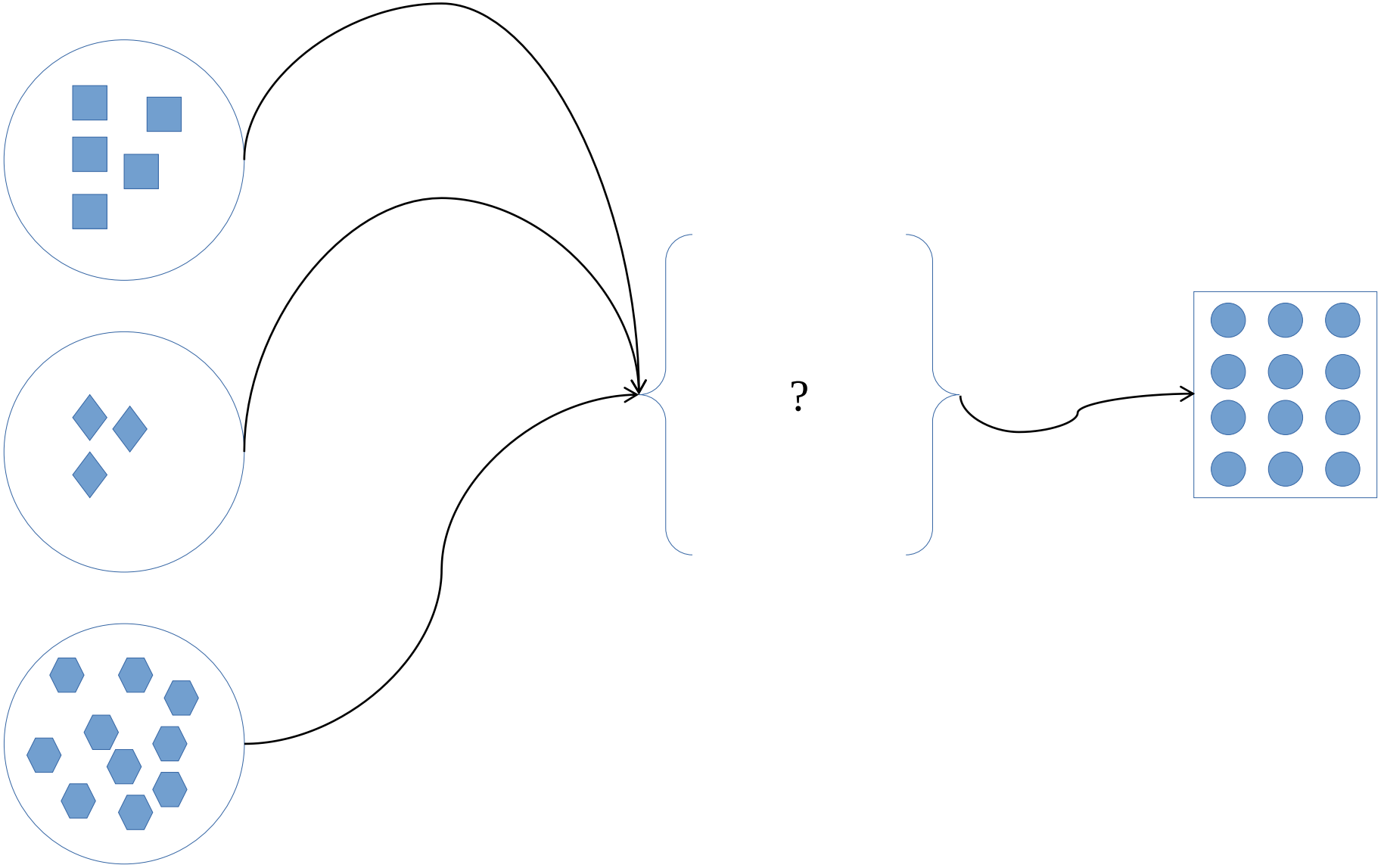
