Food and Agriculture Organization of the United Nations

>> FAO Statistics Division

FAO x UNSD Brown Bag webinar, 13th June 2023

# Essence: an integrated framework for documents retrieving and analysis

Carola Fabi, Marco Scarnò, Craig Matadeen

Essence: a FAO DataLab application

(Expert Search Semantic ENriChmEnt=>Essence)

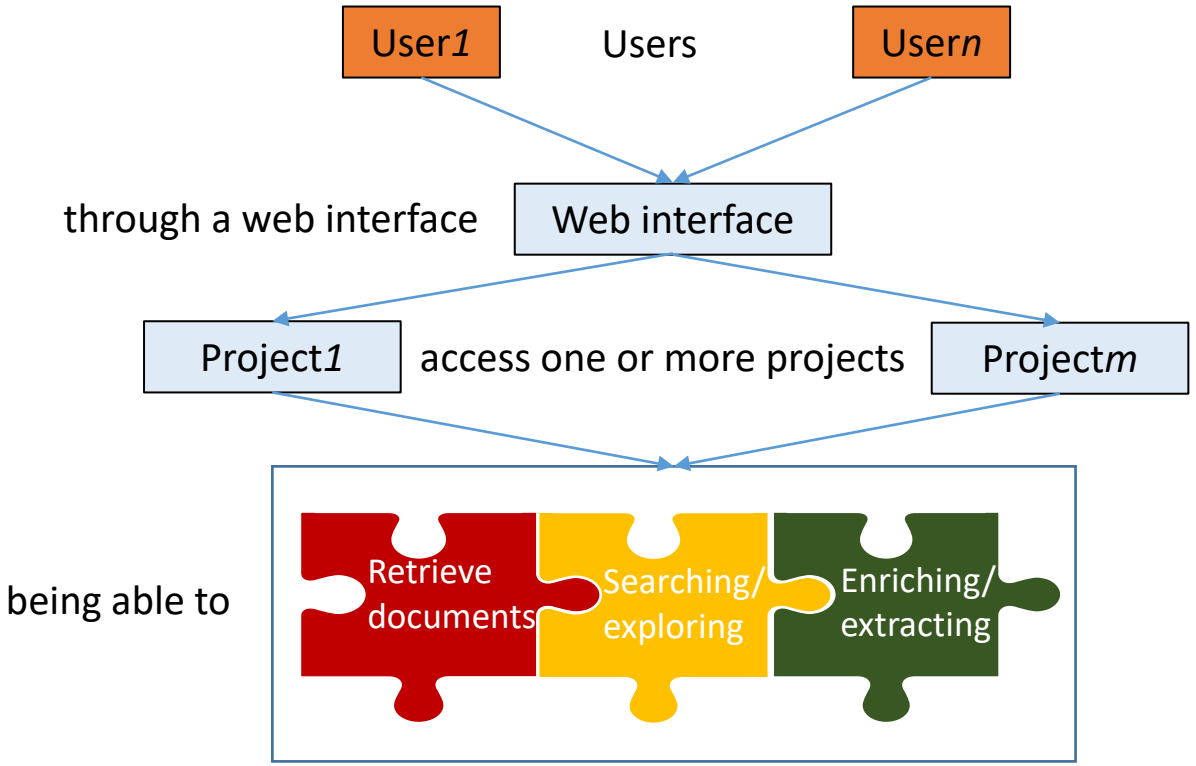
# Index

- ✓What led to developing Essence
- ✓Essence main characteristics (video)
- ✓The main elements:
  - ➢ Retrieve documents
  - ➢ Searching/browsing
  - ➢ Enriching documents
- ✓Technical details
- ✓Final remarks

