



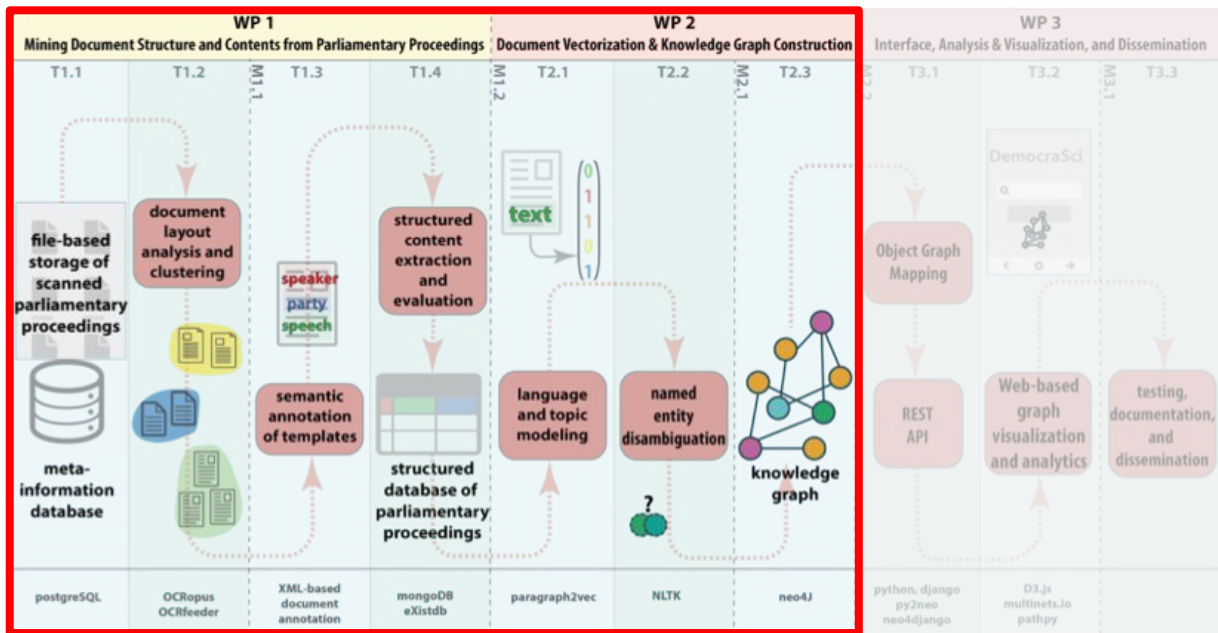
DemocraSci

Follow-up meeting, October 16th, 2018

Overview

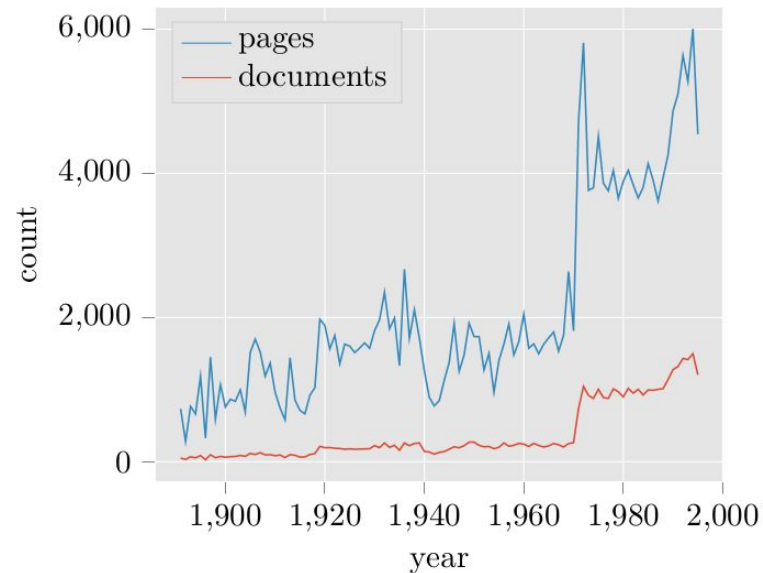
- The project is nicely structured. We have been focusing on WP1&2 so far.

