

# Workset Creation for Scholarly Analysis:

---

## *Recommendations and Prototyping Project Reports*

Prepared by:

J.Stephen Downie, Tim Cole, and Megan Senseney

Center for Informatics Research in Science and Scholarship  
Graduate School of Library and Information Science  
University of Illinois at Urbana-Champaign

CIRSS Technical Report, WCSA1215

December 2015

## Table of Contents

Executive Summary .....	2
Workset Creation through Image Analysis of Document Pages .....	4
Semantic Analysis of Documents from the HathiTrust Corpus .....	13
Distributed Metadata Correction and Annotation .....	20
EIEPHãT: Early English Print in HathiTrust, a Linked Semantic Workset Prototype .....	26

## Executive Summary

This document assembles and describes the outcomes of the 4 prototyping projects undertaken as part of the *Workset Creation for Scholarly Analysis (WCSA)* research project (2013 – 2015). The work described here was made possible through the generous support of The Andrew W. Mellon Foundation (Grant Ref # 21300666); however, any opinions, findings, and conclusions or recommendations expressed in this report are those of the author(s) and do not necessarily reflect the views of the Andrew W. Mellon Foundation.

Each prototyping project team provided its own final report. These reports are assembled together and included in this document. Based on the totality of results reported, the WCSA project team arrived the following overarching recommendations for HTRC implementation and adoption of research conducted by the Prototyping Project teams:

1. The HTRC's current workset creation tool, Workset Builder, must be redesigned and rebuilt responsive to the findings described in the Prototyping Project Final Reports. In particular, it should build on and integrate the lessons learned from the EIEPHãT and Capisco Prototyping Projects (as described below).
2. The HTRC metadata infrastructure should be transformed from MARC-based to a more Linked Open Data (LOD) compatible descriptive metadata approach – e.g., one of the LOD-based standards for bibliographic description studied, tested and analyzed as part of the EIEPHãT project. This implies the likely integration into HTRC of a metadata triple store.
3. The HTRC metadata infrastructure should be enhanced to support the augmentation of traditional bibliographic metadata with other attributes, including attributes (features) derived through the analysis of the full-text of the resources being described.
4. HTRC should adapt and implement at greater scale and within the context of the HTRC Workset Builder, the Capisco System for detecting (through full-text analysis) and tagging HT texts with "Concepts-in-Context." This will enhance subject access to HT resources.
5. HTRC should adapt and implement at greater scale within the context of the HTRC Workset Builder, the utilities for workset development that are demonstrated in the EIEPHãT demonstrator.
6. HTRC should more precisely identify and describe relationships among HT resources, between HT resources and resources elsewhere on the Web, and between HT resources and the entities relevant their creation, i.e., the individual, events, and places relevant to resource creation. This includes providing the means for scholars to explore and build worksets containing items of varying granularity (i.e., more or less granular than a single volume) and (eventually) encompassing non-textual and non-prose resources such as music and poetry.

7. In carrying out these recommendations, HTRC should assess the quality and applicability of automated systems used for concept tagging, to add links, to maintain metadata provenance and to assist in the creation of large worksets (e.g., containing tens or hundreds of thousands of items). HTRC should also solicit feedback from user communities on an ongoing basis while developing and enhancing its services.

These recommendations will inform further work by the HTRC team, e.g., as part of the new *Workset Creation for Scholarly Analysis and Data Capsules (WCSA+DC): Laying the foundations for secure computation with copyrighted data in the HathiTrust Research Center, Phase I* research project funded by the Mellon Foundation.

# Workset Creation through Image Analysis of Document Pages

## Narrative

Printed artifacts communicate historical information through three different kinds of features:

- *bibliographic features*: the paper, binding, typeface, size, quality, price, and organization of the physical book;
- *visual or graphic features*: the arrangement of text, images and white space on a page;
- *linguistic features*: the syntactic and semantic aspects of the words presented on the book's pages.

The researcher working directly with the historical artifact frequently considers its bibliographical and visual features alongside its linguistic content, either explicitly or implicitly. A researcher who discovers, for example, that a book contains an illustration, may thumb through its pages to find others, and then look through other books for related illustrations. But researchers working in large scale digital libraries like the HathiTrust will frequently only have access to digital surrogates and not the material artifacts.

Visual features of printed books that are of interest to humanities researchers are captured in the scanned page images, but may not be recorded in the cataloging record or other metadata associated with the digital file. Our prototype software application uses the visual characteristics of digitized printed pages to identify documents that contain three types of visually distinctive materials of interest to humanities researchers: illustrations, music, and poetry.

Project objectives during this initial grant period include:

- designing and building core system architecture and image analysis pre-processing components;
- development of segment classifiers for illustrations, music, and poetry;
- selection and preparation of a data set for testing
- deployment of the application in the HTRC testing environment
- initial evaluation

## Significant Changes to Personnel

There were no significant changes to personnel or management during the course of the project.

## Project Progress

The four main technical achievements for this project are:

- an extensible framework for configuring and executing image analysis workflows
- implementations of several algorithms from the research literature for image cleaning, manipulation and segmentation

- a library that encapsulates access to HathiTrust APIs and data structures for ease of use within Java applications
- an application to run on HTRC servers that reads a list of items, loads the corresponding page images and executes an image analysis workflow

Document image analysis is a mature research field with significant ongoing work and a number of commercial products. Indeed, many of these technologies have been deployed to support the analysis of a range of cultural heritage material. Despite these advances, there are no robust tools that allow scholars to approach the visually encoded information contained in document page images analogous to the widely used text analysis tools. This restricts research aimed at understanding visually constructed meaning to datasets small enough to be studied closely or to projects with sufficient funding and resources to invest in custom image analysis software. Given the quantity and variety of digitized page images in the HathiTrust digital library, developing image analysis tools that can be configured by scholars and used at scale complements the exiting text analysis services provided by HTRC.

The core technical component of our system is DataTrax, a framework for executing user-configurable image analysis workflows. These workflows are comprised of discrete image analysis tasks that take one or more well-defined inputs such as a color image or a collection of segmented glyphs and transform them into an output value such as a black and white image or the glyphs grouped into lines. The specific tasks that can be executed within a workflow are defined externally to the DataTrax framework and registered by an application. This software architecture, shown in figure 1, promotes easy reuse of existing code.

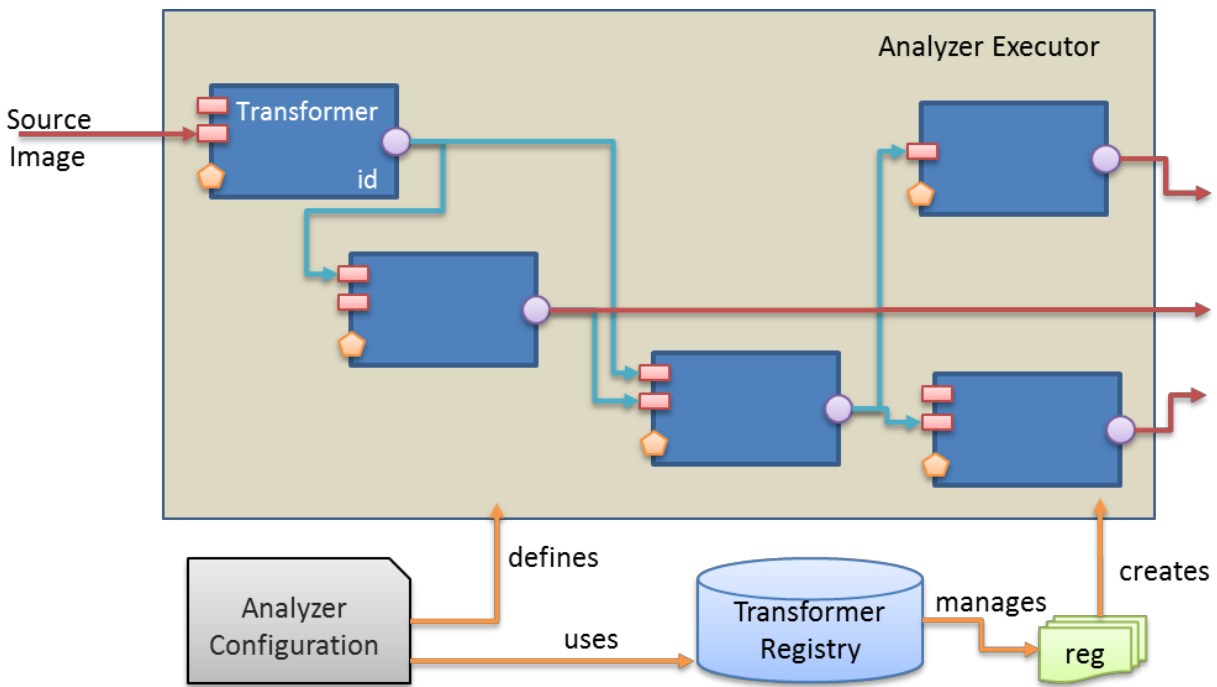


Figure 1: Overview of DataTrax Architecture

We have implemented an initial set of image analysis tasks from the research literature. These include preprocessing steps such as converting color images to black and white, image segmentation tasks such as finding all connected components (roughly glyphs on the page) and image transformations such as down sampling to create a smaller version of an image with specific characteristics. We have also created adapters that we use to integrate the widely used OpenCV library for computer vision into DataTrax. This is particularly important since it demonstrates the relative ease with which existing third-part software components can be incorporated into our application.

In addition to the core image-analysis system, the prototype development effort required that we connect this technology to the existing infrastructure and resources provided by HathiTrust. To achieve this, we have created a general-purpose software development kit (SDK) for interacting with the existing APIs and data objects provided HathiTrust and the HathiTrust Research Center. This includes tasks like communicating with the bibliographic and data REST APIs (including authentication and account credential management), resolving item data records stored in a Pairtree directory structure based on HathiTrust identifiers, and reading image data directly from the zipped item data records. This SDK meets the immediate needs of our application while providing a codebase for future projects that make use HTRC data resources without re-implementing the details of negotiating REST requests or parsing data records.

The three components, the DataTrax framework, the library of document image analysis algorithms and the HathiTrust SDK are integral to the efforts of the prototype grant but are implemented as separate libraries that can be used (and are being used) independently. This library-oriented development process will maximize the impact of our work beyond the scope of the WCSA project as we, and eventually others, use them in a range of projects.

The final technical contribution of the project is the WCSA prototype application itself. This application is the component that ties together the three libraries discussed above and implements the control logic for reading in the list of items to be processed, creating the image analysis workflow using DataTrax and interpreting the results to identify specific features.

