

Final Report/White paper

# Ethical Visualization in the Age of Big Data: Contemporary Cultural Implications of Pre- Twentieth-Century French Texts

**Award No.:** HAA-266490-19

**Date of activities:** September 1, 2019 to August 31, 2020

## Project Activities

### Project Overview

This project advanced work toward generating ethical visualizations of historical corpora comprising the European cultural imagination prior to the twentieth century without reproducing ethnocentrism. Visually representing the historical place of misrepresented peoples and locales throughout the world requires interdisciplinary collaboration focused equally on critical theory, data visualization, ethics, machine learning, and text analysis. The grant funded a two-day workshop that united top experts from relevant fields to address the conceptual and logistical challenges of visualizing French colonial historical texts: 1) how to create ethical data visualizations--and their underlying forms of training and analysis--that grapple with inherent source biases; and 2) how to computationally process non-modern, non-English languages for humanities research in a critically engaged way.

### Project Activities

Over a two-day period, we hosted eight conference sessions, each addressing a key aspect of the project's implementation plan: 1) scholarly framing; 2) metadata structure; 3) natural language processing workflow; 4) domain adaptation for early modern French; 5) data visualization ethics; 6) interface design and usability; 7) adapting open-source projects; and 8) long-term preservation. Each session was led by a domain-area expert, assisted by a scholar in an adjacent field, and was 75 minutes in length, with a 30 minute break between sessions for less formal conversations. There were two working lunches to continue conversations started during the formal sessions, as well as a working dinner to do the same.

The bulk of the first day of sessions addressed issues raised by the historical content, while the second dealt with the methodologies of data visualization and project management. Domain area experts from history, fine arts and design, journalism, computer science, machine learning, and the libraries contributed to the project.

A hired undergraduate assistant recorded all sessions and took graphical notes, which were then posted to the online Zenodo and GitHub along with the scholars' slide decks.

## Changes/Alterations to Work Plan

*Omissions and changes to key personnel:* Several of our proposed participants were unable to attend the event on the proposed date and time. We replaced some key personnel with suitable equivalent experts in their fields.

*Changes to subjects discussed:* On the recommendation of our invited experts, a new topic 'NLP workflow' was added to our workshop discussion schedule and two of the previously proposed topics were folded into one discussion topic: "input and output (i/o), ongoing maintenance, and preservation'.

*Changes to data management plan:* Due to changes in personnel and structural developments at the UNR Library System, as well as recommendations made during the planning workshop itself, the data management plan had to be altered. Instead of despoting the planning workshop materials into the ScholarWorks repository at UNR, we were advised to use Zenodo, which integrates with GitHub, provides a unique document identifier, and is more public-facing than the university's internal data repository.

## Publicizing Efforts

A Github Pages website was made as a public-facing entry into the workshop materials preserved online. Considerable effort was made to make the writing, formatting, and file access easy and accessible for non-expert audiences.

## Accomplishments

The planning stage for the project addressed two major technical goals: 1) how to mine historic corpora for named entities and their descriptors; and 2) then how to visualize them cartographically to demonstrate what nineteenth-century French imperialism looked like from a cultural and ideological perspective. In tackling these two goals, the project addressed how the tools of distant reading and visualization could be applied to a broad range of historical sources in ways that attend to cultural, geographic, and linguistic diversity. With "critical digital humanities" in mind, the workshop's sessions

melded technical concerns with humanistic ones, leading to rich discussions about identifying and preventing the visual perpetuation of pernicious narratives about historical subjects that have persisted from the past into the present.

The key findings from the planning stage of the project can be broken down into three main categories: 1) insights about the visualization of humanistic data, including pitfalls and methodological concerns; 2) technical problems related to the tools of analysis as well as the underlying nature of the humanistic data; and 3) project management revelations.