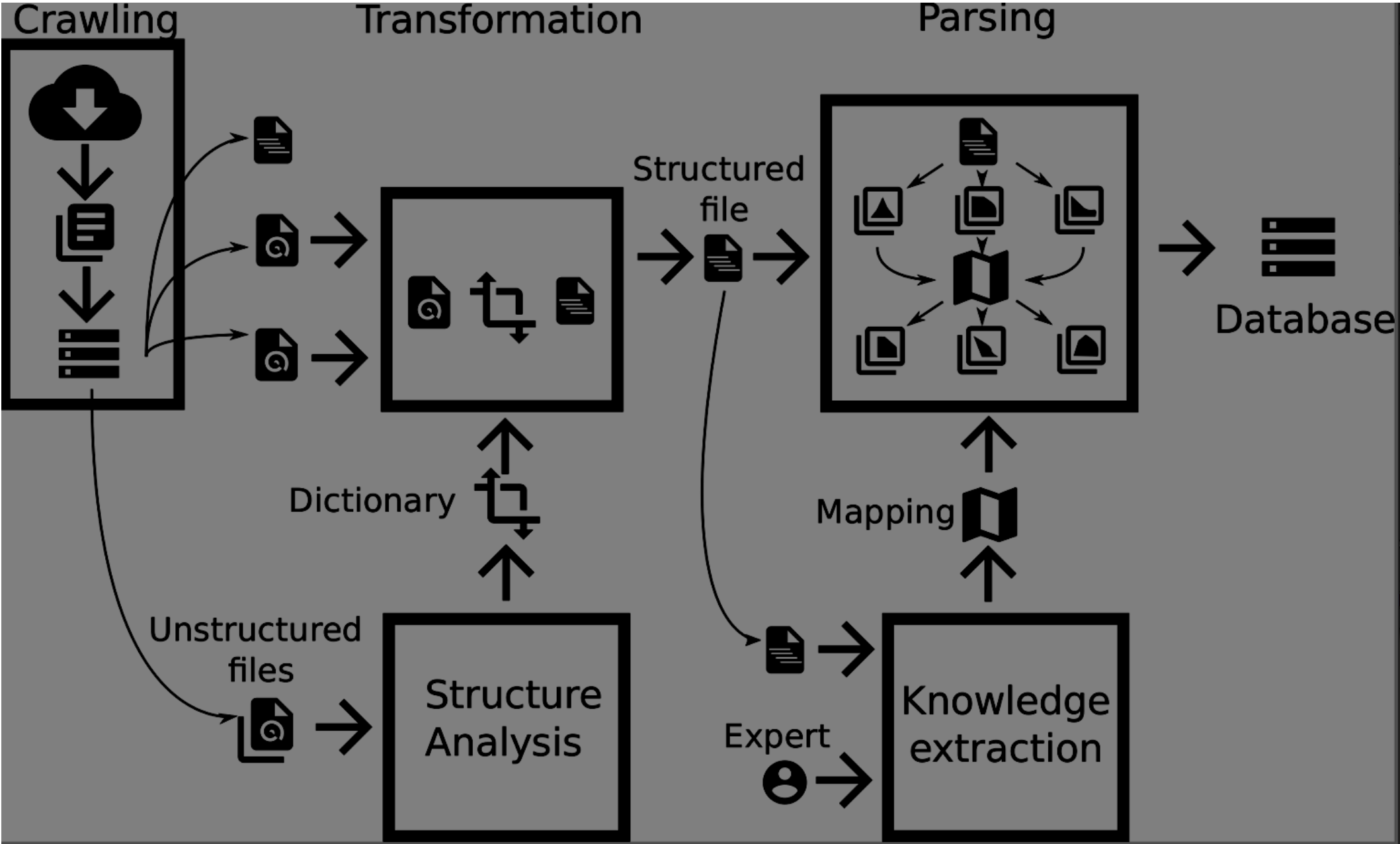
MAKI: web data knowledge extraction

