Towards a consolidated LOD vocabulary for linguistic annotations

Agenda / minutes document

|  |
| --- |
| **Important note**: Parts of the original content of this document have been migrated to <https://github.com/ld4lt/linguistic-annotation>. Also see there for minutes of earlier meetings. This document is reserved for the agenda of the next telco and for creating minutes of the current one. |

# 1. Background & Motivation

Several vocabularies currently in use, cf. <https://link.springer.com/book/10.1007%2F978-3-030-30225-2> (drafts of relevant chapters can be shared on a private basis, request via ResearchGate from <https://www.researchgate.net/profile/Christian_Chiarcos/publications>:

* Representing Annotated Texts as RDF (Chap.5)
* Chap. 6 Modelling Linguistic Annotations (Chap. 6)
* Chap. 8 Linguistic Categories (Chap. 8)

Most frequently used for linguistic annotation (in a LOD context) are

* NIF (NLP Interchange Format, <https://persistence.uni-leipzig.org/nlp2rdf/>, <https://github.com/NLP2RDF>)
* Web Annotation / Open Annotation (<https://www.w3.org/TR/annotation-model/>)

Full overview on relevant LOD vocabularies and their use now under <https://github.com/ld4lt/linguistic-annotation/tree/master/survey>

Based on a survey conducted in 2019, NIF and Web Annotation are being actively used in both academia and industry, but issues exist with respect to interoperability and expressivity. Web Annotation is a W3C recommendation and thus stable. NIF 2.0 is a stable vocabulary, as well, and referred to in W3C standards (ITS), but its development is coordinated by a single institution. More recent NIF extensions (NIF 2.1 additions for provenance) seem to be partially documented only (there is no complete definition for NIF 2.1), and updates seem to have ceased since 2016.

## 

## Proposal

* Work on harmonizing NIF and Web Annotation
* 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”).
* Publish this consensus model as persistent point of reference, e.g., as a W3C Community Report (of LD4LT or a designated, new CG)

# 2. Previous discussions (main points)

Full minutes can be found under <https://github.com/ld4lt/linguistic-annotation/tree/master/doc/minutes>

Feb, 26 2020, 14:00-15:00 CET

* discussion to be continued via LD4LT (later possibly within a new W3C CG)
* Goal: W3C CG report with designated specifications for linguistic annotations on the web

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

Jul, 09, 2020, 10:00-11:00 CET

* Focus on discussing relations with ISO
  + Suggestion to develop specifications independently from ISO, but consider publicly available information as a source of inspiration, focus on NIF+Web Annotation (JK: +1, APL: +1)
  + This means that data structures required by ISO standards should be expressible in an RDF serialization, too
  + This means that compliance with ISO standards is not a goal (but a welcome side-effect, if possible)
  + Private/proprietary information about ISO standards must not be discussed.

# 3. Current telco: 2020-07-09

Doodle poll for fixing the date: <https://doodle.com/poll/v8xqhfrgx623knpk>

Date 2020-07-09, 10:00 CEST (Berlin time)

Video conferencing software: Zoom

<https://uni-frankfurt.zoom.us/j/96129388773?pwd=ZXZlbjVtNHc0SkE3YkNaN0pyREtSZz09>

* Meeting ID: 961 2938 8773
* Password: 402092
* One tap mobile
* +496971049922,,96129388773#,,,,0#,,402092# Germany
* +493056795800,,96129388773#,,,,0#,,402092# Germany

Dial by your location

* +49 69 7104 9922 Germany
* +49 30 5679 5800 Germany
* +49 69 3807 9883 Germany
* +49 695 050 2596 Germany
* Meeting ID: 961 2938 8773
* Password: 402092
* Find your local number: https://uni-frankfurt.zoom.us/u/acb1escdE

If there are any problems with the connection, please check **here** for updates regarding the connection link or software, if nothing found, please contact Christian (christian.chiarcos) via Skype chat

## 3.1 Agenda

Overall goal for the moment is to decide about goals (*why*), approach (*how*) and to discuss possible sub-tasks (*what*), in that order