*Methodological concerns about visualizing humanistic data:*

The planning stage raised important questions about the nature of visualization, probing the inherent biases of mapping technologies specifically and data visualization more broadly. Across several sessions, we examined telling examples where visualization techniques worked at cross purposes to understanding, and we discussed how to mitigate the potential harm that visualization might cause.

One key takeaway from the discussion was the delineation of visualization for discovery versus visualization for impact, and the importance of not conflating the two. It is essential when creating a data visualization of any sort that one considers the purpose of the visualization, its frame and argument, the ethical considerations that might arise, and its potential impact on three groups: users of the visualization, subjects represented in the visualization, and people affected by those representations (for example, descendants of visualized subjects).

*Technical concerns of the outlined project:* The technical concerns that surfaced during the workshop concerned how to select and implement appropriate metadata standards, the challenges of both creating custom visualization tools and using off-the-shelf packages, how to determine which processes were most suitable for coding named descriptors in the xml text.

Over the course of the planning workshop, it became clear that the *Journal des Voyages* dataset was not as clean as promised by the third-party vendor hired prior to the start of the grant period. Moreover, the planning stage emphasized the necessity of dataset annotating in order to create a so-called gold standard set against which researchers can test the accuracy of machine learning algorithms. In the next stage of the project, the dataset must be further cleaned and annotated.

The workshop also highlighted the need to analyze the photographs, drawings, and other non-textual elements in the periodicals for key information. It became clear that the project would need higher resolution scans of images, necessitating international

collaboration with donor libraries to make new scans of original documents as microform is a highly lossy medium. The next stage of the project will use higher resolution images from the National French Library, which has recently released higher resolution scans of many of its collections..

*Project management and collaborative revelations:* During the planning workshop, the benefit of having diverse scholars conversing with one another was readily apparent. While it proved difficult to bring together colonial scholars with computer scientists, and to achieve representation of diverse ethnic, racial, national, and gender backgrounds, especially with the logistical concerns of long-distance international travel, doing so helped uncover blindspots often overlooked in digital, data-driven projects. For instance, it became clear that humanist inquiry and historical data push the boundaries of settled science in the domains of computer engineering and machine learning. Additionally, colonial scholars recognized the potential provided by digital analytical tools, but raised key concerns about the damage that could be done with the tools of aggregation and visualization at one's disposal.

Finally, we discussed the broad project management question plaguing the digital humanities: if you build it, will they come? During the final sessions of the planning workshop, we discussed how much time and effort it takes to bring a project online, but it may serve a limited audience and only remain operational for a brief duration. The brevity of operation is caused by a host of factors falling under the domain of software maintenance and preservation, including, but not limited to, updates to the platform, security patches, and eventual planned or unplanned obsolescence of the technologies used. To mitigate these concerns, we concluded to ensure that key findings are accessible in "stable" textual formats and that the data be preserved in open standards so that it can be repurposed as technologies change.

## Audiences

The primary audience for the planning workshop included domain area experts in history, fine arts and design, journalism, computer science, machine learning, and library science. As this was a planning workshop for future stages of the project, data are not yet available with respect to the geographical reach of the planning stage's findings. These findings are available online, but they are in the service of creating a larger public-facing project on the visualization of the French imperial imagination that has yet to be created.

That said, the workshop made clear the fruitfulness of interdisciplinary conversations, particularly between the computer sciences and the humanities. By having colonial scholars and people of diverse gender, racial, sexual, and cultural backgrounds

participate in discussions about data visualization, the planning workshop uncovered blindspots in the “algorithmic” disciplines with respect to the potential for harm of data aggregation and visualization, while at the same time unearthing new possibilities for scholarly inquiry that includes the perspectives of colonized, marginalized, and subjugated peoples.