We are working on both WPs at the same time. This way we can benefit from results or requirements of other stages



Data & Storage (T1.1)

- 40'000 OCRed pdf scans:
 - From 1891 to 1995, time period can be extended.
 - Protocols from national council and council of states.
 - Plus metadata on session, date, etc.
- List of politicians.



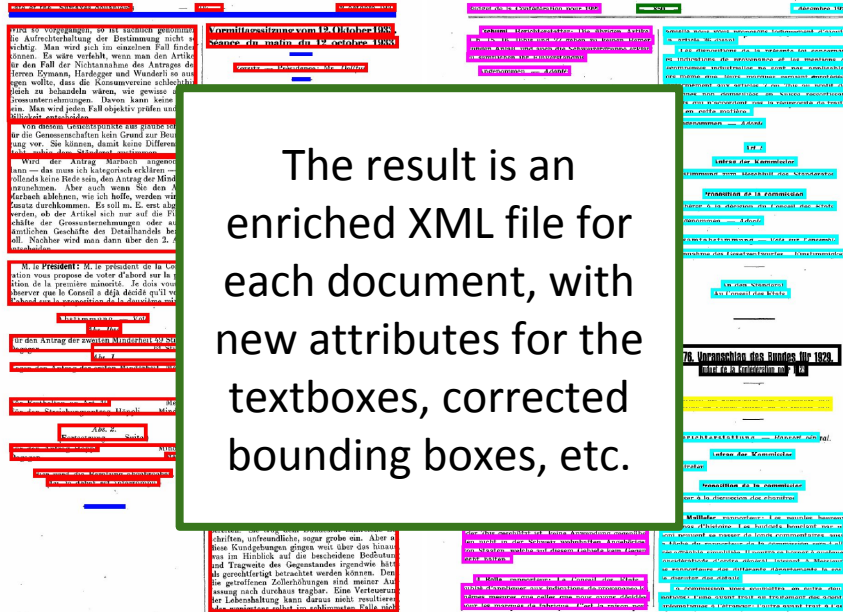
Data stored on Switch.



Dealing with the large number of files, which complicates the use of Renku, where the database will be set up.

Preprocessing of extracted XML data (T1.1-T1.2)

- OCR sufficient but problems with the extracted XMLs.



Implemented a method for:

- Correcting overlaps, mergings textlines, etc.
- Finding the margins, horizontal and vertical lines.
- Identifying page elements (needs to be improved for some).

Rule-based methods perform better than an unsupervised approach.

- There are different layouts, but normally comprised in single proceedings.

Layout analysis and annotation (T1.2&1.3)

- There are different layouts, but normally comprised in single proceedings.
 - Possible sections: laws, votes, speeches, etc.
- ↻ Analysis of the document at the level of the textboxes (paragraphs, headers, etc.) aiming at identifying different types of texts:
 - Later this will help to cluster documents according to their sections.
 - ✓ Implemented for extracting speeches, as we required them for WP2.

Structured content extraction (T1.4)

- ✓ We already extracted:
 - All speeches with corresponding speakers.
 - Enriched metadata file.
 - List of politicians and councils presidents.

- This information can be readily integrated into a MongoDB database:
 - One entry per document with previously identified elements as attributes.
 - Associated to a relational database with the list of politicians, their canton, party, etc.

Preprocessing for WP2

- ✓ For topic modeling on the extracted speeches:
 - Language identification
 - (Thus far only) for German speeches
 - Tokenization
 - POS-tagging
 - Lemmatization
 - All lower case
 - Stopword removal
 - A document are all speeches of one proceeding.
- ✓ For knowledge graph (from extracted speeches):
 - Extract mentioned people (using NER to identify person first)

Der Hund jagt die Katzen

Der - Hund - jagt - die - Katzen

Der A - Hund N - jagt V - die A - Katzen N

Der - Hund - jagen - die - Katze

hund - jagen - katze

Hund - jagen - Katze

Language and topic modelling (T2.1)

- ✓ Latent Dirichlet allocation (LDA) was applied to:
 - Two legislative periods (39th, 44th).
 - The complete time-span splitted in two sub-periods (1891 - 1970, 1971 - 1995).