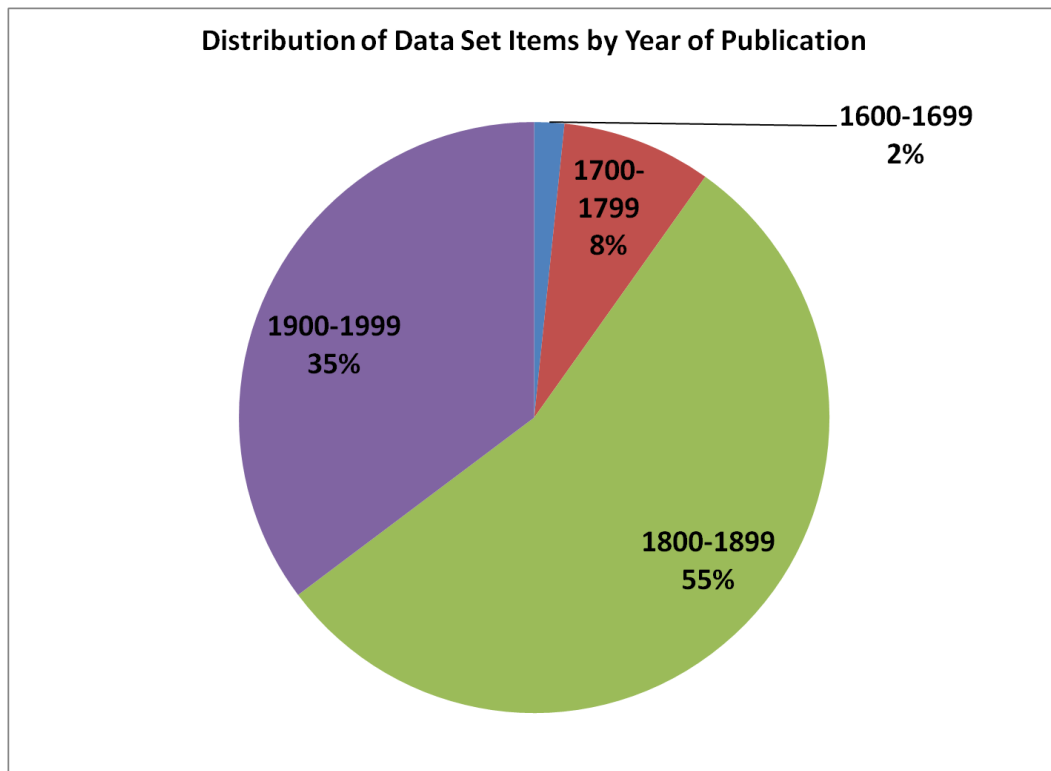
The overall system that we have developed through this work provides a framework for working with the large and diverse collection of page images housed at the HathiTrust digital library. As a result, we are now well positioned to take the next steps including partnering with experts in document image analysis to develop more refined algorithms, working on machine learning and pattern recognition tools to provide sophisticated computational analysis, and designing user interfaces that allow scholars to customize workflows, connect with an analyze workset and explore the results of their analysis.

### [Dataset Preparation](#)

A hand-curated set of 250 volume ids were selected from a broad range of LC subject headings, including subjects in literature, history, music, the arts, architecture, science, and popular

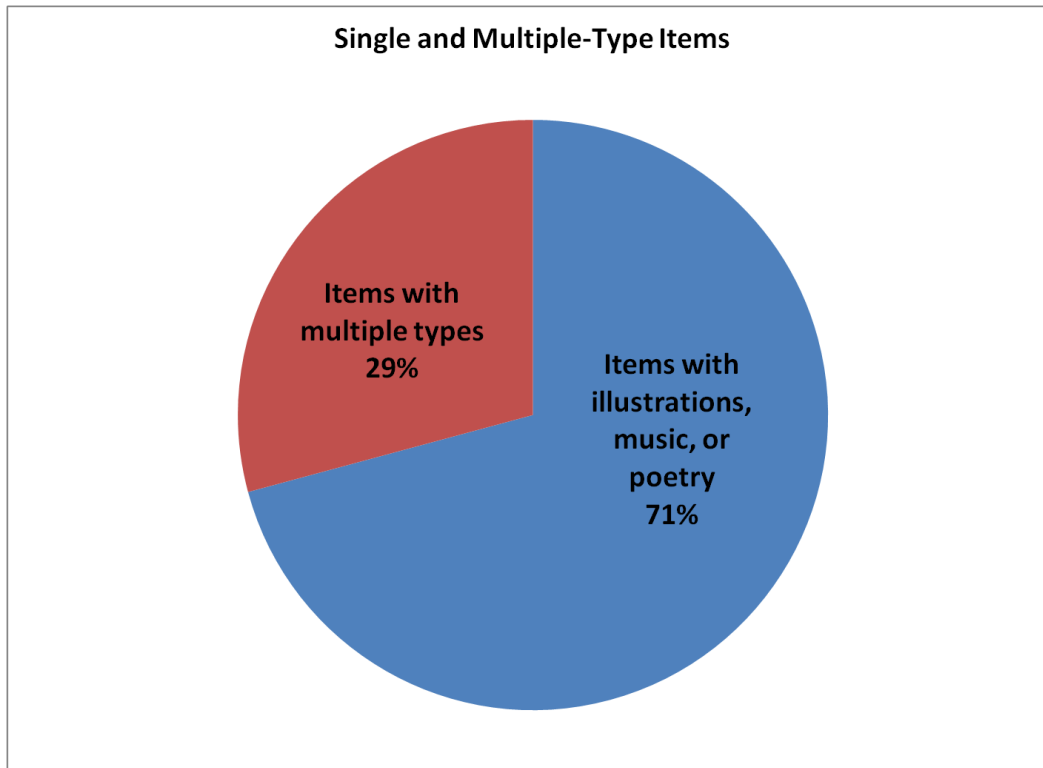
culture. These subjects were selected in order to provide the most varied kinds of books for testing.

Items were selected for the dataset with publication dates distributed over four centuries. The distribution of publication dates was weighted most heavily towards the nineteenth century (55%), followed by twentieth century items (35%). This distribution approximates that of the HathiTrust Digital Library holdings.



This curated dataset includes 122 volumes with illustrations (in sizes ranging from very small to partial page and full page), 58 volumes with music (small extracts, partial page and full page), and 128 volumes with poetry (extracts, partial page, and full page). Some volumes contain two or three of these types of materials, sometimes on the same page.





### Setbacks or Challenges

The proposed scope of work was ambitious. While we accomplished our primary goal of creating a framework for analyzing the visually salient features of printed material and deploying that framework for use on HathiTrust data sets, we were not able to analyze books to detect presence of poetry and our attempt to recognize musical scores were unsuccessful. These setbacks are largely due to the fact that recognizing pages with illustrations proved to be significantly more challenging than we anticipated.

We chose to focus on detecting illustrations at the expense of dedicating more time to music and poetry for two reasons. First, the core image analysis tools required for this task are broadly applicable and provide a solid base for future work. Second, our anecdotal experience indicates that there is a widespread interest within the community for better tools to mine large document collections for illustrations. Additionally, because we were interested in evaluating the feasibility of visual page analytics, we elected not to take shortcuts such as analyzing the generated OCR text to find illustrations.

With respect to detecting musical notation, we implemented the run length algorithm proposed by Bainbridge. Our initial implementation did not yield promising results. Rather than working to debug and revise this current work we elected to leave this task as an opportunity to collaborate with members of the optical music recognition community.

Our work on visually analyzing features of printed poetry in an earlier project informed our approach to visual text analytics for this project, which included the development of a feature

ontology for printed books and for poetry in particular (see Appendix). One of the significant challenges in detecting poetry in a diverse digital library such as the HathiTrust is the range of print conventions and styles found over works from different locations and time periods.

## List of Project Disseminations and Final Deliverables

As part of the WCSA prototype effort, TCAT has developed and released four main software components. Three of these provide core library capacities that we use extensively within the WCSA prototype application and have broader applications for the DH a library community. The fourth is the main prototype application that is responsible for leveraging and configuring the libraries in order to implement the specific image analysis tasks set forth for our project. These components are briefly summarized below.

### DataTrax Framework

DataTrax provides an extensible framework for configuring and executing image analysis workflows. We anticipate that enhancements to this framework will be ongoing as we pursue additional work on VisualPage and other related projects.

Available at: <https://github.com/tcat-tamu/DataTrax>

License: Apache 2.0

### HathiTrust SDK

The HathiTrust SDK provides a Java-based API for interacting with the HathiTrust REST APIs and content material. This SDK currently and early level prototype being developed in conjunction with the WCSA prototype grant and other projects at TCAT.

Available at: <https://github.com/tcat-tamu/HathiTrust-SDK>

License: Apache 2.0

### Document Image Analysis Algorithms

We have implemented a number of low-level document image analysis algorithms to perform basic tasks that can be combined either programmatically or using the DataTrax framework to create complex image analysis workflows. These algorithms provide a starting point for creating open source image analysis tools for use within the DH and library communities.

Available at: <https://github.com/tcat-tamu/Document-Image-Analysis>

License: Apache 2.0

### WCSA Prototype Application

The WCSA Prototype application is the main driver that reads images from the HTRC server testbed and executed the image analysis work flows and post processing. This component is tailored to use within the scope of the WCSA project and would, with additional funding, be the starting point for developing a more general purpose application for deployment.

Available at: <https://github.com/tcat-tamu/visualpage.wcsa>

License: Apache 2.0

## Appendix A: Visual Features of Printed Books

Printed books contain a variety of visual features related to typography, page design, and structured information. Some of these features contribute to the discovery of the targeted types of visual material and other features distract from that discovery by producing false positive results or by preventing the optimal functioning of page recognition algorithms, particularly in older books with less regularized typeface and layout.

Visual features of print layout include:

- One column text block
- Two column text block
- Three column text block
- Indentation of text
- Running heads
- Page numbers
- Footnotes
- Titles
- Subtitles
- Margin consistency
- Margin size

Visual features related to illustrated materials include:

- Percentage of the page occupied by illustration (full page, half page, quarter page, or smaller)
- Placement of the illustration on the page
- Arrangement of text next to or around the illustration
- Decorative borders around text and/or image
- Small ornamental designs used to separate or mark portions of the text
- Maps
- Charts
- Tables

Visual features related to music include:

- Percentage of the page occupied by musical notation (full page, half page, quarter page, or smaller)
- Placement of musical notation on the page
- Arrangement of text next to or around musical notation
- Placement of text within the musical notation (to indicate notes, scales, performance instructions, or lyrics)
- Hand drawn or typeset staves
- Multi-part musical scores

Visual features related to poetry include:

- Percentage of the page occupied by poetry (full page, half page, quarter page, or smaller)
- Placement of poetry on the page
- Arrangement of other text next to or around poetry
- Capitalization of lines of poetry
- Indentation of lines of poetry
- Separation of poetry into stanzas (groups of lines) separated by white space

# Semantic Analysis of Documents from the HathiTrust Corpus

Final project report

## Capisco: Semantic Analysis of Documents from the HathiTrust Corpus

Annika Hinze, Craig Taube-Schock, Sally Jo Cunningham, David Bainbridge  
University of Waikato, New Zealand

### 1. Description of the project and its purpose

This project developed a tool to assist scholars in identifying and selecting resources from within the HathiTrust Document Corpus. Current access to this resource is available via text-based search in full-text and metadata. Existing scholarly document search tools use purely lexical analysis, which cannot address the inherent ambiguity of natural language. For example, a simple search on the country name “Niue” will miss references to it by its traditional names (“Nuku-ta-taha”) and by its initial English name (“Savage Island”). Our Capisco System analyzes documents by the semantics of their content.

