Towards a consolidated LOD vocabulary for linguistic annotations

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

Agenda / minutes document

BTW: Currently migrating the content of this document to <https://github.com/ld4lt/linguistic-annotation>. Please do not edit here until 2020-03-25.

# 1. Preparation

## 1.1 Background

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.

## 1.2 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

Open questions:

* Would there be interest to contribute to the development or use such a vocabulary?
* Where to discuss this best?

## 1.3 Expressions of interest

In order to prepare a first joint call and to decide where to discuss this further and how to proceed, Christian C. reached out to several mailing lists and community groups in January 2020, asking for expressions of interest:

* Open Linguistics (OKF WG / Google Group)
* Open Annotation (W3C CG)
* Linked Data for Language Technology (W3C CG)
* SIG LOD (ADHO SIG)
* NLP2RDF mailing list
* Nexus Linguarum WP1 mailing list (Cost action)
* Pret-a-LLOD (H2020 RIA)
* SIGANN (ACL SIG)

Selected responses (alphabetical, incomplete, add yourself):

* Samuel Akinbo
* Vladimir Alexiev, OntoText, Bulgaria
* Julia Bosque-Gil, U Zaragoza, Spain
* Thierry Declerck, DFKI Saarbrücken, Germany
* Milan Dojchinovski, INFAI@Leipzig Uni, Germany / CTU in Prague, Czech Republic
* Daniela Gifu, Alexandru Ioan Cuza University of Iasi, Romania
* Timo Homburg, HS Mainz, Germany
* Joel Kalvesmaki, Director, Text Alignment Network
* Sebastian Nordhoff, Zentrum für Allgemeine Sprachwissenschaft, Germany
* Emilie Page-Perron, UCLA
* Antonio Pareja-Lora, UCM / ATLAS (UNED), Madrid (Spain)
* Maria Pia di Buono, L’Orientale University of Naples, Italy
* Georg Rehm, DFKI Berlin, Germany
* Élie Roux, [Buddhist Digital Resource Center](https://www.tbrc.org/), France

## 1.4 Organizational

Finding a time slot: <https://doodle.com/poll/6dh3r7v9h3mgzhgw> (closed, 26.02.2020, 14:00-15:00 CET)

Choosing conference software:

* We expect more than 10 people, so Skype and Hangouts are out.
* Google Meet: Not a free service, requires subscription by moderator
* Zoom: Basic tariff is free, but max 40 minutes; we can plan two subsequent 40 min meetings, with regular duration of 60 min
* We go for 2 subsequent Zoom meetings, links below

# 2. Minutes (Feb, 26 2020, 14:00-15:00 CET)

## Zoom links

14:00-14:40 https://us04web.zoom.us/j/805608955

14:40-15:00 <https://us04web.zoom.us/j/662012037>

In case of problems with these links: Contact me via Skype chat (christian.chiarcos)

## Participants

Please add yourself, with initials and affiliation, use your initials to mark your contributions in the minutes document

* CC - Christian Chiarcos, Goethe Universität Frankfurt, Germany
* SN - Sebastian Nordhoff, Zentrum für Allgemeine Sprachwissenschaft, Berlin
* GR – Georg Rehm, DFKI (if I can make it, time-wise)
* MD - Milan Dojchinovski, INFAI@Leipzig Uni / CTU in Prague
* JBG - Julia Bosque-Gil, University of Zaragoza (Unizar), Spain
* PL - Penny Labropoulou, ILSP/ARC, Greece
* ER - Elie Roux, Buddhist Digital Resource Center, France
* JK - Joel Kalvesmaki, Washington, DC
* DL - Dorielle Lonke, KD, Israel
* EMP - Elena Montiel-Ponsoda, Universidad Politécnica de Madrid, Madrid, Spain
* KVZ - Karolina Victoria Zaczynska, DFKI
* JMS - Julián Moreno-Schneider, DFKI
* APL - Antonio Pareja-Lora, Universidad Complutense de Madrid, Spain
* AFK - Anas Fahad Khan, ILC-CNR, Italy
* MPDB - Maria Pia di Buono, University of Naples “L’Orientale”, Italy
* FF - Francesca Frontini, Université Paul-Valéry Montpellier 3, France
* PC- Pablo Calleja, Universidad Politécnica de Madrid (UPM), Spain
* AL - Andis Lagzdiņš, Tilde, Latvia
* CF - Christian Fäth, Goethe Universität Frankfurt, Germany
* MI - Max Ionov, Goethe Universität Frankfurt, Germany
* DG - Daniela Gifu, Alexandru Ioan Cuza University of Iasi, Romania

## Agenda

1. Additions to the agenda (please add)
2. Brief problem statement
3. Introduction of participants
4. Collection of selected NIF/WA issues

*Do we actually need to do something?*

1. Discussion about how to organize future discussions

*What would be the ideal outcome?*

*Over an existing W3C CG, SIG or mailing list?*

*Creating a repository to collect sample data, documentation and literature?*

1. Possibilities for hosting the discussion

Minutes

1. Additions to agenda

None

1. Brief problem statement

(see background description above)

1. Brief introduction of participants

* Some keywords and interests that came up:
  + Interlinear glossed text
  + Representation of morphology, different serializations
  + Representation of lexical domain, bridge to corpus
  + Cultural heritage (bridge between corpus annotation and lexicography)
  + Enrichment of digital editions with ontological information (--> CLARIN)
  + Interoperability between services and tools
  + Reuse of metadata descriptions in corpora
  + Extension/exploration of aspects not currently covered by NIF
  + [general interest] Application to digital humanities
  + [general interest] Monitor progress on this line
  + Standardisation → Interest → European Commission

1. Collection of selected NIF/WA issues / Action points

Do we need to do anything? Assuming so...