- 44th legislative period (1991 - 1995)
 - 2799 documents.
 - ✓ Playing with parameters → good results with 100 topics and neglecting terms that appear in more than 10% of documents and in less than 20 documents.
 - ✓ 49 topics could be named.

LDA for 44th legislative period (1991 - 1995)

asylum	
$\lambda = 1$	$\lambda = 0$
ausländer	asylbewerber
asylbewerber	asylgesetz
flüchtling	einbürgerung
koller	ausschaffungshaft
ausländerrecht	asylverfahren
zwangsmassnahmen	asylgesuch
einbürgerung	asylsuchender
asylgesetz	anag
standesinitiative	fremdenpolizei
asylpolitik	ausweisung

drugs	
$\lambda = 1$	$\lambda = 0$
droge	drogenpolitik
drogenpolitik	betäubungsmittel
übereinkommen	drogenkonsum
betäubungsmittelgesetz	süchtigen
illegal	heroinabgabe
lette	stupéfiantes
drogenproblem	drogenabhängige
heroin	drogenszen
drogenhandel	süchtig
prävention	drogensucht

$$relevance(w|t) = \lambda \cdot p(w|t) + (1 - \lambda) \cdot \frac{p(w|t)}{p(w)}$$

given topic name(s)	number
general terms + federal elections	1
French terms	2
parliamentary investigations	3
parliamentary investigations + taxes + federal budget	4
financial center ⁴ + law	5
federal budget + research	6
parliamentary votes	7
accession to EU	8
unemployment + occupational benefit plan	9
adaption to European law	10
railway	11
world trade	12
social and health insurance for farmers	13
NEAT	14
crime prevention	15
asylum	16
UN blue helmet troops	17
armament	18
agriculture	19
AHV revision	20
genetic engineering	21
public organizations + Italian terms	22
racism criminal norm ⁵	23
Armee 95	24

LDA for 44th legislative period (1991 - 1995)

- Five topics allocated to 1993/20022799
 - Financial center + law (44%)
 - Asylum (32%)
 - Racism criminal norm (13%)
 - Federal budget (10%)
 - Acquisition of fighter aircrafts (1%)

✓ They cover the content well!

- But, “aus der Luft gegriffen”

91.3290

Postulat Seiler Hanspeter
Unterkunft und Verpflegung
von Asylbewerbern
Demandeurs d'asile.
Gîte et couvert

Diskussion – Discussion

Siehe Jahrgang 1992, Seite 627 – Voir année 1992, page 627

Frau **Goll**: Ich möchte Sie bitten, dieses Postulat abzulehnen, und zwar aus folgenden Gründen:
Ihre Sprache, Herr Seiler, ist äusserst entlarvend. Der in Ihrem Postulat angeschlagene Ton zeugt von tiefster Verachtung gegenüber benachteiligten und diskriminierten Menschengruppen und schürt in erheblichem Masse Fremdenfeindlichkeit und Fremdenhass. Ich möchte mit Ihnen Ihren Text kurz näher anschauen:

Ich bin über die Äusserungen von Frau Goll sehr verwundert, da ich mit ihr noch nie ein Wort über solche Fragen diskutiert habe. Sie hat irgend etwas aus der Luft gegriffen und mich als asozialen Typ hingestellt. Ich bitte Sie: Kommen Sie einmal zu uns; dann schauen wir einmal gewisse Sachen an. Was ich

... dass gerade Leute wie Sie mit solchen Äusserungen dazu beitragen, diesen Unmut anwachsen zu lassen. Ihrer Meinung nach sollten sich die Fürsorgeleistungen explizit und ausgerechnet «für Asylbewerber», wie Sie betonen, «in vernünftigen Grenzen halten». Was heisst denn das? Sie sug-

LDA for 39th legislative period (1971 - 1975)



