
Let's Develop an Infrastructure for Historical Research Tools

Julia Luise Damerow

jdamerow@asu.edu

Arizona State University, United States of America

Dirk Wintergrün

dwinter@mpiwg-berlin.mpg.de

Max Planck Institute for the History of Science
Germany

Robert Casties

casties@mpiwg-berlin.mpg.de

Max Planck Institute for the History of Science
Germany

Scholars conducting historical research are provided with a growing range of digital humanities tools, supporting different phases of the research process: there is software for extracting text from documents (such as [pdftotex](#), available on many Linux distributions or as part of [Poppler](#)), run OCR processes on images (for example [Tesseract](#)), tools for the creation, analysis, and visualization of datasets (for instance [Nodegoat](#), [Palladio](#), or [Visualeyes](#)), or software to work with annotations (for example [Annotation Studio](#)) or networks (such as [Gephi](#) or [Cytoscape](#)). Programming libraries are being developed to serve the needs of humanity scholars, like [Spacy](#) or [Tethne](#). There are several repositories (such as [HathiTrust](#) or the [Europeana](#)) that provide access to sources and can easily be integrated into other services through APIs. Many tools, however, work well as self-contained units that scholars can use as singular parts of their research process, but cannot easily be combined into an integrated workflow by the researcher. Existing and new tools are developed using different languages and programming frameworks depending on requirements, skillset, and preference of the original developer, making reuse and integration harder for the developer seeking to combine several tools. Moreover, since most tools are developed independently of each other,

many efforts are repeated by reimplementing functionality that is already provided by a different piece of software.

In this workshop, we would like to gather developers and programming-literate scholars to share their tool-building experiences and to present our first practical steps to create a system integrating multiple tools to work with historical documents from scan to analysis. The workshop is intended as a starting point for future exchange and cooperation for digital humanities developers.

In the summer of 2016, the Digital Innovation Group at Arizona State University (ASU) and the Max Planck Institute for the History of Science (MPIWG) started to combine their efforts in developing software for the history of science. One outcome of this collaboration is a research system that allows users to manage their documents, automatically runs OCR on uploaded files, provides an image viewer for uploaded and extracted images, and integrates document management with a multi-user Jupyter notebook server for writing analysis and visualization scripts. Rather than one big system, however, the research system is comprised of several integrated services developed independently of each other using different programming languages and frameworks.

For the Digital Humanities Conference 2017, we propose a full-day workshop with the goal to connect different tools and services to build a tool infrastructure for historical research.

The first half of the workshop will give tool developers a chance to present their software. Every presenter will be allowed 10 minutes for their presentation and 5 minutes for questions. ASU and MPIWG will present the different components of the developed research system. Specifically, we will present the following projects:

- **Collaborative Jupyter Notebooks**: a Jupyter Notebook server that allows sharing and publishing of notebooks based on Nextcloud and Dataverse.
- **DocuManager**: an environment for annotating, correcting and searching digital documents, in particular for OCR text in ALTO-XML and HOER.
- **Giles Ecosystem**: an Apache Kafka-based service to extract images and texts from documents and run OCR procedures on them.
- **Digilib**: a Java-based IIIF-compliant image server and viewer.

The second half will be dedicated to discussing how different tools can be connected and integrated, and how we can build a community around those tools.

We envision the results of this workshop to be a concrete roadmap of how different tools will be integrated. We will define interfaces and API requirements, and if possible start development work during the workshop. Second, we will develop an organizational strategy for cooperation and collaboration among different projects. To aid organization, we will provide a Jira and Confluence project that participants can use during and after the workshop to organize collaboration.

We plan on organizing a follow-up meeting at the end of 2017 at Arizona State University to review progress since the initial workshop and plan next steps. If the collaboration is successful, we hope to establish regular meetings and expand the group to connect more tools and services.

Participants/Call

We will send out a call for participation in form of a short tool presentation for the first part of the workshop. We will ask presenters to focus on the technical perspective of their tool answering the following key questions:

- What is the general workflow of the tool?
- What is the core functionality of the tool?
- What input and output formats does the tool accept? Or what interfaces does it expose?
- What license model was chosen?
- What features are still missing and what are the next development goals?
- How is maintenance and development of the tool organized?

The deadline for the call is July 1st, 2017. We plan to accept 5-10 submission for our call, based on the usefulness of the tool and the potential for integration with other tools, which would all fit the first half of the workshop.

Audience

The target audience for this workshop are developers, historians with programming background, scholars with a technical background, and generally people involved in the development of tools to support historical research.

Confirmed Presenters

- Dirk Wintergrün (Max Planck Institute for the History of Science, Germany)
- Julia Damerow (Arizona State University, USA)
- Robert Casties (Max Planck Institute for the History of Science, Germany)
- Malte Vogl (Max Planck Institute for the History of Science, Germany)