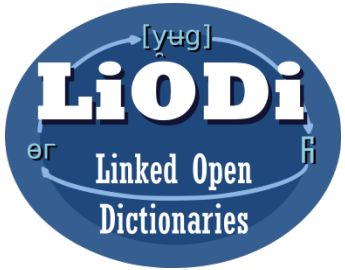
Harmonizing Linguistic Annotations
(An LD4LT workshop)

Welcome & Overview

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

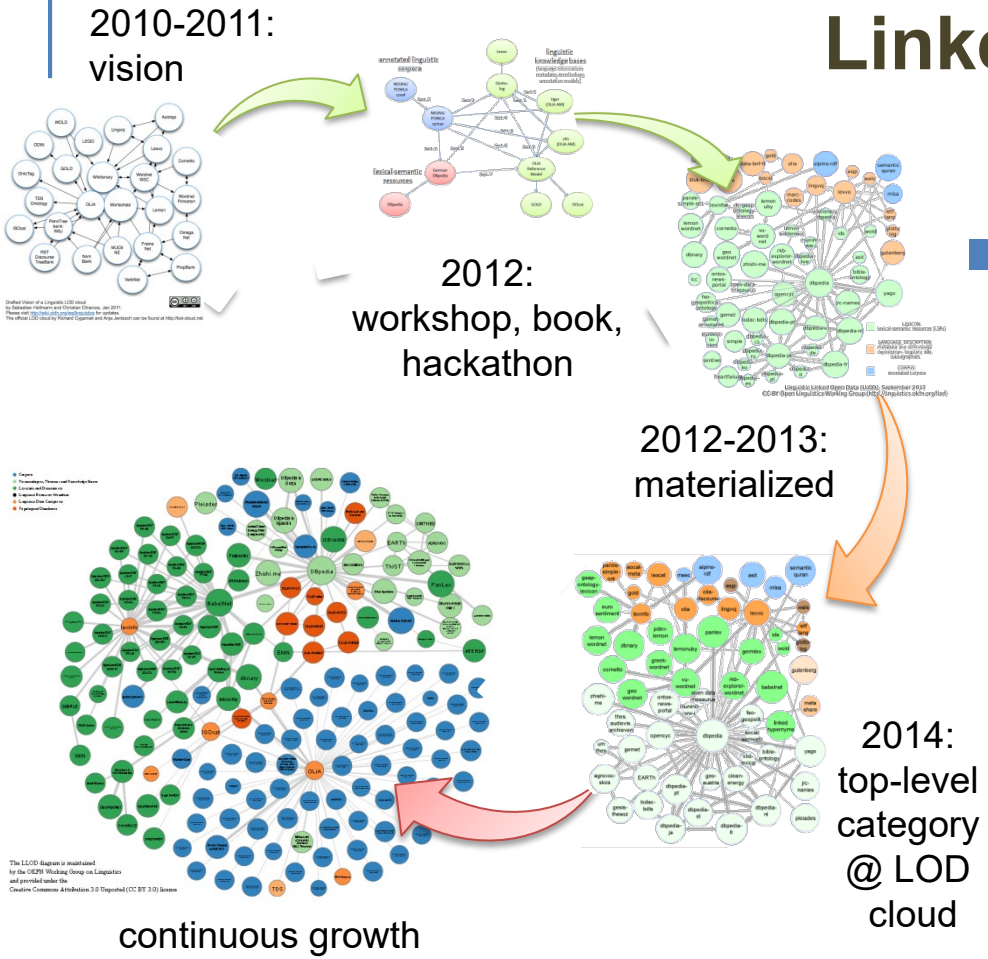


Linked Data for Language Technology (LD4LT)

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

- W3C Community Group
  - formed 2013
    - address use cases and requirements for Language Technology Applications that use Linked Data
    - ⇒ interoperability
  - language technology
    - surveying needs and requirements (esp. 2013-2016)
    - language resource metadata (since 2015)
    - linguistic annotation (intensified since 2019)

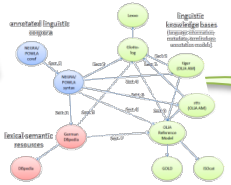
Linked Data and Language Technology



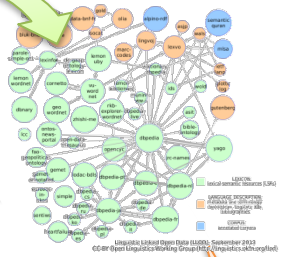
- since 2010, an LOD sub-cloud of language resources has been emerging
  - Open Linguistics Working Group
  - http://linguistic-lod.org

