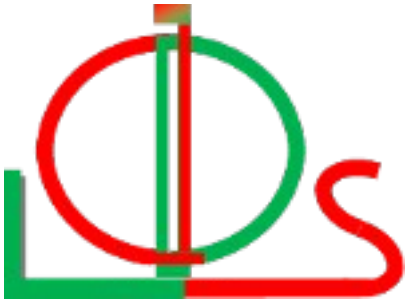
Poio API: a CLARIN-D curation project for language documentation and language typology

Peter Bouda
Centro Interdisciplinar de Documentação Linguística e Social
pbouda@cidles.eu

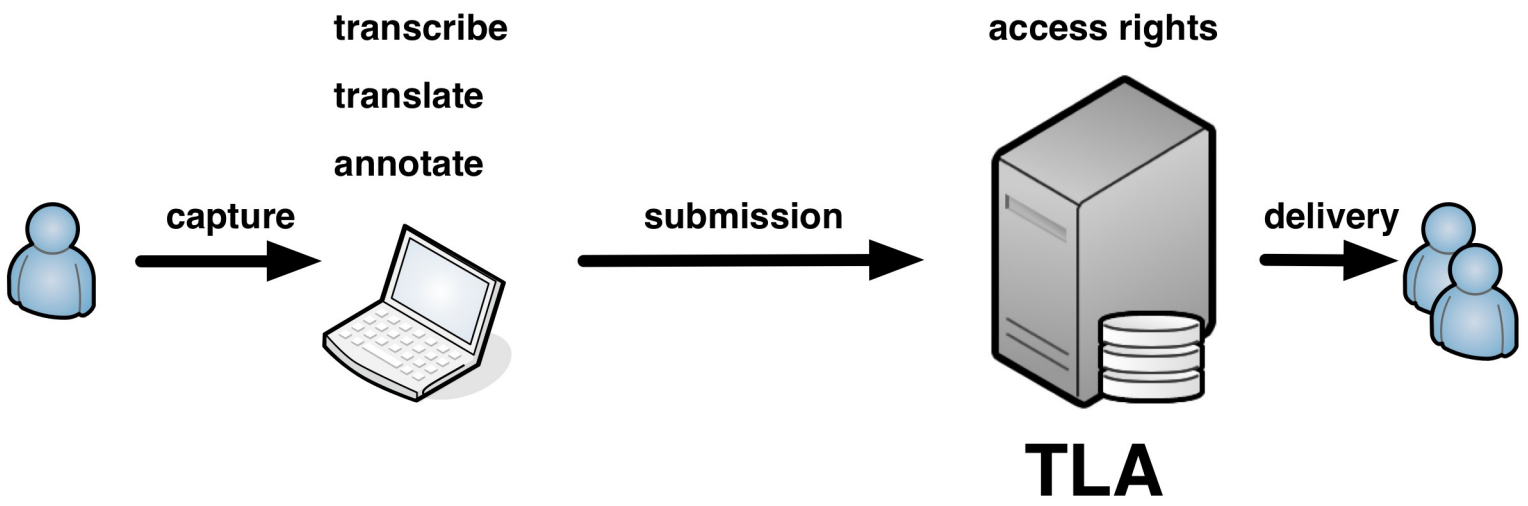


Overview

- Existing infrastructure and workflows
- Poio API and CLASS within CLARIN
- GrAF and TEI
- Poio API
- GrAF as pivot structures (IGT)
- GrAF for retro-digitization (Dictionary)

Fieldwork

Existing Infrastructure



LD tools and standards

- Elan: EAF, MPEG, WAV
- Toolbox: TXT, XML, WAV
- Arbil: IMDI/CIMDI („Component MetaData Infrastructure“)
- Praat: XML, WAV
- ...
- No standards for tier hierarchies, tier names or annotation schemes
- Efforts in ISOcat

Interlinear Glossed Text

05IH_drying_sqash_IH_JL.eaf

WIC002 Wichąwąra kiisak wamącanąkga, rookra, suura hanąąc waigopnąkga, k'orok'oros jiinąkiregi 'eegi haruce waamąxaranąąną.

