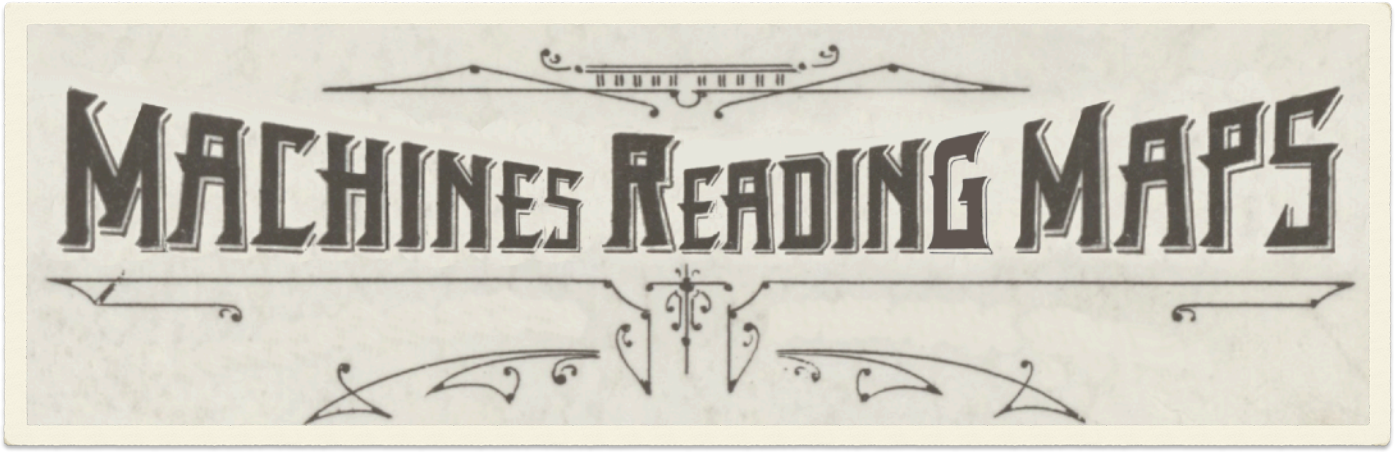


Project Newsletter

For the members of the Advisory Board

Dear Members of the Advisory Board,

This is our second newsletter on the progress of the Machines Reading Maps project. We know that it is long overdue, and we thank you for your patience. We have a lot of exciting news to share with you, and we look forward to hearing your opinions and suggestions in this final phase of the project.

Last AB Meeting

First, a practical matter. We would be delighted to have one last advisory board meeting to hear your feedback on what we have achieved so far, and your suggestions on how to make our outputs as useful and sustainable as possible. We know it's always complicated to find a time that is convenient for people across different time zones, so we thought that it would be good to start talking about dates as soon as possible. To that end, please indicate your availability on this doodle.

Executive Summary

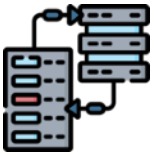
1. The *mapKurator* pipeline is being used to process samples of maps from the NLS and LC. It will be fully documented by spring 2023 in our Github repo.
2. Our bespoke 'Recogito for maps' annotation interface is integrated with mapKurator so that users can generate bounding boxes and text transcriptions on a map-by-map basis.
3. Work with our cultural heritage partners as well as new connections to colleagues who are re-using our tools continues as planned.
    - We use NLS OS maps for a case study in our main article about the methods and outcomes of *Machines Reading Maps*. We also worked closely with the NLS on a crowdsourcing experiment to annotate text on nineteenth-century maps of Edinburgh.

- We are working with the BL to evaluate the quality of *mapKurator* output and its uses in library systems.
- Work continues to create a dataset from 10% of the Sanborn collection at the LC, which will be deposited directly with the Library and documented following the principles developed during Computing Cultural Heritage in the Cloud.
- With the USC Digital Library, we have been working with student research assistants to create gold standard data for text on samples of the Sanborn and OS collections.
- There are a range of new collaborations that re-use tools and methods developed on the project, in particular we note the major collaboration with the *David Rumsey Historical Map Collection* (details below).

4. Over the last year, we have contributed to a number of conferences and workshops around the world, both in person and virtually. We have a range of outputs - from tool documentation to research articles that will be completed between this winter and next summer.
5. There have been changes to the team: Valeria and Katie accepted new permanent academic positions in the UK, and Rainer has become a freelance consultant on Digital Humanities projects.
6. Final steps focus on publishing data and articles about our work and organizing final events with project partners.

