Workset Creation for Scholarly Analysis + Data Capsules, Phase 1

Final report: January 1, 2016 – August 31, 2018

Prepared by: J. Stephen Downie, Beth Plale, Timothy Cole and Ryan Dubnicek

Progress Report:

Work on the "Worksets for Scholarly Analysis + Data Capsules: Phase 1" (WCSA+DC) project were centered around 4 main goals:

1. The deployment of a new Workset Builder tool that enhances search and discovery across the entire HTDL by complementing traditional volume-level bibliographic metadata with new metadata derived from a variety of sources at various levels granularity.
2. The creation of Linked Open Data resources to help scholars find, select, integrate and disseminate a wider range of data as part of their scholarly analysis life-cycle.
3. A new Data Capsule framework that integrates worksets, runs at scale, and does both in a secure, non-consumptive, manner.
4. A set of exemplar pre-built Data Capsules that incorporate tools commonly used by both the DH and CL communities that scholars can then customize to their specific needs.

As such, this report will focus on these four major goals, by section.

Deployment of New Workset Builder

HTRC's security policy evolved during the WCSA+DC project years, necessitating a modified plan for a new Workset Builder (WSB) 2.0 to replace a deprecated 1.0 version. With sensitivity to storing full text data in multiple locations, HTRC decided to move forward with a WSB 2.0 built on data with no restriction. Additionally, a new software review and implementation process, the HTRC Analytics Enhance Proposal (HAEP) process, was created and instated. As a result of both factors, the functions of WSB 2.0 were modularized into individual service improvements, deployed individually, with continued improvements planned for the WSB 2.0 prototype after grant-end. Three developments from WCSA+DC form the suite of tools that make up WSB 2.0: 1) a search and retrieval interface built on the Extracted Features (EF) Version 1.5 Dataset (which is released as non-restricted open data) ingested into Solr; 2) a new workset import functionality from HathiTrust's Collection Builder; and, 3) a Virtuoso RDF triple store that currently holds bibliographic records as triples for the entire HT corpus, as well as triples representing worksets from the publically accessible HTRC Analytics website (https://analytics.hathitrust.org). Each of these tools are spoken about in more detail below.

The WSB 2.0 search and retrieval interface has been successfully deployed to the HTRC development environment. It is currently undergoing beta testing and security review. The WSB 2.0 search tool is a Solr 7.4 installation that uses the unigram bag-of-words data available in the EF 1.5 Dataset. It provides new search access to both volume-level (15.7 million) and page-level (5.8 billion) metadata files to allow for workset building at different levels of granularity. Users can now also download individual EF volume or page files or a bundle of files

representing their complete workset. As a Solr 7.4 installation, advanced users can programmatically build sophisticated queries using its standard API. The use of unigram information drawn from the EF 1.5 files alleviates the security risk of exposing full-text data in a Solr index, a chief driver of the deprecation of HTRC's WSB 1.0. This prototype also enables users to browse and view page-level EF data (text tokens by sorted either by document region, token frequencies or part of speech) for each page, with a link back to the HathiTrust page viewer for public domain volumes. Because users can see page-level text features, they now have context information for reviewing search results originating from the otherwise unviewable copyright-restricted volumes. New levels of faceting, including genre and Library of Congress classification, were also implemented, allowing for new and more nuanced methods of workset creation. Additionally, the WSB 2.0 search tool supports the ability to create worksets of pages rather than volumes, a key development influenced by recommendations from domain expert partners. The WSB 2.0 beta search tool is available for exploration here: https://go.illinois.edu/htrc-wsb2-beta. In the version available at the aforementioned URL, the functionality allowing users to directly import their worksets built from the search tool into the HTRC Analytics gateway has been implemented but disabled because the tool is undergoing stress testing and formal security review. Screenshots of the import process and software are attached to this report as Attachment 2.

In addition to the Extracted Features Solr index, functionality supporting the import of HathiTrust Collections (https://babel.hathitrust.org/cgi/mb?colltype=updated) to the Analytics site has also been implemented, and is available for exploration under the "Create A Workset" heading on Analytics. The Virtuoso triple store is spoken about in detail under the next heading, but this platform will enable contextual browsing as well as serendipitous discovery, both workset building methods that have been deemed as desirable by scholars.