Data collection and storage

The general idea



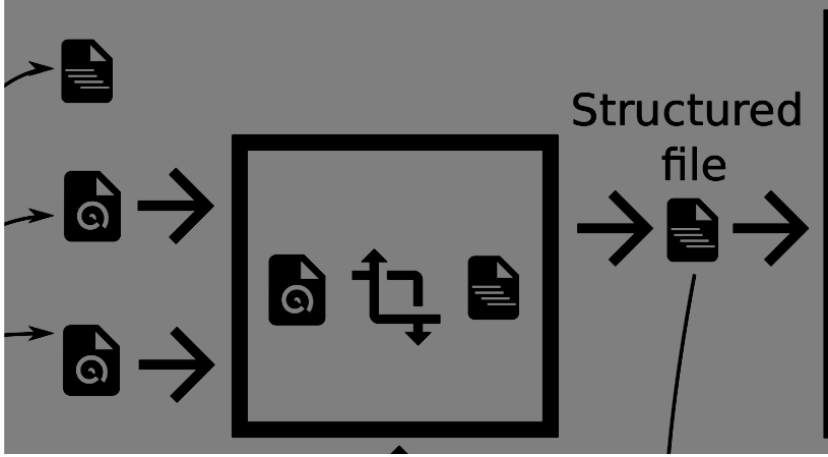
The big picture



Crawling

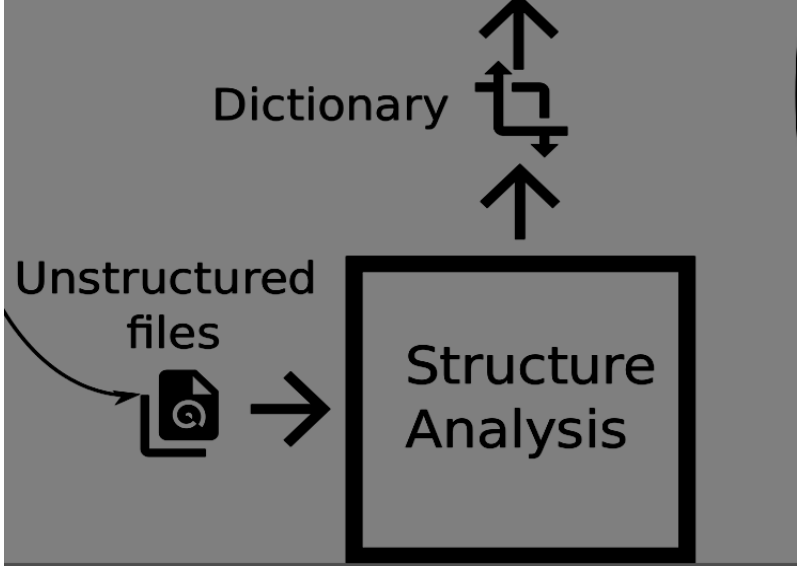
- Each web source is completely different
- Unnecessary use of Javascript
- Conte located using dynamic URLs
- Non-consistent HTML design
  - Same element under different XPATHS
  - No use of HTML identifiers
- Different languages
- Finally we have written 26 ad-hoc crawlers
- Use existing tools: Selenium, Scrapy, PhantomJS
- Try to crawl following the same steps