* + … E.g. Which issues have you come across?
    - [PL + CC] Web Annotation and NIF were discussed and considered but at the end discarded in OpenMinded. Did not consider neither WA nor NIF in the OpenMinded project. WA is an output format only, some issues with syntax. Both options have been discussed.
    - [CC]:
      * No linguistic data structures in WA
      * Hard to represent **syntax** with NIF as soon as you have multiple empty elements.
      * Gaps in morphology (NIF)
      * Does not support morphology and frame semantics

SHIFT TO OTHER ZOOM CALL

* + Issues with RDF?
    - [JK] RDF triples understood as a claim
      * Who is making that claim? How to keep provenance
      * How can we register different opinions/statements

=> general issue on RDF, keeping provenance of claims.

* + - [CC] - introduce mechanisms for provenance recording, cf. NIF 2.1 extensions
    - [ER]: word vs. sentence, complicated definitions, e.g. for “word”. Sentences are also ambiguous.
    - [CC]: should not be defined as core data structures (as in NIF), but externally defined/extensible data category inventory.
  + Issues with NIF?
    - [MD] Possible next step to collect issues and problems with NIF. Sebastian Hellman in his paper from 2013 -> how this two mechanisms can be aligned (diagram)
    - [CC] String URIs in web annotation - Develop a specification? Possibility : Web Annotation → possible extension → string URIs, cf. Sebastian Hellmann’s NIF-WA harmonization
    - [EM]: some issues with NIF and need of extensions (motivated by Lynx project)

1. Discussion about how to organize future discussions
   * preceding discussion -> Need to have this discussion, this can lead to a long list, but yes, we want to have this list
   * *What would be the ideal outcome?* (discussion, ideally with a report / consensus vocabulary in the end)
   * *Over an existing W3C CG, SIG or mailing list?* [=> 6.]
   * *Creating a repository to collect sample data, documentation and literature?* [postponed]
   * **Consensus to continue discussions. No objections to maintain this discussion or to provide a report in the end**
2. Where to host this discussion?

* Options collected beforehand:

[Linked Data for Language Technology](https://www.w3.org/community/ld4lt/) (W3C CG) - 93 participants

[Open Annotation](https://www.w3.org/community/openannotation/) (W3C CG) - 128 participants

[Open Linguistics](https://linguistics.okfn.org/) (OKF WG / Google Group)

[SIG LOD](http://digitalhumanities.org/lod/) (ADHO SIG) - last post in 2014

[Nexus Linguarum WP1 mailing list](https://www.cost.eu/actions/CA18209/#tabs%7CName:overview) (COST Action)

[SIGANN](https://www.cs.vassar.edu/sigann/) (ACL SIG)

[NLP2RDF](https://site.nlp2rdf.org/) (NIF mailing list)

* + [GR] W3C group? It would make sense to frame this with a targeted audience → European, with the goal to develop a standard + ties to the recent development in the European community (but not limited to European interested parties)
    1. [CC]: Advantages/restrictions: admits individual memberships, no entrance/membership fee required.
    2. [CC] LD4LT group still active mailing list, but biased towards language technologies, maybe not an issue here
    3. Question (during vote): would that be felt as energizing the discussion or as a hostile takeover?

[CC <speaking as a co-chair>] as energizing

* + [GR] OA transformed to Web Annotation, but OA still active?
    1. [CC] Does still exist, but not active/not much happening there
  + [FF] SIG LOD (ADHO SIG) still active
    1. [CC] recently rebooted, but using this → discussion centered on the digital humanities aspect
  + [CC] Open Linguistics: focus too broad, no obligation to provide a deliverable
  + [CC] ACL SIG: requires ACL membership, SIGANN not very active
  + Form new W3C CG?
    1. [GR] : Starting something new has the advantage that gives the impression that this has the potential to be continued as a COST action/EU project/other initiative, etc (in contrast to reviving an old group)

Vote among different possibilities to host the discussion. Results:

* Linked Data for Language Technology CG (LD4LT): 14 votes
  + CC, MI, ER, MD, JBG, AL, APL, CF, AFK, FF, JK, PC, PL, MPDB (GR: for the moment)
* New W3C CG: 5 votes
  + MD, AL, JBG, MI, JMS (GR: in the longer perspective)
* Other options: 0 votes

Discussion:

[GR] The LD4LT mailing list is there, we can use it and, if after some time, we come up with a new name and we have a clear goal to create something new, then we can open a new group (suggestion)

[CC] +1 to that, possible that this will even become necessary  
**\*No objections to this plan\***

**Consensus:**

* **Use LD4LT for the moment, create a designated W3C Community Group if/when necessary**

Action Points:

* [CC] People should join to the LD4LT group until March 15, we start the discussion then, in case of difficulties, contact me

**Discussion to be continued on the LT4LT mailing list, announcement there on March 15**

Appendix: LD4LT description (revised 2020-03-20):

This group aims to consult with current and potential users of linguistic data to assemble use cases and requirements for Language Technology Applications that use Linked Data. The results will be used to guide future interoperability, research and development activities spanning the language technology and linked data domains. Potential users are companies and public bodies involved in natural language processing, language resources, content management, the language services and localisation industry and other applications of content analytics techniques used in search, recommender systems, sentiment analysis and terminology management. The group does engage with users through surveys, international events and training activities organized in conjunction with partners from academia or industry, resp. designated research projects and networking efforts (esp., EU or other multi-national projects). We identify use case and requirements priorities, technology gaps and interoperability roadblocks. We work towards community group reports that describe our findings and/or solutions to the challenges identified in our work.