WORDS	Wichąwąra	kiisak	wamącanąkga,	rookra,	suura	hanąąc	waigopnąkga,	k'orok'oros	jiinąkiregi	
MORPH	*wicąwą-ra*	*kiisak*	*wa-mąące-nąk-ga*	*rook-ra*	*suu-ra*	*hanąąc*	*wa-gigop-nąk-ga*	*k'orok'oros*	*jiinąk-ire-gi*	
GLOSS	squash-DEF	half	OBJ.3PL-cut-POS.NTL-CONT	inside-DEF	seed-DEF	all	OBJ.3PL-hollow.out-POS.NTL-CONT	be.hollowed.out(OBJ.3SG)	become-SBJ.3PL-TOP	
TRANS	*Cut the sqash in half, scoop out the inside, and when it is hollowed out it may be sliced crosswise.*									

WIC003 S~ooga waamąxaraanąkga, mąąhį huuną hošogara hikisge waamąxaranąkgiži nąąksikšjąra hižą wookiaxurucnąkga cąąkeja, hotakaceja hoicga taawus ważunąąna

WORDS	S~ooga	waamąxaraanąkga,	mąąhį	huuną	hošogara	hikisge	waamąxaranąkgiži	nąąksikšjąra	
MORPH	*šooga*	*wa-hamąxara-nąk-ga*	*mąąhį*	*huuną*	*hošoga-ra*	*hikisge*	*wa-hamąxara-nąk-giži*	*nąąksik-šjąą-ra*	
GLOSS	be.thick(OBJ.3SG)	OBJ.3PL-slice-POS.NTL-CONT	knife	handle	thickness-DEF	resemble(SBJ.3SG&OBJ.3SG)	OBJ.3PL-slice(SBJ.3SG)-POS.NTL-TOP	stick-be.hard(OBJ.3SG)	
TRANS	*The slices should be thick, or equal to the thickness of a knife handle. Use a sturdy stick and pick up circles of squash, or lace them through with a heavy stick. This stick may then be hung outside*								

WIC005 Z~eegųgi, 'ee wii hotakacra wirokų hi'ųnąąną.

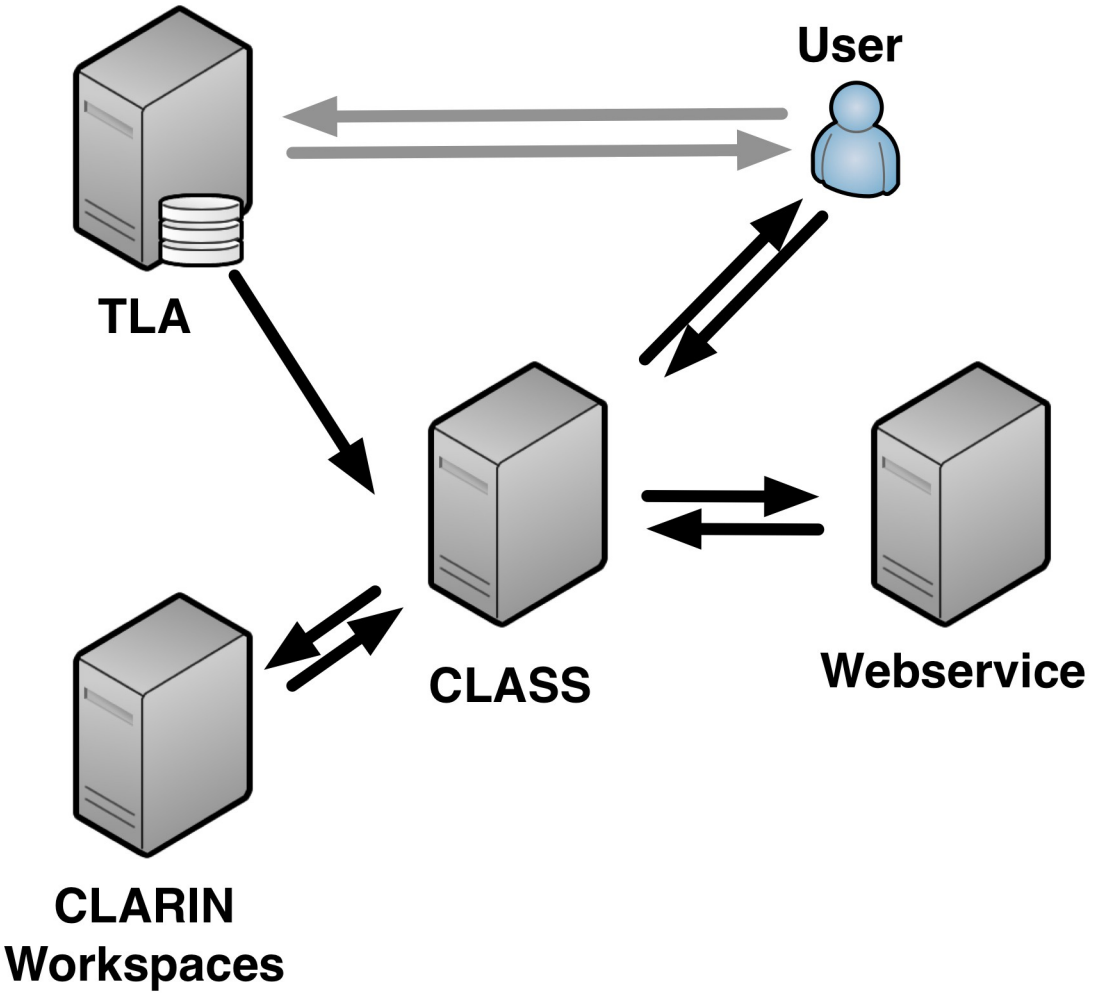
WORDS	Z~eegųgi,	'ee	wii	hotakacra	wirokų	hi'ųnąąną.
MORPH	*žeegų-gi*	*'ee*	*wii*	*hotakac-ra*	*wa-hirokų*	*hi-'ųų-nąą-ną*
GLOSS	thus-TOP	3EMPH	sun	warm.place-DEF	OBJ.3PL-utilize(SBJ.3SG)	APPL.INST-do/make-POT-DECL
TRANS	*That way, it will utilize the hot rays of the sun to advantage.*					