Traditional access to the digitized document collections is available primarily via string-based search in the documents’ full-text and metadata. Such a text-based search identifies documents purely according to lexicographical analysis. Most research questions and areas of scholarly interest, however, can rarely be described by simple textual keywords and instead, they encompass larger concepts. Relevant sources remain undetected unless the right keywords are found. Easy identification of appropriate keywords is further hindered when different languages are involved and when an area contains sources from diverse fields that do not share a common vocabulary. Further problems are introduced through the inherent ambiguity of natural language, e.g., synonyms and homonyms. In all these cases, false negatives (i.e., missed documents) and false positives (i.e., unrelated documents that have to be manually identified and eliminated) have significant adverse effects on the scholar’s research.

To facilitate scholarly work on the HathiTrust document set, a clustering of documents by semantic similarity could open up a wealth of further opportunities. We suggest analyzing documents not purely by their text but rather by the semantics of their content. A semantic search approach offers the potential to overcome the shortcoming of lexical search, but—even if an appropriate network of ontologies could be decided upon—it would require a full semantic markup of each document. Our project developed the conceptual design and initial implementation of a new framework that affords the benefits of semantic search while minimizing the problems associated with applying existing semantic analysis at scale. Our approach avoids the need for complete semantic document markup using pre-existing ontologies by developing an automatically generated Concept-in-Context (CiC) network seeded by a priori analysis of Wikipedia texts and identification of semantic metadata. Our Capisco system analyzes documents by the semantics and context of their content. The disambiguation of search queries is done interactively, to fully utilize the domain knowledge of the scholar. Our method achieves a form of semantic-enhanced search that simultaneously exploits the proven scale benefits provided by lexical indexing.

The project was executed in close collaboration with two humanities scholars from the areas of Māori & Pacific Studies, and Historical Anthropology. The research team like to acknowledge their collaboration with humanities scholars Tom Ryan (Cultural Studies and Historical Anthropology) and Rangi Matamua (Māori & Pacific Development Studies), University of Waikato, New Zealand. The scholars did not only drive this project with research questions based on their scholarly practice, but also provided ongoing input and feedback during the development process.

## 2. Changes to personnel or project management since initial proposal

No changes were made to the proposed personnel.

## 3. Summary of project progress and significant accomplishments

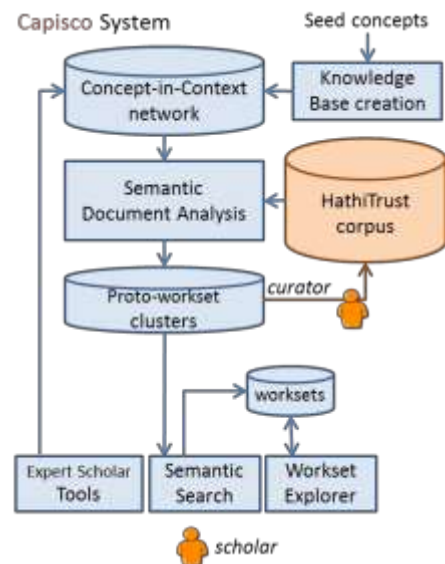
This section describes the project approach, its research contributions and each of the components that were developed.

### Capisco Approach

Capisco's semantic analysis is executed in two steps:

**Knowledge Base creation:** The various meanings of words are encoded in an automatically generated Concept-in-Context (CiC) network seeded by *a priori* analysis of Wikipedia texts and identification of semantics. Our CiC network encodes, for example, that the term "apple" refers to a *fruit* in the context of *nutrition* and to *computers* in the context of *IT*.

**Semantic Document Analysis:** We then analyze which concepts and contexts appear in each document in the corpus with the goal of assigning a set of semantic meanings to each document. The documents are then clustered by semantic concepts forming so-called *proto-worksets*.



### Support for Scholars

Two tools are provided for scholarly search and exploration of worksets.

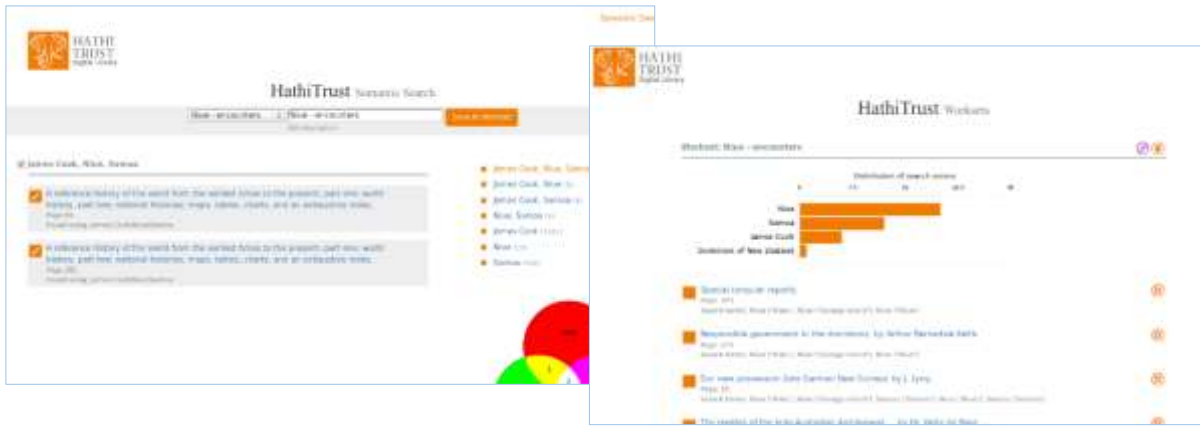
**Semantic Search:** This tool allows scholars to search the corpus using concepts instead of keywords. For example, instead of having to search separately for any mentioning of "Niue", "Nuku-ta-taha" or "Savage Island", the scholars search for the concept *Niue*. The results are ordered



by semantic clusters.

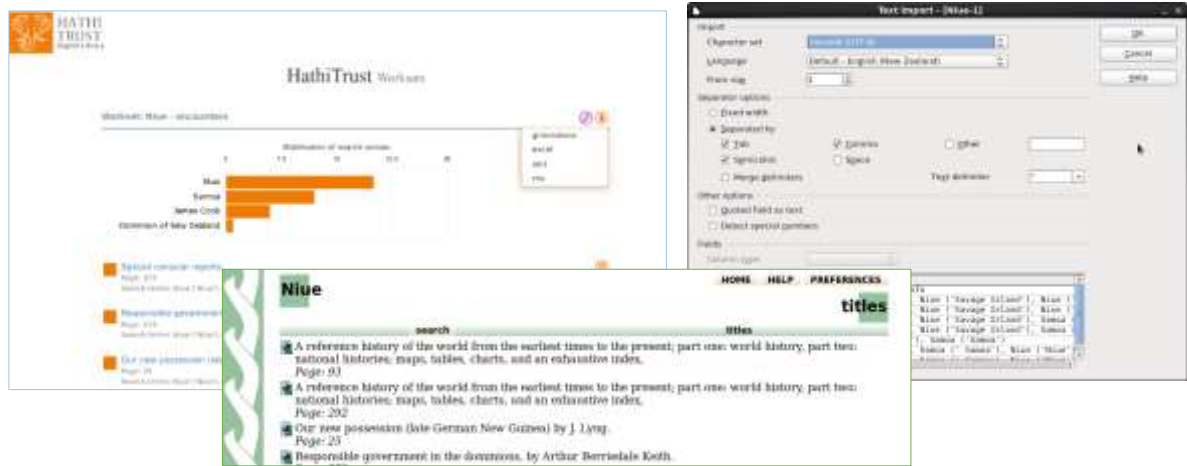
**Workset Explorer:** The results of the semantic search are presented not only as a list of documents, but also as graphic representation of clusters (proto-worksets). From here scholars can easily explore,

create and manipulate worksets, and select them for integration into formal worksets.



## Interoperability

Worksets created by a scholar or group of scholars can be exported and joined into existing data sets created from HathiTrust data or other external sources (e.g., Greenstone Digital Library, Excel, XML, CVS). The latter allows the scholar to incorporate worksets into a range of familiar bibliographic tools.



## Expert Scholar Tools

High quality semantic analysis requires manual adjustment of concepts such that marker terms and key concepts from a scholar's field are well represented (e.g., adding the name "Nuku-ta-taha" for concept *Niue*). We provide five expert tools to both explore and enrich the knowledge base.

**Synonym browser:** The Synonym Browser allows scholars to explore synonym words for a given concept (e.g., "Niue" and "Savage Island" for concept *Niue*); it provides a list and a graph view.





**Concept Browser:** The Concept Browser allows scholars to explore all links between concepts, contexts and terms. The scholar can interactively walk through the network as an expanding graph.

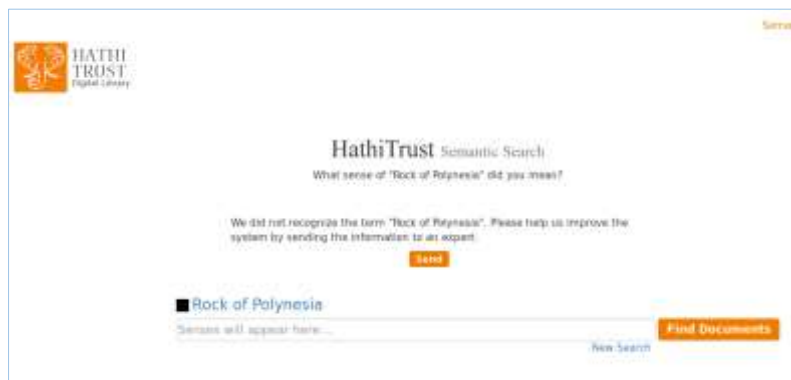


**Context Browser:** The Context browser shows synonyms and their context, for a selected concept (e.g., the concept *Niue*). They can be viewed through two different facets: by terms and by context.



**Synonym Adder:** For a given concept and a given context, the scholar can add new synonyms to an existing term. This provides the scholar with a first step into easy tailoring of the knowledge base. The Synonym Adder is integrated with the Context Browser (see above).

**Basket of Knowledge:** When a scholar uses terms or phrases that are currently not recognized through the semantic analysis, these can be submitted as a request into a *Basket of Knowledge*. This request is then submitted to expert scholars (knowledge workers), who then can include the terms and concepts into the Knowledge Base (the Concepts-in-Context network). Every scholar is assumed to be an expert in their specialization—submitting a request may therefore be effectively a ‘note to self’.



## Curation: Corpus Enrichment

Integration of the semantic links between concepts and documents as annotations or mark-up, e.g., into HathiTrust metadata, is possible via the curator export function of semantic information.



## Contributions of the project to Humanities Scholars

The Capisco System provides a number of benefits for scholarly search and workset creation:

- More documents are found through semantic search (fewer missed matches)
- Complex searches are made easy through avoiding repeated lexical searches
- Exclusion of results that match at word level but do not match desired semantics
- Workset creation and exploration by semantics (discover more about selected documents)
- Incorporation of scholarly knowledge through expert tools
- Semantics export into HathiTrust metadata (semantic enrichment of metadata)
- Interoperability through workset export
- Use cases of worksets: Small Nations (Niue) and Maori Astronomy
- Parallelizable software for non-consumptive document analysis (respects copyright)

## 4. Explanation of any setback or challenges

The current Capisco version is a proof-of-concept implementation, which fulfills the research aims for this project.

However, in order to transfer the software into a production system, a number of challenges need to be addressed to transfer the insights of this project into production software.