Transformation



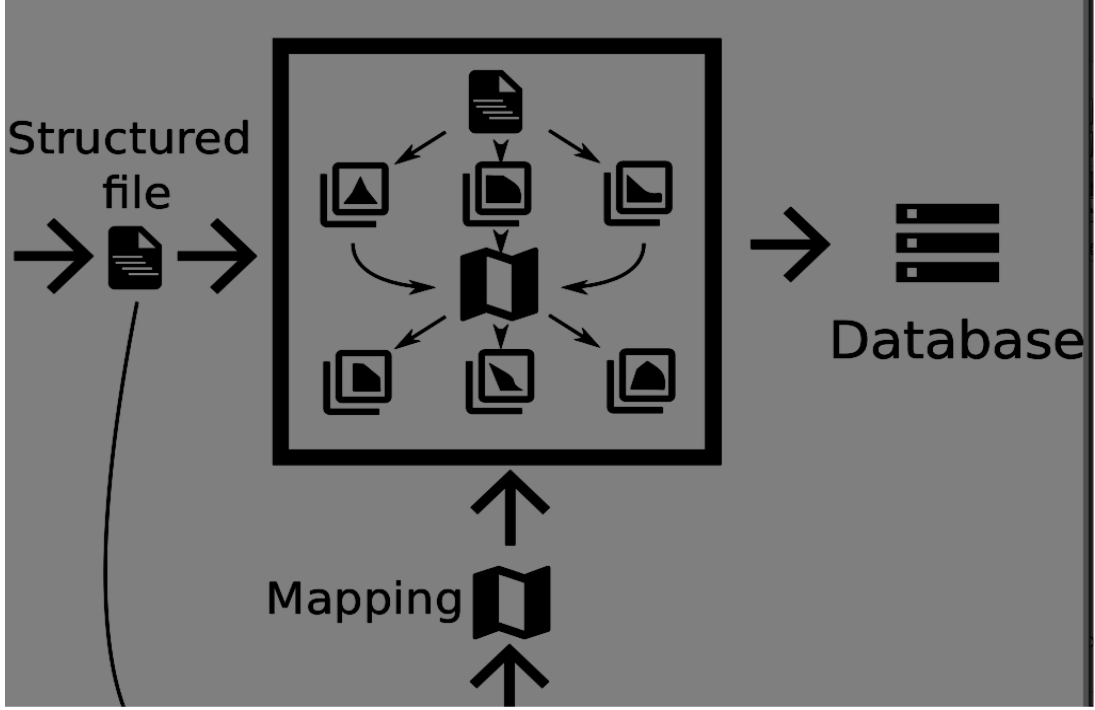
- Most online databases are displayed in a key-value way
- However, large portions of data are incosistent with this approach
- We transform crawled elements into a ready-for-parsing key-value file
- Before that we define a key-value transformation schema

Structure analysis



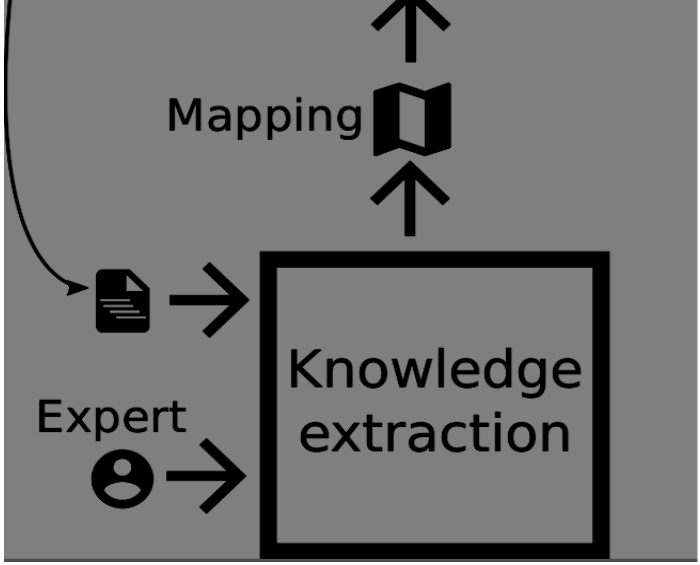
- Find the key/value structure in the document
- Frequency analysis, synonyms, similarity matrix, tokenization
- Work independently of the language or the number of existing templates

Parsing



- Transform the original input file into a defined entry in our database
- Mapping dependant process
- Different input sources and/or database schemas need different mappings
- Use existing expert knowledge to define parsing

Knowledge extraction



- The expert is really the person who should define how to transform the existing data
- The knowledge extraction is the cornerstone of our approach
- Any user can parse a large amount of data only defining a mapping file

Database

- Final delivered product
- Schema design is crucial for the final solution
- Modifications in the database schema invalidate existing mappings
- However, changing mappings is less expensive than modifying parsers

The extraction cycle

- The full process is an iterative cycle
  - Crawl
  - Extract existing structures
  - Create mapping
  - Parse
  - Check database output
  - Repeat until the output is good enough

Conclusions

- MAKI: is a set of tools for web data knowledge extraction
- New approach:
  - Expert driven
  - Reusable tasks
  - Supports structured and non-structured data
  - Pipeline solution
  - Machine-assisted
  - Multiple sources
  - Multilingual
  - Specific vocabulary
  - Non domain experts driving the process
- Real case study building knowledge using various European procurement entities