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: 2020-11-26

Doodle poll for fixing the date: https://doodle.com/poll/2bvb78z42tpsa5fm

Date 2020-11-26, **11:00 CET** (Berlin time)

Link for participation (NEW): <https://meet.google.com/oig-osrm-jkn>

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 (please add or comment)

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)
* Introduce and discuss relation with Nexus Linguarum
  + <https://nexuslinguarum.eu/the-action/join-us>
* Summary of last telco
  + Approve tentative consensus on relation to ISO
* Update on survey (vocabularies and requirements)
  + Verify if NIF 2.0 or WebAnnotation are sufficient to address the requirements of linguistic annotations on the web
  + If not, how and where best we want to suggest extensions to.
* Discussing the general approach
  + 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 ;)
  + Identify sub-tasks
  + Scheduling calls

## 3.2. Participants (please add yourself with initials and -- optional -- affiliation)

CC - Christian Chiarcos, Goethe University Frankfurt

SM - So Miyagawa (sorry, I will be a bit late)

JBG - Julia Bosque-Gil

RS - Ranka Stanković

MD - Milan Dojchinovski

JK - Joel Kalvesmaki ([Text Alignment Network](http://textalign.net) / [Government Publishing Office](http://gpo.gov))

FK - Fahad Khan (can only stay till 11:40-45)

GS - Gilles Sérasset (Univ Grenoble Alpes)

DG - Dagmar Gromann

PL - Penny Labropoulou

APL - Antonio Pareja-Lora

GV - Giedre Valunaite Oleskeviciene

MPdB - Maria Pia di Buono

Excused:

MR - Mike Rosner

## 3.3 Minutes

### a. Nexus Linguarum

COST action: Scientific network on investigating linguistic data science, focus on LLOD technologies

Funded by EU (= travels, meetings, training schools, research visits; no personnel), participation not restricted to participants from Europe

<https://nexuslinguarum.eu/>; <https://nexuslinguarum.eu/the-action/join-us>

Involves a Working Group on Linked Data based language resources, with a task on data modelling.

CC: suggestion to conduct telcos in a conjoint fashion with this task. Benefits: Some infrastructural support, better outreach. Already now, many LD4LT participants are active in Nexus. Entails that joint publications are partially attributed to Nexus.

No objections => LD4LT annotation telcos will be listed on Nexus calendar

### b. ISO / last telco

approve/revise tentative consensus from last telco:

|  |
| --- |
| Develop specifications independently from ISO   * consider publicly available information as a source of inspiration   + data structures required by ISO standards should be expressible in an RDF serialization, too   + compliance with ISO standards is not a goal (but a welcome side-effect)   + private/proprietary information about ISO standards should not be discussed. |

No objections

=> focus on NIF + Web Annotation (or other RDF vocabularies) as a basis

### c. NIF vs. WA (vs. ISO)

* Update on survey (vocabularies and requirements)
  + Goal of survey: decide on which vocabulary to build on
  + <https://github.com/ld4lt/linguistic-annotation/blob/master/survey/required-features.md> (so far, mostly compiled by MD and CC)
  + Verify if NIF 2.0 or WebAnnotation are sufficient to address the requirements of linguistic annotations on the web
    - review <https://github.com/ld4lt/linguistic-annotation/blob/master/survey/required-features-tab.md> (compiled from required-features.md)
      * Brief overview over the first 10 features
      * This high-level overview is largely unpractical, possibly better to discuss features in-depth.
      * Suggestion: use github issues
    - Suggestion (CC; AFTER TELCO):
      * After the selection of candidate vocabularies is concluded, prepare a discussion of requirements and coverage of (up to) about, say, 10 features per telco and to move on to the next 10 features, then [we need to have more frequent telcos, then]
      * Iterate until we can decide for a “host” vocabulary as a starting point
  + If NIF or WA are insufficient in their current form, decide how and to which vocabulary we want to suggest extensions to.
    - No conclusions yet, consensus to broaden the band-width of vocabularies covered in the survey first
      * See Overview document (<https://github.com/ld4lt/linguistic-annotation/tree/master/survey>) for candidate vocabularies
      * Text Encoding Initiative
      * LAPPS Interchange Format: LIF
      * NAF RDF: to be checked
      * ?LMF, models from terminology: only to the extent they do not fall under the scope of “lexical resources” (=> should be discussed on OntoLex, then)
  + DG: consider terminological data categories as provided originally in ISOcat (now available in TermWeb <https://datcatinfo.termweb.se/termweb/app>)
    - CC: problem of TermWeb is that there are no machine-readable URIs - discussed with Sue Ellen Wright: would be needed
    - PL: CLARIN Concept Registry: to be revived
    - FK: lexinfo?
    - CC: we should provide vocabulary to link with *any* such resource but leave unspecified which system of data categories we refer to. We have seen some fluctuation among them. Minimal requirement would be to have URIs.
  + JK: Can I suggest introducing a question mark to the two documents, where there is reasonable uncertainty, and needs discussion
    - CC: in required-features.md, this is done (but not systematically and in comments only), for <https://github.com/ld4lt/linguistic-annotation/blob/master/survey/required-features-tab.md>, we would need to adjust the extractor script (<https://github.com/ld4lt/linguistic-annotation/blob/master/survey/required-features.py>)
    - JK: I will work on a way to develop required-features.md into the table + other things I think could be helpful

### d. Discussing the general approach

* Suggestion
  + Continue requirement analysis / survey

=> decide about the starting point

* + Elicit use cases

=> Use these use cases to develop the vocabulary

* + Decide on what to standardize
    - Core vocab
    - API
    - Serializations
    - => Best Practice guidelines
  + Decide on procedure
    - Sub-task after sub-task or working in parallel?
    - Top-down (model-driven) or bottom-up (use case-driven)?
    - Defining sub-tasks
* Request
  + We’re looking for volunteers for co-moderators

Summary of earlier discussions:

|  |
| --- |
| *Consensus so far*   * Work on harmonizing NIF, Web Annotation and other vocabularies * Publish this consensus model as persistent point of reference, e.g., as a W3C Community Report (of LD4LT or a designated, new CG) * Confirmed: 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. |

### e. Next steps

* Scheduling calls

Suggestion (via email): For narrowing down a regular 6-weekly slot, please document your time and day preferences in this document (in CET ;) so that we can include these in the discussion of future meeting slots. If there is a significant number of participants from time zones outside Europe (or Africa), we can consider having telcos at regularly alternating times.

Next call in early January, coordination by Doodle

* Naming the child ;)

Suggestions welcome

* Identify sub-tasks & use cases
* TODOs:
  + look into <https://github.com/ld4lt/linguistic-annotation/blob/master/survey/required-features.md>, either request edit access (mail to CC) or create an issue
  + Think about use cases

# 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?