Linked Data and Language Technology

2010-2011:
vision

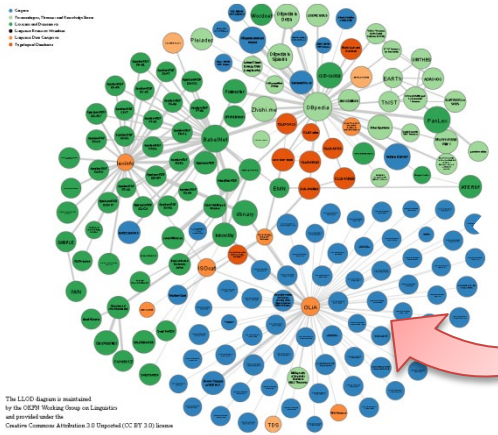


2012:
workshop, book,
hackathon



2012-2013:
materialized

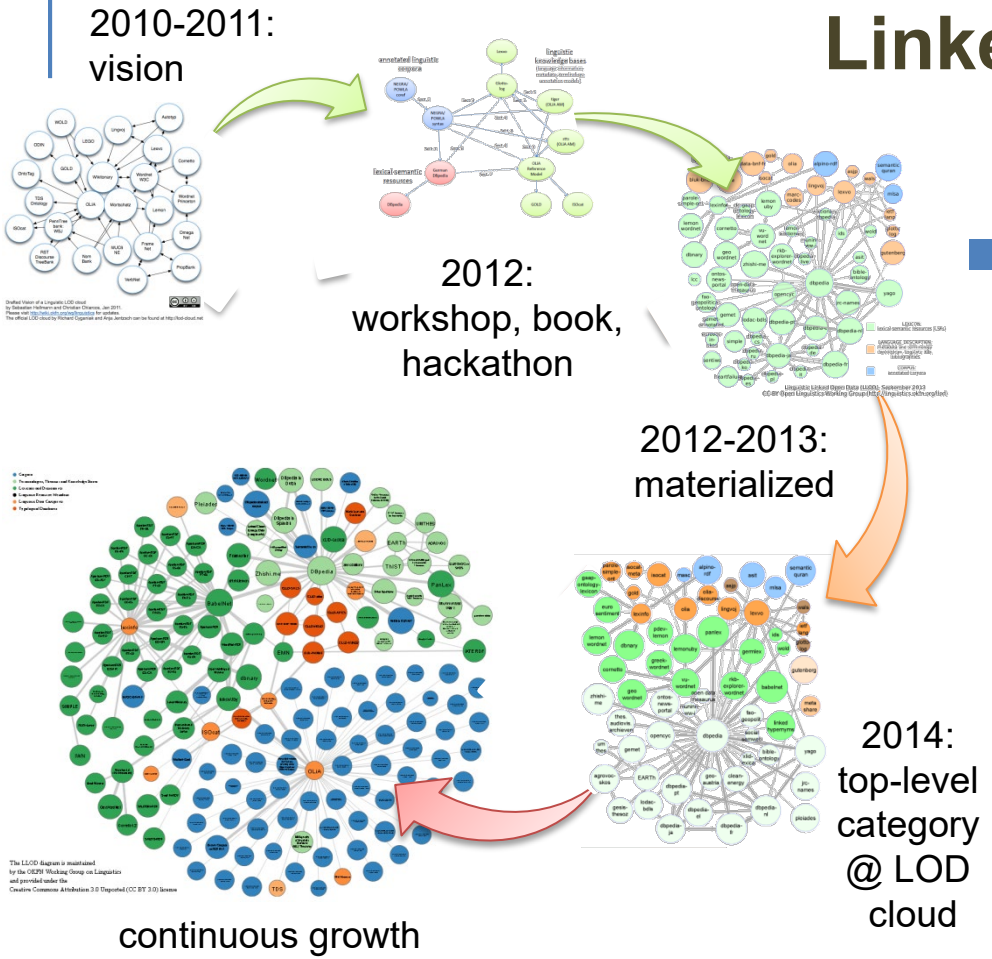
2014:
top-level
category
@ LOD
cloud



continuous growth

- since 2016, it has particularly grown with respect to lexical resources
  - W3C CG Ontology-Lexica (OntoLex)
    - one vocabulary, many use cases

