e.g. via GitHub repo
* Before discussing next steps, sub-tasks, etc., create an overview document over weaknesses of NIF, WA and LAF
  + **TODO@Christian+Milan**: “survey preparation”. => This overview document will be the basis for the next telco
* 6 week rhythm for telcos

# 2. Original agenda

(We deviated from this.)

Overall goal today is to decide about approach (*how*) and to discuss possible sub-tasks (*what*)

* Additions to the agenda (please add)
* Overview over problem and context
* Introduction of new participants (if any)
* Discussing the suggested approach

|  |
| --- |
| *Proposal*   * Work on harmonizing NIF, Web Annotation and other vocabularies, using Web Annotation as a basis * Extend the consolidated model both wrt. genericity and explicitness (cf. LAF-based vocabularies above) and support for use cases currently not sufficiently covered (be it from language technology, knowledge engineering, computational lexicography or philology). * Develop a minimal consensus vocabulary that complements Web Annotation with NIF functionalities and generic linguistic data structures; can be an extension of Web Annotation or as a revision of NIF (“NIF 3.0”) or a stand-alone vocabulary (module) that can be used with NIF and WA. * Publish this consensus model as persistent point of reference, e.g., as a W3C Community Report (of LD4LT or a designated, new CG) |

* Introducing the GitHub repository

<https://github.com/ld4lt/linguistic-annotation>

* What to standardize

|  |
| --- |
| *Proposed areas of activity* (=> sub-tasks)   * (Linguistic) Data structures (e.g., *token*) ? * (Linguistic) Data categories (e.g., *noun phrase*) ? * Fragment IRIs and selectors: How to access strings (etc.) in a web document ? * Protocols and API? (possible inspiration: [DTS](https://distributed-text-services.github.io/specifications/)) * Serializations (default format[s], embedding in markup languages) * Anything else? |

* General approach:
  + Sub-task after sub-task or working in parallel?

|  |
| --- |
| *Depends on participants’ interests* |

* + Top-down or bottom-up?
* Top-down: compare main vocabularies and generalize, then extend to other vocabularies

=> priority is to collect problems and to select and to compare vocabularies

* Bottom-up: choose one “base” vocabulary (say, WebAnnotation or NIF) and identify use cases that motivate certain extensions

=> priority is to collect use cases, problems, sample data

|  |
| --- |
| *Proposal*   * We can do both. With NIF and Web Annotation, we have a very good starting point for top-down modelling. At the same time, we need to start collecting sample data and applications to test this model on.   => additional sub-task: elicit/collect use cases |

* Defining sub-tasks

|  |
| --- |
| *Proposal*   * Moderation / overall coordination: 2 persons   + Organize regular telcos, maintain GitHub and documentation * Standardization subtasks: >= 1 person each   + Report on subtask at telcos, organize task discussion (e.g., via sub-telcos, if necessary), maintain documentation * Use case collection: >= 1 person   + Report on use case collection at telcos, maintain GitHub * Other? Could be, for example   + community proxies (say, to TEI standoff SIG)   + any tools to be implemented ? (say, NIF 2.0/2.1 converter, fragment IRI converter)   + duties include progress report at telcos |

* Next steps
  + Naming the child ;)
    1. Please send proposals to LD4LT mailing list (and mention “consolidating linguistic annotations” in the subject line)
  + Assigning sub-tasks
    1. suggestions for the nomination/election process?
  + Next call
    1. 6 week rhythm ? [Doodle for finding a date]

# 2. Participants

CC - Christian Chiarcos

APL - Antonio Pareja-Lora

GS - Giulia Speranza

FF - Francesca Frontini

JBG - Julia Bosque-Gil

JK - Joel Kalvesmaki

JMS - Julián Moreno Schneider

MD - Milan Dojchinovski

PC - Pablo Calleja  
PL - Penny Labropoulou

PMC - Patricia Martín Chozas

TD - Thierry Declerck

# 3. Minutes

* Overview over problem and context  
    
  **Focus**  
    
  APL> What shall we be looking at, what are we aiming at? “Same as always ?” (i.e., replicating ISO standardization processes?)

