Eetu Mäkelä
Mikko Tolonen
Risto Turunen
Leo Lahti
Krista Lagus
Maria Valaste
Veronika Laippala
Katri Tegel
Aleksi Kallio
Eetu Mäkelä
Pasi Koste
Erno Liukkonen

WP 4.3 SUBSETTING AND EVALUATING DATA

FIN-CLARIAH

RATIONALE

DARIAH-FI gathers together datasets of interest to a wide community of researchers, but which have not originally been created for research.

Such datasets contain biases, confounders and noise which need to be understood, evaluated and handled when using the data for research. In the case of large datasets, each research use is also often interested in only a subset of the data.

We need to develop tools by which researchers are able to robustly query and examine the large datasets to extract the subsets that cover their particular interest. Further, these tools should make it possible to examine the representativeness and character of both the subset as well as the whole of the data.

DELIVERABLES AND TIMETABLE

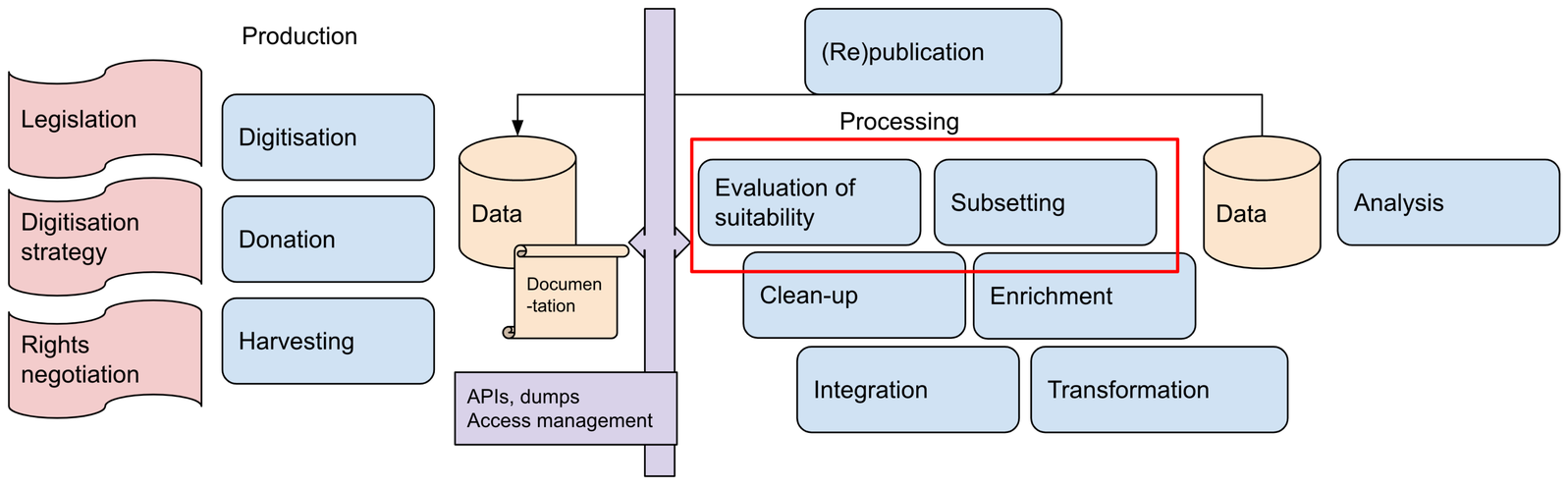
Tools will be developed iteratively, so that first iterations completed within DARIAH-FI will be ready Q3/Q5, after which they will be improved until Q8

Quarter	1	2	3	4	5	6	7	8
D1 Subsetting tool			1					2
D2 Evaluation tool					1			2

CURRENT STATUS

Prototypes for both tools have been developed in prior projects. Work in DARIAH-FI is pending availability of technical personnel.