Linked Data and Language Technology



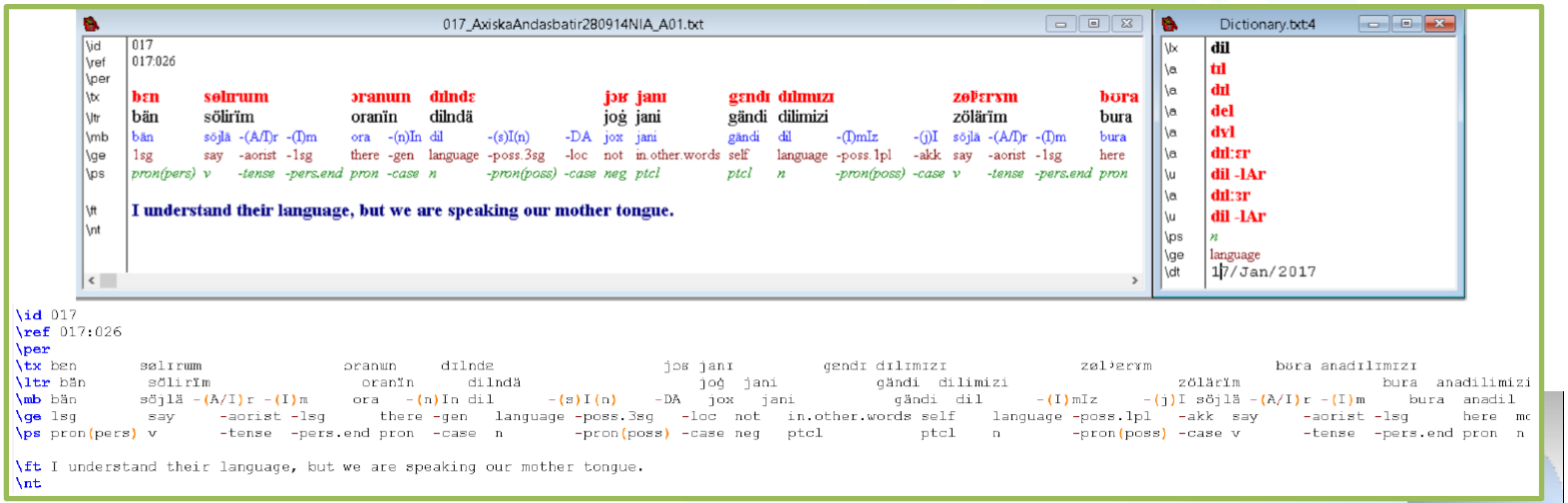
- linguistic annotation remains a problematic area
  - some resources (blue)
    - few successful user stories
    - concurrent standards
    - USP: interoperability?
      - not in the current situation

What is Linguistic Annotation, Actually?

- Different communities of practice annotate very different things

Interlinear Glossed Text (Toolbox)

https://software.sil.org/toolbox/



What is Linguistic Annotation, Actually?

- Different communities of practice annotate very different things

Exmaralda

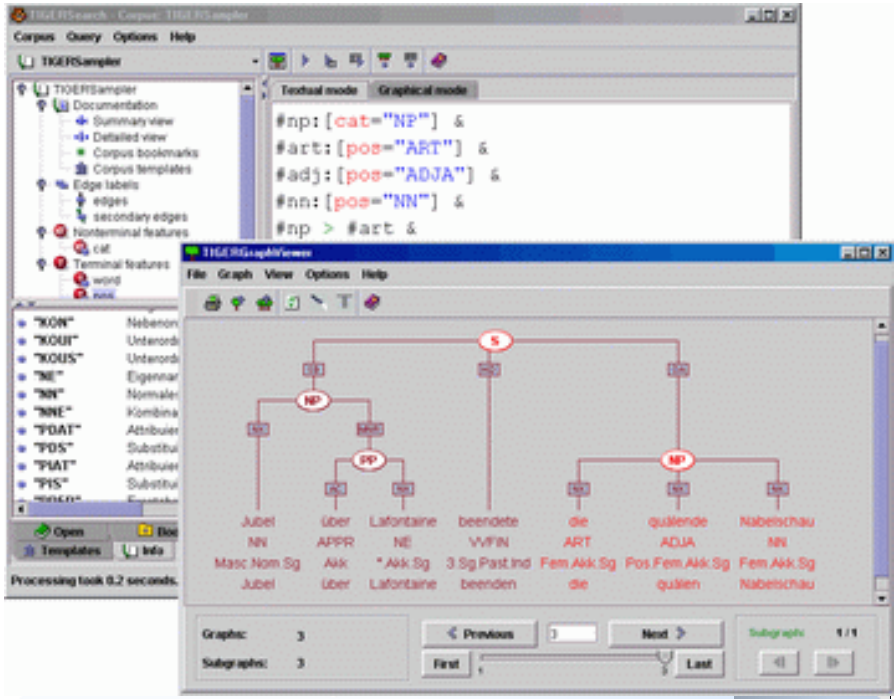
layered annotation over multimodal content (annotating dialogue, guestures, etc.)

http://exmaralda.org

What is Linguistic Annotation, Actually?