CC> We are focusing on developing an RDF vocabulary, not from scratch, but grounded in existing standards  
Suggested approach: Go over RDF vocabularies + generalize them + extend them with the requirements for linguistic annotation, specially LAF. Ideally -> avoid the development of something new.  
  
**Bridges to other initiatives, e.g. ISO**  
FF> Are we focusing on formats and/or vocabularies?

CC> We are focusing on developing an RDF vocabulary, possibly API specifications, but not necessarily a format

FF> Concerning vocabularies there is a very big need and we might want to build a bridge to other initiatives. ISO is working on categories for annotation, but it’s a closed effort. <https://www.iso.org/committee/297592/x/catalogue>.   
Apart from LAF & co, maybe another group of standard which might be of interest is the CQLF: <https://www.iso.org/standard/37337.html?browse=tc> <https://www.iso.org/standard/72103.html?browse=tc>

CC> Is there a way to access the (draft) documents of ISO?   
MAF, SynAF...are they based on LAF, correct?

TD> They were published before LAF, the focus in the discussion (internal) was the representation as graphs.

APL> In general, LAF is referenced in the documents, there is declared compliance, but in practice it is not so straightforward. XML is driving the model for this particular type of annotation. Recommended approach within ISO TC37 is now to provide UML models.  
LAF was distributed by Nancy Ide at that time.

SynAF -> comprehensive papers by the SynAF editorial board (links: see below).   
  
**ISO documentation**  
**\*\*Todo-Thierry\*\***: check his records for more info on the documentation/drafts.

APL > FDIS vs. final standard -> only minor changes/proof-reading amendments

CC> Create a private repository and place the documents there? (GitHub - see the link below)

APL> For the SynAF papers, yes (see the following links: <https://hal.inria.fr/hal-00765413/document>, <https://hal.inria.fr/hal-01085219/document>). MAF - there is a HTML version published (by Éric de la Clergerie - <http://atoll.inria.fr/~clerger/MAF/html/index.html>)

CC [after telco]> LAF: <https://www.cs.vassar.edu/~ide/papers/LAF.pdf> (no right to store)

FF> We can try to contact ISO, involve people who work with these standards in the discussion.

CC> Would be good, but discussion should be grounded in material accessible to everyone involved in this discussion

* Introduction of new participants

Giulia Speranza (LOD modeling for archeology)  
Patricia Martín Chozas, Pablo Calleja (Prêt-à-LLOD project)

* Discussing the suggested approach

|  |
| --- |
| *Proposal*   * Work on harmonizing NIF, Web Annotation and other vocabularies, using Web Annotation as a basis * Other points postponed |

**General approach**MD> Agree, but we need to revisit the goal/scope of these formats when deciding what exactly we want to cover with our extensions (direct extension vs. standalone module). Proposed process: detect gaps/issues + decide what we want to cover with extensions + decide whether direct extension or standalone module.

CC> Agreed. We also need to take into account missing features. E.g. String URIs missing in WA; linguistic annotation missing in WA.

**Collection of issues/survey**  
CC> Proposal: first, collect issues within NIF and WA.

Christian + Milan + Thierry -> expressions of interest to be involved in this.   
The dynamics of the discussion (*Sub-task after sub-task or working in parallel?*) can be decided once there are preliminary results in this step.

