IR 4 OpenGov: finding the needle in 22M pages of Open Government data!

A collection of FAIR Dutch
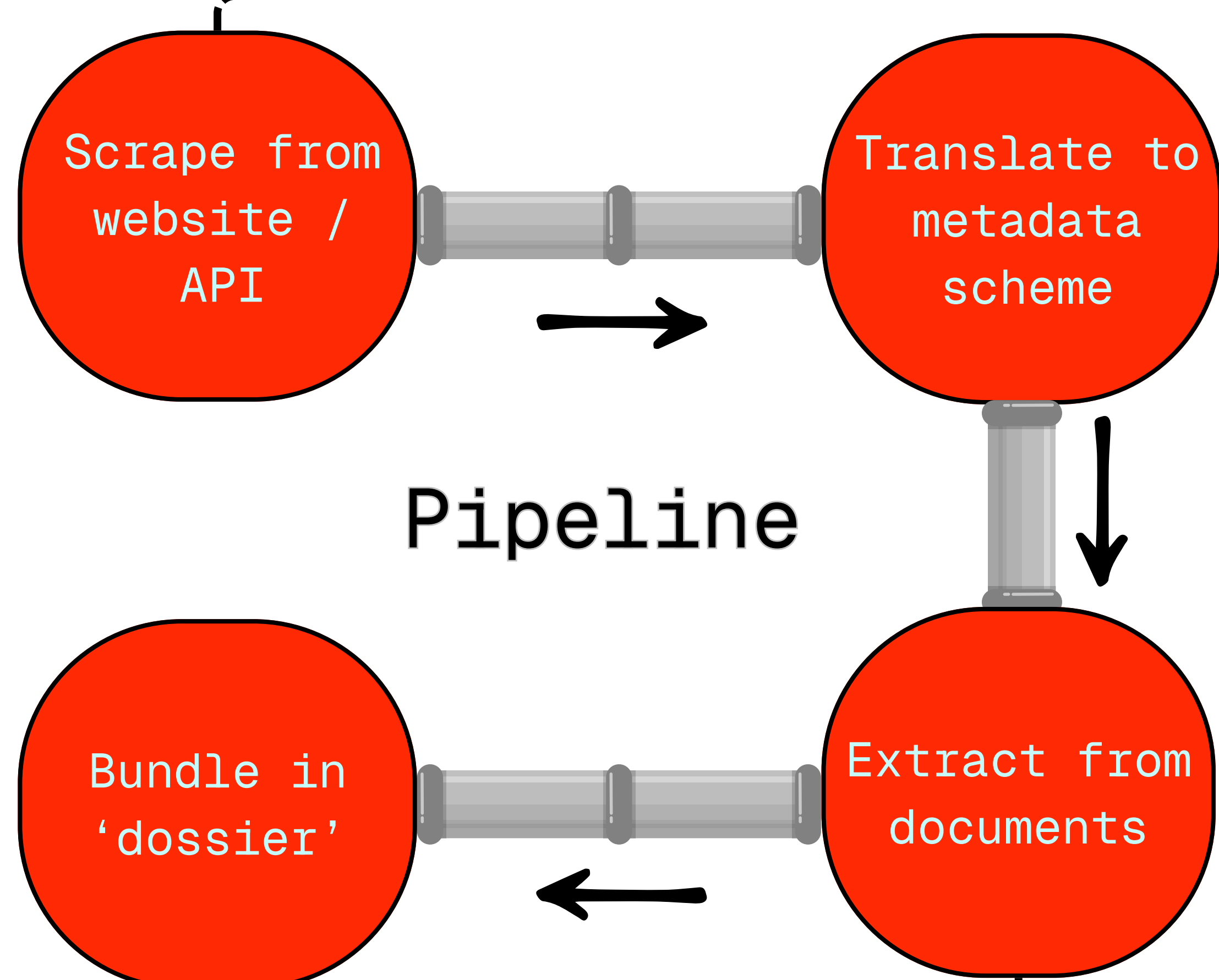## Freedom of Information Act documents

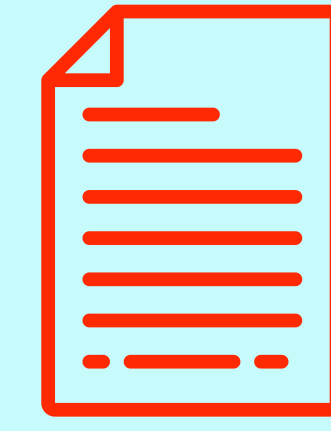Ruben van Heusden, Maik Larooij, Jaap Kamps, Maarten Marx

Maik Larooij (larooij@opengov.nl)
IR @ OpenGov Lab x IRLab, UvA

The Dutch 'Wet open overheid' (Woo)

Dutch governments are obliged to publish information regarding policy in both a **passive** (on request) and **active** (self-initiated) manner. In total, there are **17 mandatory categories of information**. Examples include transcripts of **parliamentary meetings**, decisions on **information requests**, complaints, **research articles** and **covenants**.
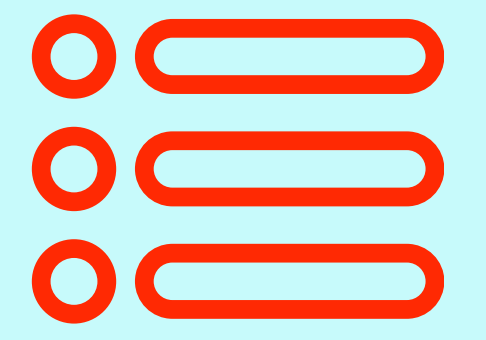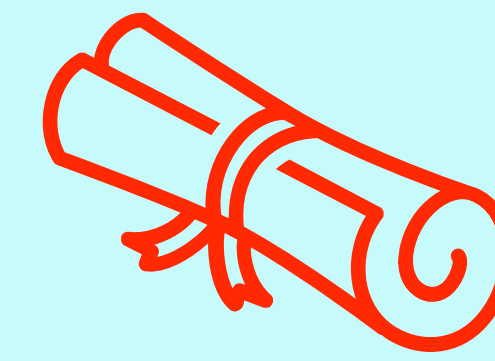
Pipeline