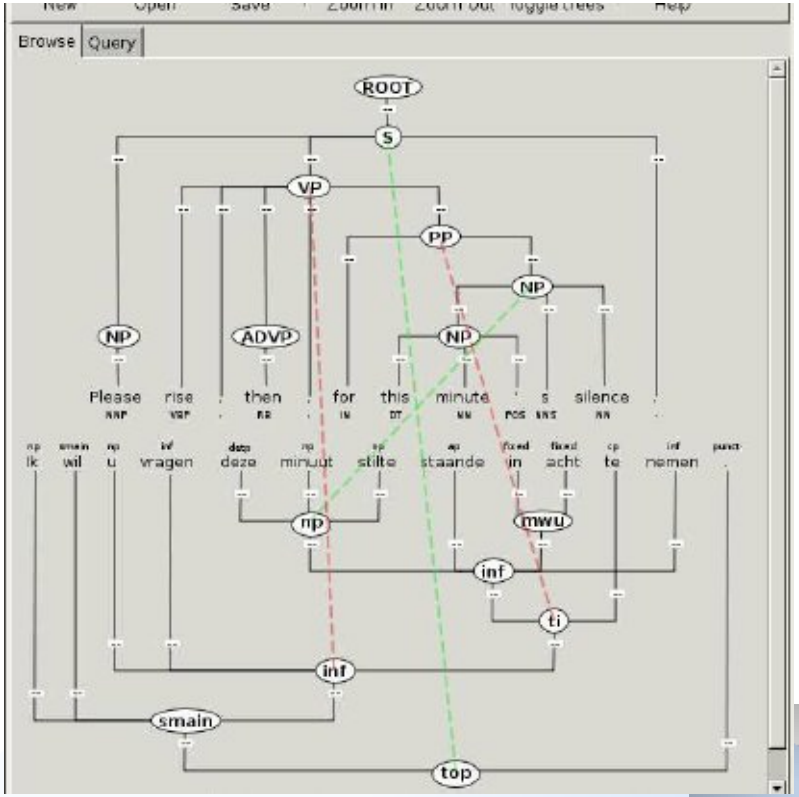
- Different communities of practice annotate very different things

Phrase Structure Syntax
(TIGERSearch/ Annotate)



What is Linguistic Annotation, Actually?

- Different communities of practice annotate very different things



TreeAligner

syntax annotation
for two aligned texts

https://www.ling.su.se/english/nlp/tools/stockholm-treealigner

What is Linguistic Annotation, Actually?

- Different communities of practice annotate very different things

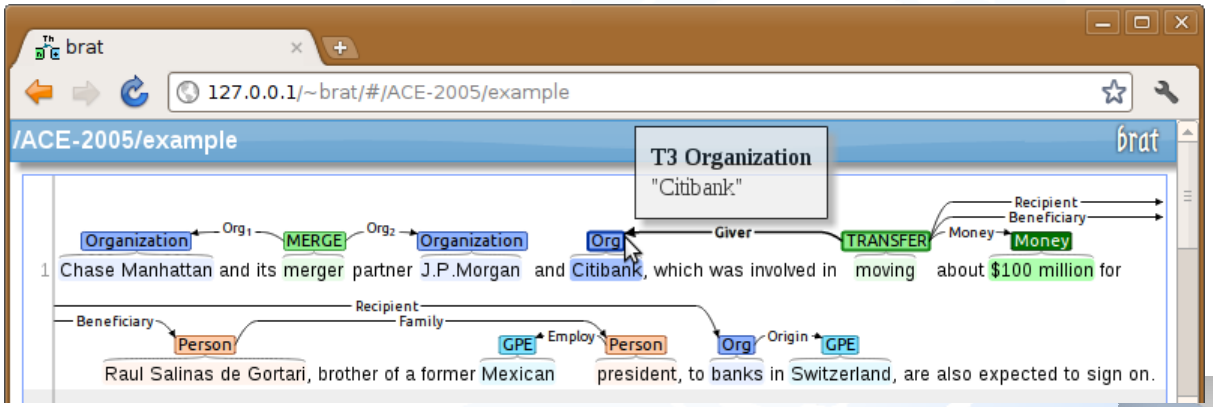
Recogito
Georeference
(Entities)

http://commons.pelagios.org/2018/03/you-can-now-do-everything-in-recogito/

What is Linguistic Annotation, Actually?