Creation of Linked Open Data Resources

Creating infrastructure and software to leverage Linked Open Data (LOD) to improve HTRC services has been a key piece of the WCSA+DC project. LOD presents novel and potentially more efficient ways of searching and retrieving HathiTrust volumes and pages, especially at scale. To these ends, the WCSA+DC project teams deployed a Virtuoso RDF triplestore, populated with BIBframe XML records for each volume in HathiTrust, generated from MARC records. Using HTRC's workset model (http://doi.org/10.5334/johd.3), worksets were implemented as RDF objects in the triplestore, with lists of included volume IDs along with workset-level metadata (e.g. creator, creation date, and creator-submitted description of the workset). This proof-of-concept triple store allows for eventual increased incorporation of contextual browsing and serendipitous discovery into HTRC's information seeking model. Additionally, Virtuoso's store of workset objects will enable the eventual implementation of search by and search within workset queries, which will allow a new form of workset building previously unsupported by HTRC. To simplify interactions with the triple store, an API was developed on top of pre-canned SPARQL queries that allow for browsing of worksets and volume metadata within worksets, and discovery of worksets containing specified volumes. The triple store has been actively connected to HTRC's Analytics Gateway page since September of 2018, and has stored all of the worksets created by users since then.

In collaboration with Key Research Partners at the Oxford e-Research Centre, led by Dr. Kevin Page, LOD has also been leveraged to develop a proof-of-concept, cross-corpora workset builder, that enables users to create worksets combining material in the HathiTrust and Early

English Books Online Text Creation Partnership (EEBO-TCP). This is implemented using LOD and federated SPARQL queries over HTRC RDF triples, EEBO RDF created in Oxford from the EEBO-TCP TEI headers, and 'bridging' triples reconciled using external authorities (e.g. VIAF) and entity reconciliation. Prior to development, this work required an extensive survey and analysis of existing bibliographic ontologies, including MADSRDF/MODSRDF, Bibframe, schema.org, BIBO (http://bibliontology.com/), and FaBiO (https://sparontologies.github.io/fabio/current/fabio.html), regarding their suitability for building and parameterising worksets. Of these, only Bibframe was primarily developed specifically for library-centered use cases in mind and, thereby it was chosen for implementation. Demos of the proof-of-concept WCSA+DC-EEBO workset builder are available for viewing here: https://uofi.box.com/s/llzupxdlb7txrjydj5nu4f9wtg1q7c1k

The team at Oxford e-Research Centre also collaborated with HTRC to develop a model to characterize the information-seeking needs of users in large-scale digital libraries, and evaluated that model against the workset model to assess the ability to meet identified user needs for workset building (the focus of papers presented at the *ACM Joint Conference for Digital Libraries* in 2017, http://dx.doi.org/10.1109/JCDL.2017.7991583, and 2018, http://dx.doi.org/10.1145/3197026.3203886). Further extensions were made to this model to afford for LOD resources, which are detailed in a paper presented at the *ASIS&T* 2018 annual meeting, with the proceedings forthcoming.

A New Data Capsule Framework

The Data Capsule framework is a controlled compute environment for conducting computational analysis of restricted data while also protecting the data from unintended uses or uses prohibited by law, policy or licensing agreement. The Data Capsule framework, implemented as a set of policies and technologies that together enable controlled access and use of the copyrighted texts of the HathiTrust, has made considerable progress as a result of this award in its availability, stability, scalability, and usability. It is actively serving a growing group of researchers with analysis access to HathiTrust as as part of the production software release 4.0 of the HTRC.

The primary areas of contribution are in customized capsule solutions, enhanced user experience, scale of capacity and functionality. Each of these three areas is in response to one or more proposed activities in the original proposal. For instance, enhanced user experience of the Capsule requires that a researcher have their workset available to them in their Capsule. It further does not limit the workset to a specific size - a workset can be anywhere from tens of volumes to a few millions. The individual contributions are described in more detail in the paragraphs below.

Customized Capsule solutions

With the new availability of Capsule activity using the copyrighted content of HathiTrust, we conceptualized and built several classes of Capsules:

- Demo Capsule: is a smaller Capsule that has access to HathiTrust public domain content only. Results cannot be released from the Capsule.
- Research Capsule (public domain): customizable Capsule up to 4 cores and 16GB memory; derived data release is allowed pending review. Access to HT public domain only by default. Researcher must agree to term of use;

- Research Capsule (copyright): above plus additional information required on intent of use; review/approval is required to access full HT corpus. This option is currently limited to HT members only.