These were discussed during the project presentation (1 April 2015) and have been begun to be addressed after the project finished. They are listed as current limitations below.

### Current limitations

The semantic-enhanced search in Capisco provides user interfaces geared towards Humanities Scholars. The underlying software is divided in three main functions, some elements of which are proof of concepts.

1. CiC knowledge network creation: The knowledge network creation is a lengthy process but only needs to be executed once.

2. Disambiguation & indexing: Disambiguation and indexing was explored using two approaches: one building on a pre-existing software (faster but lower quality) and another one with improved semantic analysis quality. For this second software package, we explored semantic quality but not yet automation and scalability in a data-heavy environment.
3. Semantic-enhanced search of documents: The search function relies on the higher-quality indexing process in order to ensure best results.

In order to explain the details of current ongoing investigations into scalability and performance, we list the software elements of the Capisco backend. This part of the software (1.+2. above) is structured into three main pieces.

- Sentence parser: This component attempts to break blocks of text into what it can best identify as sentences.
- Context identification: This component attempts to identify the context (or context) of the body of text. This is the most computationally expensive component because it attempts to reconcile between context and concepts from a very large number of ambiguous terms. Performance can be dramatically improved if unambiguous context is provided as a parameter along with the body of text being analyzed.
- Disambiguation: This component disambiguates terms based on context/concept relationships. These relationships can be strong (if they are bidirectional) or weak if they are unidirectional.

Ongoing research is dealing with the computational complexity of the context identification. Among other issues, this complexity is increased by the noise in Wikipedia (reducing the performance of high-quality automatic identification of context). The various pieces outlined above are currently independent processes that are combined sequentially. This is currently done manually due to the investigative nature of the current research as this allows flexible testing of combinations of processes as part of both analysis and synthesis.

### Future research directions

A project like this always opens possibilities and avenues of future research. We outline here some of these that are particularly relevant to digital humanities.

Performance:

- Incremental indexing after knowledge base extension
- Improving query performance for very large corpora

Functional:

- Integration of semantic search into scholarly workflow,
- In-depth semantic workset analysis for scholars
- Scholarly expert tools for knowledge base manipulation

## 5. List of project disseminations and final deliverables

### Current project disseminations

- Annika Hinze, Craig Taube-Schock, David Bainbridge, Rangi Matamua, J Stephen Downie: “Improving access to large-scale Digital libraries through Semantic-enhanced Search and Disambiguation”, *Proceedings of the International Joint Conference of Digital Libraries*, June 2015
- Sally Jo Cunningham, Annika Hinze, David Bainbridge, Craig Taube Schock, Thomas Ryan: “Building heritage document collections for Pacific Island nations using semantic-enriched search”, *Proceedings of the Samoa III Conference*, March 2015

Further publications are in preparation.

### Final project deliverables

The following list gives an overview of the proposed project deliverables as per contract and the references to the respective deliverables.

1. Final report: this document
2. Well-documented source code for software for semantic clustering, concept browser, semantic search interface, work set explorer, integration tool, and curator tool:  
The software has been uploaded onto github ([github.com/HTrustProject/SemanFinal2015](https://github.com/HTrustProject/SemanFinal2015))
  - Capisco backend: knowledge network (CiC network) creation
  - Capisco indexer: disambiguation, document indexer
  - Capisco frontend: semantic search interface, workset explorer, synonym browser, concept browser, context browser, synonym adder, basket of knowledge)
3. Sample output demonstrating enhancements for workset creation:  
Sample outputs have been documented
  - in this report
  - in the two peer-reviewed publications, and
  - in the demo video (available at [youtu.be/2LiW\\_4X\\_6iU](https://youtu.be/2LiW_4X_6iU))
4. New worksets:
  - We created smaller worksets for Niue Cultural Heritage and Samoan Cultural Heritage.
  - The Niue Cultural Heritage workset features in the demo video.
  - Larger worksets are in preparation (available after scalability has been addressed).
5. Copies of or links to any scientific publications:
  - The copies are attached and further publications will be made available as they are published

## 6. Projected vs actual expenses

There are no differences between projected and actual expenses for the project.

# Distributed Metadata Correction and Annotation

*Workset Creation for Scholarly Analysis Prototyping Project*

## Distributed Metadata Correction and Annotation

University of Maryland  
Principal Investigator: Trevor Muñoz

### Project Description and Purpose

As part of the Workset Creation for Scholarly Analysis (WCSA) project led by the HathiTrust Research Center (HTRC), the Maryland Institute for Technology in the Humanities (MITH) proposed to develop a set of services and interfaces that would allow scholarly research teams to pull metadata records from the HathiTrust APIs, correct and annotate these records using standardized vocabularies, gather corrections and annotations from other application instances, and export them in formats suitable for publication as linked data. MITH also proposed to produce a demonstration of an index service that would allow research groups to register their data publications in order to make them available to other groups through a discovery interface.

The purpose of this project was to prototype simple tools that could support correction and enhancement of HathiTrust metadata by research teams who are building worksets and collections from the larger Hathi corpus for their own projects. Presently, teams of researchers might maintain a spreadsheet or custom database of metadata corrections and annotations that would be used internally for search and analysis. This approach requires limited infrastructure and can be effective for individual projects in the short term, but when it is adopted by many projects over many years it becomes extremely inefficient. MITH's prototype demonstrates how a distributed system of data publishing, annotation, and sharing might be created around the HathiTrust by leveraging existing and emerging linked data standards such as [Open Annotation](#) (OA), and [CSV on the Web](#) (CSVW).

### Summary of Project Accomplishments

MITH accomplished all project goals within the period of performance. Over the course of two cycles of iterative design and development the MITH team created an initial prototype application, tested this prototype with a group of scholarly target users, radically simplified the prototype design for metadata correction based on user feedback, and finally created a demonstration version of an accompanying web application for registering, discovering, and publishing enhanced sets of HathiTrust metadata.

MITH's partners in developing the distributed metadata correction and annotation prototype were members of the Foreign Literatures in America (FLA) project research team. The FLA project aims to challenge conceptions of "American literature" that turn upon the American citizenship of an author, since historically it is clear that foreign authored works—as well as works by immigrant authors who wrote in many languages and were not citizens of the United

States—have constituted an important part of the literatures and cultures of the U.S. Many of the primary materials that are the focus of the project are included in the holdings of the HathiTrust Digital Library, but discovery and analysis of these volumes can be challenging because of incomplete and inconsistent metadata. Transliterations of personal names and translations of titles can vary widely for a single work, as can the bibliographic fields used to identify translators, editors, and writers of introductions—all information that is essential to this kind of study of reception history. The non-standard and inconsistent encoding of places and dates of publication in library metadata also raises problems for the study of regional and temporal variation in the reception of foreign authors, which is an important focus of the FLA project.

### Initial Prototype

The first phase of project work, which led to the initial prototype, focused on those portions of the overall workflow dealing with acquisition of HathiTrust metadata through data correction and enhancement. (Support for modeling, publishing, and distributing such changes was not part of this initial phase.)

Members of the FLA team built worksets of materials related to their research using the existing collection building tools provided by the HathiTrust Digital Library. The first version of the MITH team's prototype utilized a software library originally created by Travis Brown for the Princeton Prosody Archive to download and reformat bibliographic metadata for volumes contained in the sample public worksets created by FLA. This metadata was loaded into an instance of the Open Refine software running on a central server. Interactions with the Open Refine application were brokered by a web application created by MITH to provide functionality to log in different users for tracking provenance information about various data changes.

### User Testing

The MITH team conducted a small user study of this initial prototype with the FLA team members who had created worksets. FLA team members were asked to interact with metadata for volumes in their worksets loaded in the Open Refine application. MITH staff provided a basic tutorial and walk through for using Open Refine. During the testing, a basic “speak aloud” protocol was used in conjunction with screencasting software to record both participants' impressions as well as on-screen actions. Users were asked to review metadata records, and, where appropriate, attempt to make corrections or additions based on their existing subject knowledge as well as any project-specific conventions (e.g. related to transliteration of certain names).

The results of user testing were clear—even with a very small sample. Participants reported being confused by the presentation of information from different metadata fields as the information appeared in the Open Refine interface. This presentation resulted from constraints in transforming MARC metadata, which, generally speaking, contains series of nested fields into flat rows and columns in Open Refine's spreadsheet-like interface. Participants were also

uncertain how to add additional data through Open Refine, which may have been due to insufficient training on the program’s basic functionality unrelated to this test. Also, at this stage, the MITH team had not determined a solution for modeling added data so participants could not be given much specific guidance. User testing focused attention on the challenges—but also the vital importance—of moving from data representations suited to library use-cases (MARC serializations and/or complex RDF) to representations of data that researchers found more approachable—tabular, or spreadsheet-like formats.

### Second Prototype

Given the user feedback and technical challenges arising from the architecture of the initial prototype (discussed further below), the MITH team decided to revise its prototype. The second prototype developed from work on the demonstration of an index service for corrected and enhanced metadata. Consequently, this prototype radically minimized the tooling related to acquisition and basic correction/enhancement of metadata.

The MITH team proposed a process for distributed metadata correction and annotation that uses existing HathiTrust Digital Library and HathiTrust Research Center tools as well as existing “off-the-shelf” applications such as Open Refine. New software developed for this project serves to create workflows for using such tools in concert, supported by open standards.

This workflow entails:

- Users create worksets through existing mechanisms provided by HathiTrust Digital Library and HathiTrust Research Center. Leveraging these functions means that basic user account management, authentication for protected resources, and large-scale full-text search are all already provided
- To support downstream components of the workflow, MITH proposed that HTRC could make workset metadata available for download in CSVW format. Presently, bibliographic metadata for materials in worksets can be downloaded in Comma Separated Value (CSV) format. For the purposes of demonstration, MITH used the HathiTrust Bibliographic API to download metadata and make it available in CSVW format (see <https://umd-mith.github.io/fla-metadata/>). CSVW format allows creators to provide a schema for CSV that associates columns of tabular data with semantic predicates thereby facilitating processing of tabular data into other linked data standards such as RDF.
- Users register this original metadata with the prototype web application (<https://github.com/umd-mith/csvwww>) MITH created for this project. This registration step captures basic administrative information—such as the source of data—used for basic provenance tracking. Though it falls outside the scope of the current prototype, this function could also support harvesting and discovery of enhanced metadata by HTRC or other research teams.
- In the second prototype, users are free to edit or enhance metadata however they see fit, using any application or custom programming. (For the purposes of testing, MITH

continued to use Open Refine to perform data corrections). The only requirement is that changed data be saved back to the original CSVW document.

- Users upload changed data back to the prototype web application which tracks changes to the data and creates a log. When enhanced or corrected data is uploaded users are asked to describe changes—in the manner of a “commit log.”
- MITH’s prototype application captures changes to data and accompanying commit messages and incorporates this information as annotations in OA format. These linked data annotations are henceforth included as part of the CSVW metadata and could be acted upon by future applications to display changes to metadata or potentially to merge data sets edited by different groups of users.

The second prototype was presented and demonstrated at HathiTrust UnCamp 2015, however, there was insufficient time in the period of performance for a second round of user testing with the original FLA team of data creators. Testing and refinement of a system based on the second prototype would be an important component of any future development.

