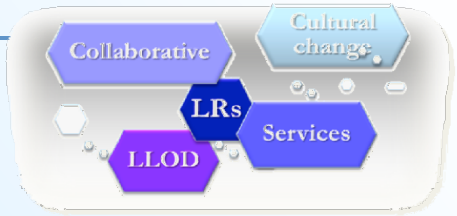
Data and Metadata of Language Resources as Linked Data on the Web: Examples

Christian Chiarcos

chiarcos@uni-frankfurt.de

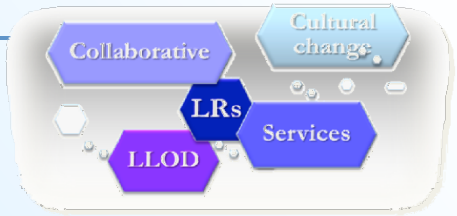
Linked Data **for** Linguistics



- Representation and modelling
- Structural interoperability
- Integrating distributed resources
- Conceptual interoperability
- Dynamic Import
- Ecosystem, infrastructure and community

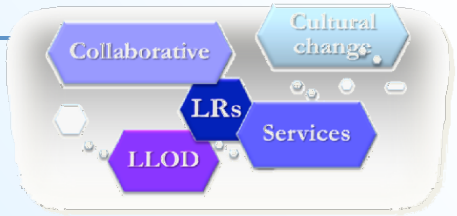
Chiarcos et al. (2013)

Representation and modelling



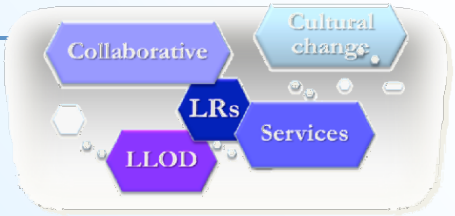
- Different linguistic subcommunities have developed representation standards, e.g.,
  - LMF: Lexical Markup Framework (Francopoulo et al. 2009)
    - lexical-semantic resources
  - GrAF: Graph Annotation Framework (Ide and Suderman 2007)
    - annotated corpora
  - based on labeled directed acyclic graphs (feature structures)
- RDF data model: labeled directed graphs
  - Uniform formalism for *different* resource types
  - Sublanguages (e.g., OWL) to define specific vocabularies

Structural interoperability



- With different language resources represented in RDF, we can combine both sources of information freely
  - cross-resource queries with SPARQL
  - Given a corpus with WordNet sense annotations
    - "Retrieve all sentences that describe locations"
    - i.e., sentences containing a token annotated with a WordNet sense that is a hyponym of "location"
- Difficult to realize with GrAF or LMF

Integrating distributed resources

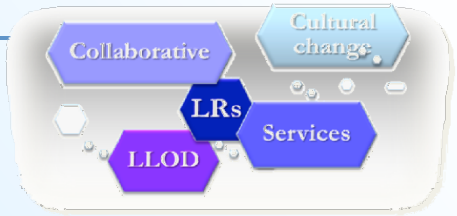


- SPARQL supports nested subqueries to run on different repositories

```
SELECT ?token {
  service <http://wordnet.rkbexplorer.com/sparql> {
    rkbWN:synset-land-noun-2
       wn20:containsWordSense ?sense .
    ?sense rdfs:label ?synonym .
  }
  ?token powla:hasString ?synonym .
}
```

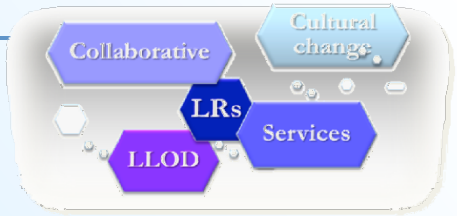
- No physical integration of resources in a single data base required
  - Easy to link to centralized repositories of reference terminology, etc.