Enhanced user experience

We developed a novel software package called HTRC Workset Toolkit which makes it easier for researchers to import/export their workset to/from their Capsule, and makes it easier to connect their analysis tools to the HathiTrust collection. The HTRC Workset Toolkit works with the JSON-LD description of Worksets. The researcher's Capsule additionally now comes enriched with pre-installed sample data, with the Voyant data exploration tool, and with a rich set of other user requested analysis tools and packages, such as Anaconda, Mallet, R, InPho TopicExplorer, and number of popular Python libraries like GenSim, numpy, scipy, pandas and nltk. These tools and packages were included after collaboration with WCSA+DC Key Research Partners. Finally, a researcher interface to their Capsule has been upgraded to use encrypted clientless VNC and SSH connections, which allow users to seamlessly and securely access data capsules in different computing environment without installing specific VNC or SSH clients.

Scale Capsule functionality

Through contributions of hardware by University of Illinois, the Data Capsule service can now can provision Capsules from a pool of 120 cores and 640GB memory. Researchers can now spec out a single Capsule with up to 4 cores and 16GB memory through the standard web interface, with additional resource needs exceeding the capacity can be requested and addressed through special handling. The data capsule threat model has been evaluated against the multi-server hosting environment for verified security and reliability.

Topic modeling is notoriously computationally intensive. Through a combination of hardware upgrades and software refactoring, we were able to increase the capacity of topic modeling inside a research Capsule to approximately 500 volumes per GB of RAM allocated. This enabled analysis of up to 8,000 volumes in a 16GB capsule. With an average volume size of approximately 150,000 words, we can now analyze up to 2.4 billion words in a single Capsule.

Exemplar Data Capsules

Progress on goal number four, the creation of a set of exemplar Data Capsules to meet needs of computational linguistics (CL) and digital humanities (DH) users, was slightly modified as the Data Capsule service evolved. Three default Capsule formats were instated, as mentioned above, while HTRC, in conjunction with key research partners at University of Illinois, Brandeis University and University of Waikato, moved forward with identifying standard software packages and tools that could be included with all Data Capsules. This model allows a user access to a number of domain-specific tools and test data within every Capsule.

Computational Linguistics domain

Serving as expert users in the area of Computational Linguistics (CL), the team at Brandeis University, led by Prof. James Pustejovsky, was tasked with showing a proof-of-concept integration of the LAPPS Grid / Galaxy platform and workflow, a project on which Pustejovsky is a Principal Investigator along with WCSA+DC Advisory Board member Prof. Nancy Ide (Vassar College), within the HTRC Data Capsule environment. The Brandeis team successfully achieved this, integrating the LAPPS Grid natural language processing (NLP) tools and cloud platform using the Galaxy web front-end in a Docker container that can be installed within an off-the-shelf Data Capsule. Analysis tools included support basic text processing (sentence split,

tokenization, parts-of-speech tagging) as well as information extraction (entity recognition, relation extraction) and linguistic analysis (syntactic parsing, anaphora resolution). The LAPPS Grid tools were then integrated with the HTRC Workset Toolkit, a library written for retrieving and interacting with HT texts and metadata within the Data Capsule.

With software installed and running, the LAPPS Grid NLP tools could be evaluated against HT data, as well as modified and evolved to include common functionalities used by Digital Humanities researchers, such as entity and relation extraction. As part of this process, datasets were developed for evaluation and tweaking of NLP tools with regard to HT data and success of the newly incorporated DH tools.

Digital Humanities domain

Prof. Ted Underwood (University of Illinois) served as a Digital Humanities (DH) domain expert, and was tasked with using the new Data Capsule infrastructure for his projects on gender in fiction and character in biography. The centerpiece of Prof. Underwood's work on the grant was research on characterization in nineteenth- and twentieth-century fiction, published in the Journal of Cultural Analytics as "The Transformation of Gender in English-Language Fiction" (available here: http://culturalanalytics.org/2018/02/the-transformation-of-gender-in-english-language-fiction/). This research required natural language processing on connected text, so it couldn't be accomplished with HTRC's EF Dataset, and was previously impossible to explore due to copyright restrictions on much of the fiction in HathiTrust. This piece on character was well received, with journalistic coverage in *Smithsonian* (https://www.smithsonianmag.com/arts-culture/what-big-data-can-tell-us-about-women-and-novels-180968153/), *The Economist* (https://www.economist.com/prospero/2018/03/08/machines-are-getting-better-at-literary-analysis), and *The Washington Post* (https://www.washingtonpost.com/news/posteverything/wp/2018/07/30/how-computational-analysis-is-teaching-us-to-read-in-new-ways/?utm_term=.aa84c7f01b39). Parts of that research will also be used in a forthcoming book from Dr. Underwood, *Distant Horizons: Digital Evidence and Literary Change* (Chicago: University of Chicago Press, 2019).

