Towards a IIIF-based corpus management platform
Joke Daems, Sally Chambers, Christophe Verbruggen, Tecle Zere

To cite this version:

HAL Id: hal-01571618
https://hal.archives-ouvertes.fr/hal-01571618
Submitted on 3 Aug 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License
Towards a IIIF-based corpus management platform

“Towards a IIIF-based corpus management platform”

<table>
<thead>
<tr>
<th>COLLECTED DATA</th>
<th>Goals</th>
</tr>
</thead>
<tbody>
<tr>
<td>TXT</td>
<td>- Collect data from different possible datastreams</td>
</tr>
<tr>
<td>PDF</td>
<td>- Generate/extract high-quality textual data</td>
</tr>
<tr>
<td>scan</td>
<td>- Search through data to create relevant subcorpus</td>
</tr>
<tr>
<td></td>
<td>- Export data for subsequent digital text analysis</td>
</tr>
</tbody>
</table>

**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 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 the QR code.

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 (individual files or batch upload)
5. The platform shall ensure storage, long-term preservation and backup of their input text corpora
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
8. The platform shall enable advanced search functionalities (including full text, metadata, faceted search, etc.) throughout personal collections or public collections.
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.
10. The platform shall use the Dublin Core metadata standard for all document metadata.
11. The platform shall include OCR capabilities for uploaded documents as a background process. Users will be notified when pdfs have been processed.
12. The platform shall include ways to manually and automatically improve the OCR output
13. The platform shall enable export of search results, collections, and individual data (metadata and/or documents)
14. The platform shall enable export of files in the original format and metadata as CSV and Dublin Core.
15. The platform shall include security and user management functionalities: CAS + DARIAH-EU
16. The platform will be compatible with different languages and writing systems (taking into account regional and diachronic variation)
17. The platform will be compatible with the current ICT infrastructure at Ghent University

**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

**Sharing digital arts and humanities knowledge**

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