Conceptual interoperability



- Resources should specify which vocabulary (e.g., for annotation) they use and how it is defined
  - By reference to community-maintained terminology repositories, e.g.,
    - GOLD (Farrar and Langendoen 2010)
    - ISOcat (Windhouwer and Wright @ LDL-2012)
    - OLiA (Chiarcos 2008)
  - Can be used, e.g., for disambiguation
    - e.g., *land* as a noun, but not as a verb

Dynamic import



- Linking resources with URIs

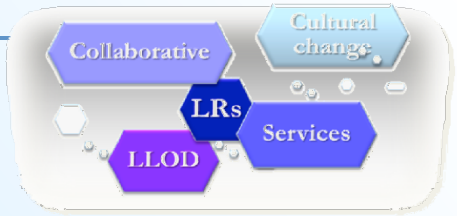
  => resolved on-the-fly to enrich with up-to-date background knowledge

    e.g., every update in a lexical resource is available to every resource linked with it

  => updates are available at query time

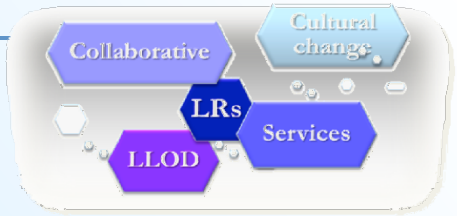
- Inconsistencies can be avoided through versioning

Ecosystem, infrastructure and community



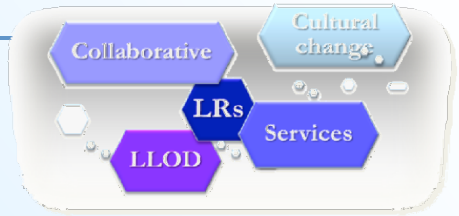
- RDF and related standards are maintained by an active and relatively large community
  - Different fields of application
    - Libraries, GeoData, BioMed, ...
  - Established W3C standard and technological infrastructure
  - Linguistically relevant resources already provided
- RDF facilitates distributed development, re-using data, and, indirectly, interdisciplinary cooperation

Towards LLOD



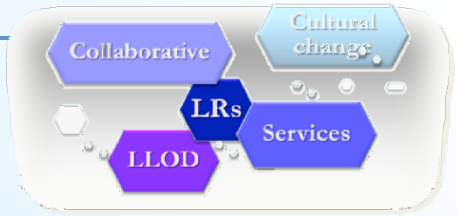
- There are independent motivations to provide data in RDF
  - generic data model
  - generalization over heterogeneous DB schemes
  - generalization over heterogeneous terminologies
  - connect existing resource portals
  - a conceptual view on annotations

Towards LLOD



- There are independent motivations to provide data in RDF
- Side-effects: Data can be
  - linked with other RDF resources
    - e.g., lexicon – linguistic terminology
  - queried (federation)
    - even if stored in physically separated repositories

Towards LLOD

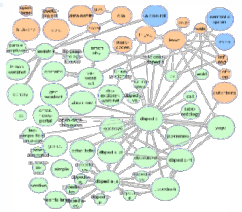


- There are independent motivations to provide data in RDF
- Side-effects: Data can be
  - linked with other RDF resources
  - queried (federation)

⇒ Linguistic Linked Open Data (LLOD)

An on-going effort orchestrated by the Open Linguistics Working Group of the Open Knowledge Foundation

