# OntoLex-FrAC
## Standardizing the Corpus-Lexicon Interface

Christian Chiarcos, U Augsburg
Anas Fahad Khan, ILC/CNR
Maxim Ionov, U Cologne

Elena Simona Apostol &
Ciprian-Octavian Truica,
Uppsala University

Besim Kabashi, FAU Erlangen-Nuremberg
Katerina Gkirtzou, Athena Research Center

## Motivation: Language Resource Interoperability

### RDF and Interoperability

- ✓ **Resource Description Framework (RDF)**
  Standard for machine-readable data on the web
  Provides uniform access to and integraton of heterogeneous data, *regardless of backend technology*
- ✓ **Linguistic Linked Open Data (LLOD)**
  Use RDF technologies to share, access and link language resources on the web

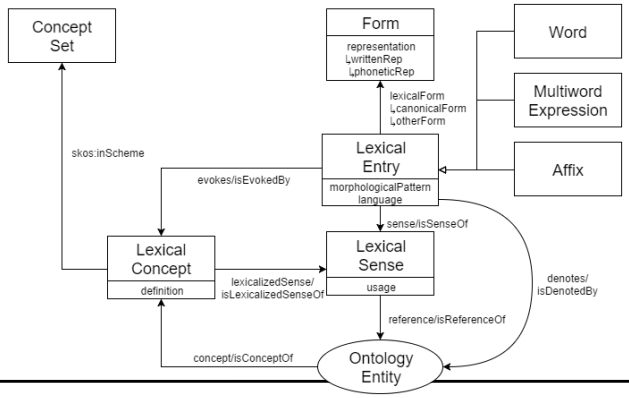
### LLOD Standards

- ✓ **W3C standards** for formats, access and wrapper technologies (HTTP, URI, RDF, SPARQL; JSON-LD, CSVW, R2RML, RDFa, ...)
- ✓ **Community standards** for language resources (OntoLex; LexInfo, OLiA; NIF, CoNLL-RDF, ...)

=> **FAIR data:** accessible, interoperable, re-usable

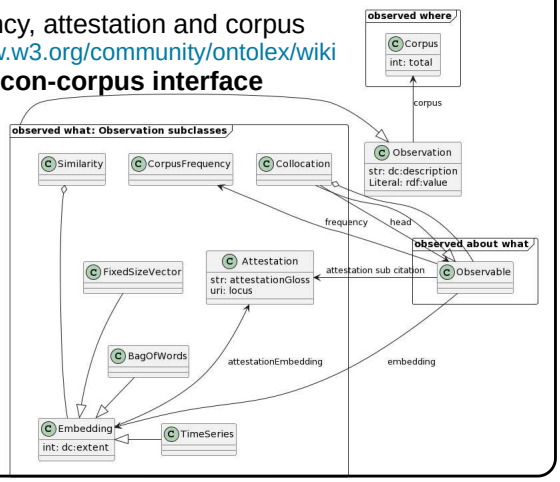
## OntoLex (2016)

Community standard for lexical resources on the web
https://www.w3.org/2016/05/ontolex/
**widely adopted for machine-readable dictionaries**




## NEW: OntoLex-FrAC

frequency, attestation and corpus
https://www.w3.org/community/ontolex/wiki
**lexicon-corpus interface**




## A Solution for UniDive WG2?

OntoLex and OntoLex-FrAC formalize lexical resources and their linking with corpus information so we can

- use **off-the-shelf technology** (e.g., R2RML, TARQL, SPARQLAnything, GRDDL, Fintan) to
- expose legacy data from **almost any source** as RDF,
- **retrieve, enrich, link, merge, query and transform** these graphs using web technologies and resources, and
- export to shallow, **easy-to-process formats** (e.g., CSV) on demand (using SPARQL SELECT)

as previously demonstrated for

- 3000+ bilingual dictionaries for 430+ languages (Chiarcos et al. 2020)
- merging WordNets and morphologies (Racioppa & Declerck 2019)
- OntoLex exports of UniMorph, Universal Derivations, Universal Dependencies and various structured dictionary formats (e.g., https://github.com/acoli-repo/LLODifier)

## A Challenge for WG1!

OntoLex and OntoLex-FrAC formalize lexical resources and their linking with corpus information, *but*
only from the perspective of the dictionary
`frac:attestation` -> `frac:locus` ...
**but what does that point to in the corpus ?**

**competing solutions for corpora in RDF !**

- NLP Interchange Format (NIF)
- Web Annotation (Open Annotation)
- CoNLL-RDF (for tabular data)
- Ligt (for interlinear glosses)
- TEI+RDFa (for digital editions, inline XML)
- TEI+Web Annotation (for digital editions, standoff)
- TEI+GRDDL/XSLT (for digital editions, native XML)
- Linguistic Annotation Framework (LAF) / POWLA (OWL2/DL rendering of LAF)

To be addressed by W3C Community Groups LD4LT and BPMLOD – shall/can we involve UniDive ?

Curious? Join our calls! ... or our **Day of W3C language technology community groups at LDK-2023, Vienna, Sep 12**
**W3C CG Ontology-Lexica**: https://www.w3.org/community/ontolex (OntoLex and OntoLex-FrAC)
**W3C CG Linked Data for Language Technology**: https://www.w3.org/community/ld4lt/ (harmonizing vocabularies for corpus annotations)
**W3C CG Best Practices for Multilingual Linked Open Data**: https://www.w3.org/community/bpmlod/ (current practices, e.g., for annotation)
Until April 2024, the coordination between these W3C CGs takes place via **Cost Action Nexus Linguarum**: https://nexuslinguarum.eu/