Screenshot of a Sanborn Map of Los Angeles annotated manually by the students at USC as part of our gold standard

1. Using mapKurator



First, we are very pleased to report that the project's machine learning pipeline (*mapKurator*) has been improved and optimised by the team at UMN, and we are now able to process digitised maps at scale detecting and identifying text on maps and transforming it in what we call *maptext*, i.e. the textual information that appears on map (as place names or part of indexes and legendas) converted into searchable (and likeable) structured data.

Currently, there are two types of output that are available from the *mapKurator* pipeline, depending whether the input has been georeferenced or not.

A) For collections where digital maps have sheet-level metadata and are georeferenced (like some OS maps at the NLS), *mapKurator* produces:

- Bounding polygons with pixel and geo-coordinates
- Predicted transcriptions
- Optional: links named entities to OpenStreetMap or another knowledge base of choice

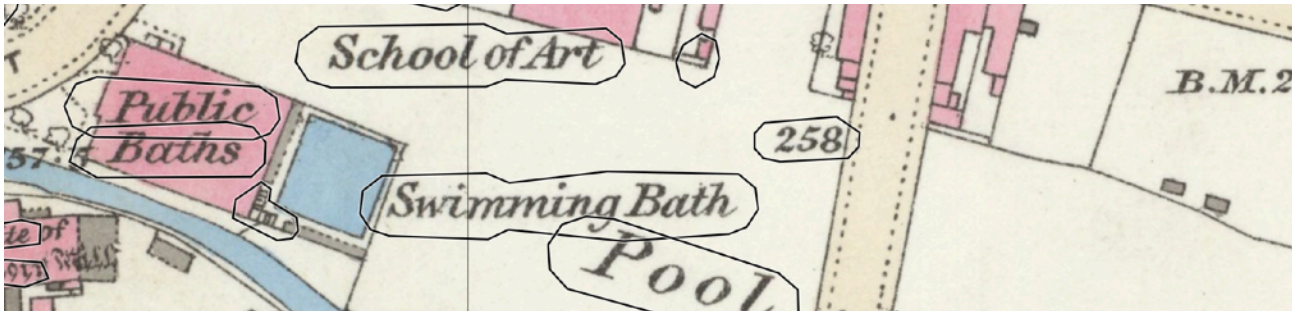
We know that the relationship between labels and place coordinates is not always unproblematic, and that the placement of a label may not be the most accurate reference for a place's position. We also are so far not able to automatically link words in phrases together. However, we also think that this is a wonderful achievement that creates an unprecedented rich and interesting textual corpus, and adds considerable value to the maps, making them more searchable but also easy to study and analyse.

B) Working with non-georeferenced maps, on the other hand, meant dealing with limitations but also addressing new opportunities. When processing non-georeferenced sheets, *mapKurator* produces only pixel, not geographic, coordinates for each label's bounding polygon. However, leveraging the proximity between labels that could be automatically linked to OpenStreetMap data using *mapKurator*'s entity linking module, the output includes a predicted, centre point for each of the sheets. This rough geo-location of the map was tested on 100 maps and evaluated against manually geo-referenced maps prepared by the LC maps team (see just below for details). To sum up, in the case of non-georeferenced collections, *mapKurator* output includes:

- Bounding polygons for all text on maps
- Predicted transcriptions
- Pixel coordinates for each label
- Predicted map centre point

Producing even an approximate geo-location for a large number of digitised maps is an incredible opportunity to improve map metadata for our library partners, and make maps held in collections easier to find.

MapKurator is currently available in a public Github repository, and complete documentation for the library will be available by next spring: https://github.com/machines-reading-maps/mapkurator-system



Output of *mapKurator* on a OS 25 inch map of Coventry.

2. New version of the annotation interface prototype: 'Recogito for maps'

The integrated version of *Recogito* and *mapKurator* that we developed for this project has also evolved since our last AB meeting. Rainer has added several new functionalities, ranging from new tools (including ordered groups and circular selection) to new visualisation modes (including visualising the annotations by "checked or unchecked", "grouped or not", label type "category", and "editing progress").

Recently, the version of *mapKurator* in the prototype has also been updated, and it now offers not only the automatically generated bounding boxes, but also predicted transcriptions.