## Evaluation

The planning workshop was in itself an evaluation of our visualization project’s intellectual merits, practicability, appropriateness of its scope and audiences, and technical considerations. Collectively, the workshop participants verified that the project was considered important, relevant to academic communities, and feasible. The question of which open source tools might be adapted for use on the project was not sufficiently resolved, and will need to be pursued in the next stage of the project. In terms of evaluating our proposed activity, the most interesting outcome was that several expert participants raised questions about the necessity and/or dominance of a cartographic visualization in our proposed next stage visualization project. Instead, they suggested other visualization types might offer additional insights, such as network charts.

## Continuation of the Project

The planning workshop was held as a preliminary activity to building an interactive visualization for exploring the *Journal des Voyages* dataset and ultimately creating the finalized corpus. We will seek Office of Digital Humanities Stage 2 funding to continue the project.

*Collaborative partnerships:* The workshop initiated collaborative relationships between the University of Nevada, Reno and The Bancroft Library and the Information School at the University of California, Berkeley, CNRS/LORIA at Université de Lorraine, the School of Fine Arts at the Universidade Federal do Rio de Janeiro, Swinburne University of Technology, The Alan Turing Institute at the British Library, and LIRIS at the National Institute of Applied Sciences. These partnerships will continue through the next stage of project activities, as all participants have expressed interest and willingness to contribute to the project going forward. Additional partnerships are currently being negotiated as well.

Project decisions for next steps across domains

1. [Scholarly framing](#). (Digital humanities & French colonial history)

**Guiding Question:** What would a postcolonial distant reading and visualization of the corpora look like?

**Considerations:** Historiography, theoretical focus, scholarly and source biases

**Goal:** Defined intellectual focus

**Discussants:** Christopher Church (lead) & Charles Tshimanga-Kashama

**Discussion summary:** This wide-ranging discussion covered both potential analytic methods and their implications for scholarly inquiry into the French colonial world, with a special focus on how technical decisions either enable or disable avenues of humanistic study. As an overture for the conversations to follow, it interrogated both the ethical considerations and the data to be used in such an analysis, raising concerns with respect to the data storage location, the impact of algorithms and confirmation bias, the role of computer vision, and the potential for visual analysis of non-textual elements contained in the data set.

Principal to the conversation was addressing the best way to visualize an “imagined geography” that served for the French an ideological as well as a political and economic purpose. Not only did preliminary data analysis speak to Eric Hobsbawm’s idea that symbols become more potent after their practical use is over, particularly in light of a heightened focus on the American West, but it also reinforced the necessity of “decentering” Europe in any analytic narrative about the *Journal des Voyages*, which was a purposefully eurocentric publication about colonized peoples.

**Decision:** The key decisions made with respect to scholarly framing were

1. to encode the data to allow analysis to distinguish between people and places within and without the French empire;
  2. to decide upon the types of named entities that will be encoded, ensuring that people are not solely represented by “eurocentric” sources;
  3. to explore the emotional affect attached to named entities, while maintaining a critical eye to what that says about the French gaze;
  4. to create a narrative apparatus using select colonial sources that contradict and/or reshape the historical arguments contained in eurocentric source material; and
  5. to build a partnership with the National French Library (BnF) and seek funding related to a more involved collaboration.
2. [Metadata structure](#).
- Guiding question:** What are the metadata needs for realizing this analysis?

**Considerations:** XML-tags [article (beginning & end), title (beginning & end), image, issue number, issue date, page number], library standards, interoperability

**Goal:** Metadata structural framework

**Discussants:** Mary Elings (lead), Teresa Schultz & Elena Azadbakht

**Discussion summary:** In this session, we discussed the essential though often understated role metadata play with respect to digital humanities projects, particularly the importance of building metadata schema suited to a particular disciplinary context. With the goals of the project in mind, we discussed several potential metadata standards and how well suited they might be to the disciplinary scope of a digital history project.

Additionally, this session addressed the different types of required metadata to describe the assets within the collection, spending time distinguishing what the smallest unit of the collection might be (whether a particular page, article, journal issue, or year) and how we might associate all related assets within a database structure.