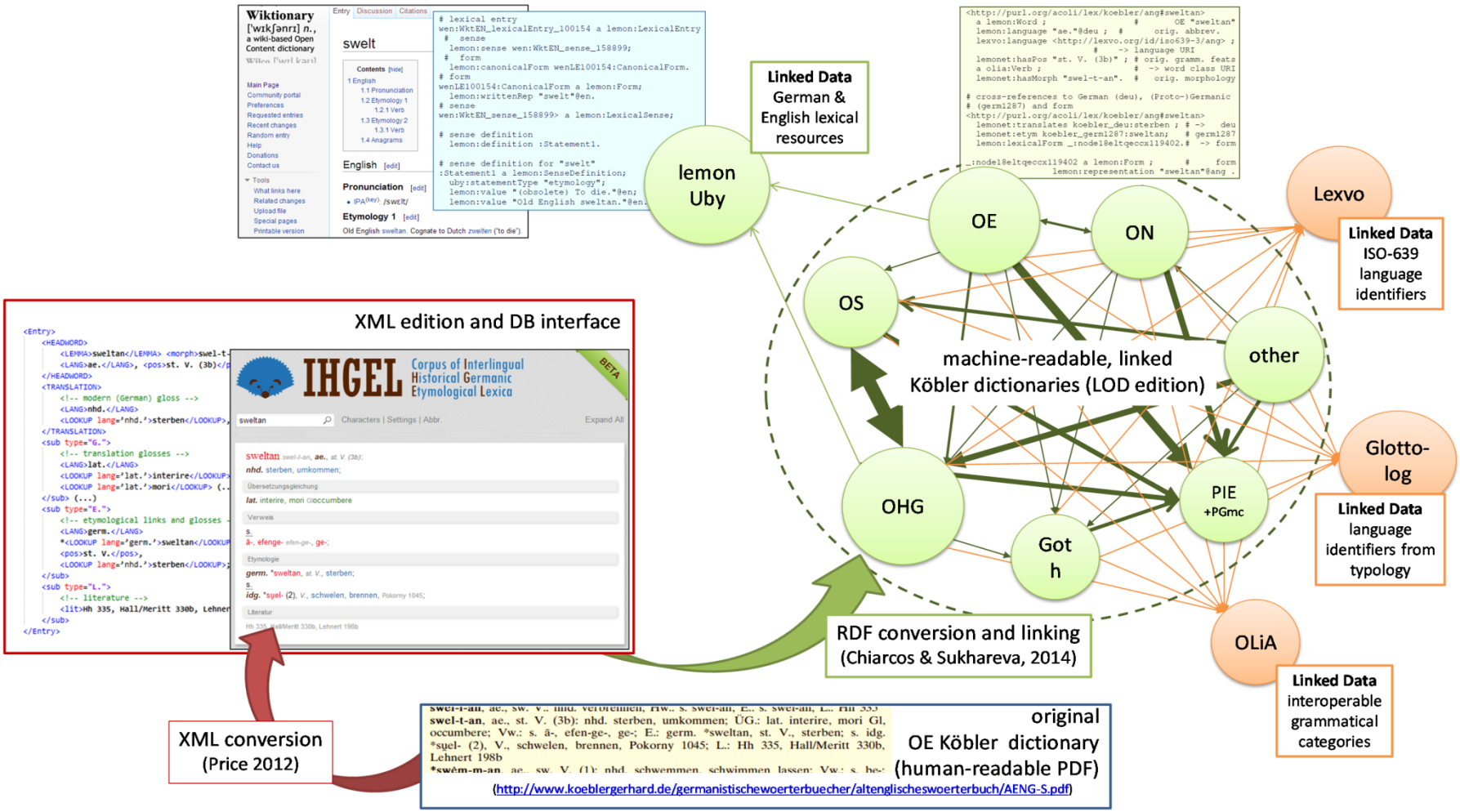
Building the Cloud: Examples



- Each data provider has different incentives to use Linked Data and/or RDF
- Concepts of RDF and Linked Data have been brought up to solve open problems in different subcommunities of linguistics and neighboring fields
- Examples
  - Corpora — Cassidy (2010), Chiarcos (2012)
  - Machine-readable dictionaries
  - Term and data bases — Chiarcos (2010)
  - Etymological dictionaries — Chiarcos & Sukhareva (2014)

  → combined queries — Burchardt et al. (2008), Rehm et al. (2008), Chiarcos & Götze (2007)

