Government Gazette Text Mining, Cross-Linking and Codification - 3gm

Google Summer of Code

Marios Papachristou
GFOSS - Open Technologies Alliance

github.com/eellak/gsoc2018-3gm

Mentors: D. Spinellis, A. Zavras, S. Kapidakis

Problem & Project Statement

- Codification (Wikipedia): In law, codification is the process of collecting and restating the law of a jurisdiction in certain areas, usually by subject, forming a legal code, i.e. a codex (book) of law.
- Done by hand!
- Automate it!

Working / Finished

1. Document Parser
2. **Named Entities** for Legal Acts (e.g. Laws, Legislative Decrees etc.) encoded in regular expressions.
3. **Topic models** for finding Government Gazette Issues that have the same topics using LDA and NMF.
4. Optionally trained Word2Vec Model for further usage
5. **MongoDB** Integration
6. Fetching tool for documents (automated process)
7. OCR tool for converting legacy (<1999) documents with Google Tesseract 4.0
8. Digitize GG documents from 1976 - today
9. Project **Wiki**
10. etc.

In Progress

Heuristic methods for detecting amendments. For example (taken from Greek Government Gazette):

Amendment

Μετά το άρθρο 9Α του ν. 4170/2013 **προστίθεται** *άρθρο 9ΑΑ*, ως εξής:

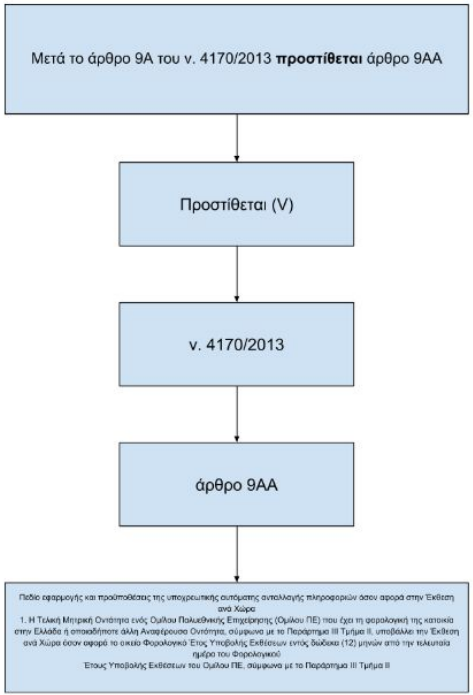
Main Body / Extract

Άρθρο 9ΑΑ

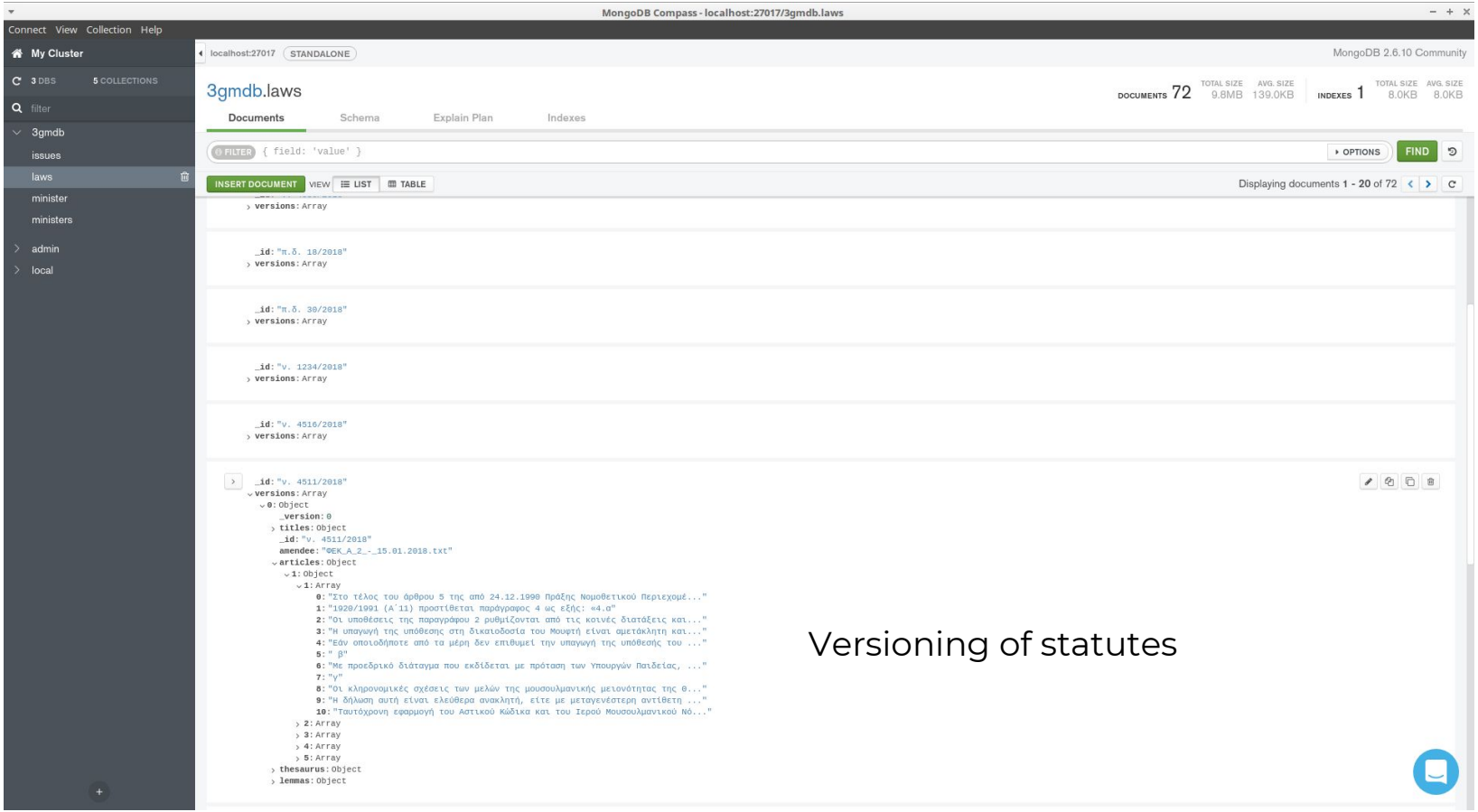
Lorem Ipsum.....

Results in a database query that adds an article on a MongoDB

database



Database Structure



Versioning of statutes

Challenges

1. Government Gazette Issues may not always follow guidelines
2. Improving heuristics
3. No substantial NLP progress in Greek
4. Documents not in plaintext format

Thank you!

Repository: github.com/eellak/gsoc2018-3gm

Contact: papachristoumarios [at] gmail [dot] com

5862 31C0 05F4 8C71 A5B6 A1CE D6BC 45BD E0DC 0EDA