**Decision:** Having concluded that the collection would be divided into units at the article level and collected by both issue and year, we resolved to:

1. explore Function Requirements for Bibliographic Records (FRBR), an entity-relationship model created by the International Federation of Library Associations and Institutions. FRBR would be the most viable method for describing the resources within our collection, and it is more user friendly than XML-based standards like The Encoding Initiative (TEI).
3. [Natural language processing \(NLP\) workflow.](#)

**Guiding question:** Which natural language processes are most appropriate for this corpora?

**Considerations:** Language models, named entity recognition, RNN, dialects, semantic analysis, part of speech tagging, OCR requirements, analytic depth

**Goal:** Roadmap for training and employing NLP model

**Discussants:** Claire Gardent (lead), Ludovic Moncla & David Bamman

**Discussion summary:** During this session, we discussed the technical and conceptual requirements necessary for performing natural language processing on the corpora and the implications of differing approaches. We addressed three approaches to the visualization of text: 1) document-based visualization, which

provides a network-based bird's eye view of multiple documents in order to identify key topics and distinguish between relevant and irrelevant assets; 2) location-based visualization, which allows for the geographic representation of events and generates a map or cartographic visualization from the documents in the corpora; and 3) event-based visualization, which yields a network view of the various entities within a collection of texts.

We discussed the importance of fidelity in annotating texts, regardless of the method employed, as errors can propagate from stage to stage as the project progresses. With respect to the naming of entities, disambiguation is frequently the largest hurdle, which is compounded by the necessity for historical gazetteers for place names that may have changed from their historical antecedents. Further, machine learning models are based on probability and frequency, so the most egregious errors may occur with those textual elements that diverge from their most frequent use. Oftentimes in humanistic study, these divergences are the most telling and important.

**Decision:** We decided that we would employ two of the three methods discussed during the session, focusing our efforts on:

1. at least two visualizations of the corpus:
    - a. an event-based visualization that looks at the relations between textual entities and correlates them to locations on a physical map,
    - b. an event-based network visualization that identifies strength of relationships between textual entities
  2. ensuring that the text is never obscured by the visualization apparatus, thus allowing both the researcher and the user to quickly and easily access the original source material.
  3. an annotation workflow that includes an inter-annotator agreement analysis.
4. [Domain adaptation \(for early modern French\)](#).  
**Guiding question:** What are the best practices when creating in-domain (early modern French) data for natural language processing and machine learning?  
**Considerations:** Domain mismatch, validation, annotation tools, training a model (illustrated with NER)  
**Goal:** Understanding the risks of out-of-the-box tools and how to overcome them by training new models with your own annotated data  
**Discussants:** David Bamman (lead), Katie McDonough & Claire Gardent

**Discussion summary:** Building on the previous session, this session addressed the applicability and accuracy of natural language processing across

language domains. Most methods for natural language processing, including name entity recognition, are trained using the 1998 Wall Street Journal, which consequently yields highly accurate results for modern, journalistic American English (rates as high as 100% for tokenization and 98% for part of speech tagging). However, once those trained models are applied to other language domains (historic English, foreign languages), the accuracy declines precipitously to 40% to 60% accuracy, well below the threshold for making reliable analytic conclusions.

Therefore, it is essential to create trained models within the language domain being studied, most preferably using proximate texts from the materials to be algorithmically analyzed. This means a great deal of discipline-specific work and human hours to create these datasets.

We also discussed different methods for annotating such a dataset, as well as the various NLP methods that could be used.

