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.

Nov 26, 2020, 11:00-12:00 CET

- Consensus on relation with ISO approved
    - develop specifications independently from ISO
    - focus on NIF + Web Annotation (or other RDF vocabularies) as a basis
- Participants approved association with COST Action Nexus Linguarum (https://nexuslinguarum.eu/; https://nexuslinguarum.eu/the-action/join-us; Involves a Working Group on Linked Data based language resources)
    - "Association": Future LD4LT meeting on linguistic annotation will also be featured on internal Nexus calendar and events will be co-organized with Nexus; Nexus may sponsor travels and training, not limited to participants from the EU
- Started requirement analysis and methodology for that
    - https://github.com/ld4lt/linguistic-annotation/blob/master/survey/required-features-tab.md
    - Overview over first 10 features
        - high-level overview is largely unpractical, possibly better to discuss features in-depth.
        - Live review during the meeting seems to be ineffective

Dec 04, 2020, 12:00-13:00 CET

- Preparatory meeting for organizing a face-to-face and/or virtual meeting at LDK-2021
- See separate notes under https://docs.google.com/document/d/1jlSt7NqzkLHW6txP6_SniohAoICW3FBSvuUTumBdRXs/edit?usp=sharing, archived under https://github.com/ld4lt/linguistic-annotation/tree/master/doc/meetings/ldk-2021
    - Proposal submitted Dec 06, 2020, approved Dec 18, 2020 by LDK workshop chairs, Sara Carvalho and Renato Souza
    - To be held as a half-day (4h) workshop on June 17th, 2021 (W3C Community Group day) details tba.

3. Current telco: 2021-06-01

Date 2021-06-01, **11:00 CET** (Berlin time)

Link for participation: https://meet.google.com/bzr-ipib-wvv

Agenda:

- Status update and wrapup
- Preparing the LDK workshop (Sep 4, 2021)
- Telco organization & moderation

Participants

- CC Christian Chiarcos
- TD Thierry Declerck
- JK Joel Kalvesmaki
- FM Francisco Mambrini
- MD Milan Dojchinovski

Minutes

- Status update and wrapup
    - Basic idea:
        - Provide the functionality of NIF, but with linguistically adequate data structures
        - Provide the functionality of Web Annotation, but -"- (and less verbose)
        - Provide the functionality of the Linguistic Annotation Framework (+ accompanying ISO standards), but in RDF
    - Goal:
        - Accommodate practical use cases and requirements from downstream applications, e.g., web technology in general (Web Annotation!), applied NLP (web services!), corpus linguistics (querying!), and digital humanities (TEI!)
        - Eliminate the need of consumers and providers of linguistic annotations on the web to implement divergent protocols
        - Develop a specification that is maximally backward-compatible with both Web Annotation and NIF to facilitate migration of existing technology
    - Actions so far:
        - Analyzed requirements and the extent to which vocabularies fulfill them: https://github.com/ld4lt/linguistic-annotation/blob/master/survey/required-features-tab.md
        - This is progressing slow, however, general consensus that no existing vocabulary accommodates *all* requirements
    - Getting more focused
        - More frequent calls (4 weeks), 2 moderators (thanks, Thierry!)
        - Publish outcomes of requirement analysis and survey as a paper
            - Semantic Web Journal?
                - MD: Nexus could (probably) cover costs of the Open Access publication
            - Deadline end-2021 or Jan 2022
                - Alternatively in an LREC paper?
            - Have designated sub-telcos on this topic.

        - Multiple series of sub-telcos with smaller group of attendants between those (i.e., every other 2 weeks); results will be reported at overall telco
            - requirement analysis => paper discussion
                - **TODO@Christian**: prepare
            - Fragids => see https://github.com/Arithmeticus/writing-fragids
                - **TODO@Joel**: coordinate
- Preparing the LDK workshop (Sep 4, 2021)
    - Submission: https://raw.githubusercontent.com/ld4lt/linguistic-annotation/master/doc/meetings/ldk-2021/submitted_proposal.pdf
    - TD co-organizer
    - Bridget probably no longer available: **TODO@CC**: confirm
    - Agenda: https://docs.google.com/document/d/1jlSt7NqzkLHW6txP6_SniohAoICW3FBSvuUTumBdRXs/edit?usp=sharing
    - Background presentations
        - CC: WA
        - MD: NIF
        - TD: ISO
        - FK: TEI
        - JK: Fragids
        - DTS?: **TODO@FM**: reach out
    - Discussion
    - Use Case discussion
        - FM: LiLa confirmed
        - Other see agenda
        - More topics welcome
    - **TODO@Christian**: consolidate agenda with LDK time frame
    - **TODO@Thierry**: publish agenda at LDK website
- Telco organization & moderation
    - TD joining as co-moderator, more frequent calls, but more focused discussions/concrete joint work in designated sub-telcos
        - Every 4 weeks overall telco, with subtelco(s) in every other 2 weeks (in between)
    - **TODO@CC**: organize (Doodle/invite) overall telco in 4 and paper telco in 2 weeks
    - **TODO@JK:** organize (Doodle/invite) fragids telco in 6 weeks
    - Next general telco: Thierry on SynAF?