## Challenges Encountered

In developing a prototype workflow for distributed correction and annotation of HathiTrust metadata for this project, the MITH team encountered a number of challenges. The most significant challenges are listed here:

- Metadata for HathiTrust materials are currently available in MARC format. Linked data representations of HathiTrust metadata are under development but were not available to work with during the period of performance. The in-process state of metadata in linked data formats challenged the ability of the MITH team to precisely target annotations representing corrections or enhancements to metadata. In the prototype, MITH elected to use fragment identifiers tied to CSVW serializations of HathiTrust metadata to provide referents for annotations. Additional processing of CSVW data would be required to translate such annotations into forms usable for updating HathiTrust data directly.
- The complete information represented in bibliographic metadata is often overwhelming to non-library researchers interested in data for their own projects. In practice, this means that more prescriptive choices to be made in serializing HathiTrust metadata for correction and enhancement by researchers. Flattening complex hierarchical structures and eliminating irrelevant fields should be two important considerations but further work is needed on the usability of various metadata serializations for different use cases.
- The area of data correction and enhancement suffers from both immature tools and tools that may be reaching their end of life without clear prospects for upgrades or support. CSVW is an important emerging standard that offers great promise for translating between tabular and graph-based representations of data but it is an emerging



standard—currently only one implementation exists for “distilling” CSVW to RDF. Likewise, tools to support the production and consumption of Open Annotation linked data have not been adapted to this data pipeline use case. There are also newer entrants into this class of tools—for example, Dat, a version-controlled, distributed data tool (<http://dat-data.com/>), but it is too early to judge their potential application. Finally, the MITH team remains concerned about the sustainability of Open Refine, the most user-friendly data correction and enhancement tool available, now that it is once again a community-maintained project with relatively low levels of active development.

## Project Dissemination and Deliverables

All software source code produced for this project is available on GitHub at the following locations:

<https://github.com/umd-mith/csvwww>

The main prototype for this project, `umd-mith/csvwww`, is a web application that lets implement a workflow for distributed metadata correction and annotation. Users load in data sets in CSVW format, and publish changes to them using Web Annotation standards.

<https://github.com/umd-mith/fla-metadata>

A set of sample metadata derived from the HathiTrust Digital Library Bibliographic API and formatted as both CSV and CSVW documents.

<https://github.com/umd-mith/hathitables>

A command line utility for generating CSVW formatted data for worksets built through either the HathiTrust Digital Library or the HathiTrust Research Center collection build. Used to create the sample metadata found at: <https://github.com/umd-mith/fla-metadata>

<https://github.com/umd-mith/hathilda>

The underlying Python library used by the `umd-mith/hathitables` utility. This library contacts the HathiTrust Digital Library Bibliographic API, downloads volume metadata, and serializes this metadata as JSON-LD.

All software source code is licensed for redistribution and reuse under Open Source Initiative (OSI) approved licenses.

Principal Investigator Muñoz presented on the Distributed Metadata Correction and Annotation prototype at the HathiTrust Research Center UnCamp 2015, held March 30-31, at the University of Michigan. Presentation files (as well as interim and final reports from this project) may be found in the University of Maryland institutional repository at:

<http://hdl.handle.net/1903/14717>

## Changes to Personnel and Project Management

The key personnel of the Maryland team changed substantially between the initial proposal and the period of performance.

Original Principal Investigator Travis Brown and MITH Software Architect James Smith both left the University of Maryland to pursue other opportunities shortly after the prototyping grant was awarded. Brown was to have been responsible for overall direction of the project and the development of project software, data models, and data sets. Smith was to have provided consultation on data modeling and conducted reviews of software code.

Trevor Muñoz, Associate Director of MITH, assumed Principal Investigator duties. Muñoz was responsible for the overall direction of the project and consulted on data modeling. Development of the first iteration of project software was led by Raffaele Viglianti, MITH's Research Programmer. Development of the second iteration of project software, as well as of project data models, and data sets was led by Ed Summers subsequent to his joining MITH as Lead Programmer in September 2014.

As initially projected, Stephanie Sapienza, MITH's Project Manager, scheduled and managed project meetings and tracked progress toward deliverables. Sapienza also led a key user testing session to evaluate the first iteration of the project prototype with Humanities Consultant, Dr. Peter Mallios, and members of the FLA research team.

## Projected vs. Actual Expenses

Budgeted expenses for this project were \$39,690. Grant award was \$36,690. Actual expenditures amounted to \$36,689.01.

Actual expenses varied from projections in the initial project budget in the following ways: 1) Personnel salaries were redistributed after the departure of Brown and Smith (described above); 2) Due to personal circumstances, Peter Mallios's time was curtailed due to personal circumstances and funds projected for his stipend were reallocated; 3) Expenditures for domestic travel for Muñoz to present on the project at the HathiTrust UnCamp 2015 were not included in the original budget; 4) Demand for Amazon Web Services computing resources was less than anticipated and the project did not make use of funds for this purpose.



Principal Investigator

06-15-2015

Date

# EIEPHãT: Early English Print in the HathiTrust a Linked Semantic Worksets Prototype

## Final report

Principal Investigator: Kevin Page, Oxford e-Research Centre, University of Oxford

Co-Investigator: Pip Willcox, Bodleian Libraries, University of Oxford

*The EIEPHãT project – Early English Print in HathiTrust, a Linked Semantic Worksets Prototype – demonstrates the use of Linked Data for combining, through worksets, information from independent collections into a coherent view which can be studied and analyzed to facilitate and improve academic investigation of the constituents.*

## 1. Background

Early English Books Online Text Creation Partnership (EEBO-TCP) is a partnership with ProQuest and more than 150 libraries and universities, led by the universities of Michigan and Oxford, to generate highly accurate, fully-searchable, XML-encoded texts corresponding to books from ProQuest’s Early English Books Online database.

The EEBO corpus consists of the works represented in the English Short Title Catalogue (STC) I and II (based on the Pollard & Redgrave, and Wing short title catalogues), as well as the Thomason Tracts and the Early English Books Tract Supplement. Together these trace the history of English thought and learning from the first book printed in English in 1473 through to 1700. From 2000-2009, EEBO-TCP Phase I successfully converted 25,000 selected texts from the EEBO corpus into TEI-compliant, XML-encoded hand-transcribed texts, which were made freely available to the public in January 2015.

## 2. Project objectives

The EIEPHãT project focuses on the potential symbiosis between two datasets as an illustration of how the Semantic Web (including RDF, SPARQL and ontologies) can provide a technological foundation for applications grounded in scholarly investigation (figure 1). The first dataset is EEBO-TCP, as described above; and the second is a custom HathiTrust dataset consisting of all materials described in their metadata as being in English and published between 1470 and 1700, with variable bibliographic metadata inherited from the contributing institutions.

The key objectives of EIEPHãT were to generate an RDF export of metadata extracted from EEBO-TCP which, when combined with RDF from the HathiTrust Research Center using appropriate ontologies, could be employed to create prototype interfaces through which a scholar might perform seamless investigations over the combined linked dataset.

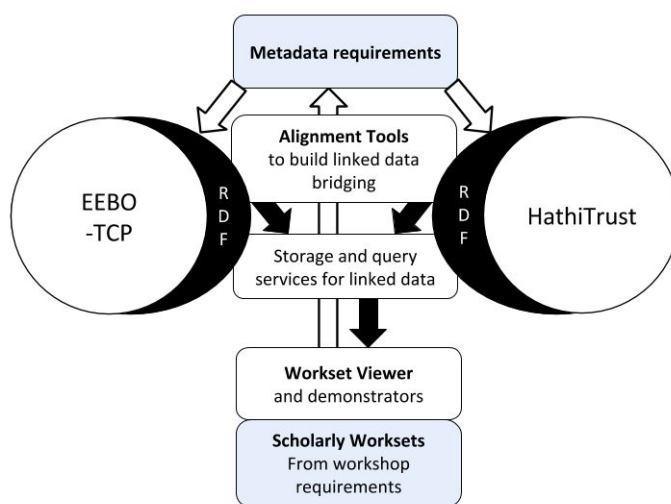


Figure 1. EIEPHãT overview

### 3. Project personnel

The EIEPHãT project commenced in June 2014 and was completed in April 2015. The principal investigator of the project was Dr Kevin Page of the University of Oxford e-Research Centre, with Pip Willcox, Curator of Digital Special Collections at the Bodleian Libraries, as his co-investigator. Research effort was concentrated within the Oxford e-Research Centre, with Dr Terhi Nurmikko-Fuller joining the team in September 2014 as the primary researcher; John Pybus, a senior researcher in the Centre, contributed to initial requirements analyses, and valuable input was made by David Weigl during the latter stages of the project and during deployment of the SALT software (section 4.4), of which he is the core developer.

The project team reported to a Technical and Scholarly Steering Committee for advice and guidance, which met 14 times over the duration of the project (approximately fortnightly) and was formed of experienced members of the e-Research Centre and Bodleian Libraries: Professor David De Roure (Director of Oxford e-Research Centre), and the Bodleian Libraries' Michael Popham (Head of Digital Collections and Preservation Services), Dr Christine Madsen (Head of Digital Programmes), and Judith Siefring (Digital Editor).

### 4. Project activities and accomplishments

The project has undertaken a number of research activities, often in parallel, in pursuit of the above objectives. We report on these outcomes in the following subsections.

#### 4.1 Consultative workshop: validation of workset requirements and utility

On 12 June 2014, the EIEPHãT team organized a consultative workshop with academics and other potential audience members for the tools created by the project. From the outset we had planned that scoping and prioritization of the project's technical developments were to be informed by actual scholarly questions guided by the community, and iterated in consultation with them.

To this end, we invited 12 colleagues from 5 universities (Bath Spa, London, Oxford, Oxford Brookes, and Reading), two libraries (the Bodleian and the British Library), and a publisher (Oxford University Press) to the Oxford e-Research Centre for a day of discussion, formal and informal. As well as our Technical and Scholarly Steering Committee members, librarians, digital curators, and publishers, our guests included academics across career stages, from students to professors.

EIEPHãT was put into its wider context before the group broke out into smaller discussion groups to consider the questions we had raised about how the tools might best work for them. The groups also considered the sample academic questions outlined in the project proposal, those highlighted in the SECT<sup>1</sup> project, and came up with research questions of their own that were refined during the closing plenary discussion. Attendees also filled in a questionnaire, to give us quantitative data to guide our decisions. The day's discussions influenced softer aspects of the project as well as its motivating technical questions, for example the need for an interface for refining worksets that humanists would be at home using. Several of the academics have kept in close touch with the project. Two (from Bath Spa and Reading) returned for a consultation with Dr. Nurmikko-Fuller in the autumn. One of these is now incorporating more digital humanities and curation training in his undergraduate teaching. He also came to the EEBO-TCP hackfest (March 2015) and was the academic lead on one of the prize-winning teams.

---

<sup>1</sup> <http://blogs.bodleian.ox.ac.uk/eebotcp/files/2012/05/SECTWorkshopSummaryReport.pdf>

A more detailed report of the workshop can be found as an appendix to this report.

