DATA MINING THE PARLIAMENT

A DATASET OF ANSWERS

IN SEARCH OF QUESTIONS

„OPEN DATA“ MADE IN GERMANY – THE THEORY

[sic]

Transparenter Staat

Die digitale Berichterstattung über den Bundestag und seine Sitzungen sowie über öffentliche Ausschusssitzungen und Anhörungen (z. B. in Streams) wollen wir ausbauen. So bald wie möglich werden wir Bekanntmachungen wie beispielsweise Drucksachen und Protokolle in Open Data tauglichen Formaten unter freien Lizenzbedingungen bereitstellen.

(aus dem Koalitionsvertrag der großen Koalition, 2013, S. 152)

Lol, § 5 Abs. 2 UrhG

„OPEN DATA“ MADE IN GERMANY – IN PRACTICE

Parliamentary document system: pdok.bundestag.de

- No API
- No Bulk-Download
- PDF-Dokuments
- No* metadata with information about connection between related documents

* Exists in separate system with other document identifiers, and even harder to script (come on, it‘s 2017, stop it with the server-side state already)

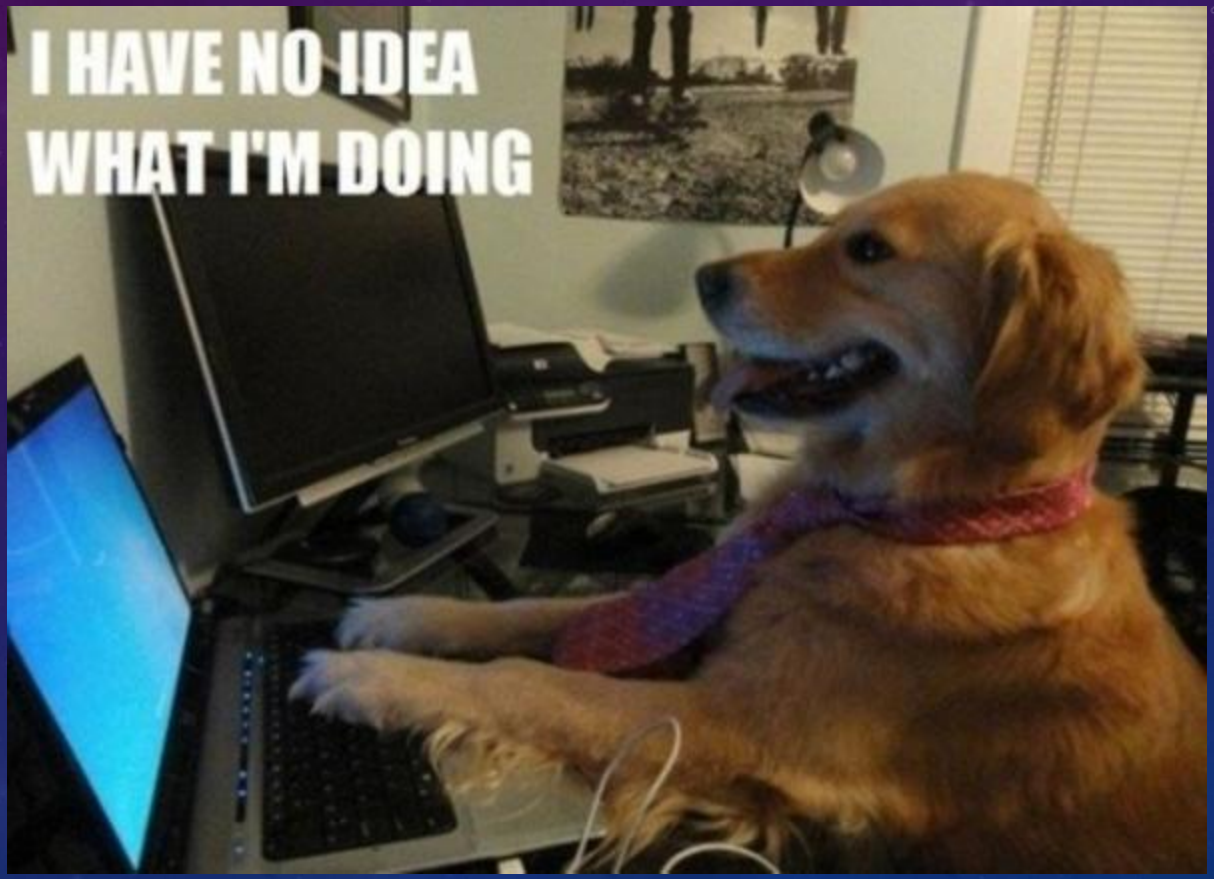
THERE ARE SOME UPSIDES:

- PDFs with text track and very good OCR for all documents, dating back to 1949
- Predictable URLs
- Decent HTML to extract metadata using site scraping
- A very friendly tech team („You want to scrape us? Go right ahead, just please do it at night.“)

THE RESULT

- ~ 500 LoC Python (github.com/malexmave/pdok-mirror, AGPLv3)
- ~ 19 000 „kleine Anfragen“
- ~ 10 000 draft laws
- ~ 4 000 Plenars protocols
- ~ 130 000 documents in total
- In total: 76 GB as .pdf and .txt (<3 `pdftotext -layout`)
- A backup of our democracy at Archive.org, for good measure

WHAT‘S NEXT?



YOU GOT A COOL IDEA?
YOU KNOW HOW TO IMPLEMENT IT?

- Get the dataset using:
  - IA client: ia download --search 'collection:deutscherbundestag' --glob '*.pdf' (needs internet archive account and "pip install internetarchive" – takes quite a while to download)
  - HTTPS: https://more.velcommuta.de/34c3/bundestag/ (as .tar.gz)
  - Collect it yourself: github.com/malexmave/pdok-mirror (please be polite to the server and only run this at night)