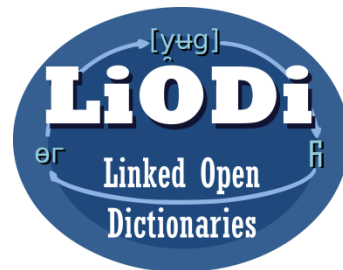


LD4LT Discussions on Linguistic Annotation (so far)

Christian Chiarcos

Applied Computational Linguistics (ACoLi)





chiarcos@informatik.uni-frankfurt.de



RDF and Annotation: A brave new world?

Not quite (yet):

Concurrent, incompatible vocabularies

-  Web Annotation (mostly for bioinformatics and DH)
-  NLP Interchange Format (mostly for NLP web services)
-  Ligt (morphology, not supported otherwise)
-  POWLA (generic LAF data structures)

Prospects on information integration recognized already during the 2000s

Hampered by incompatibilities

⇒ Consolidation initiative

W3C Community Group „Linked Data for Language Technology“

+ supported by Nexus Linguarum, WG 1, T1.1

Linked Data for Language Technology



<https://www.w3.org/community/ld4lt/>

- two active lines of discussion
 - language resource metadata (METASHARE OWL)
 - consolidate linguistic annotations on the web (Web Annotation + NIF + ...)

- address use cases and requirements for Language Technology Applications that use Linked Data
 - ⇒ interoperability

LD4LT Harmonization Initiative

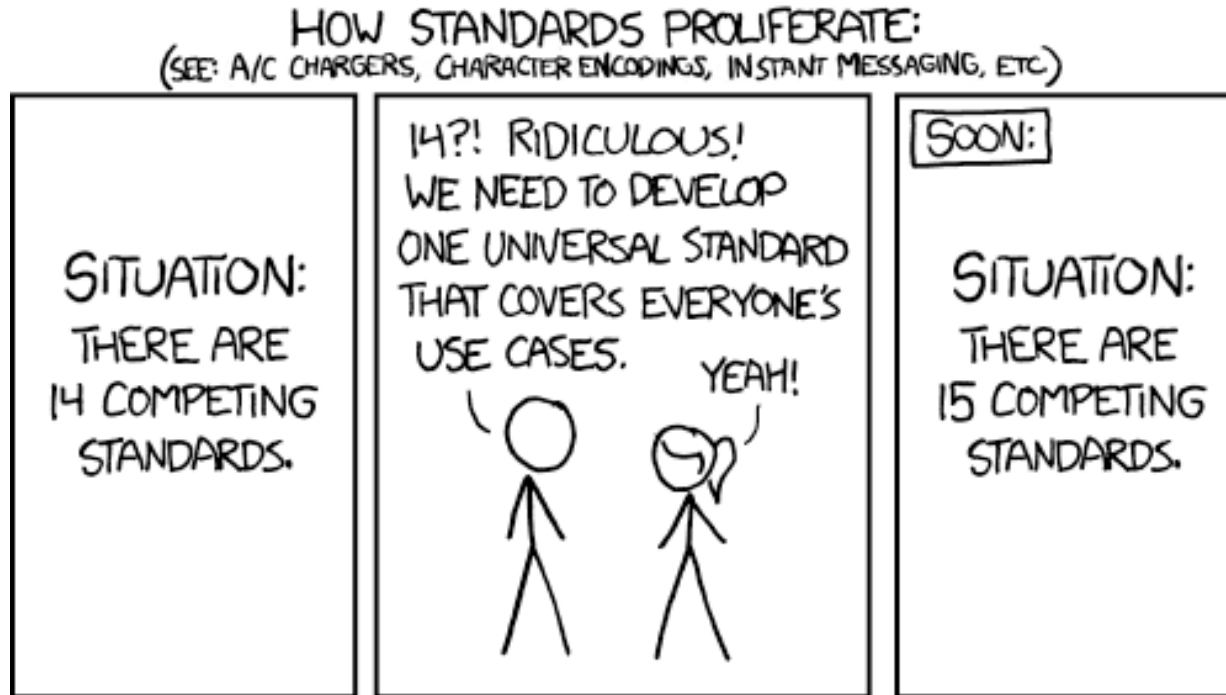
- ❑ Establish *one* RDF vocabulary for annotations on the web
 - API specifications
- ❑ Guidelines/specs for linguistic annotations on the web
 - publishing, processing, exchanging, accessing
- ❑ Largely compatible/building on existing standards
 - detect and compensate gaps
 - compatible with or easily upgradable from existing implementations