05IH_ED_01_IH.eaf

ED1002 wąąkšik hit'e raažra Mąąšųųrukąnąbįga hįgaire

WORDS	wąąkšik	hit'e	raažra	Mąąšųųrukąnąbįga	hįgaire
MORPH	*wąąkšik*	*hit'e*	*raaš-ra*	*mąąšųų-rukąnąp-įk-ga*	*hį-hige-ire*
GLOSS	Indian/person	speak	name-DEF	feather-be.shiny(OBJ.3SG)-DIM-PROP	1E.U-say.to-SBJ.3PL
TRANS	*my Hocank name is Shining Feather*				

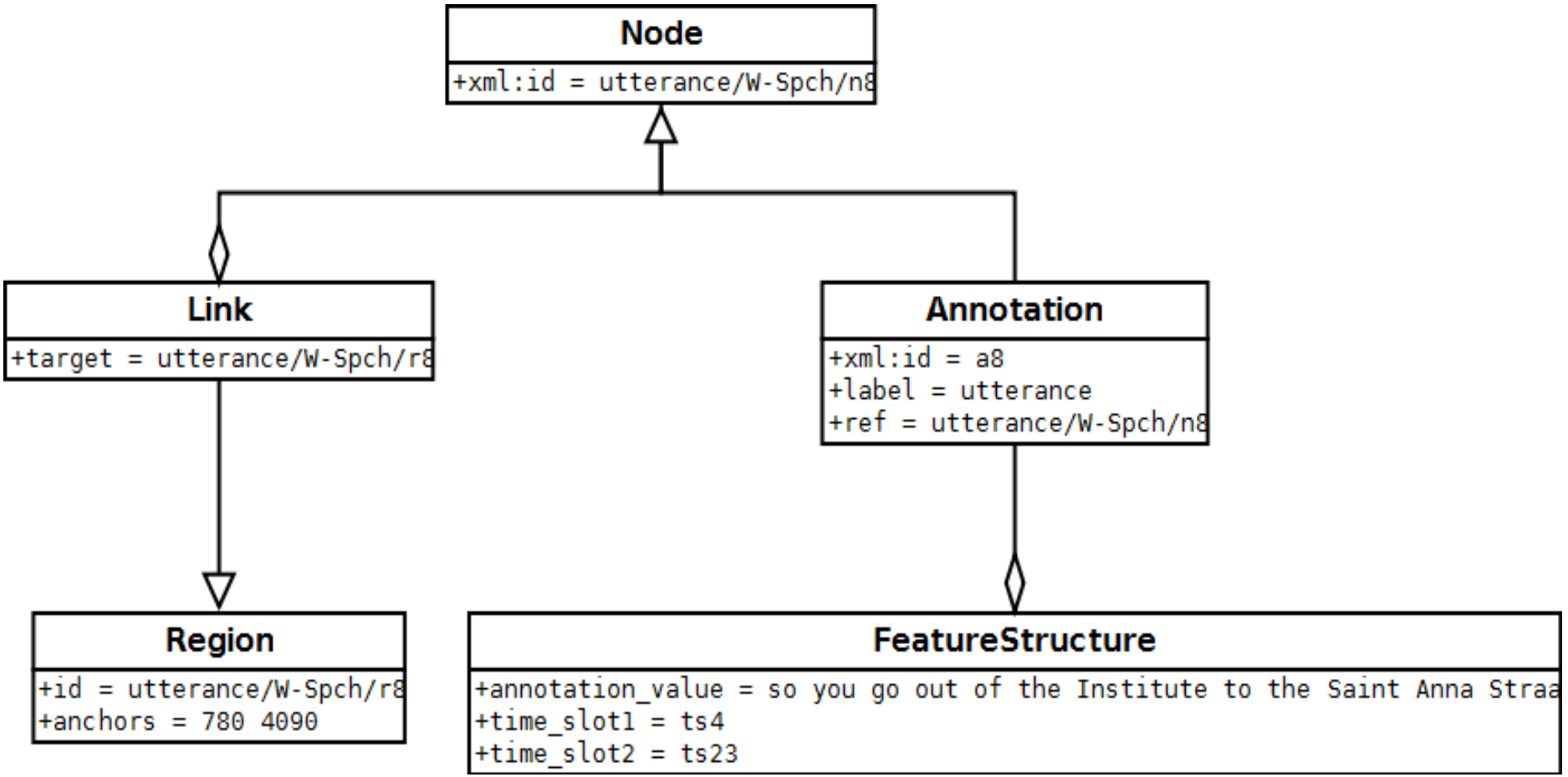
CLARIN



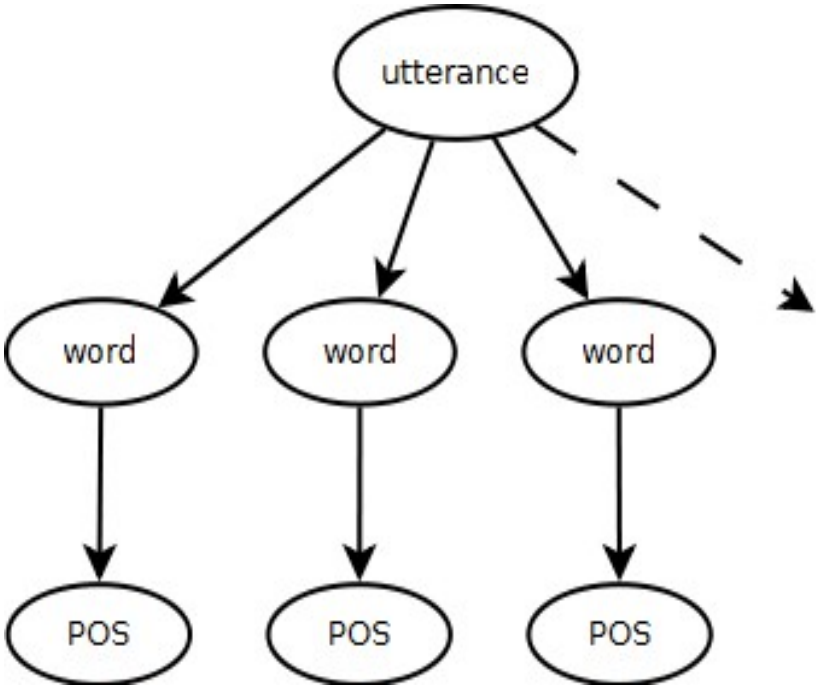
GrAF

- GrAF: Graph Annotation Framework
- ISO 24612: Language resource management - Linguistic annotation framework (LAF)
- Started as stand-off version of XCES
- API and representation as data structures, not a file format
- GrAF/XML as XML representation
- Used for the MASC of the ANC
- Nodes, edges, regions, annotations, feature structures

GrAF entities



GrAF structure



GrAF-XML

```
<node xml:id="words..W-Words..na23">
   <link targets="words..W-Words..ra23"/>
</node>
<region anchors="780 1340" xml:id="words..W-Words..ra23"/>
<edge from="utterance..W-Spch..n8" to="words..W-Words..na23"
        xml:id="ea23"/>
<a as="words" label="words" ref="words..W-Words..na23"
       xml:id="a23">
   <fs>
       <f name="annotation_value">so</f>
   </fs>
</a>
```