You can view an NLS OS map processed using the in-browser Recogito *mapKurator* feature here: http://recogito.uksouth.cloudapp.azure.com/document/l9ma03duhcw0zt
NB: the current version does not yet produce transcriptions, but this feature will be live shortly.

Our online annotation platform was crafted initially to allow us to produce gold standard data, but it has proven to be extremely successful during our workshops, and has shown great potential for the critical manual and hybrid machine-human annotation of maps. We have nearly completed an article that focuses on this aspect of our work on theorizing and practicing map annotation, which will be submitted to *Historical Methods* later this spring.

Code for this bespoke version of Recogito is openly available on Github: https://github.com/machines-reading-maps/mrm-recogito-ui. This interface will be hosted on a VM at the Turing until November 2023. We are exploring other options for making this version of *Recogito* a publicly available tool, which we hope to discuss with you in March.

Output of *mapKurator* on a OS 25 inch map of Coventry.

3. Collaborations with Cultural Heritage Partners and New Connections

During the course of the project, we have worked with our partners at the NLS, BL, and LC to refine the scope of our work using *mapKurator* to automatically process maps and improving the functionality of the *Recogito* platform.

In early spring 2022, the NLS invited us to be part of their very successful series of events for volunteers around historical maps.

National Library of Scotland

Our manual annotation interface was used by volunteers to annotate labels of an historical map of the city of Edinburgh. This event generated 21,950 annotations from approximately 70 volunteers, completing the task of transcribing the entire city map in less than a week. The Alan Turing Institute featured the project and this event in a blog post last March.

This event pushed us to add new functionalities to the manual interface (such as the different visualisation options) and to produce further documentation (targeting a non-academic audience of annotators). It was also a terrific opportunity to experiment with the use of controlled tags, and to get a direct insight on what users thought of our tools, especially thanks to the interactions on the dedicated forum that was moderated by the NLS with our help. The data collected during the event will become a useful tool for the training and evaluation of machine-generated annotations, and is already the cornerstone of a new historical gazetteer of Edinburgh. The data is documented and available to download openly here.

Next, for our case study using maptext as research data, we are using NLS maps to examine the representation of 'antiquities' on 25-inch maps of four counties, two in England and two in Scotland. The data collected through *mapKurator* will be linked, where possible, with modern-day heritage data to allow us to explore change over time in the documentation of historical sites, in particular those classified as 'Roman', between the mid- and late-nineteenth century and today. We are currently applying for further funding to expand this small experiment to a larger corpus, with a deeper investigation of how Victorian archaeological and historical methods shaped map content, and what we can learn about changing approaches to periodizing the landscape.

We are writing up this case study alongside the LC work and our overall contributions from this project in an article we plan to submit this summer to *Digital Scholarship in the Humanities (DSH)*.

British Library

While we have not been able to work directly with the map collections at the BL that we had originally planned, we are delighted that the collaboration with David Rumsey (see below) has made it possible to explore a) the quality of *mapKurator* outputs for a small selection of early modern maps and b) to open up conversations with BL colleagues about ways that maptext can play a role in library infrastructure in order to improve map discovery. We are actively planning activities around these two goals for spring 2023.

Library of Congress

As part of our collaboration with the Library of Congress, we tested the automatic georeferencing method developed by Zekun, Jina, and Yao-Yi on 100 maps from the newly available, yet not georeferenced Sanborn collection. To evaluate the results, LC staff manually georeferenced 1005 maps.

Thanks to this information, we are able to assess the performance of map*Kurator*. We observed that for most of the maps, the predicted map center is within 2.5 km from the ground-truth map center. The errors mainly came from street name changes over time, uncertain name placement for linear geo-features (e.g. streets), multiple occurrences of similar names, and imperfect prediction from the upstream text spotter. Although the results are far from perfect, they are also completely machine-generated, and could become a useful starting point for further georeferencing in the future. Approximate map centre is now part of the standard output that *mapKurator* offers when processing non georeferenced maps. More details about this work can be found in the slides we presented at the LC.We are now processing about 10% of the very large corpus of the Sanborn Fire Insurance Map Collection held at the Library of Congress, including all of California, New Orleans, and a few other major metropolitan areas as well as a random sample.

USC Library

It has been a pleasure to work closely with staff and students at USC's Digital Library to help prepare gold standard data for the LC Sanborn and NLS OS maps used in project experiments. We aim to annotate the text on 55 Sanborn and 13 OS maps. Working with student research assistants, even across an ocean, has been a great experience. We use Github Discussions to introduce students to map annotation and troubleshoot issues that come up along the way. One major output from this piece of our work has been the detailed annotation guidelines that combine expert knowledge in digital annotation best practices, humanistic approaches to data curation, and the needs of computer scientists.

