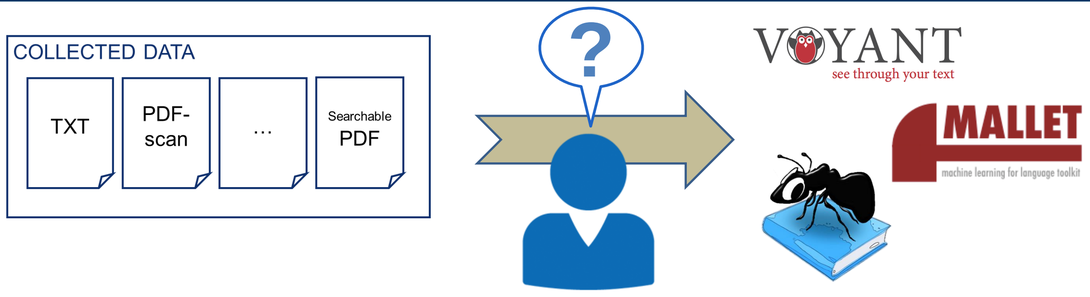
Towards a IIIF-based corpus management platform

Joke Daems, Sally Chambers, Christophe Verbruggen, Tecle Zere

Towards a IIIF-based corpus management platform

"I want to perform digital text analysis"

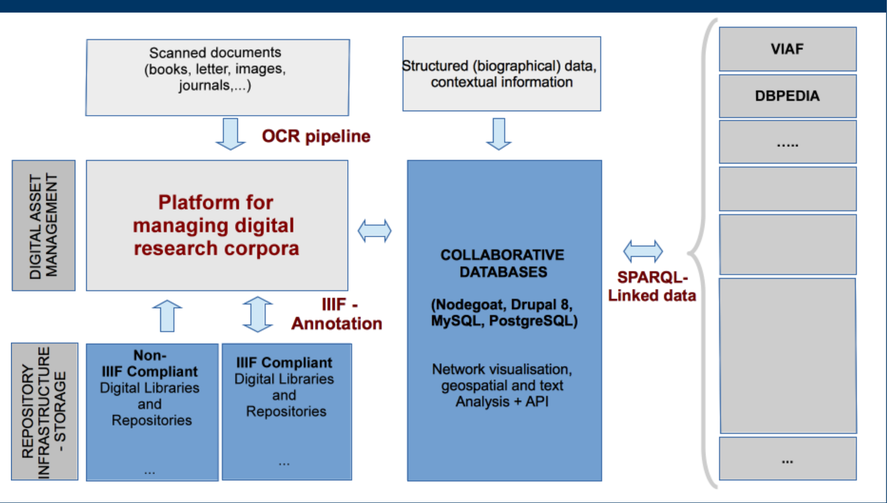


Goals

- Collect data from different possible datastreams
- Generate/extract high-quality textual data
- Search through data to create relevant subcorpus
- Export data for subsequent digital text analysis

The envisioned solution

- Import through various datastreams
- OCR pipeline (ingestion + improvement)
- Collaborative addition of metadata and annotations
- Extend the International Image Interoperability Framework to textual data
  - → interoperable
  - → sharing without exchanging
  - → international standard
  - → multilingual data



Testing existing solutions against key requirements

Legend:
+ this feature is fully implemented in the existing solution
- this feature is not implemented in the solution
± this feature is partially implemented in the solution but needs further fine-tuning according to our specific needs
? = this feature is not currently implemented, but could be implemented after further discussion of our specific needs.

For a description of the different solutions + link to their respective websites, please scan:

	dLab	digirati	4SCIENCE	slandora	PICCL	ResourceSpace	TextGrid	OPEN SEMANTIC SEARCH
1. The platform shall consist of one or more open source solutions	-	±	+	+	+	+	+	+
2. The platform shall consist of a modular infrastructure, enabling plug-ins of external tools and services.	+	+	+	+	+	+	+	+
3. The platform shall enable researchers to import their documents in pdf and txt format and preserve the original format (individual files or batch upload)	+	±	+	+	+	+	+	+
4. The platform shall enable researchers to upload additional versions of the same document in different file formats	+	±	+	+	?	+	+	tbd
5. The platform shall ensure storage, long-term preservation and backup of all imported text corpora.	+	±	+	+	?	+	+	tbd
6. The platform shall enable sound data presentation functionalities, including a user friendly interface with a PDF + IIIF compatible viewer	±	+	+	-	±	±	-	-
7. Researchers can define and build their own collections and sub-collections consisting of documents in the platform. Collections are by default private but can be shared with other researchers or made public to anyone with access to the platform	±	±	±	?	±	±	+	tbd
8. The platform shall enable advanced search functionalities (including full text, metadata, faceted search, etc.) throughout personal collections or public collections.	±	+	+	+	?	+	+	tbd
9. In shared collections, the platform shall enable collaboration between researchers by means of simultaneous improving of textual data, editing of metadata, or adding comments to the documents and collections.	?	±	?	±	?	+	±	tbd
10. The platform shall use the Dublin Core metadata standard for all document metadata.	+	+	+	-	?	+	+	tbd
11. The platform shall include OCR capabilities for uploaded documents as a background process. Users will be notified when pdfs have been processed.	+	±	±	+	+	+	?	tbd
12. The platform shall include ways to manually and automatically improve the OCR output	?	-	?	±	+	?	?	tbd
13. The platform shall enable export of search results, collections, and individual data (metadata and/or documents)	+	±	+	+	?	+	±	tbd
14. The platform shall enable export of files in the original format and metadata as CSV and Dublin Core.	±	±	+	-	+	+	?	tbd
15. The platform shall include security and user management functionalities: CAS + DARIAH-EU	+	±	+	±	?	+	+	tbd
16. The platform will be compatible with different languages and writing systems (taking into account regional and diachronic variation)	+	±	+	+	+	+	±	tbd
17. The platform will be compatible with the current ICT infrastructure at Ghent University.	+	±	+	+	?	+	+	tbd

Next steps

- Further evaluation of existing solutions (input welcome!)
- Collaboration with other institutions (e.g.: Huygens ING)
- Development of first version towards the end of 2017
- Testing the platform on data of running pilot projects
- Further development & customisation of the platform

Daems, J., Chambers, S., Verbruggen, C., and Zere, T. (2017) *Development of a IIIF-based digital corpus management and text analysis platform* - a poster for *Digital Humanities Benelux 2017*, Utrecht, The Netherlands 3-5 July 2017.

Sharing digital arts and humanities knowledge

http://be.dariah.eu | @DARIAHBe | http://www.ghentcdh.ugent.be/ | @GhentCDH