## 4.2 Technical analysis of metadata requirements and ontology selection

In light of the domain motivations articulated at the scholars' workshop, a technical analysis of the metadata requirements to support scholarly worksets bridging the two datasets was undertaken. In collaboration with the WCSA team, we surveyed both the addressable resources and the schema expressivity of ontologies that could potentially be used to parameterise these types of workset, including MODSRDF, Bibframe, Schema.org and FRBRoo. A full description of this work can be found in Nurmikko-Fuller *et al* <sup>2</sup>.

We then identified parsable information structures and terms in the EEBO-TCP Text Encoding Initiative (TEI) data which were suitable for parameterizing worksets and, informed by the survey, selected ontology elements used to encode this EEBO-TCP metadata, which would be compatible (or at least, for our purposes, comparable) with the RDF structures being generated for HathiTrust metadata by the WCSA team. The resultant ontology collection - the Early English Books Online Ontology, or EEBOO - includes selections from MODS, Bibframe, and PROV, along with custom elements to encode additional structures (e.g. dates); and in addition encodes worksets using the WCSA workset ontology or, alternatively, the Research Object (RO) model. EEBOO can be found in the project github repository.

## 4.3 Tools for generation of custom RDF derived from EEBO-TCP

Several tools and scripts were re-used or developed to generate RDF triples, conforming to EEBOO, from the derived from EEBO-TCP TEI headers. These tools and configurations can be found in the project github repository.

First a set of python scripts was developed to process the TEI P5 XML, then the Karma Data Integration Tool<sup>3</sup>, from the University of California Information Sciences Institute, was used to map the EEBO-TCP data structures into the EEBOO ontology and output RDF. Particular attention was paid to dates encoded within strings in EEBO-TCP, as an illustration of rich semi-structured data that can be extracted into structured RDF (there are numerous other similar metadata elements that could form the basis of extension work). These were matched using regular expression rules and parsed for encoding using corresponding temporal extensions to EEBOO. We matched 98 distinct expressions of author date using 75 regular expression, and 1,510 distinct expressions of publication date, parsed into 54 types. Finally RDF links to author records in VIAF and the Library of Congress, and multimedia pages for texts in JISC Digital Books and HathiTrust, were generated and added to the EEBOO triples.

We took 24,926 EEBO-TCP Phase 1 records with 22 distinct types of information in the headers, including 6 different ID numbers and 3 main types of date (publication date of historical work, author associated historical date(s), XML publication/editing dates). EEBOO incorporates 7 of these datatypes, and extends into subcategories for author names and different types of date. After processing the EEBO-TCP headers, EEBOO contains 713 unique places, 6,489 unique expressions of Person of which 3,588 have VIAF and LoC IDs.

---

<sup>2</sup> Nurmikko-Fuller, T., Page, K., Willcox, P., Jett, J. Maden, C., Cole, T., Fallaw, C., Senseney, M., Downie, J.S. 2015. Building Complex Research Collections in Digital Libraries: A Survey of Ontology Implications. In Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (Knoxville, USA, June 2015).

<sup>3</sup> <http://www.isi.edu/integration/karma/>

#### 4.4 Alignment of instance data between EEBOO and HathiTrust using ‘SALT’

The Semantic Alignment and Linking Tool (SALT) was developed at the Oxford e-Research Centre to simplify the process of confirming alignments between datasets, prioritising potential matches to bring them to the attention of the scholar or expert by using both string and semantic-contextual distances. SALT is designed on the basis of assisting, not completely automating, the process of making a scholarly judgement during alignment. SALT accesses data held in RDF triplestores using a semantic configuration for the (potentially multiple) schemas used; it then writes alignments and match decisions back to a triplestore enabling layering of assertions and generation of provenance. SALT was configured to access both the EEBOO and HathiTrust custom dataset RDF, then used to align author names between the two (Figure 2).

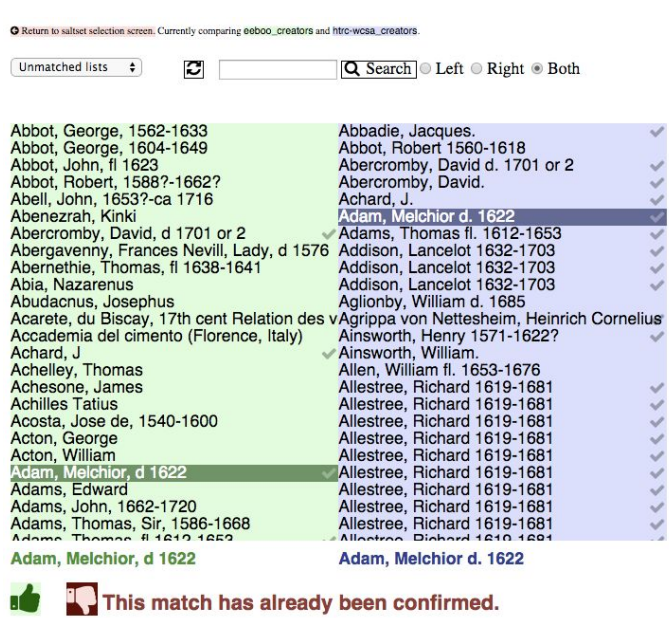


Figure 2. Aligning EEBOO and HathiTrust data in SALT

#### 4.5 Provisioning an RDF triplestore populated with EEBOO and HathiTrust RDF

Having collected and generated the RDF triples as outlined in the previous sections, we populated these datasets into a Virtuoso<sup>4</sup> RDF triplestore which could provide the necessary SPARQL query interface<sup>5</sup> upon which the user interfaces (described in the following section) are built. The triplestore was initially configured on several development workstations, then migrated to a server for final demonstration, and ultimately contained named graphs for EEBOO, alignment and match decisions from SALT, a copy of the WCSA-produced graph of the HathiTrust custom data set, and worksets as generated by the prototype UIs. While the named graphs are all currently all co-located within one Virtuoso instance, the design of the tools in Section 4.6 is such that they could be transparently distributed across different hosts and institutions should this be desired.

The EIEPHät Virtuoso triplestore contains 1,137,502 triples, consisting of: 251,725 entities, 66 distinct classes, 214 distinct predicates. 287,581 distinct subject nodes and 294,677 distinct objects. A breakdown by named graph yields: 468,022 EEBOO triples, 341,949 triples aligning EEBOO and HathiTrust, 309,257 triples from the WCSA produced HathiTrust custom dataset RDF, 11,168 SALT match decisions, and 1,354 triples for workset descriptions.

#### 4.6 Implementation of prototype architecture and user interfaces for workset construction and viewing

The data extraction, ontological, and alignment outputs outlined in the previous sections enable us to make SPARQL queries to the triplestore that semantically match and return works from both EEBOO and the HathiTrust custom dataset, and which build provide the types of investigative construction requested at the

<sup>4</sup> <http://virtuoso.openlinksw.com/>, <https://github.com/openlink/virtuoso-opensource>

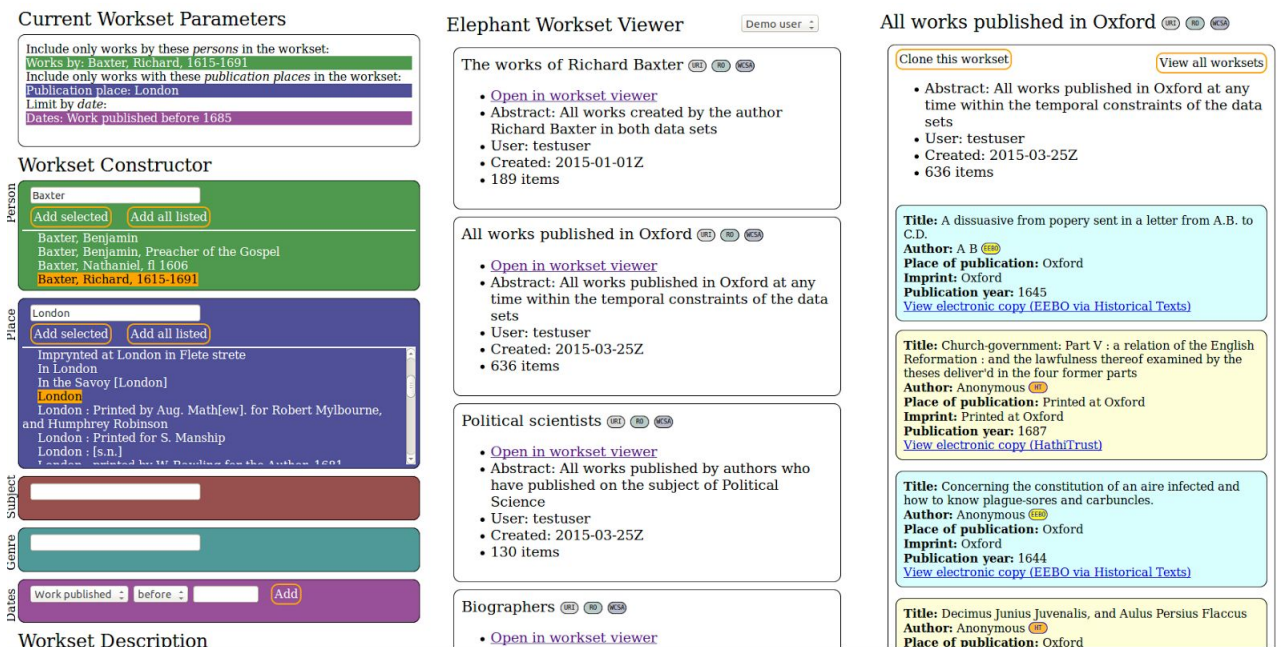
<sup>5</sup> The SPARQL interface is at: <http://eeboo.oerc.ox.ac.uk/sparql> ; please email [kevin.page@oerc.ox.ac.uk](mailto:kevin.page@oerc.ox.ac.uk) for access.

consultative workshop. The following are paraphrases of SPARQL queries that will successfully return results from the EIEPHãT architecture and datasets:

- Find all the works, appearing in both datasets, which have been written by Richard Baxter.
- Find works in both datasets that have been published in Oxford.
- Find works which were published outside of London (where the bulk were published).
- Find works from both datasets that were published outside of London in the mid-to late 1600s.
- Find all the works contained in the two datasets for authors who have at least once published on the subject of “Political science”<sup>6</sup>.
- Find all the works contained in these two datasets for authors who have at least once published works which are categorised as “biography”.

SPARQL queries such as these provide the basis for our prototype user interfaces: a workset *constructor* that allows a scholar to specify a number of attributes by which a workset is assembled; and a workset *viewer* that enables the scholar to inspect the contents of a workset and follow links to further resources (e.g. the EEBO-TCP text or images).

The workset constructor (figure 3) defines a workset using parameters selected by the user which are then, in the background, assembled into the SPARQL queries used to create a workset. The interface is automatically populated with valid potential attributes that are themselves retrieved from the triplestore using SPARQL queries, using ontological terms asserted to have equivalent meaning between the datasets. The workset viewer (also figure 3) then retrieves RDF workset contents, record metadata, data links, and multimedia links (to the JISC Historic Books collection or the HathiTrust Digital Library). The underlying workset data can be retrieved according to the draft WCSA workset ontology, or using the Research Object model, and small ‘beans’ are used throughout the interface to illustrate where an element has an underlying EEBOO or workset RDF URI.