New Collaborations

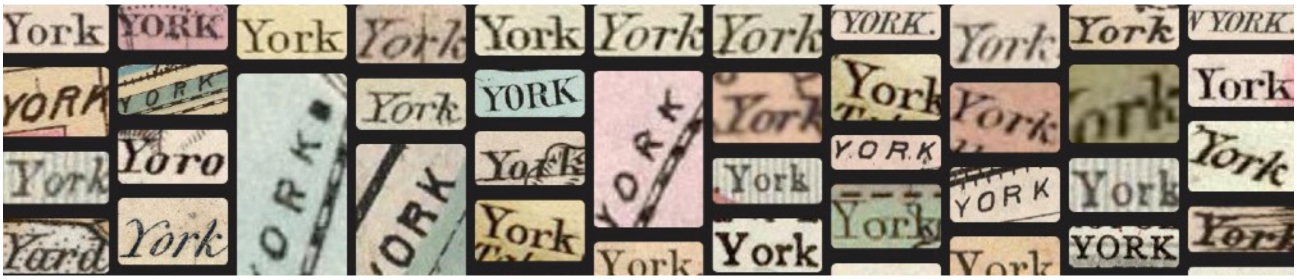
This year has been brightened by many interesting collaborations with other projects or institutions that have expanded and enriched our project, showing us new potential applications for our research.

Our largest collaboration has surely been the one with David Rumsey. After reading about our project, the owner of the large and beloved online map library contacted us, and expressed an interest in applying some of our methods to his vast and diverse collection (in terms of mapping traditions and languages of the labels).

We agreed enthusiastically, and saw this new collaboration as an excellent way to test our tools and methods and make them more robust. We also wanted to keep our work, and the admin, for the main MRM project separate, so both the US and UK sides of the project asked for (and obtained) a no cost extension for the main project, so that we could keep working, in parallel, on both fronts. The project has been extended to different dates in the US (October 31, 2023) and UK (March 31, 2023).

For this spin off of the project we received additional funding, and had the wonderful opportunity to collaborate with the Luna Imaging team in order to integrate the machine-generated annotations with the DR search and Rainer's annotation interface. We think that the results are stunning, and are able to show, in practice, what it means to be able to search about 60.000 maps by their text. We look forward to presenting the outcome of this collaboration in April, at a conference and workshop held in Stanford, in collaboration with Stanford Libraries.

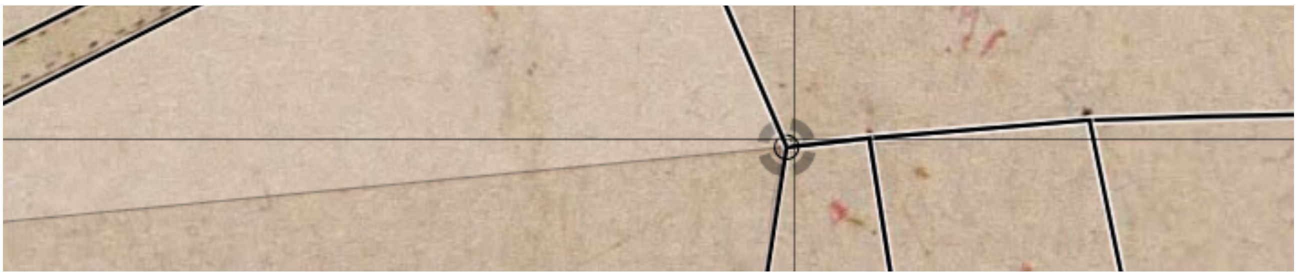
Detail of the search results for the word "York" in the Rumsey Map Collection. Visualisation by Luna Imaging.