Mostly similar topics but with a different focus:

- Development aid for Africa and Asia instead of Eastern Europe
- Drugs on alcoholism instead of hard drugs

alcoholism (drugs)	
$\lambda = 1$	$\lambda = 0$
bundesanwaltschaft	alkoholismus
bundesanwalt	droge
alkoholverwaltung	bundesanwalt
lehrer	straflos
strafe	händler
alkoholismus	bundesanwaltschaft
jugendliche	bestrafung
konsum	betäubungsmittel
bestrafen	betäubungsmittelgesetz
droge	strafandrohung

drugs	
$\lambda = 1$	$\lambda = 0$
droge	drogenpolitik
drogenpolitik	betäubungsmittel
übereinkommen	drogenkonsum
betäubungsmittelgesetz	süchtigen
illegal	heroinabgabe
lette	stupéfiant
drogenproblem	drogenabhängige
heroin	drogenszen
drogenhandel	süchtig
prävention	drogensucht



Some different topics

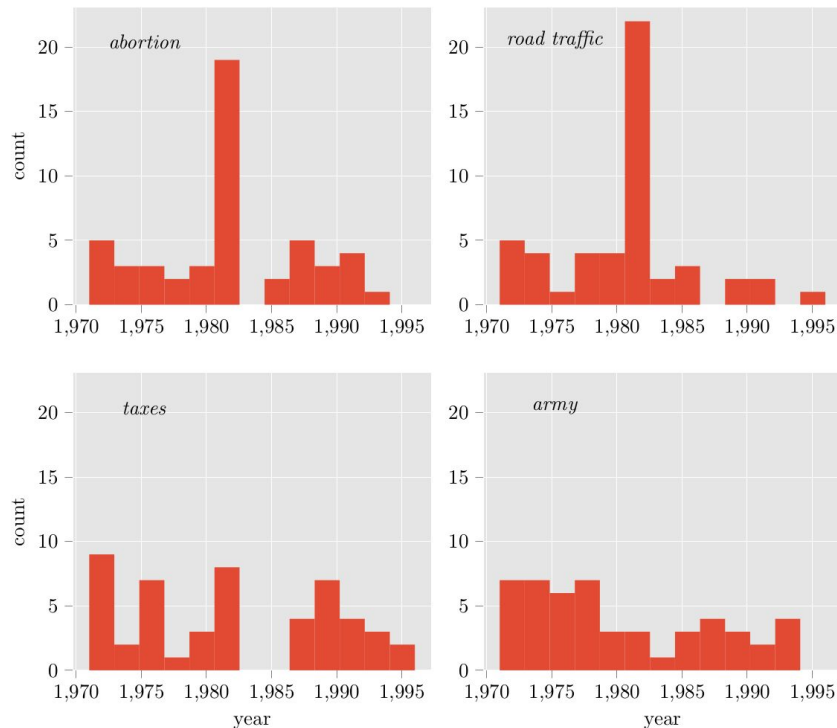
- 39th: electricity from nuclear energy
- 44th: UN blue helmet troops

LDA for 1971 until 1995



Dynamic evaluation

- By plotting years of top 50 documents for each topic as a histogram
- Abortion law in 1982
- Introduction of “Autobahn Vignette” in 1985



Language and topic modelling (T2.1)



Summary:

- Obtained good overview of corpus.
- LDA proved successful for topic identification.



Future work:

- Advanced dynamic topic modeling
- Approaches to determine hierarchy of topics (general and sub-topics)
- Embedding techniques for dealing with different languages
- Revisit after identifying sections and types of documents on WP1
- Enrich database with extracted topics
- Other ideas?

Open issues

- Named entity disambiguation (T2.2)
 - Are the following databases available?
 - Semantically enriched and disambiguated database of delegates of the Swiss Parliament from Prof. Simon Hug.
 - Structured database in PI's group for the disambiguation of committees, members of government, parties and fractions, cities and regions.
- Knowledge graph (T2.3)
 - ↻ We are currently investigating on how to set this up
 - Following example of LegisGraph project
 - Using neo4j and py2neo