3. Past telco: 2021-01-15

Date 2021-01-15, **11:00 CET** (Berlin time)

Permanent link for participation: 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 telcos
- LDK meeting plans
- Technical issues:
    - Survey table
- Requirement analysis (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
BA - Bridget Almas, Alpheios Project
JK - Joel Kalvesmaki (ICOR, CUA [fellow] / TAN / GPO)

IK - Ilan Kernerman, KD - Lexicala
JM - John McCrae
TD - Thierry Declerck
JBG - Julia Bosque-Gil
PL - Penny Labropoulou

Move your name up (or add it above) to confirm participation ;)

Other regular participants:

SM - So Miyagawa
RS - Ranka Stanković
MD - Milan Dojchinovski
FK - Fahad Khan
GS - Gilles Sérasset (Univ Grenoble Alpes)
DG - Dagmar Gromann
APL - Antonio Pareja-Lora
GV - Giedre Valunaite Oleskeviciene
MPdB - Maria Pia di Buono
MR - Mike Rosner

3.3 Minutes

(feel free to add your [planned] contributions before or during the call)

a. Planned meeting at LDK

- Telco Dec 04, 2020, 12:00-13:00 CET
- Preparatory meeting for organizing a face-to-face and/or virtual meeting at LDK-2021
- See separate notes under https://docs.google.com/document/d/1jlSt7NqzkLHW6txP6_SniohAoICW3FBSvuUTumBdRXs/edit?usp=sharing, archived under https://github.com/ld4lt/linguistic-annotation/tree/master/doc/meetings/ldk-2021
    - Proposal submitted Dec 06, 2020, approved Dec 18, 2020 by LDK workshop chairs, Sara Carvalho and Renato Souza
    - To be held as a half-day (4h) workshop on June 17th, 2021 (W3C Community Group day) details tba.
    - **Update Jan 2021**: LDK to be moved to Sep. We signalled tentative support for the idea to move the LDK meeting along with it -- requires discussion.
    - overview/intro call in early summer?
        - TD: SynAF RDF
        - BA: DTS
        - JK: lex. Morph. Data with standoff?

            - TD: ONtoLex morph restarting
        - IK: reduncancies with OntoLex, etc. ? to be discussed
            - Suggestion: discuss there
    - Suggestion: both a longer call in early Summer and a meeting at LDK
        - TODO@CC: Doodle for finding slot
    -

b. Technical issues: Update on survey table

- Old table under https://github.com/ld4lt/linguistic-annotation/blob/master/survey/required-features-tab.md
- Novel extractor by JK => https://raw.githubusercontent.com/ld4lt/linguistic-annotation/master/apps/parse%20survey.xsl-output.html
- Novel features:
    - Visually mark uncertainties
    - Access to comments from survey table
    - Score aggregation
    - Direct link with requirements document (https://github.com/ld4lt/linguistic-annotation/blob/master/survey/required-features.md)
- Not deployed on the web yet

c. Update on survey: NIF vs. WA

- Main document: https://github.com/ld4lt/linguistic-annotation/blob/master/survey/required-features.md
- Goal of survey: decide on which vocabulary to build on
- https://github.com/ld4lt/linguistic-annotation/blob/master/survey/required-features.md
- 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)
        - Nov 2020: Brief overview over the first 10 features
        - This high-level overview is largely unpractical, possibly better to discuss features in-depth.
        - ? github issues instead
        - ? 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
    - New vocabularies, A feats until end of Feb / next telco
        - TEI? - CC + JK
        - SynAF(-RDF) - TD [+ISO standards]
        - LAPPS? - [PL, but LATER]
        - NAF(-RDF) - ?? [TD to ask Antske Fokkens / Piek Vossen, TODO: reminder]
- If NIF or WA are insufficient in their current form, decide how and to which vocabulary we want to suggest extensions to.
    - Nov 2020: 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
- Data categories
    - CC: suggestion to focus on structures that can be annotated in the first step, then look into categories and values for annotation
    - From last time (updated)
        - ISOcat (now available in TermWeb https://datcatinfo.termweb.se/termweb/app -- no machine-readable content)
        - PL: CLARIN Concept Registry: only used within CLARIN, not extended upon public request
        - FK: lexinfo?
        - CC: OLiA? 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.

d. Discussing the general approach [postponed]

- 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)
> - 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

  Move to 4-weekly, Thu 11:00 CET
  For next time: conflict with EUROLAN => Doodle poll to confirm week

- Naming the child ;)
  Suggestions and comments welcome [we collect]
    - JK email 2021-01-11: "Linguistic Annotations", prefix "la:"
    - CC (in this doc) 2021-01-11: "Web Annotation for Language Technology", prefix "walt:"

- JK: alt:, Annotation for Language Technology
- CC (in this doc) 2021-01-11: “NIF 3.0”, prefix “nif:”

- Identify sub-tasks & use cases
    - Fragment identifiers? “Literature FragIDs” (?)
        - JK, BA, CC interested in pursuing this discussion
        - Can run in parallel with general annotation telcos, and then be integrated

- 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) - 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?