DH2019 Lunch session - Researchers & Libraries working together on improving digitised newspapers

Please consider this document as a starting point and feel absolutely free to add, change and modify to your needs, questions or other ideas you have.

Background

During the DH2019 conference several presenters discussed their wishlist on digitized newspapers and other heritage sources. With the presence of both researchers as well as collection holders, this would be the perfect starting point to 1) jointly work on a common wishlist, and 2) explore possible collaborations and discussion who would take up these requirements.

Some links

- https://periodica.github.io which contains links to
    - https://groups.google.com/d/forum/digital-historical-periodica/ - joining this group is the easiest way to get everyone together to continue the conversation
    - https://www.zotero.org/groups/704613 - share articles you'd like others to know about or discuss
- DH2019 for Historical Newspaper Geeks - an overview of sessions at the Digital Humanities 2019 conference
- There was some discussion of IIIF for newspaper collections - see https://github.com/IIIF/iiif-stories/issues?q=is%3Aopen+is%3Aissue+label%3Anewspapers and https://iiif.io/community/groups/newspapers/
- Tim Sherratts GLAM Workbench https://glam-workbench.github.io/
- Europeana Newspapers http://www.europeana-newspapers.eu/

Possible topics / actions

- Mapping of pixel size of newspaper image to physical size using catalogue data (linked to presentation by Eetu Mäkelä)
- Estimation of OCR quality
- Process description of digitisation projects
- Choices behind digitisation, what is taken into digitisation, possible policies
- How to access the data through other means than Interfaces
- Calls focused on really specific topics
- Allowing researchers to access/download .txt format of OCR in bulk
- Sharing ad hoc scripts with researchers (like Chronicling America for instance)

Next steps

- Format as used within Impresso project for joined telcos/Skype calls to discuss 1 topic in 1 hour (suggestion Maud)
- Joined paper at Conference 'Digitised newspapers - a new Eldorado for historians?' 23-24 April 2020, Lausanne https://impresso-project.ch/news/2019/06/12/WS5-CfP.html ?
- Panel / Workshop DH2020 ?
- Consider adding implementation of wishlist items (if reasonably implementable) ;)
- Manifestos from the researcher and GLAM perspective?

DH2019 Papers related to this topic

References

Ames (2019) *Digital Scholarship and Data Provenance at the National Library of Scotland.* https://zenodo.org/record/3269291#

Claeyssens (2016) *The Ideal Corpus. Towards a Critique of Large Digital Libraries from a Digital Humanities Perspective*. DHBenelux http://www.dhbenelux.org/wp-content/uploads/2016/05/65_Claeyssens_FinalAbstract_DHBenelux-2016_short.pdf

Prescott, Hughes (2018) *Why Do We Digitize? The Case for Slow Digitization.* https://www.archivejournal.net/essays/why-do-we-digitize-the-case-for-slow-digitization/

Prescott (2018) *Searching for Dr. Johnson: The Digitisation of the Burney Newspaper Collection* https://brill.com/view/book/edcoll/9789004362871/B9789004362871_006.xml

Participants

Name	Institute	Contactdetails
Jussi-Pekka Hakkarainen	National Library of Finland	jussi-pekka.hakkarainen@helsinki.fi
Clifford Wulfman	Princeton University	cwulfman@princeton.edu
Jani Marjanen	University of Helsinki	jani.marjanen@helsinki.fi
Jana Keck		

Maud Ehrmann	EPFL (CH)	maud.ehrmann@epfl.ch
Ina Serif	University of Basel	ina.serif@unibas.ch
Melodee Beals	Loughborough (UK)	M.h.beals@lboro.ac.uk
Lorella Viola	Utrecht University (NL)	l.viola@uu.nl
Torsten Roeder	Leopoldina (DE)	torsten.roeder@leopoldina.org
Estelle Bunout	C2DH (LU)	estelle.bunout@uni.lu
Steven Claeyssens	National Library of the Netherlands (KB)	steven.claeyssens@kb.nl
Martijn Kleppe	National Library of the Netherlands (KB)	martijn.kleppe@kb.nl
Mila Oiva		
Mia Ridge	British Library / Living with Machines	mia.ridge@bl.uk
Daniel Wilson, Tim Hobson, Daniel Van Strien, David Beavan	Living with Machines	
Sinai Rusinek	OMILab, Israel/Hameorer	sinai.rusinek@
Neil Fitzgerald, Rossitza Atanassova		neil.fitzgerald@bl.uk rossitza.atanassova@bl.uk
Sarah Ames (participating remotely!)	National Library of Scotland	sarah.ames@nls.uk