Why we use GrAF

- No inline markup
- Radical stand-off approach
  - Easier to share and manage data
  - Preferred solution to archive cultural heritage
  - Ideal for sparse annotations
- Existing code: Java and Python
- API vs. XQuery
- The beauty of annotation graphs

Poio API

- Think of GrAF as an assembly language for linguistic annotation; then Poio API is a libray to map from and to higher-level languages
- Subset of GrAF to represent tier based annotation
  - Interlinear glossed text (IGT)
- Filters and filter chains for search
- Plugin mechanism for file formats
  - Mapping semantics: tiers and annotations to nodes and edges
- Meta-data for additional information (tier types etc.)
- Efforts to map between TEI and GrAF
  - Poio API supports IGT, next step is dictionaries and lexica
  - Retro-digitized dictionary data at University of Marburg are published as GrAF files
  - We want to publish as TEI

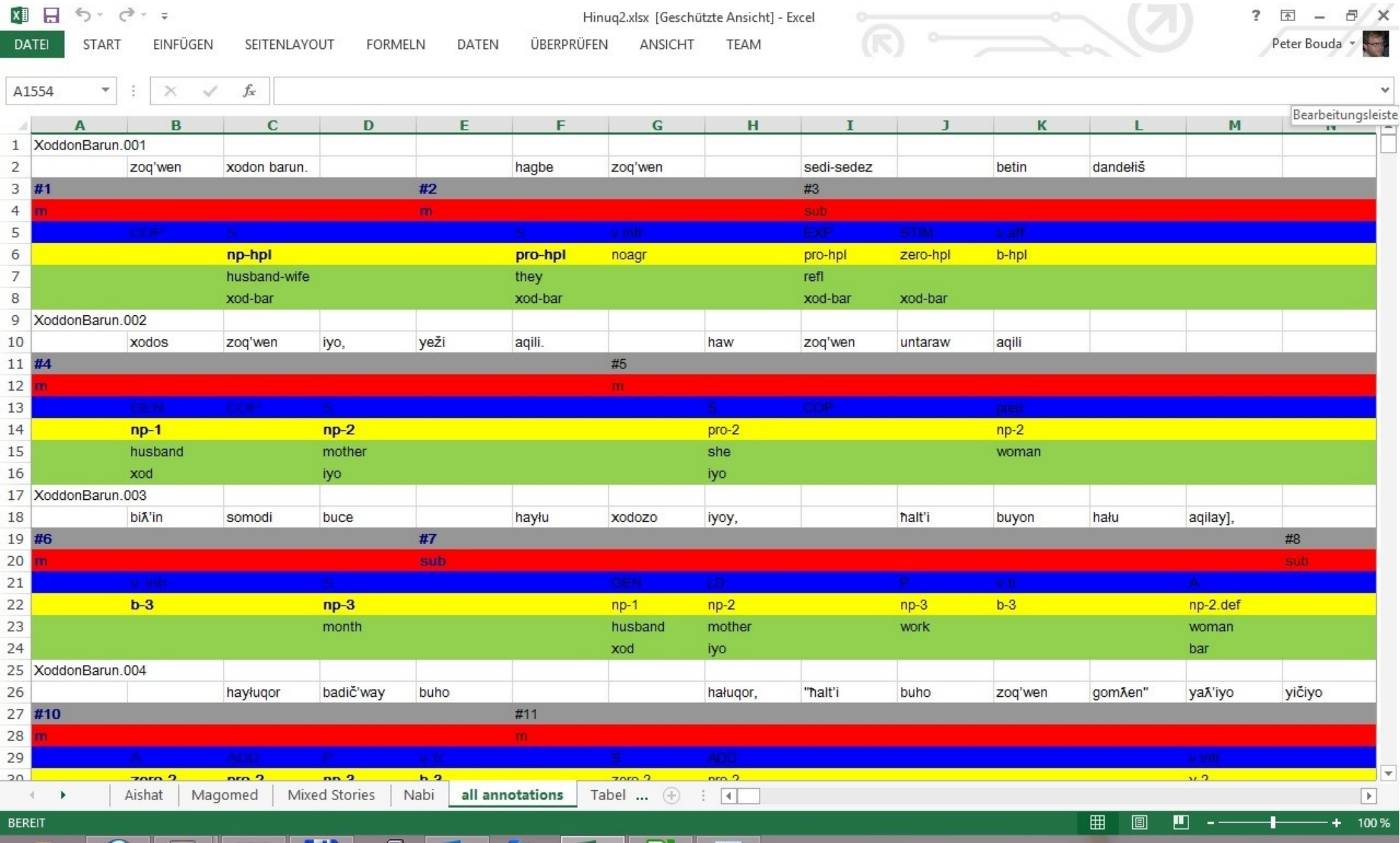
A basic converter in Poio API

```
parser = poioapi.io.wikipedia_extractor.Parser("Wikipedia.xml")
writer = poioapi.io.graf.Writer()

converter = poioapi.io.graf.GrAFConverter(parser, writer)
converter.parse()
converter.write("Wikipedia.hdr")
```