Related Activities: OntoLex-FrAC

https://www.w3.org/community/ontolex/wiki/Frequency,_Attestation_and_Corpus_Information

- frequency, attestation, corpus information
 - W3C Community Group *Ontology-Lexica*
 - *pointers from lexical resources into corpora*
 - *annotation with dictionary references* (OntoLex)
 - *requires* a vocabulary for annotating a corpus with lexical links, but does not provide it
 - Instead, it refers to external vocabularies *such as* NIF or Web Annotation
 - It would be better to provide a concrete recommendation

What do we want to do?



Approach so far

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

- Pilot survey: WA / NIF / both ?

- ❑ 2018-2019

- ❑ H2020 project Pret-a-LLOD

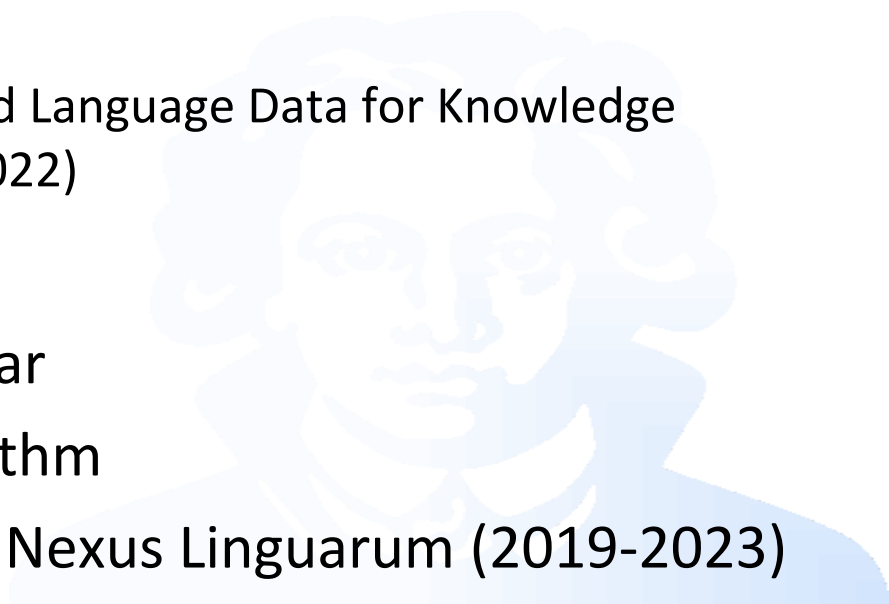
- Ready-to-use Multilingual Linked Language Data for Knowledge Services across Sectors (2019-2022)

- Series of telcos

- ❑ since 2019, somewhat irregular

- ❑ *aiming* for a more regular rhythm

- ❑ joint activity with Cost Action Nexus Linguarum (2019-2023)