# What led to the development of Essence

- 2018-2019: *first project-based steps in text analysis*
  - Fill data gaps on *Food Loss and Waste* in a database, i.e. download, select, extract and organise unstructured data from papers
- 2020 (establishment of the Data Lab): *storing and explore*
  - Use of an open source tool (OpenSemanticSearch - OSS) to download and store general texts (e.g. tweets+articles)
  - Produce metrics on the use of statistics (evidence base) in policy documents i.e. text mining and extracting indices on "sparse" scripts
- 2021: *personalization and specific features*
  - New features added to the OSS web interface to meet the Lab needs (manage downloads, execute scripts, etc.)
  - Inclusion of the different applications (scripts) within the same framework
- 2022: *develop an internal framework, Essence*
- 2023: *Use of Essence in real projects*

# Essence characteristics



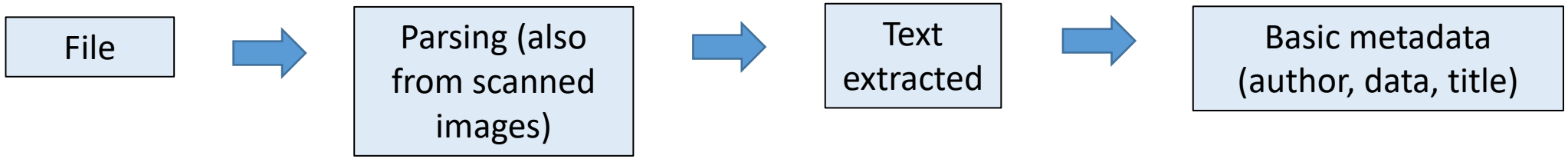

- manage a high volume of textual data
- download documents searched/queried from the web
- automatic identification of relevant information (through models, characteristics of texts, etc.) pattern and metadata extraction
- use of a semantic search engine to navigate the collected information, enriched with filters (facets)
- give a structure to unstructured data
- a framework from which it could be possible to build dashboards based on the extracted data, on specific synthesis of the data, etc.

# Essence by images (and sound)

# Main elements: *retrieve documents*

built-in functions:

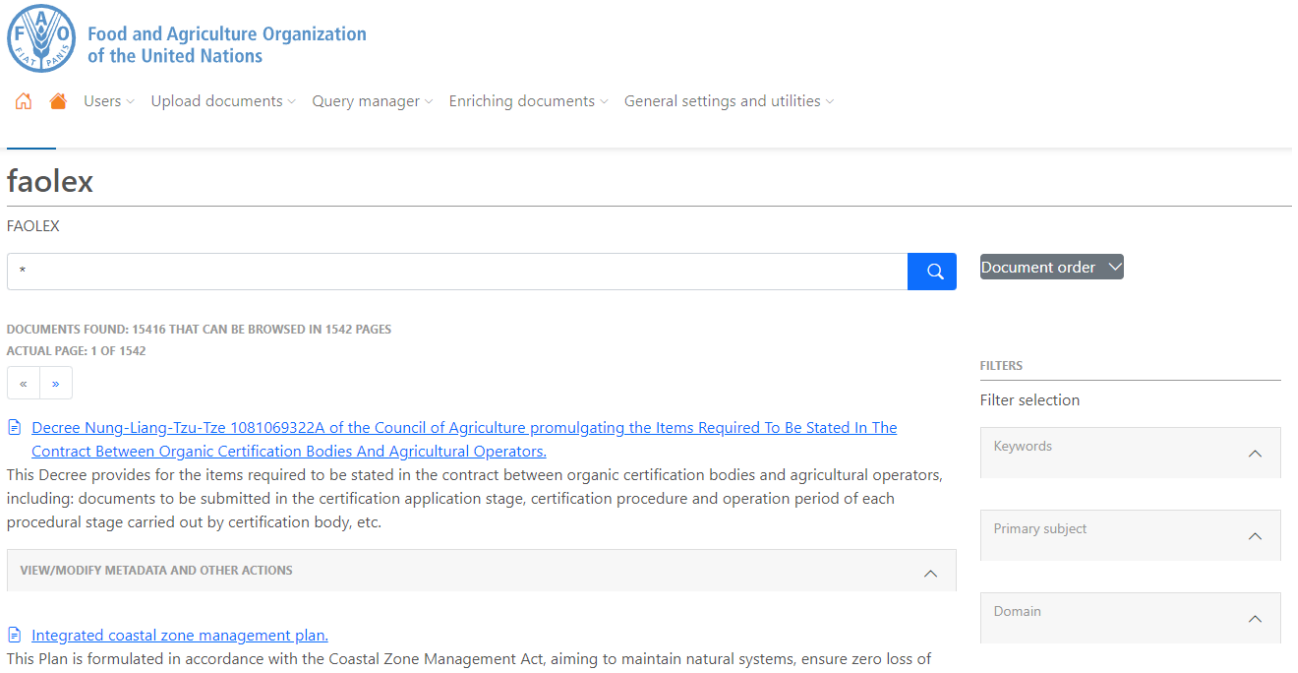
- Automatic interactions with API or search engines (google, FAO repository, etc.)
- Files uploaded by the users (from their local repositories or by a provided list of URLs)