Linked Germanic Etymologies



TECHNISCHE UNIVERSITÄT DARMSTADT ACoLi digital humanities JOHANN WOLFGANG GOETHE UNIVERSITÄT FRANKFURT AM MAIN

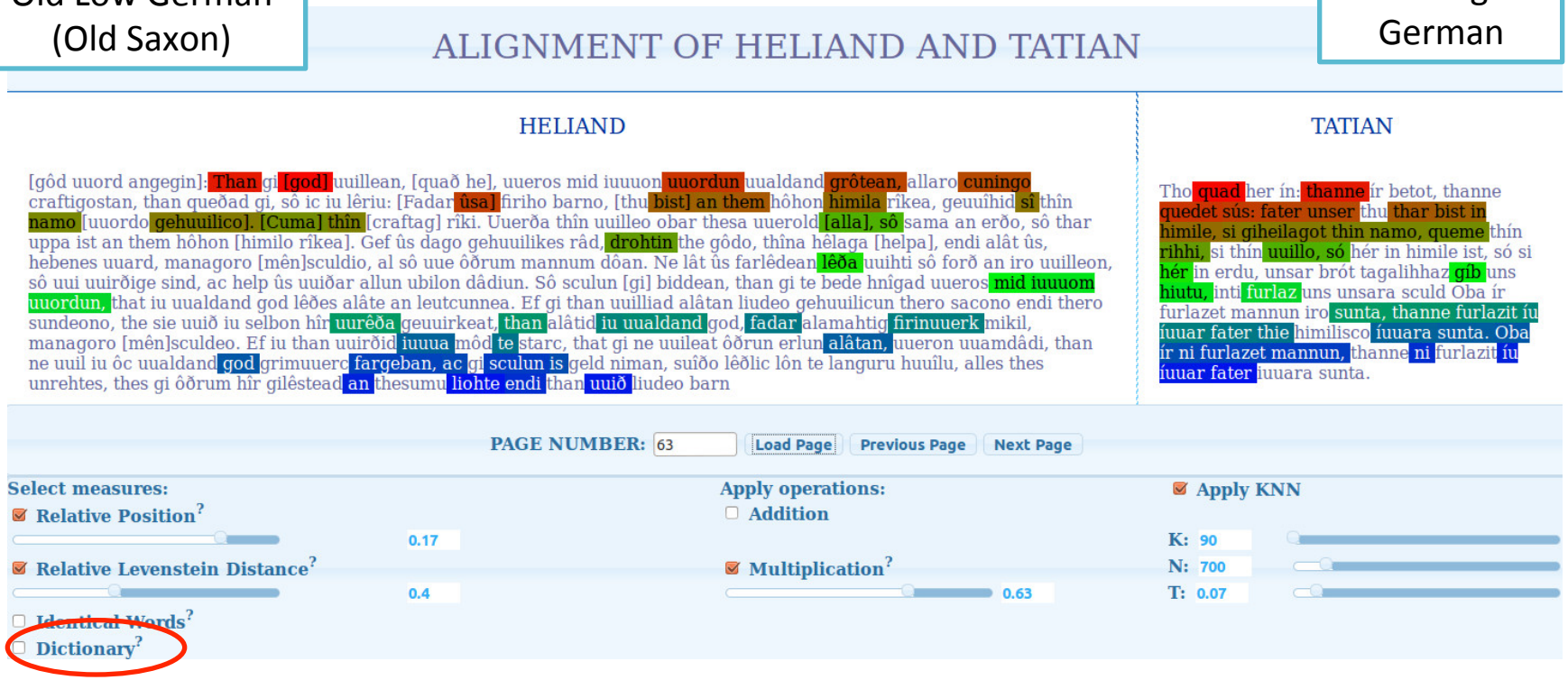
Linked Germanic Etymologies

Application:

Aligning early medieval gospel harmonies

Old Low German (Old Saxon)

Old High German



ACoLi

digital humanities

JOHANN WOLFGANG GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Linked Germanic Etymologies

Conversion of etymological dictionaries to RDF

- **Linkability**: representation of relations within and beyond lexicons
- **Interoperability**: (meta)data representation through community-maintained vocabularies (lexvo, Glottolog, OLiA, *lemon*)
- **Inference**: filling the logical gaps of the original XML representation
  - *Symmetric closure of cross-references*