Approach so far

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

■ Survey of requirements and features

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

	NIF				Web_Annotation	ISO and derivatives					TEI
	NIF_2.0	NIF_2.1	CoNLL-RDF	Ligt		POWLA	LAF	MAF	SynAF	SemAF	
A.1 RDF serialization	+	+	+	+	+	+	-	-	-	-	
A.2 Extent of standardization	(+)	(+)	(-)	-	+	(+)	+	+	+	+	
A.3 Documentation	+	+	+	(-)	+	(+)	(+)	+	-	-	
A.4 IRI fragment identifiers for strings	+	+	-	-	(+)	(-)	-				
A.5 Explicit selectors	+	+	-	-	+	(+)					
A.6 Explicit context strings	+	+	-	-	-	-					
A.7 API specifications for web services	+	+			+						
A.8 Assign data categories	(+)	(+)	(+)	(+)	(-)		(+)	(+)	(+)	(+)	
A.9 Compatible with Web Annotation vocabulary	(+)	(+)	(+)	(+)	+	(+)	(+)	(+)	(+)	(+)	
A.10 Compatible with NIF 2.0 core vocabulary	+	+	+	+	(-)	(+)	-				
A.11 Compatible with ISO standards	-	-	-	-	(+)		+	+	+	+	

Approach so far

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

■ Survey of requirements and features

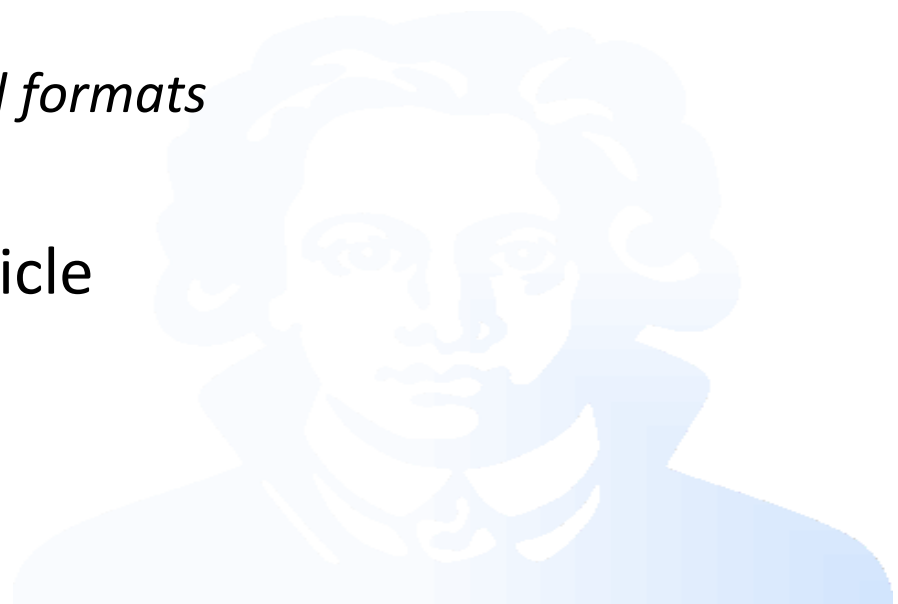
<https://github.com/ld4lt/linguistic-annotation/tree/master/survey>