The 'Recogito for maps' annotation platform has also been an excellent way to get in touch with other projects that were interested in experimenting with our tools. The new opportunity to also generate predicted transcriptions one sheet at a time will surely open new doors, and generate even more interest in possible collaborations.   There is already a growing community of people using the interface experimentally:

1. After the 2021 Linked Pasts workshop, we were contacted by a volunteer-led project, the Old Leith. Its members asked to use our map-specific version of Recogito to do collaborative annotations of historical maps of the Leith area. This collaboration was the first test for our platform, and was extremely useful in spotting early bugs and scale issues.
2. IN-ROME, a recently started ERC project led by Barbara Borg at SNS Pisa. The project will use the annotation interface to extract vector boundaries from maps of the Catasto Gregoriano. To make the process of tracing polygons more efficient, IN-ROME will contribute additional features (such as automatic snapping of the cursor to neighbouring annotation corners) to Recogito's existing drawing tools.
3. Living with Machines spin-off work linking street names on maps to OS Open Roads vector data and UK microcensus data about residential streets.

Detail of a screenshot showing one of the new Recogito tools developed for the IN-ROME project.



4. Conferences, travel, and other outputs

We have been very busy in the past year presenting our project and sharing the enthusiasm about our achievements. We have been invited to several conferences, in different fields, and we have always received very encouraging feedback. The academic venues where we have presented include (but are not limited to):

- Datafication in the Historical Humanities, German Historical Institute June 2022 (Washington, DC/virtual)
- Digital Humanities conference, July 2022 (Tokyo, Japan/virtual)
- Spatial Humanities, September 2022 (Ghent, Belgium/virtual)
- ArcheoFoss 2022 (Rome, Italy)
- Digital Classicist 2022 (London, UK)
- OBTIC Seminar, June 2022 (Paris, FR)
- Linked Pasts, December 2022 (York, UK)
- AEOLIAN Network Final Workshop Keynote, February 2023 (virtual)
- Voltaire Foundation and Maison française d'Oxford, June 2023 (Oxford, UK)

But the highlight of the year, in terms of communication, has surely been the opportunity for the team to meet in person (for the first time!), and present our work to our partners in Washington DC and in Los Angeles. Last summer, the European half of the project flew over to meet the US team. Together, we were honoured to present the state of Machines Reading Maps at the Library of Congress, in two separate events: one for library staff and one for the general public.

Similarly, we enjoyed the amazing hospitality of USC Libraries. While in LA, we presented our work to the university's research staff, but we were also able to deliver a very well attended workshop for students. Students from both USC and UMN are now involved (and very much engaged!) in the creation of the gold standard annotations.

Summary of key outputs:

- Documented code for *mapKurator* & integrated Recogito + *mapKurator* annotation interface
- Our open **documentation** about using Recogito for maps and its *mapKurator* integration is growing and diversifying, thanks to the feedback we received from our first users.

- Annotation guidelines developed for the gold standard work are a major contribution to DH methods.
- Datasets
  - Sanborn: LC-specific documentation and dataset available via LC catalog
  - OS: dataset available via NLS Data Foundry with data card and model card. Data will also be shared via HuggingFace.
- Although our timetable for formal publications has been delayed by the many changes in the lives and careers of the team members, we aim to produce one **collaborative article that details the methodology of the project and discusses the results of our case study** (to be submitted to *DSH*).

Historical Methods article about theory and practice of map annotation.

5. Team Updates

The project has experienced some delays, but the disruption had some very good reasons! First, as we mentioned, we extended our time to make room for the work on the Rumsey collections. Second, three of the members of the team have in the meantime changed affiliation (and, in some cases, city). Our Turing-based cell, Katie and Valeria, have accepted permanent positions as Lecturers (Assistant Professors) in two UK universities with excellent presence in Digital Humanities (Lancaster and Sheffield, respectively). Because of funding constraints, Valeria had to officially leave the project, but she is still involved with the Rumsey collaboration (that had more flexible funding) and working on the major project publications. Katie, on the other hand, will be able to lead the project until its conclusion. Rainer also had big news, as he recently decided to become a freelancer.

6. Final project steps

- Process and prepare documentation for LC Sanborn dataset (Feb-May)
- Process and analyze output for OS experiment on antiquities (Feb-May)
- Document and publish gold standard datasets as LC and NLS resources (March-June)
- Document mapKurator and mapKurator + Recogito code to facilitate re-use (Feb-May)
- Organize mapKurator data correction campaign & maptext for library infrastructure workshop with BL (March)

- Submit Historical Methods article (March)
- Organize final Recogito + mapKurator workshop and open tool up for public use (May)
- Establish secure institutional home for Recogito + mapKurator interface (May)
- Submit *DSH* article (June)
- Project post mortem (September)

That's all from us. We would like to thank you for your support throughout the project. We look forward to seeing many of you at our next meeting.

The *Machines Reading Maps* Team