The screenshot displays three main components of the EIEPHãT interface:

- Current Workset Parameters:** A list of filters including 'Works by: Baxter, Richard, 1615-1691', 'Include only works with these publication places in the workset: London', 'Limit by date: Work published before 1685', and 'Dates: Work published before 1685'.
- Workset Constructor:** A form with sections for Person (Baxter, Benjamin, Nathaniel, Richard), Place (London), Subject, Genre, and Dates. It includes 'Add selected' and 'Add all listed' buttons for each section.
- Elephant Workset Viewer:** A 'Demo user' interface showing three worksets:
  - The works of Richard Baxter:** 189 items, created 2015-01-01Z.
  - All works published in Oxford:** 636 items, created 2015-03-25Z.
  - Political scientists:** 130 items, created 2015-03-25Z.
  - Biographers:** 130 items, created 2015-03-25Z.
- All works published in Oxford:** A list of individual works with details such as title, author, place of publication, and year. Examples include:
  - Abstract:** All works published in Oxford at any time within the temporal constraints of the data sets.
  - Title:** A dissuasive from popery sent in a letter from A.B. to C.D.
  - Title:** Church-government: Part V : a relation of the English Reformation : and the lawfulness thereof examined by the theses deliver'd in the four former parts.
  - Title:** Concerning the constitution of an aire infected and how to know plague-sores and carbuncles.
  - Title:** Decimus Junius Juvenalis, and Aulus Persius Flaccus

Figure 3. (Left to right) Workset constructor; workset viewer (overview and specific)

<sup>6</sup> It is of particular note that this returns results across both datasets since our EEBO-TCP import contained no genre or topic information; this associated is entirely inferred from the semantic links.

Both web applications are written in Python, using the Flask framework, and both rely heavily on the semantic information encoded in RDF and queried using SPARQL. Figure 4 shows the overall composition of the final EIEPHãT architecture. The online prototype<sup>7</sup> has been populated with exemplar worksets based on the scholarly questions listed above.

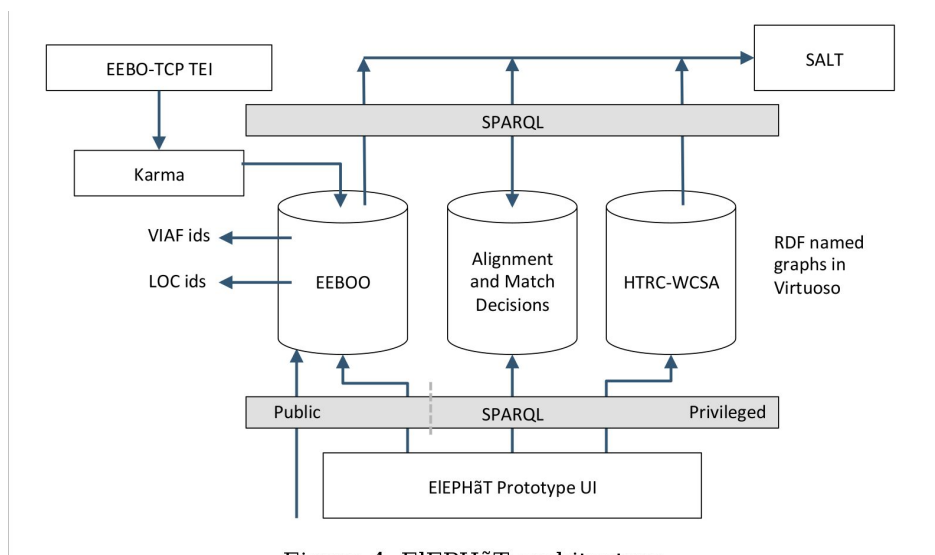


Figure 4. EIEPHãT architecture.

## Summary

The EIEPHãT prototype has demonstrated the utility of using RDF and Linked Data to bring together HathiTrust content with ‘boutique’ collections such as EEBO-TCP, providing a template for how this can be achieved. While the project has highlighted the potential complexity of generating structured metadata from new sources, we believe it also shows how RDF provides a common ‘middle ground’ that can decouple the process of aligning items, from the preparatory re-encoding (or restructuring) of external corpora that are to be incorporated or linked. EIEPHãT also highlights that, using RDF, this can be beneficially performed in a piecemeal or iterative fashion - an obvious future addition being the RDF encoding of other EEBO-TCP headers including remaining date fields and imprints. In linking between HathiTrust and EEBO-TCP we have also demonstrated how either – or both – of these might be further linked using tools such as SALT. In this sense, we have only scratched the surface of the information held here, not to mention the possibilities offered by the content of the texts; while EIEPHãT was conceived and implemented as a limited prototyping project, we believe the methods it has proved can, and should, be more widely deployed in both scope and ambition.



Kevin Page  
University of Oxford e-Research Centre  
Oxford, 15th June 2015

<sup>7</sup> <http://eeboo.oerc.ox.ac.uk/>



## Dissemination

### Publications

Nurmikko-Fuller, T., Page, K., Willcox, P., Jett, J. Maden, C., Cole, T., Fallaw, C., Senseney, M., Downie, J.S. 2015. *Building Complex Research Collections in Digital Libraries: A Survey of Ontology Implications*. In Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (Knoxville, USA).

### Talks

*Making Links: Connecting humanities resources for scholarly investigation*, International Conference on Digital Humanities in Japan, JADH2015, Kyoto University, September 2015.

*Trunk to tail: linking EIEPHãTs through the Semantic Web*, Centre for Digital Scholarship seminar, University of Oxford, June 2015.

*EIEPHãT: Early English Print in HathiTrust, a Linked Semantic Worksets Prototype*, HathiTrust Research Center Uncamp, University of Michigan, March 2015.

*Reborn Digital: coding text*. COST Action IS1310: Reassembling the Republic of Letters, March 2015.

*Early English Print in the HathiTrust—“Elephant, That Wonder in Nature”*: A Linked Semantic Workset Prototype.’ Chicago Digital Humanities and Computer Science Colloquium, Northwestern University, October 2014. Invited paper.

*Stories at Scale: linking data, linking ideas in EIEPHãT and Cultures of Knowledge*. Scientia Quantitatis: Quantitative Literaturwissenschaft in Systematischer und Historischer Perspektive. Schloss Herrenhausen, Hannover/Deutsches Literatur Archiv, University of Stuttgart, October 2014. Invited paper.

*“The Author’s Drift”*: scholarship, scale, and society. Research and/as Engagement. University of Edinburgh, September 2014.

*Exercising ourselves “in the Analysis of many examples”*: corpora, collaboration, and communication, a digital humanities masterclass at the Institute of Advanced Studies, University of Western Australia, August 2014

### Outreach and teaching

EIEPHãT consultative workshop, Oxford, June 2014 (see section 4.1 and attached report).

The EIEPHãT team attended the EEBO-TCP Hackfast at the Bodleian Weston Library in March 2015, and made the EEBOO triplestore available to participants. During the event we produced an initial alignment between EEBOO and the Oxford University Press ‘Early English Authorities’ data.

At the Digital Humanities at Oxford Summer School in July 2015, EIEPHãT will be disseminated in two workshops (Digital Approaches Medieval and Renaissance studies; and Humanities Data: Curation, Analysis, Access, and Reuse) and will form the basis of examples throughout the week in a third workshop, on Linked Data.

The Oxford Illinois Digital Libraries Placement Programme (OIDLPP) will host a student in the Oxford e-Research Centre between July and August 2015, who will work on a project extending the EIEPHãT data to include information from EEBO-TCP imprint fields.

## Deliverables

1. Report on consultative workshop (attached as an appendix).
2. Software to produce RDF from EEBO-TCP and support alignment with HathiTrust and external entities (code to be uploaded to github repository). Including Python scripts, regular expressions, Karma documentation/configuration (as described in section 4.3).
3. Documentation/configuration of SALT for use with EEBOO and the HathiTrust custom dataset (to uploaded to the github repository; as described in section 4.4).
4. Software implementing the workset constructor and viewer interfaces (to be uploaded to the project github repository; as described in section 4.6).
5. This final project report.

The github code repositories for EIEPHãT will be populated during summer 2015 under the project space at: <https://github.com/oerc-elephant>

## Explanation for variance of costs

The attached financial summary shows an overspend in travel expenses compared to the original, relatively small, travel budget. This is due to payment of expenses towards presenting the accepted JCDL paper (see Publications, above) at conference; and of underestimated costs for travelling to the final WCSA reporting meeting (this was originally budgeted for a single day event in Chicago; additional costs were incurred due to length and location of the 3 day event in Ann Arbor including the HathiTrust UnCamp, and coincidental timing with a major UK holiday resulting in high flight costs). The overspend in travel was compensated by lower than budgeted personnel costs, maintaining the overall budget and levels of effort and delivery.



# **Appendix I**

## **Report on the consultative workshop**

A MEMORY OF ELEPHANTS



CONSULTATIVE WORKSHOP REPORT

12 June 2014

University of Oxford e-Research Centre



# Consultative Workshop Report

## Summary

From the EIEPHaT project's inception we wanted its outputs to be useful to stakeholders inside and outside the project. To this end, we organized a consultative workshop with representatives of the scholarly community as one of the first activities of the project. We invited participants from a variety of academic backgrounds, including editors and book historians, and colleagues from libraries and publishers.

The day was arranged to exchange knowledge. Kevin Page, Judith Siefring, and Pip Willcox introduced participants to the [HathiTrust](#), to [Early English Books Online Text Creation Partnership](#) (EEBO-TCP), Linked Data, the Semantic Web, and some outcomes of the [Sustaining the EEBO-TCP Corpus in Transition](#) (SECT) project. We designed breakout and roundtable discussions to gather qualitative data, and a questionnaire to gather quantitative data. These data, and continuing correspondence with participants from the day, informed the decisions we made as to what to develop in the project.

The starting point for many of the participants' scholarly questions was metadata, particularly authors, dates, and imprint information. This guided the areas of our technical development throughout the project. The workshop led to subsequent correspondence and meetings with participants, particularly following the employment of the project's Research Associate, Terhi Nurmikko-Fuller. The workshop was a successful illustration of the importance of scholarly community building and engagement.

## Project Overview

The EIEPHaT project demonstrates the use of Linked Data for combining, through worksets, information from independent collections into a coherent view which can be studied and analyzed to facilitate and improve academic investigation of the constituents.

The project has been a joint endeavour between the University of Oxford e-Research Centre, where software development was undertaken, and the Bodleian Libraries, which brought expert understanding of the collections. The project began with a consultative workshop with scholars from the field, and progressed over 9 months with fortnightly software sprints, each planned and evaluated by a Technical and Scholarly Steering Committee comprising experienced members from the e-Research Centre and Bodleian.

The project produced software to expose the necessary metadata from the individual collections so that it can be used to parameterize aggregate worksets, in accordance with the model being developed as part of the primary WCSA award. This presented a number of challenges both in the structuring and alignment of the metadata internally and with external resources, and the development of software to investigate and capture alignment, links, and analysis, which the project also developed.

The technical challenges were motivated and scoped through focus on a particular area of academic research: the study of early English books (1473-1700). This is a field in which the project team has significant familiarity and expertise.