- Scrape from website / API
- Translate to metadata scheme
- Extract from documents
- Bundle in 'dossier'

(OCR) text extraction

Quality check: FAIRiscore

Redaction detection

Reference extraction

Woo-verzoek
Techsector Nijmegen

ID	nl.gm0268.2i.2025.1
Type	2i (Woo-request)
Publisher	Gem. Nijmegen
Request date	2024-09-16
Decision date	2025-01-14
Theme	infrastructure

- Request
- Decision
- Inventory
- Released (682)

An example dossier (bundle of documents) of information category 'Woo-request'

22M pages
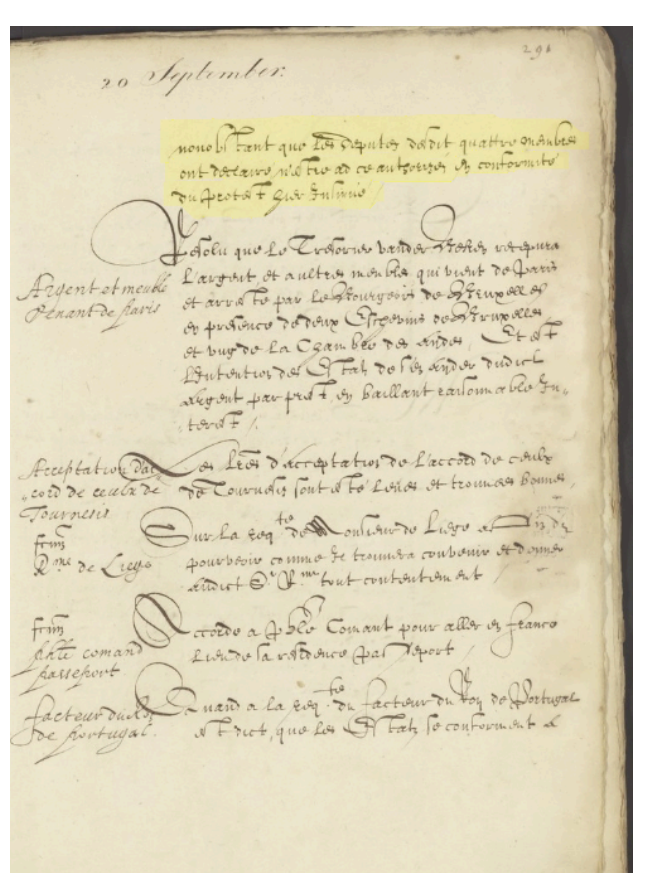
6.5M docs

12 categories

1000+ organizations

Oldest document from 1577

1 search engine

WooGLe
wooverheid

Some of the **diverse** types of information we encountered!

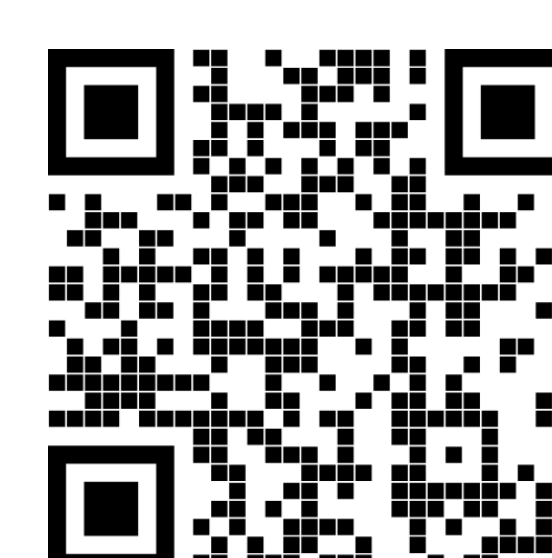### What can/did we do with this data?

- **Extract** information: entities, timelines, topics, summaries...
- **Find** information: semantic search, RAG, conversational search...
- **Monitor** information: machine-readability, accessibility, quality, redaction detection
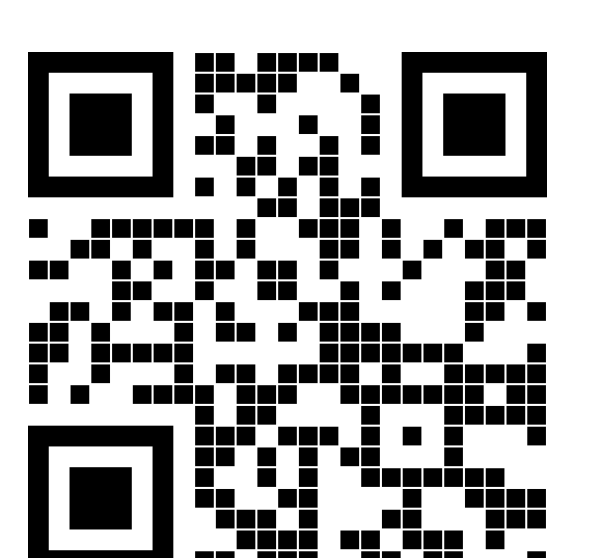
Future directions / challenges

- What do **citizens** want to know?
- Sensitive context: how can we use documents that are **not public yet**?
- LLMs: hallucinations → **grounding**, explainability, bias

DANS dataset

WooGLe

OpenGov Lab

Invitation: what do you think are interesting future directions?