* Introducing the GitHub repository

<https://github.com/ld4lt/linguistic-annotation>

CC> If we create a survey document, place it in the /survey folder.   
There is also the private repository option -> related to the ISO discussion

* What to standardize [postponed]

|  |
| --- |
| *Pre-telco: Proposed areas of activity* (=> sub-tasks)   * (Linguistic) Data structures (e.g., *token*) ? * (Linguistic) Data categories (e.g., *noun phrase*) ? * Fragment IRIs and selectors: How to access strings (etc.) in a web document ? * Protocols and API? (possible inspiration: [DTS](https://distributed-text-services.github.io/specifications/)) * Serializations (default format[s], embedding in markup languages) * Anything else? |

* General approach (postponed)
  + Sub-task after sub-task or working in parallel?
  + Top-down or bottom-up?

[To discuss after we have some preliminary results for the survey document]

* Defining sub-tasks (partially postponed)

|  |
| --- |
| *Proposal CC*   * Moderation / overall coordination: 2 persons   + Organize regular telcos, maintain GitHub and documentation * Other sub-tasks postponed |

**On moderators**  
TD> Will this be something under the umbrella of LD4LT? Because, if so, maybe the chairs of LD4LT could assist in moderating…  
Volunteers -> contact CC via email. It would be good if we have in this role people who also have connections with another community

* Any other business
  + Wikipedia - contribute and translate
    1. <https://en.wikipedia.org/wiki/Linguistic_Linked_Open_Data> [now en,de,sr,fr; JBG: Spanish version is on its way -> Patricia and myself]
    2. <https://en.wikipedia.org/wiki/Web_annotation>
    3. <https://en.wikipedia.org/wiki/Linguistic_Annotation> => <https://en.wikipedia.org/wiki/Text_annotation#Linguistic_annotation>
    4. <https://en.wikipedia.org/wiki/Overlapping_markup>
* Next steps
  + Next call
    1. 6 week rhythm ? [Doodle for finding a date]

6 weeks proposal -> +2 votes, no objections

[end of regular telco]

|  |
| --- |
| **Main results**   * Discussion underlined the importance of ISO standards for future discussion (both LAF and domain-specific standards)   **=> TODO**: find a way to provide access to these documents or related documentations (e.g., publications), ideally via GitHub repo   * Before discussing next steps, sub-tasks, etc., create an overview document over weaknesses of NIF, WA and LAF   **=> TODO:** “survey preparation”.  This overview document will be the basis for the next telco   * 6 week rhythm for telcos |

# 3. Follow-up discussion on survey preparation

Participants: Christian, Milan, Julia

Contributors: Christian, Milan, others? (Thierry?)

**Context:**CC> Personal feeling: WA more widely used (e.g in digital humanities), but NIF less known (from the outside). NIF has a lower level of standardisation in comparison to WA.   
 **Tentative goal:**MD> The idea is to align NIF and WA or to, on the basis of NIF, extend WA?  
CC> Personal view: Engage the WA community into using NIF, maybe “sell” the outcome of the discussion as extension to WA, and so, attract additional NIF users. As for the relation between the new vocabulary and Web Annotation, we can think of the relation between RDFS and RDF.   
CC> To be discussed in the regular telco  
  
**Next steps for survey**  
MD> Maybe start by making a list of people/companies who we know are using NIF   
CC>   
- Effort started last year (NIF used by Semantic Web Company, Leipzig, U Bielefeld, DFKI, other direct Leipzig collaborators).   
- Probably Leipzig knows better about who exactly is using NIF nowadays.   
- In contact with Penny wrt. to the OpenMinted project, in the end they went for WA and they have survey that lead that discussion.

CC [after telco]> We provide NIF-based vocabularies, esp. CoNLL-RDF (for TSV annotations in NLP) and Ligt (for interlinear glossed text). Both are used by external (academic) partners, but CoNLL-RDF and Ligt only build on 4-6 NIF concepts and neither the NIF ecosystem nor NIF String URIs.

MD> Start collecting the issues + also analyse LAF data model (to have a sense of its coverage)

CC> - As for shortcomings of NIF -> references + comparison in the LDL book (\*\***CC To-do\*\***: send chapters)  
  
**Logistics:** Next meeting in ~1 week. Thursday 30.04 - 10 CEST. Skype