The prototype built integrated worksets drawing resources from the HathiTrust and from the Early English Books Online Text Creation Partnership (EEBO-TCP) collection, which is co-led by the Bodleian Libraries and the University of Michigan Library. EEBO-TCP provides an interesting and complementary corpus to HathiTrust, being focussed upon high quality images and accurate transcriptions of items which are often found in libraries' special collections. Preliminary investigation showed a significant intersection between the collections, indicating fruitful potential avenues for investigation of alignment and linking through worksets, including editions that span the collections, and networks of common authors, publishers, and locations.

### Area of Scholarly Study: Early English Print

We elected to focus our prototyping work on the study of, and in relation to, early English books, from the beginning of print in English (1473) to 1700. While there were many generic technical challenges to investigate and overcome in developing a workset demonstrator, scoping our work to early English print brought a number of specific benefits, while enabling us to illustrate the generalizable benefits of our approach, particularly the use of Linked Data.

### Motivation

The technical challenges of the project are best viewed through the lens of use cases motivated by actual scholarly questions. Such questions enabled the scoping and prioritizing of technical requirements according to needs which can be trialled, evaluated, and refined in an iterative and incremental manner—as opposed to a monolithic, predetermined design. To develop these questions, we needed to consult researchers, students, and librarians who use these texts “in the wild”.

Identifying workset problems from a scholarly perspective in turn helped identify the metadata required to parameterize that workset. This helped prioritize the elements of RDF encoding and alignment between collections and external entities, which enabled effective use of these parameters.

### Participants

We were delighted to benefit from the expertise of colleagues from eight institutions, and from a range of academic and professional backgrounds, including book historians, crowd-sourcing and citizen science experts, digital curators, digital humanists, early modern historians, editors, literary scholars, special collections librarians.

- Jacqueline Baker, Oxford University Press
- James Baker, British Library
- Rebecca Bullard, University of Reading
- Stephen Gregg, Bath Spa University
- Ben Higgins, University of Oxford
- Rupert Mann, Oxford University Press
- Jane Potter, Oxford Brookes University
- Kirsty Rolfe, University College London
- Matthew Symonds, University College London
- Victoria van Hyning, University of Oxford
- Sarah Wheale, Bodleian Libraries
- Elizabeth Williamson, University of Oxford

Pip Willcox (EIEPHãT Co-investigator) led the workshop. Kevin Page (EIEPHãT Principal Investigator), and David De Roure and Judith Siefring, from the project’s Technical and Scholarly Steering Committee, presented talks and, with Michael Popham, took an active part helping to facilitate the workshop.

### Workshop Programme

Following refreshments, kindly provided by the Oxford e-Research Centre, and introductions, Kevin Page introduced participants to Linked Data and the idea of worksets (“The EIEPHãT from above”), and Pip Willcox gave an overview of the HathiTrust and EEBO-TCP corpora (Feeding the EIEPHãT). Kevin presented the work of the project (The EIEPHãT enclosure), and Judith Siefring, from the project’s Technical and Scholarly Steering Committee, presented some findings of the SECT project, offering examples of the types of scholarly questions the project could address (What the EIEPHãT shouldn’t forget). The presentations that accompanied these talks appear as appendices to this report, as does the hand-out prepared for participants.

After lunch, where informal discussions continued, David De Roure gave a wider context to the EIEPHãT project, describing other projects across the University of Oxford, particularly at the Oxford e-Research Centre (The EIEPHãT from

a distance). Judith Sieftring and Pip Willcox facilitated breakout discussion groups (Herding EIEPHaTs) with notes taken by Michael Popham and David De Roure respectively, before more refreshments, and a roundtable discussion (Talking EIEPHaTs). Participants were asked to complete a questionnaire (see appendix) before the day ended.

### Results

Comments were collected from breakout groups and reported back and further discussed during the roundtable discussion. An outline of the discussions with results of the questionnaire appear below.

Discussion was both broad-ranging and detailed, as participants exchanged ideas and generated new possibilities and directions for scholarly investigation. The researchers' varied perspectives provided insight into the potential they saw for Linked Data, allowing us to take ideas forward directly from the community, translating these into use cases.

### Recommendations

Of paramount importance to all areas of research was the discoverability of texts, through searching accurate metadata, and through less immediately obvious links, such as networks of printers and authors. For this prototype project, then, it was clear that use cases concentrating on imprint information—authors, dates, and places of publication—was the natural starting point.

Information gathered from information in the imprint or colophon is augmented in EEBO-TCP and the HathiTrust by the research published in Pollard and Redgrave's and Wing's Short Title Catalogues, the Thomason Tracts, and the Early English Books Tract Supplement. Additional metadata includes genre (for example, "biography"), and the ability to search by explicit as well as inferred metadata is important.

While participants were content to be guided through example use cases, it was unanimously agreed that in order to use resources to produce worksets themselves, human-friendly interfaces were vital.

### Future Work

Many ideas for future work came out of the workshop's discussions, formal and informal, as well as a scholarly community willing to continue advising on any future work.

Addressing further enrichment and annotation of metadata content is beyond the scope of current work. The possibility of coordinating HathiTrust and EEBO-TCP metadata with scholarship as it develops, including through book trades indexes, will be facilitated by describing it through Linked Data. EEBO-TCP title pages are marked up separately, though not richly, in the TEI XML: this could be developed, including programmatically. The physicality of the digital could assist with this, for example using OCR to map the spaces on a page.

Networks are a rich seam of research which Linked Data can help to trace and be used to visualize. These include networks of people—authors, printers, booksellers, bookbinders—and works, as well as more abstract abstract concepts, such as the movement of ideas and language.

When Linked Data is applied to HathiTrust and EEBO-TCP full texts, the potential will be extraordinary. Useful worksets could be created around occurrence of particular terms, such as "Whig" and "Tory". Work could be undertaken on flows of ideas, as well as piracy, plagiarism, and co-citation (actual or falsified). Linking data across languages would enrich this still further. Linking metadata and full-texts with historical databases and encyclopaedias would create a gold standard of research resource.

*Pip Willcox*

*Bodleian Libraries, University of Oxford*

*Oxford, 15 June 2015*

## Appendix 1: Notes from the Discussions

Notes from the breakout groups were taken by David De Roure and Michael Popham. Additional notes were taken during the roundtable discussion by Pip Willcox, and have been collated and re-ordered here.

### A) Stationers' functions

Distinguishing between trade functions of members of the Stationers' Company, for example people who print, bind, and sell books, is currently difficult in EEBO.

### B) Prioritizing text or container?

EEBO-TCP silently prioritizes text over its container (typically a book). HathiTrust leans towards prioritizing container over text.

### C) Releasing data without releasing content

Releasing data *about* texts is useful, even where content is restricted, for example, pages per book, or illustrations per page.

### D) Seeing through the eyes of a machine

When a human views a book, the heft, thickness of paper, weight and size of a book are taken in. Combining this information with physical information more easily provided by a machine will be a powerful tool, potentially leading to new insights about book production.

### E) Comparing texts

HathiTrust content is most useful when there is more than one version of a text, or where the output of a printer can be considered together.

### F) Variant spelling

Comparing modern and original editions of a book might be used to overcome the difficulties of variant spelling.

### G) Recognizing anonymous and pseudonymous publications

Anonymous and pseudonymous publication are commonplace in this period. Programmatically identifying features or style, for example, meter and form, could help identify individuals or groups. Scepticism was expressed regarding how much machines can be taught to recognize, and whether such results would be accepted by the academic community (cf. authorship debates).

### H) Recognizing genres

Teaching a machine to recognize genre would necessitate revisiting implicit scholarly assumptions. Could texts be marked up semantically to facilitate this?

### I) False or absent metadata

Where metadata is derived from deliberately falsified title pages or colophons, can it be annotated as such, then used to cross-search other occurrences or co-occurrences of the same information? Similar use could be made of metadata derived from extra-textual sources.

### J) Understanding limitations

This data will enable scholars to answer new questions, but it has limits: these must be clearly stated. Where feasible, resources could be linked to each other to provide more data.



**K) Human limitations of metadata**

When we search, we find what the cataloguer was interested in. Resources are one filter, but another is the cataloguer's interests, and individual library policy and practice.

**L) Re-considering working practices**

"What kinds of new questions can I generate?" For example, a computer can count numbers and letters but not see meaning, for example irony. Researchers need to learn new working protocols.

**M) Tacit knowledge**

Working with, and particularly teaching, digital methods forces tacit knowledge to be made explicit.

**N) Multidisciplinarity**

Economic historians bring different tacit knowledge from literary scholars, and different again from computer scientists. Working collaboratively to create, research, teach, and learn with digital resources will seed knowledge across and between disciplines.

**O) Funding**

Being forced to articulate methods and methodologies has an unintended outcome of improving research funding applications.

**P) Using whitespace**

Where OCR text is not of high quality (as is frequent with early modern text), can whitespace be used to identify printing practice or to match copies of the same work?

**Q) Cost and marketing**

Number of pages, illustrations, paper size, amount of whitespace could be used to infer the cost of publications. Similarly, blank pages (for example, in almanacs) could be used to find publications where the reader could interact with the text.

**R) Correcting text**

Could corrections be made programmatically? Remembering there are errors in human-transcribed texts too, could potential errors be highlighted for machine or human correction?

**S) Contextual corrections**

Could machines be taught to correct contextually, cf. predictive text.

**T) Supporting scholarly editing**

Accessing different states and editions of a text is of primary importance to editors. Could these be programmatically brought together, for human consideration and editing?

**U) Errata and reception**

Combining printed lists of errata with metadata about ownership, provenance, and annotation, could inform understanding of textual reception.

**V) Date of publication**

Particularly during times of war or political contention, month as well as year of publication is of interest. Could this be added to metadata where it is known or can be derived from the text?

**W) Networks**

Uncovering networks of printers, writers, and booksellers is useful. Could intellectual and temporal networks also be mapped, e.g. publications around events or contentious topics. These might be found by tracing phrases and other information on title pages, as well as through close textual reading.

**X) Contentious words**

Co-occurrence of particular words and phrases would make an interesting workset, for example “Whig” and “Tory”.

**Y) International, multilingual linking**

The growing interest in translation studies could be supported by linking encoded texts with collections in other languages, for example in the HathiTrust, or the Medici Archive. Tracing networks of people and ideas across languages is of increasing interest to scholars.

**Z) Linking across genres of resources**

Linking data between disparate genres of resources, for example HathiTrust and EEBO-TCP with the *Oxford Dictionary of National Biography*, would provide rich research, between authors, networks, style, language, and ideas.

## Appendix 2: Questionnaire Data

We received eight completed questionnaires from the workshop, whose results appear below.

	<i>Very often</i>	<i>Often</i>	<i>Occasionally</i>	<i>Never</i>	No an	Total
How often do you use EEBO-TCP material in your own work?	3	2	3	0	0	8
How often do you use HathiTrust materials?	0	0	3	5	0	8
How often do you search digital collections?	7	1	0	0	0	8
	<i>Almost always useful and reliab</i>	<i>Usually useful and reliabi</i>	<i>Not normally useful or reliab.</i>	<i>Rarely useful or reliable</i>		
How useful and reliable do you find digital collections?	1	7	0	0	0	8
	<i>Yes</i>	<i>Maybe</i>	<i>No</i>			
Would you find it useful to be able to search across the EEBO-TCP and HathiTrust materials together?	5	3	0	0	0	8
Based on what you have learned today, would you be interested in the research outputs of the EIEPHaT pr	8	0	0	0	0	8