PLACE IN THE RESEARCH DATA FLOW



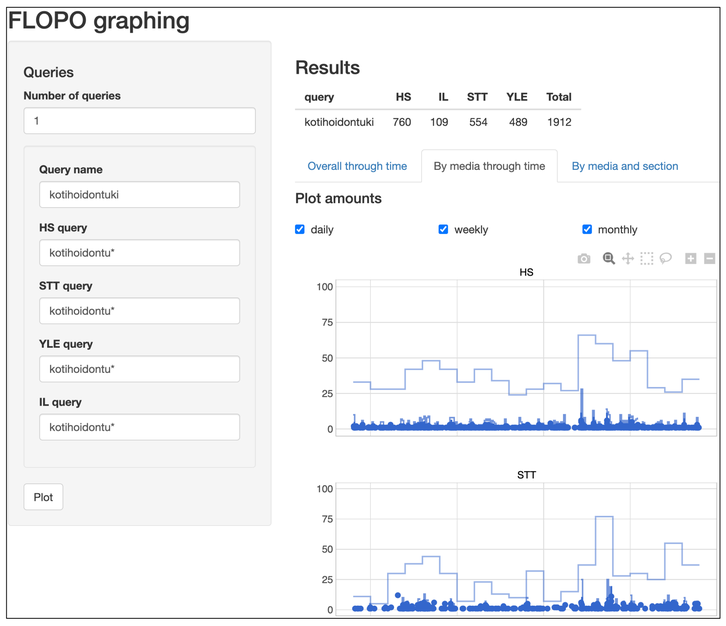
SUBSETTING TOOL

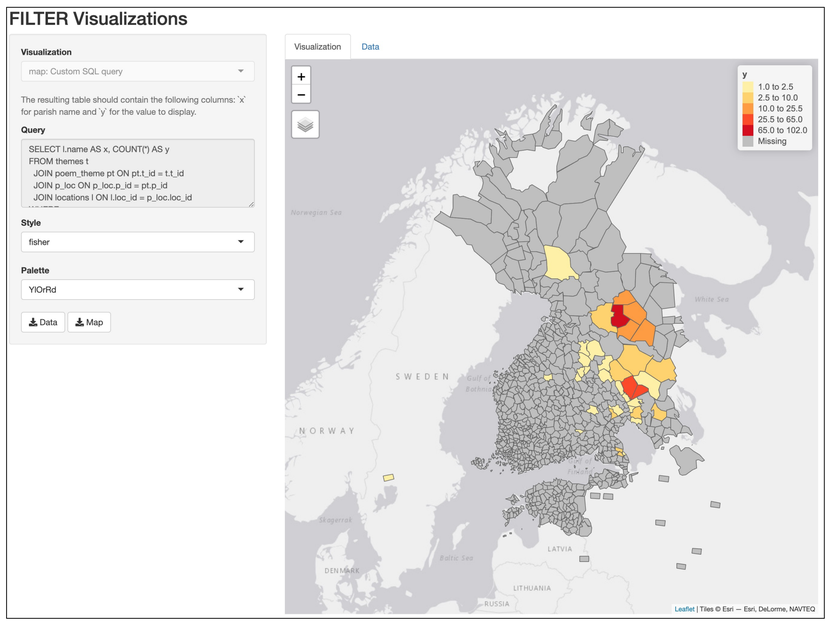
- Octavo tool developed in prior research – custom server built on top of a Lucene index with a REST API
- Validated in many projects with many different kinds of data and use cases:
  - Allows querying by combinations of content and metadata: “Give me all articles published by STT that mention alcohol, but which are not in the foreign department and which are not stub versions.”
  - Deals with noisy content (OCR noise, linguistic variation): “Let me discover and filter the ~200 different forms in which Väinämöinen appears in folk poetry, after which return all poems containing them.”
  - Allows attaching rich metadata also to the content: “Find all sentences where any word classified as an adjective is applied to the lemma phrase ‘Juha Sipilä’.”
  - Can return results on various structural levels (sentences, paragraphs, sections, threads, whole documents) as well as with rich content and document metadata

The primary task in DARIAH-FI WP4.3 will be to create a production version of this tool. Questions include paring down prototype functionality to a core, maintainable subset, as well as discovering whether we can move from the custom Lucene backend to a more standard, scalable and maintainable one (e.g. Elasticsearch). A separate question is whether we also need to expand to support vector-search (e.g. Weaviate), or more robust structural queries (e.g. SQL).

EVALUATION TOOL

Multiple visual subsetting and result set evaluation tools built on top of Octavo have been prototyped in projects. In DARIAH-FI WP4.3, functionalities and best practices from these need to be unified and developed into a production version / base codebase for producing such.





UNIVERSITY OF HELSINKI

DARIAH-FI

KIELIPANKKI The Language Bank of Finland

FIN-CLARIAH