# Main elements: *searching/exploring*

✓Semantic search engine (with an advanced syntax)

✓Filters (facets) on categorized documents



# Main elements: *enriching/extracting*

✓Supervised approaches:

- The system learns what are the relevant documents for the user
- The system learns how to classify automatically a document

✓Unsupervised approaches:

- Integrations of SKOS vocabularies
- Document similarities

✓General utilities:

- Automatic translations of texts into English
- Automatic summaries (thorough a BERT engine)
- Manual (collaborative) enrichment of metadata

# Technical details: *supervised approaches*

- ✓Document relevance:
  - ➢ Logistic regression on the TF-IDF matrix (1,2 ngrams)
  - ➢ Use of combined approaches (keywords+model) as for FLW

- ✓Document classification:
  - ➢ Many multilabel methods derived from the scikit-learn Python package are already included and can be executed (and tested) from the web interface; these are based on:
    - ➢ Lemmas (whose type can be selected by the user)
    - ➢ Ngrams (with variable dimensions)
  - ➢ Possibility to include:
    - ➢ Other methods (to be executed from the web interface)
    - ➢ Externally estimated models

# Technical details: *searching/exploring*

A PHP interface that queries SOLR; where

Solr → … a document database that offers SQL support

In particular Solr:

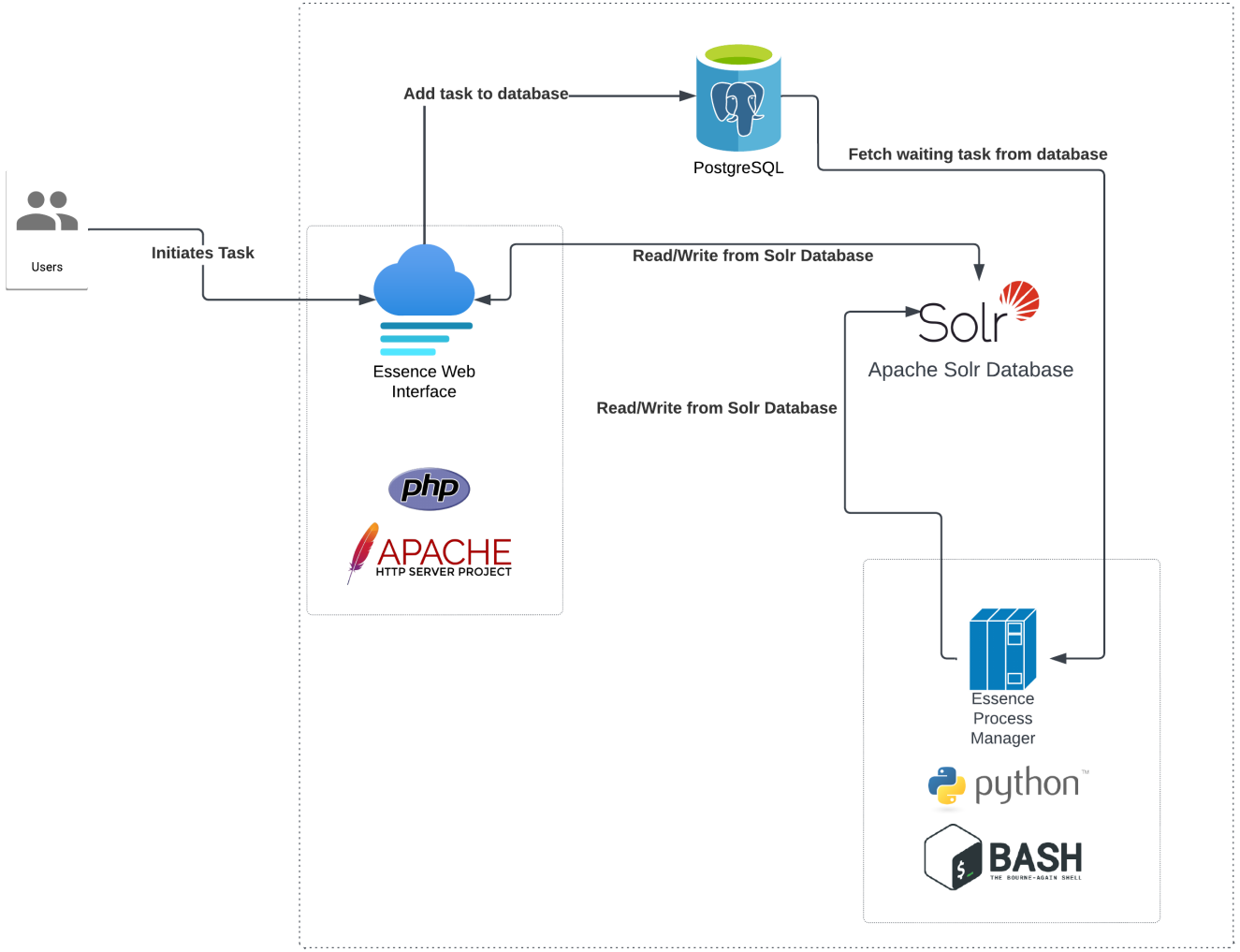
- is a de-normalized bag of documents with fields that aren't necessarily consistent across the collection of documents
- is indexed in a sophisticated way allowing to search across all fields
- permits to search in the stored objects with a powerful (and semantic) language