**Decision:** In our natural language processing workflow, we resolved to:

1. restrict our algorithm to a maximum of five categories: people, physical places, metaphoric places, organizations, and events. Tracking more than locations in the text, will avoid replicating the militaristic ulterior motives inherent to the corpora themselves.
  2. use the BiLSTM system of neural networks as a machine learning method,
  3. use BERT contextual embeddings, a Python library published by Google,
  4. employ the BRAT rapid annotation tool for creating the annotated data necessary for NLP training.
  5. use three datasets in the project:
    - a. development data that is annotated,
    - b. test data very similar to the corpus to evaluate the accuracy of the machine learning algorithm, and
    - c. the actual corpus itself as the data that will be processed.
5. [Ethical data visualization.](#)

**Guiding question:** What ethical challenges arise in visualizing this data in this context?

**Considerations:** Source and methodological biases, media and design biases, visual rhetoric, ethical visualization, colonial relations, subaltern narratives, audience, agency

**Goal:** *Understanding the risks of intending to produce objective visualizations*

*and how to overcome them by employing ethical visualization practices, formats, and interactions*

**Discussants:** Katherine Hepworth (lead), Karel van der Waarde & Doris Kosminsky

**Discussion summary:** The discussion started with emphasis on the power structures at work in colonial narratives and in visualizations, and how these power structures leave distinct, traceable, and measurable visual evidence. The group discussed the limitations inherent in the problematic narratives in the source material of *Journal des Voyages*, and ways to potentially mitigate them. By visualizing the imagined geography represented within this source material, we run a real risk of perpetuating these narratives. We therefore need to make conscious efforts to offer multiple views and contrasting perspectives on our source data.

Discussion then moved to the importance of understanding the ways visualizations amplify and reinforce biases inherent in primary source material by carrying arguments and narratives through rhetorical design elements, compositions, and navigation patterns. The discussion turned toward data availability, the notion of gatekeeping, and finally moved on to the importance of finding, understanding, and targeting those audiences that have an interest in your subject as well as those who have a need that is likely to result in them regularly using your site.

**Decision:** We resolved to:

1. Use visualizations to explore the research question: how does *Journal des Voyages* contribute to the cultural and ideological momentum necessary for the French empire to maintain itself without an emperor?
2. Research our potential audiences in detail, and understand them and their needs before visualizing. We will look for potential audiences in:
  - a. Research communities
  - b. Lay communities in the Francophone French diaspora
3. After production, pitch our visualizations project to those audiences to encourage continued use
4. Research potential unintended audiences (such as groups promoting hate speech or xenophobia) and incorporate dissuasion of involvement by those groups.
5. First product: Design and release multiple interactive visualizations - one cartographic and one network - with guided pathways that support the

argument we intend to communicate through the visualizations (following ethical visualization principles). It should have:

- a. A compassionate user interface designed to mitigate potential for harm within the visualization;
  - b. Bilingual translation
  - c. An associated site with supplementary text and ‘show your work’ documentation
  - d. Guided pathways could be created by using interaction effects such as scrollytelling, tooltips, restricted zoom levels, heavily annotated views.
  - e. Consult museum design researchers with regard to guided pathways
6. Second product: Jupyter notebooks containing a second level of functionality for the same visualizations, that allows for exploratory interaction with the data. We acknowledge this is a risky undertaking with such a sensitive corpus, and would only allow users to access this functionality after going through a series of disclaimers and agreement interactions (ie an agreement to do no harm with the data).
  7. Final product: A data repository accessible after signing a memorandum of understanding that commits users to doing no harm with the data and understanding its limitations. This would come last, after reflection on the two other processes. The repository would contain:
    - a. base text data;
    - b. annotation of text data (our annotation);
    - c. images;
    - d. image in situ with text
  8. Use IIIF technologies to replace the images in locations (current dataset contains low quality bitmap images, BnF has released high quality greyscale versions of them, and may have some color illustrations available)
  9. Ask BnF if the higher quality color scans are available

**Further considerations:** Submit data and paper to *Journal of Humanities Data* about the thinking around our data cleaning decisions. A second paper discussing the ethics of the visualization project would also be valuable. Participants also suggested possibly submitting one or both of these papers to IEEEVis4DH.

