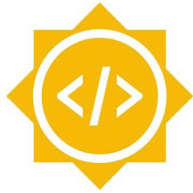


# Government Gazette Text Mining, Cross-Linking and Codification - 3gm

Google Summer of Code



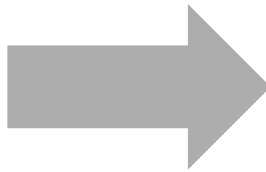
[github.com/eellak/gsoc2018-3gm](https://github.com/eellak/gsoc2018-3gm)

Marios Papachristou  
GFOSS - Open Technologies  
Alliance

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:

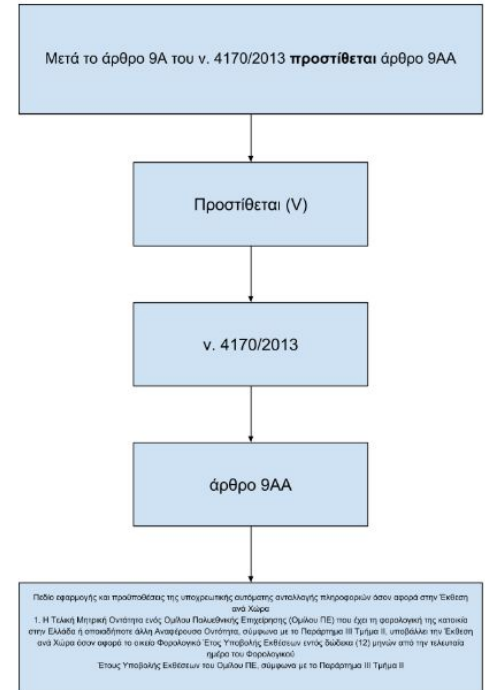
Main Body / Extract

Άρθρο 9ΑΑ

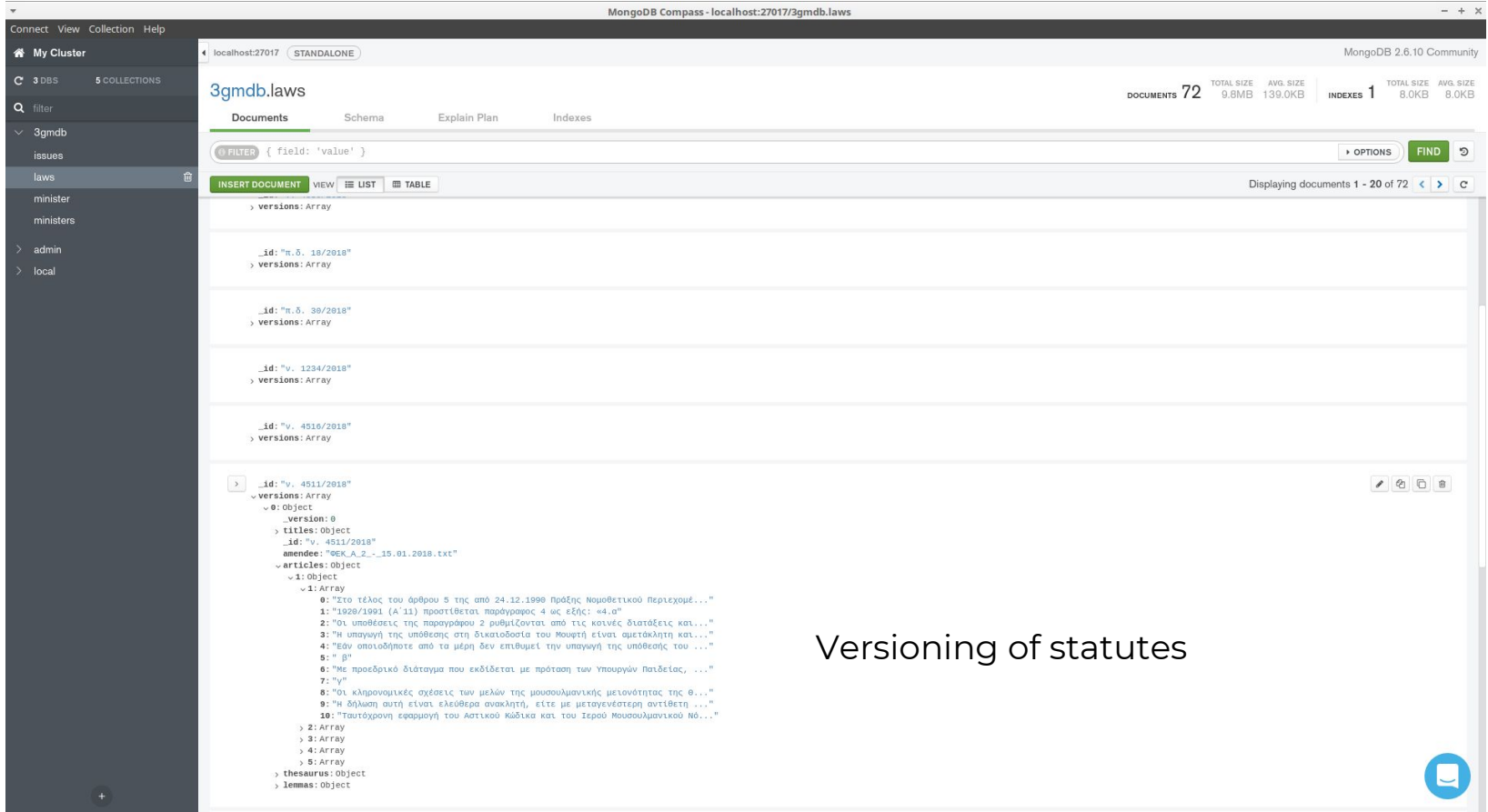
Lorem Ipsum.....