- Different communities of practice annotate very different things

Relations and Entities (Brat)

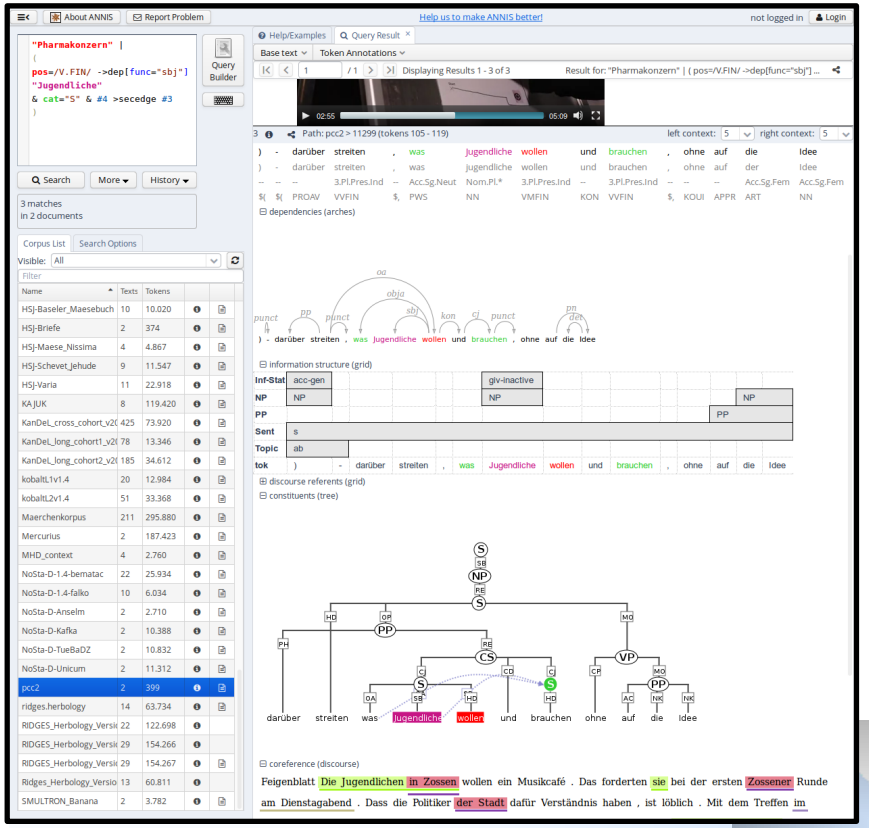


https://brat.nlplab.org/manual.html

What is Linguistic Annotation, Actually?

- Different communities of practice annotate very different things

All at once ?
Multi-Layer Annotation



https://corpus-tools.org/annis/

LD4LT Workshop @ LDK-2021

09:00 **Welcome**

09:10 - 10:30 **Background: Linguistic Annotation on the Web**

- *W3C Standard Web Annotation* Christian Chiarcos
- *NLP Interchange Format* Milan Dojchinovski
- *Text Encoding Initiative* Fahad Khan
- *ISO TC37 standards* Thierry Declerck
- *Text Fragids* Joel Kalvesmaki

10:30 - 11:00 **Break**

LD4LT Workshop @ LDK-2021

11:00 - 11:40 Discussion

- *QA: What is missing? What is unclear? Where are problems?*
- *Summary of LD4LT Discussions on Linguistic Annotation*

11:40 - 12:30 Use Cases, Experiences, Extensions

- *Linking Latin* Francesco Mambrini
- *Distributed Text Services* Christian Chiarcos
- *Interlinear Glossed Text* Maxim Ionov
- *Discourse Research* Giedre Valunaite Oleskevicienė
- *Transforming Language Resources* Christian Fäth

12:30-13:00 Brainstorming & Next Steps

LD4LT Workshop @ LDK-2021

- primary goals
  - provide and collect background information
  - present and discuss use cases and requirement
  - initiate and plan future discussions
- we record background presentations
  - for future reference
  - if you don‘t want to be recorded, please switch off camera and join the discussion via chat

LD4LT Workshop @ LDK-2021

- held in conjunction with
  - 3rd Conference on Language, Data and Knowledge (LDK-2021)
  - Face-to-face meeting of the W3C CG Ontology-Lexica
- organized in cooperation between
  - W3C CG LD4LT
  - Cost Action Nexus Linguarum, T1.1

LD4LT Workshop @ LDK-2021

09:00 **Welcome**

09:10 - 10:30 **Background: Linguistic Annotation on the Web**

- *W3C Standard Web Annotation* Christian Chiarcos
- *NLP Interchange Format* Milan Dojchinovski
- *Text Encoding Initiative* Fahad Khan
- *ISO TC37 standards* Thierry Declerck
- *Text Fragids* Joel Kalvesmaki

10:30 - 11:00 **Break**