6. [Mapping and interface design usability.](#)

**Guiding question:** Which end-user considerations are critical and optimal for interactive mapping?

**Considerations:** User experience design, graphic user interfaces, data transparency, legibility, cultural geography, geographic power dispersion

**Outcome:** Necessary and preferred usability features

**Discussants:** Katie McDonough (lead), Doris Kosminsky & Karel van der Waarde

**Discussion summary:** The discussion started with considering the importance of aesthetics and affect to ethical visualization of data. The affect created by the Visualization of Fear exhibit by artist Kader Attia, was raised as an effective balance of triggering emotion without overwhelming the audience's capacity to explore the historical content. The unconscious influence on affect of aesthetic decisions was discussed, using the design of a related project, Mapping of Slave Revolt, by Vincent Brown and Elizabeth Maddock Dillon as an example. In this project, the visualizations are rendered with a hand-rendered, period-appropriate feel. Some archival research by some participants suggests that hand-rendering may have considerable influence on making visualizations feel relatable and personally impactful. Period-appropriate aesthetics in the visualization can provide extra reinforcement of historical context, though they should be used with caution in the user interface itself.

The power structures inherent in the corpus were discussed in rounding out the session. The *Journal des Voyages* was discussed as a tool of governmentality: it taught readers what to imagine about themselves, perhaps more so than about the places depicted, by using geographical coverage, emotional language, and relationships between entities described. In terms of image content, the vast majority of the depictions are of bodies. There are very few maps in the *Journal des Voyages*, and the ones that are depicted are most commonly in the regular profiles of government departments.

The discussion concluded with exploration of what it might mean to design a compassionate user interface in the context of this corpus. The importance of sketching, hand-drawn aesthetics, choice architecture, and static visualizations was emphasized for creating a compassionate user interface that is graceful in the ways it utilizes human factors for understanding.

**Decision:** We resolved to:

1. Use polygons for storing locations wherever possible (ie for cities, states, regions)

2. Use malleable visualizations (either as notebook outputs or D3.js) in order to make them adaptable for different audiences - potentially as our 'training wheels off' second visualization product
3. Frame the Journal des Voyage corpora as a technology of government/governmentality central to ideas of national identity and culture in France

**Further considerations:** Opportunity for writing about:

1. the corpus as a tool of governmentality, as a theoretical means to explore the phenomena of exceptionalism in the cultural imagination of governing republics, and
2. a short article surveying the compassionate tech initiatives that could inform development of a compassionate user interface for interaction with the corpus.

7. [Adapting and integrating existing open source projects.](#)

**Guiding question:** Which of the available open-source projects may meet project needs?

**Considerations:** Open-ONI, OpeNER, Apache, NERC-Fr, Palladio, Voyant Tools, D3

**Goal:** Shortlist of open source tools to adapt

**Discussants:** Ludovic Moncla (lead), Mary Elings & Elena Azadbakht

**Discussion summary:** During this session, we reviewed several open-source projects, paying particular attention to the research activities and experience of our grant participants. In discussing the available tools, we attended to four main project needs: 1) browsing and sharing the document collection; 2) annotating the corpus; 3) processing the corpus using machine learning, geoparsing, and text mining techniques; and 4) visualizing and exploring the corpus.

**Decision:** We resolved to adapt tools for the following purposes in building our project:

1. Browsing Corpus  
**OpenONI, The Online Newspaper Initiative**  
which provides a function set for loading, modeling and indexing data
2. Annotating Corpus  
**BRAT**  
A server-based tool used to annotate the training and verification data for natural language processing  
**Pelagios and Recogito**

A semantic annotation tool for texts and images that can identify and map places

#### **PERDIDO Geoparser**

A flexible geoparser that could be adapted to use a manually annotated gazetteer of historic place names

### 3. Natural Language Processing

#### **Spacy**

A flexible python library for natural language processing capable of performing most of the needed project tasks

### 4. Visualization

#### **D3, Data-Driven Documents**

A Javascript library that will enable us to make custom visualizations that are web-based and interoperable across browsers