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

database



# Database Structure



The screenshot shows the MongoDB Compass interface for a database named '3gmdb.laws'. The 'Documents' tab is active, displaying a list of documents. The selected document is expanded to show its structure, including a versioned array field.

Database: 3gmdb.laws  
Documents: 72 (Total Size: 9.8MB, Avg. Size: 139.0KB)  
Indexes: 1 (Total Size: 8.0KB, Avg. Size: 8.0KB)

Filter: { field: 'value' }

Displaying documents 1 - 20 of 72

Document Structure:

```
{
  "_id": "v. 4511/2018",
  "versions": Array
    [
      {
        "_id": "n.5. 18/2018",
        "versions": Array
          [
            {
              "_id": "v. 1234/2018",
              "versions": Array
                [
                  {
                    "_id": "v. 4516/2018",
                    "versions": Array
                      [
                        {
                          "_id": "v. 4511/2018",
                          "versions": Array
                            [
                              {
                                "_version": 0,
                                "titles": Object
                                  {
                                    "_id": "v. 4511/2018",
                                    "amendee": "ΦΕΚ Α 2_- _15.01.2018.txt",
                                    "articles": Object
                                      {
                                        "1": Object
                                          {
                                            "1": Array
                                              [
                                                "0: \"Στο τέλος του άρθρου 5 της από 24.12.1990 Πράξης Νομοθετικού Περιεχομέ...\"",
                                                "1: \"1929/1991 (Α'11) προστίθεται παράγραφος 4 ως εξής: «4.α\"",
                                                "2: \"Οι υποθέσεις της παραγράφου 2 ρυθίζονται από τις κοινές διατάξεις κατ...\"",
                                                "3: \"Η υπαγωγή της υπόθεσης στη δικαιοδοσία του Μουφτή είναι αμετάκλητη κατ...\"",
                                                "4: \"Εάν οποιοδήποτε από τα μέρη δεν επιθυμεί την υπαγωγή της υπόθεσής του ...\"",
                                                "5: \"β'\"",
                                                "6: \"Με προεδρικό διάταγμα που εκδίδεται με πρόταση των Υπουργών Παιδείας, ...\"",
                                                "7: \"γ'\"",
                                                "8: \"Οι κληρονομικές σχέσεις των μελών της μουσουλμανικής μειονότητας της θ...\"",
                                                "9: \"Η δήλωση αυτή είναι ελεύθερα ανακλήσιμη, είτε με μεταγενέστερη αντίθετη ...\"",
                                                "10: \"Ταυτόχρονη εφαρμογή του Αστικού Κώδικα και του Ιερού Μουσουλμανικού Νό...\"
                                              ]
                                            "2": Array
                                              [
                                                "2: Array",
                                                "3: Array",
                                                "4: Array",
                                                "5: Array"
                                              ]
                                            "thesaurus": Object
                                            "lemmas": Object
                                          ]
                                        ]
                                      ]
                                    ]
                                  ]
                                ]
                              ]
                            ]
                          ]
                        ]
                      ]
                    ]
                  ]
                ]
              ]
            ]
          ]
        ]
      ]
    ]
  ]
}
```

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/qsoc2018-3gm](https://github.com/eellak/qsoc2018-3gm)

Contact: papachristoumarios [at] gmail [dot] com

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