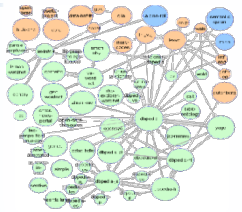
Linked Germanic Etymologies

	West Germanic					other		reconstr.	
lexicon	OE	OHG	OS	OLF	OFr	ON	Got	PGmc	PIE
entries (XML, in K)									
	25	24	9	2	13	12	5	9	7
triples (RDF, in M)									
	1.2	1.6	.6	.2	.6	.7	.4	.2	.2
lemon:Words & links (in K)									
OE	25					1			
OHG	2	26	7	2	3	1			
OS	1	4	9	1	2	1			
ON	1				1	14			
Got	1	1			1	1	6		
PGmc	5	3	3	1	2	4	2	8	
PIE	2	1	1	1	1	1	1		8
German	16	23	8	4	10	12	7	6	3
English		10	4	2	5		9		2
symmetric closure of etym. links (triples *per lang.* in K)									
	+11	+14	+11	+5	+9	+8	+5	+21	+9
links to (L)LOD data sets (triples *per data set* in K)									
OLiA	24	22	8	2	12	11	5	8	7
lexvo	132	186	82	21	68	82	49	14	15
Glottolog	15	11	8	3	7	11	6	9	13