* Additions to the agenda (please add)
* Introduction of new participants (if any)
* Summary of last telco
* Update on survey (vocabularies and requirements)
* Discussing the suggested approach
  + General goals
  + What to standardize
  + Sub-task after sub-task or working in parallel?
  + Top-down or bottom-up?
  + Defining sub-tasks
* Next steps
  + Naming the child ;)
  + Assigning sub-tasks
  + Next call

## 3.2 Participants (please add yourself)

CC - Christian Chiarcos

APL - Antonio Pareja-Lora

RS - Ranka Stanković

MD - Milan Dojchinovski

FK - Fahad Khan

JBG - Julia Bosque-Gil

TD - Thierry Declerck

JM - John P. McCrae

JK - Joel Kalvesmaki

EMP - Elena Montiel-Ponsoda

PMC - Patricia Martín-Chozas

## 3.3 Minutes

* Additions to the agenda (please add)
* Introduction of new participants (if any)
* Summary of last telco

|  |
| --- |
| * 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 |

* Update on survey over vocabularies and requirements
  + <https://github.com/ld4lt/linguistic-annotation/blob/master/survey/required-features.md>
    1. Overview over part (A) LLOD compliancy
    2. **TODO@all**: please think about additional requirements
    3. **TODO@next time**: parts (B) to (F): other requirements
  + <https://github.com/ld4lt/linguistic-annotation/tree/master/doc/iso>
* Relation with ISO
  + Access to proprietary content via <https://github.com/ld4lt/linguistic-annotation-restricted> ?
    1. Not for ISO (German copyright permits a 15% excerpt of text for scientific research)
  + Subsequent discussion about ISO involvement (TD, APL, CC, others)
    1. We cannot legally share ISO drafts nor standards within the group, ISO contributors are held to disclose conceptual models such as UML diagrams. In a public discussion, it is impossible to stay in sync with ISO developments
       1. ISO standards are actively being developed, e.g., a current trend to intensify ties with the TEI and to focus on XML serializations
       2. For the moment, we focus on a specification independently from ISO. This could be standardized at some later point (by ISO or otherwise), but ISO compliancy is not a priority right now
       3. Because of the different serializations envisioned, ISO and LD4LT specifications can co-exist as they address different user groups
       4. In order not to reinvent the wheel, we should consider results (publicly available documents) of the ISO standardization process as a source of inspiration. This would be published papers, and published papers only, even if they do not reflect the latest development of the respective ISO standard.
          1. APL: we can also use and provide summaries/excerpts of ISO-internal documents.
          2. APL: Also use the Data Category Registry: <http://isocat.tbxinfo.net/rest/user/guest/workspace>.
    2. Publicly available papers about ISO papers are collected in <https://github.com/ld4lt/linguistic-annotation/tree/master/doc/iso>, **TODO@all**: please add to that
    3. Tentative consensus (tbc):
       1. JK: draft in turtle, avoid any XML issues (CC: +1)
       2. CC: develop specifications independently from ISO, but consider publicly available information as a source of inspiration, focus on NIF+Web Annotation (JK: +1, APL: +1)
          1. This means that data structures required by ISO standards should be expressible in an RDF serialization, too
          2. This means that compliance with ISO standards is not a goal (but a welcome side-effect, if possible)
          3. Private/proprietary information about ISO standards will not be discussed.
* Next call
  + 6 weeks [Doodle for finding a date]
* Summary: Discussing the suggested approach (largely postponed)

|  |
| --- |
| *Consensus so far*   * Work on harmonizing NIF, Web Annotation and other vocabularies, using Web Annotation as a basis * Publish this consensus model as persistent point of reference, e.g., as a W3C Community Report (of LD4LT or a designated, new CG) * 2020-07-09 (tbc): develop independently from ISO, take their publications as an inspiration, no (current) plans for ISO compliance/standardization   *Open*   * Depending on survey results and original goals of these formats, decide what to cover and whether to provide extensions as direct extensions or a standalone module   *Other points from the original proposal*   * 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. |

# 4. For future discussions

* 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 |

* Naming the child ;)
  + Please send proposals to LD4LT mailing list (and mention “consolidating linguistic annotations” in the subject line)
* Assigning sub-tasks
  + suggestions for the nomination/election process?