□ still incomplete

- *add statistics on features and formats*
- to be added: TEI, ISO

□ partially fed into a draft article

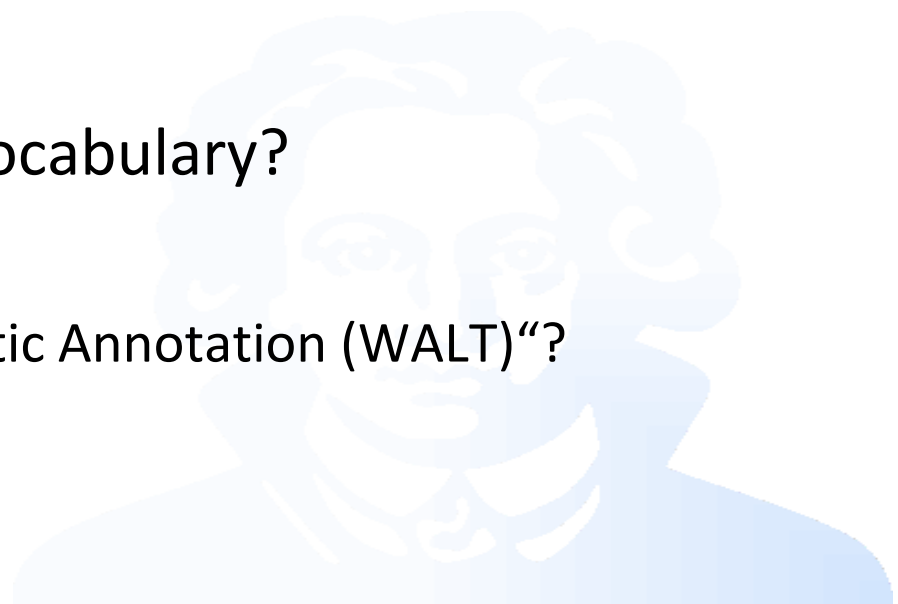
- Khan et al. (ms), TITLE



Approach so far

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

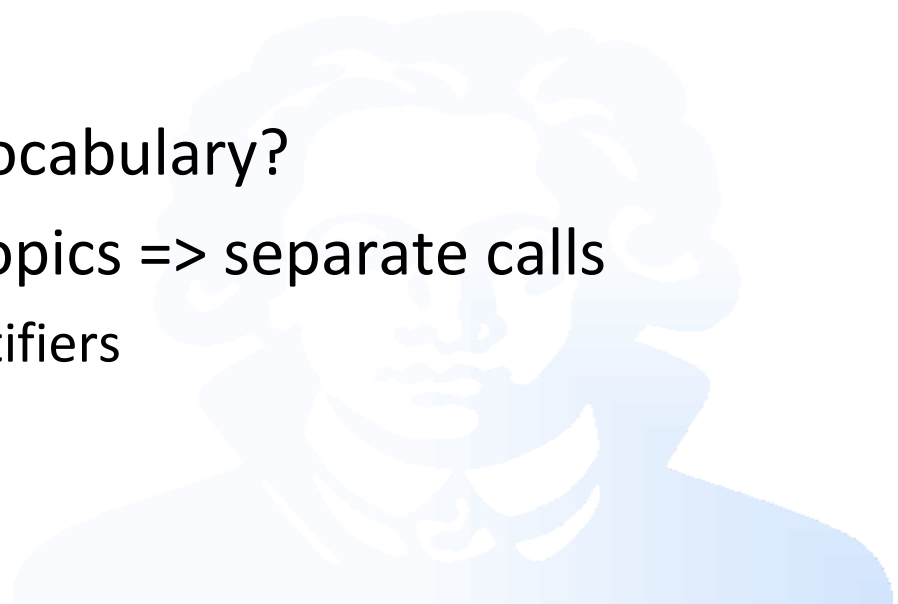
- Survey of requirements and features
- Now, decide how to develop common specifications
 - ❑ Extending an established vocabulary?
 - ❑ Which one?
 - „Web Annotation for Linguistic Annotation (WALT)“?
 - „NIF 2.0“ ?
 - „LAF-RDF“ ?



Approach so far

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

- Survey of requirements and features
- Now, decide how to develop common specifications
 - Extending an established vocabulary?
 - Deeper discussion of sub-topics => separate calls
 - suggested for fragment identifiers



Approach so far

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

- Survey of requirements and features
- Now, decide how to develop common specifications
- Need help, feedback and additional use cases ;)
 - ❑ This can have a similar impact as OntoLex had on lexical resources
 - Since the publication of the vocabulary in 2016
 - If developed with an eye on usage, usability and compatibility

Materials

- LD4LT mailing list & wiki
 - ❑ <https://www.w3.org/community/ld4lt/>
 - ❑ https://www.w3.org/community/ld4lt/wiki/Main_Page#Use_Cases
- GitHub, incl. archive
 - ❑ <https://github.com/ld4lt/linguistic-annotation>