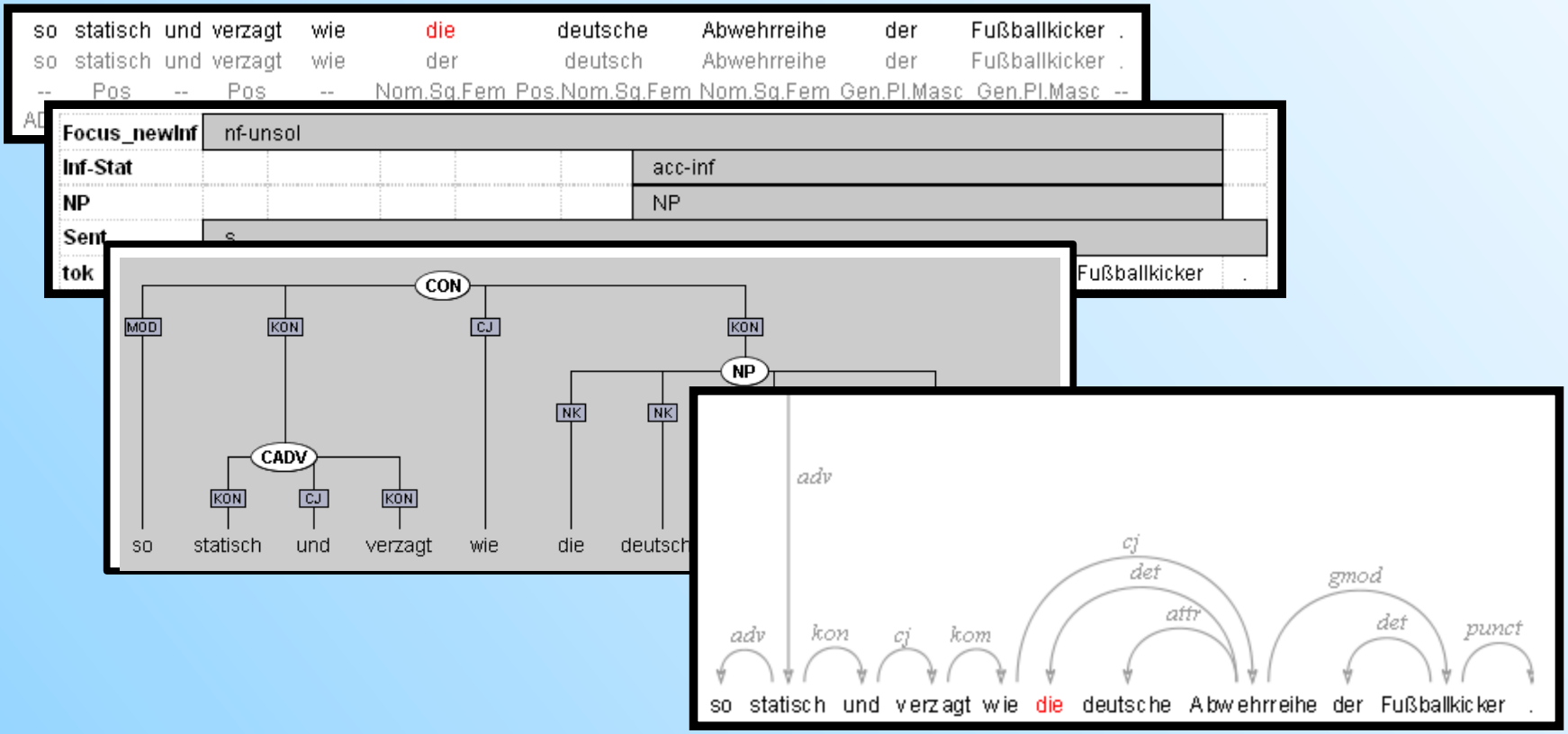
A few critical remarks

- existing vocabularies insufficient
  - lemon: etymological relations between lexical entries ?
  - glottolog/lexvo/ ISO-693: reconstructed languages ?
  - OLiA: Old Germanic inflection paradigms ?

Use Case I: Corpora

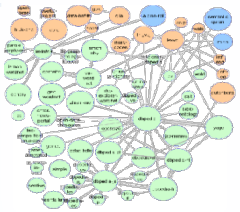


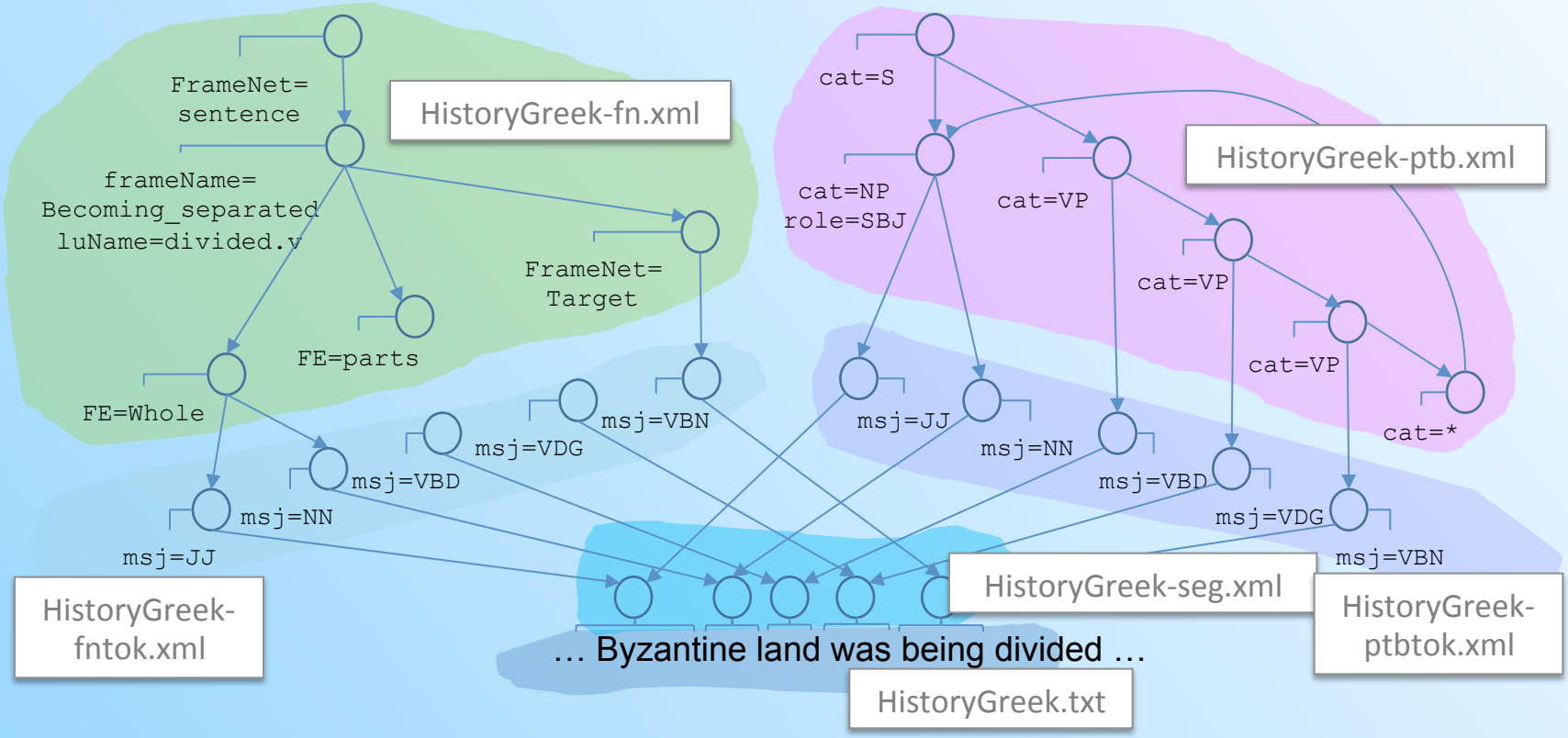
Analyses produced by different researchers / NLP tools use different representation formalisms