Saskia Scheltjens	Rijksmuseum	
Lotte Wilms	KB National Library of the Netherlands/ LIBER Digital Humanities Working Group	lotte.wilms@kb.nl
Mikko Tolonen	University of Helsinki / Helsinki Computational History Group / NewsEye project	mikko.tolonen@helsinki.fi
Nanette Rissler-Pipka (joining remotely - on my way home :-)	Karlsruhe Institute of Technology / University of Siegen	nanette.rissler@gmail.com
Hannu Salmi	University of Turku / Oceanic Exchanges project / COMHIS project	hansalmi@utu.fi
Meghan Ferriter (remote participant)	Library of Congress	mefe@loc.gov
Peeter Tinits (missed the meeting but interested in followups)	University of Tartu / National Library of Estonia	peeter.tinits@gmail.com
Matthias Arnold (missed the meeting but interested in follow-ups)	University of Heidelberg	arnold@uni-hd.de
Tuula Pääkkönen (missed the meeting, interested in follow-ups)	National Library of Finland	tuula.paakkonen@helsinki.fi
Hui Li	Shanghai Library	lhjulie@gmail.com
Alba Irollo (missed the meeting but *strongly* interested in follow-ups)	Europeana Foundation	alba.irollo@europeana.eu

Digital source criticism

Tool criticism
Living with Machines project -- survey
Impresso

Working group of periodicals in the German DH: https://dhd-ag-zz.github.io/

https://www.newseye.eu/blog/news/the-newseye-case-studies-a-first-dig-into-the-newspaper-corpuses/

Tim Sherrat Jupyter Notebooks for using Trove newspaper collections

Big Questions: (FEEL FREE TO ADD)

How, practically and pragmatically, to collaborate?

Impresso community call: format where the technical expert explains how the data was processed, and the users, researchers can ask their questions directly, during lunch. They are willing to open it up in some way for others to participate

Are there common standards, frameworks we should be using/aware of?

Prioritization of wish list items. Be aware of € & ⏰ constraints.

Impacting funders (digitisation, post-processing, enrichments) and what it could do to serve researchers.

Copyright topics (what , to whom, how. DSM and TDM possibilities from there)

Operationalisation: how to transpose the humanists' research question in the digital libraries context, discuss on how to best formulate the general research question to make the most of the digitised source material

Mediation role of the libraries: what is in the data, etc.

Multilanguage tools (or at least some documentation if a tool could be applied in another lang).

Libraries as partners, not data partners

Diversity of research needs

Does the digital humanities research need to be reproducible?

How to reintegrate the research projects output into the institutional digital collections?

What are the roles?

Common methodologies?

data critique, corpus critique

How does GLAM infrastructure need to change to support DH and data science methods? What kinds of data could we exchange that'd make digitised newspapers and periodicals easier for other researchers to find and use? How porous does the infrastructure need to be?

Linked open data ontologies, interoperability
Linked open data identifiers

Resistance to ontology building in the 'fit everything in' sense

When and how to collaborate on standards - how much work does a project need to do to understand their own requirements before they consider other needs and use cases?

What's the minimum viable shared standard that lets people meaningfully share but also allows them to do specialist things in their own projects?

OCR, article extraction, software options/accuracy

Two manifestos:

1. From researchers
2. From “institutional collaborators” (not merely providers)

Why not simply use Impresso for everything?
How to do it best?

Who has to do what? Do the libraries need to take over all the research outcomes, store the research projects output? what has the academic world to carry: for instance sharing their processed data, e.g. lemmatised corpora of newspaper articles

where are the university libraries?

audience of the libraries? researchers vs. wider audience ? how complex can the interfaces be?

how to make that exchange more fluid between researchers and libraries (and others)?

Saskia Scheltjens (Rijksmuseum) pointed to Dutch initiative of heritage institutes (Network Digital Heritage) that aims to collaborate and link digital collections, see https://www.netwerkdigitaalerfgoed.nl/en/

Tim Sherratt at TROVE: http://timsherratt.org/ see the GLAM workbench he created at https://glam-workbench.github.io/ more on this in this nice 8 minute lecture https://www.youtube.com/watch?v=ONnxd-1KJFI

IIIF Newspaper Community Group https://iiif.io/community/groups/newspapers/

What are the possible institutional roles in projects

Include GLAMs on grant applications to make sure there's hard money to make things happen!

Don't call GLAMs 'data providers'

Libraries shouldn't build shiny interfaces that aren't needed. Researchers need access to the data, to APIs

Why are there so few user studies available for catalogues and newspaper interfaces? Can we share any that we have?

Do we want a SIG?