## 8. [Input and output \(i/o\), ongoing maintenance & preservation.](#)

**Guiding question:** *Which inputs and outputs are critical and optimal? What are the requirements for proper software maintenance and long-term archiving?*

**Considerations:** Open and/or widely-used data formats [TXT, CSV, XML, PDF, PNG], consuming the dataset, research outcome formats, interoperability, maintenance costs, data longevity

**Goal:** Necessary and preferred data formats, importation, and exports, software maintenance and management plan

**Discussants:** Teresa Schultz (lead) & Doris Kosminsky

**Discussion summary:** The discussion centered around currently feasible and standard practices in maintenance and archiving. Currently, every major data repository is using Amazon Web Services (AWS) for data storage, which presents issues with respect to corporate longevity as no repositories currently implement decentralized storage such as SWARM. To be accessible after archiving, data must be FAIR: findable, accessible, interoperable, and reusable. Our options for storing and sharing data include UNR's ScholarWorks, which is expensive and very prescriptive, Zenodo, which gives a DOI to any contribution but only stores zip archives, Dataverse, and ICPSR. For preserving the visualization and its accompanying code, Binderhub is most suitable and better than Docker for preserving the environment of a platform.

**Decision:** We will give the project a DOI via Zenodo storage.

**Further considerations:** A paper about preservation issues we encounter would be beneficial to other DH projects.

## Long Term Impact

*Long-term impact:* This workshop, and intellectual activity leading up to it, have resulted in several related projects.

- *Ethical and Effective Visualization*  
<https://kathep.github.io/ethics/>  
 A method for ethically visualizing data that was developed preparatory to the workshop. It has since been updated due to the contributions at the workshop participants.
- *Ethical Data Visualization: Taming Treacherous Data*  
<https://github.com/cmchurch/DHSI-ethical-dataviz>  
 A 1-week workshop held at the Digital Humanities Summer Institute at the University of Victoria, Canada, in 2018 and 2019, teaching fellow humanists about the critical issues related to data processing and visualization in the digital humanities.
- *Ethical Data Visualization: Taming Treacherous Data*  
<https://github.com/kathep/DHDownunder-ethical-dataviz>  
 A 2-day workshop held at DHDownunder at the University of Newcastle in 2019, teaching fellow humanists about the critical issues related to data processing and visualization in the digital humanities.
- *Visualizing Empire*  
<https://unr-ndad.github.io/empire/>  
 The website hosting the documentation and description of the workshop.

The materials for each of these activities are made publicly available via Github repositories under open source licenses, and in some cases, as Github Pages websites. The main long-term impact of these activities is preparation for the next stage of the project, namely creating a visualization to explore the *Journal des Voyages* corpus. The additional long-term impacts of these activities are that: 1) approximately 70 scholars have been trained in our considered approach to ethical visualization; 2) several government organizations, startups, and research groups are currently using the Ethical Visualization for Impact method, greatly expanding the scope and reach of our project beyond what we expected.

## Award Products

The grant period produced several different products, foremost among them a comprehensive plan for future project stages that takes into account the advice from the diverse range of scholars participating in the planning workshop.

The award also contributed to the drafting of a peer-reviewed paper that will be submitted to *Digital Scholarship in the Humanities* outlining a method for computationally processing non-modern, non-English languages for humanities research in a critically engaged way.

All materials from the planning workshop have been collected into an online repository, accessible via GitHub and deposited at Zenodo with a document object identifier. These materials include recordings of all discussions and copies of all slide decks from the planning workshop itself, as well as a summary of the key findings from the grant period. They also include the dataset of the *Journal des Voyages* corpus produced through a combination of OCR scanning with keyed re-entry. Alongside the *Encyclopedie* dataset hosted by Stanford University, these data will ultimately be used to create a training model for natural language processing and named-entity recognition in non-modern French.