- could apply automatically many data enrichment/transformation to each newly loaded text

# Technical details: *retrieve documents*

We are able to scrape documents from the following list of sources

- Arxiv
- Asian Development Bank
- Core AC UK
- EBSCO
- FAO Publications
- Google Scholar
- Inter American Development Bank
- International Labour Organization
- International Monetary Fund
- OECD
- Poverty Action
- Relief Web
- RIMISP
- Save The Children
- Unpaywall
- World Bank
- World Food Programme

# Technical details: *general flowchart*



# Final remarks

- ✓Essence is a framework because:
  - ✓it gives us the possibility to link together many sparse «pieces» that were developed by many people in many years
  - ✓it is a place in which we can easily study, test and add new features concerning text analysis

- ✓Essence is also a tool because:
  - ✓it offers services to the team and to external divisions

# *A final… note*

[redacted] (LEGN) | Scarno, Marco (ESS); Matadeen, Craig (ESS); [redacted] 25/05/2023

**Re: Account activated to automatically enrich legal documents**

Cc Matadeen, Craig (ESS); [redacted]

Messaggio inoltrato in data 25/05/2023 17:37.

Dear Marco,

First of all, thank you very much for your efforts. This application looks very useful, promising and time-saving.

I've just submitted a 110 page policy and I got the estimations in 3 minutes. This is incredible! The abstract needs modification or at least a double-check but guessed domains, primary subject, MCK and most of keywords are in the proper place.

I am pretty sure that I am gonna use this application frequently, and let you know if I encounter any errors.

Thanks again!

Best regards,

[redacted]