Standoff XML: GrAF, MASC corpus

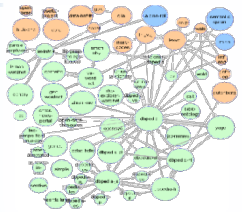
(Ide & Suderman 2007, Ide et al. 2008)





node (segment) — edge (relation, hierarchical) — edge (relation, non-hierarchical) — label attached to a node or an edge (annotation) — filename — GrAF file (layer id)

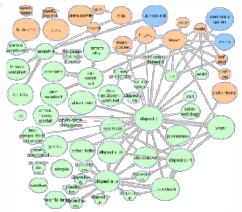
Use Case I: Corpora



Developmental trajectory

- proprietary/ad hoc formats
- XML
- directed acyclic graphs
  - standoff XML
- RDF
  - to store, query and integrate standoff annotations
    - Burchardt et al. (2008), Cassidy (2010), Chiarcos (2012), Rubiera et al. (2012), Hellmann et al. (2012), etc.

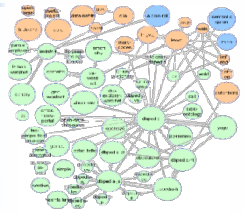
Use Case II: Lexical Resources



Similar trajectory

- proprietary formats
- XML, e.g., TEI dictionary
  - primarily for printed dictionaries
  - application in, e.g., NLP Budin et al. (2012)

Use Case II: Lexical Resources

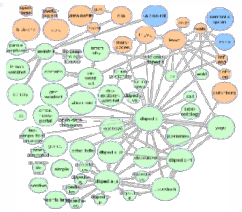


Similar trajectory

- proprietary formats
- XML, e.g., TEI dictionary
- generic data model, LMF
  - feature structures/directed acyclic graphs
  - linearized in XML/UML
  - **but**: semantics of links needs to be externally specified

(Francopoulo et al. 2009)

Use Case II: Lexical Resources

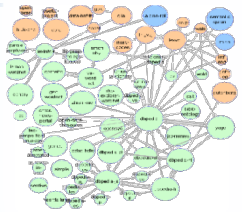


Similar trajectory

- proprietary formats
- XML, e.g., TEI dictionary
- generic data model, LMF
- RDF, e.g., lemon
  - Buitelaar et al. (2013), Fiorelli et al. (2013), Moran & Brümmer (2013)

=> links to language resources other than lexicons

Use Case III: Bring them together



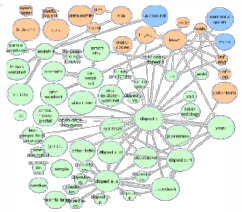
e.g.,

- words in a corpus and lexical concepts
  - at the moment, there exists no other unified formalism capable to express the following query

WordNet	syntax	WordNet
all hyponyms of *land*	modified by	a pertainym of *Byzantium*

- lemmas in a lexicon and grammatical features from a repository of linguistic reference terminology

Use Case IV: Term Bases



- linguistic terminology
  - Ontologies of Linguistic Annotation (http://purl.org/olia)
- language identifiers and descriptions
  - Glottolog (http://glottolog.org)

=>

  - ontology-based querying
    - via query rewriting for corpus information systems

(Rehm et al. 2007, Chiarcos & Götze 2007)

  - ensemble combination architectures
    - merge output of different NLP tools on a conceptual basis

(Chiarcos 2010, Pareja-Lora 2010)