A parser for CSV files

```
class CsvParser(poioapi.io.graf.BaseParser):

    def get_root_tiers(self):
        pass

    def get_child_tiers_for_tier(self, tier):
        pass

    def get_annotations_for_tier(self, tier, annotation_parent=None):
        pass

    def tier_has_regions(self, tier):
        pass

    def region_for_annotation(self, annotation):
        pass

    def get_primary_data(self):
        pass
```

Example: Analysis of CSV data



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	XoddonBarun.001													
2		zoq'wen	xodon barun.			hagbe	zoq'wen		sedi-sedez		betin	dandełiš		
3	**#1**				**#2**				#3					
4	m				m				sub					
5		COP	S			S	v.intr		EXP	STIM	v.aff			
6			**np-hpl**			**pro-hpl**	noagr		pro-hpl	zero-hpl	b-hpl			
7			husband-wife			they			refl					
8			xod-bar			xod-bar			xod-bar	xod-bar				
9	XoddonBarun.002													
10		xodos	zoq'wen	iyo,	yeži	aqili.		haw	zoq'wen	untaraw	aqili			
11	**#4**						#5							
12	m						m							
13		GEN	COP	S				S	COP		pred			
14		**np-1**		**np-2**				pro-2			np-2			
15		husband		mother				she			woman			
16		xod		iyo				iyo						
17	XoddonBarun.003													
18		biƛ'in	somodi	buce		hayłu	xodozo	iyoy,		ħalt'i	buyon	hału	aqilay],	
19	**#6**				**#7**									#8
20	m				**sub**									sub
21		v.intr		S			GEN	LD		P	v.tr		A	
22		**b-3**		**np-3**			np-1	np-2		np-3	b-3		np-2.def	
23				month			husband	mother		work			woman	
24							xod	iyo					bar	
25	XoddonBarun.004													
26			hayłuqor	badič'way	buho			hałuqor,	"ħalt'i	buho	zoq'wen	gomƛen"	yaƛ'iyo	yičiyo
27	**#10**					#11								
28	m					m								
29		A	ADD	P	v.tr		S	ADD					v.intr	
30		**zero-2**	**pro-2**	**np-3**	**b-3**		zero-2	pro-2					v-2	

Example: Analysis of CSV data

http://nbviewer.ipython.org/urls/raw.github.com/pbouda/notebooks/master/Diana%2520Hinuq%2520Word%2520Order%2520TEI.ipynb

Retro-digitization of dictionaries

- From scan to .doc to XML to DB to GrAF
- Radical stand-off approach for unsupervised collaboration
- Dictionaries as cultural heritage texts
- GrAF as primary publication format
- Connectors to brat and TEI

Analysis of the data

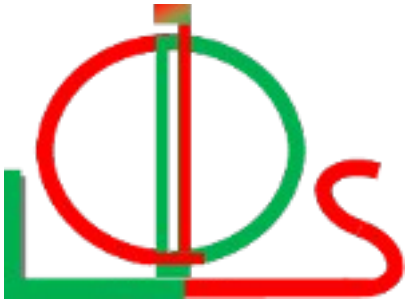
- Spanish as pivot language, subset of bodypart terms
- Converting GrAF to networkx graph
- Nodes are heads, translations, etc.
- Head and translation connected via edges if they appear in one entry
- Merge of graphs
- Count of paths of length 2 between spanish heads
- Python writes JSON graph, visualized with D3.js

D3 visualization

http://www.peterbouda.eu/bodyparts/index_bodyparts.html

Thank you for your attention!

pbouda@cidles.eu



Links

Clarin curation project:
http://de.clarin.eu/en/discipline-specific-working-groups/wg-3-linguistic-fieldwork-anthropology-language-typology/curation-project-1.html

Poio:
http://media.cidles.eu/poio/

GrAF:
http://www.xces.org/ns/GrAF/1.0/