Prof. Underwood was also able to explore additional projects during the WCSA+DC grant period, including a similar analysis of "characters" in biography, focusing on comparing the "characters" in biographies to fictional characters, as well as a project on book reviews, for which he was awarded a fellowship at the National Humanities Center. This project produced several interesting results that are still being written up by the project team, and, with the project on gender in fiction, served as a useful pilot of the enhanced Data Capsule framework, informing the inclusion of a number of standard DH tools (mentioned fully under the "A New Data Capsule Framework" heading) as well as technical resources for off-the-shelf Data Capsules. Prof. Underwood's project on book reviews is currently ongoing, and has employed the WSB 2.0 prototype to identify his workset, with analysis occurring within the Data Capsule environment. This exciting project will be one of the first to use HTRC services from workset building stage to the final analysis and publication phase.

Concept tags

The University of Waikato team, led by Assoc. Prof. Annika Hinze, was tasked with testing and implementing a prototype of their Capisco concept tagging system within a Data Capsule to test and potentially develop a workflow for eventual wider release of this tool as a standard inclusion in the Data Capsule. This process involves *seeding* and *tagging* of concepts. Seeding

represents the process of disambiguating terms that are flagged from the tokens of the full text because they appear in a Concepts-in-Context (CiC) network the Waikato team generated from Wikipedia data. The seeds generated are then semantically compared to each other to disambiguate into cogent concepts represented by the text data. This process is detailed, as well as a comparison of various seeding strategies, in a paper presented at the *ACM Joint Conference for Digital Libraries* 2018 meeting, and available online here: https://doi.org/10.1145/3197026.3203874.

Once seeds are successfully disambiguated, they were evaluated and improved, then added as tags to pages and volumes. These tags enable eventual implementation of search and retrieval, as well as results filtering, by concept. This has been illustrated in a small, diverse test corpus, with a sample query (searching for the concept 'Bank') of this set available for exploration in the WSB 2.0 prototype here: https://solr1.ischool.illinois.edu/solr-ef/index.html?solr-col=dbbridge-fict1055-htrc-configs-storeall&solr-key-q=0132F306CB4EB6FCAD97EB51280D12984#search-results-anchor. As with other project work, links to code repositories and resultant publications are available in the "Additional Information" section of this report.

Setbacks or Challenges:

Given the nature of HathiTrust's data, security is a primary concern of HTRC, and influences every new service or tool in both the design and deployment phases. The work on WCSA+DC is no different, with security being a challenge to overcome for each piece of the project. For HTRC staff and developers, security concerns and policy help shape how services work and can be implemented. This is true for both WSB 2.0 and the enhanced DC environment, as well as in partner software. Security's impact on design of WSB 2.0 is spoken about in detail under the Progress Report section on WSB 2.0. Security continues to influence and evolve services on search, retrieval and analysis of worksets. Similarly, the original design and threat model of the DC framework requires a hosting environment that is made up of physical servers. That is, Capsules, which are VMs, will not run inside a VM, only on a physical server. This security limitation currently precludes our scaling of the framework to cloud hosted environments where computation is provided in the form of VMs. In spite of this, we have successfully scaled capacity of Capsules to utilize resources of multiple physical servers and now run both larger (more cores) Capsules, and more Capsules at the same time.

Scale is also an ever-present concern when working with the nearly 17 million volume HathiTrust collection. Partners at Brandeis, Waikato and Illinois all reported that scaling up their work to many tens of thousands of volumes can present difficulties while working within the DC and WSB 2.0 tools. Specifically, parallelizing and executing more compute-intensive methods within the DC presents longer processing times than traditional data analysis on dedicated servers or hardware. Similarly, bibliographic metadata of such a large corpus creates the possibility of generating very large extractions of entities and relationships that in turn stress the ability of users to contextualise and manage the information through, for example, LOD-enabled browsing. The addition of external corpora will only increase this challenge. An LOD approach to this type of browsing theoretically supports such large and varied relationships and entities, but rigorous and continued user engagement and testing should be part of the development process to ensure user needs are met in the most effective way. This process is also key to ensuring that each distinct technology for workset building (Solr indices, EF data, LOD) is thoughtfully mapped to user requirements and continually improved. Generally, challenges of scale are necessary side effects of working with a massive digital library with data under some form of copyright restriction. However, future work, discussed in more detail in the following

section, will focus on continuing to find novel ways of working with scale issues within a secure compute environment to improve analysis capabilities, experiences and results.

Lastly, typical logistical challenges surfaced as the project progressed. A number of staffing changes, both within HTRC and Key Research Partners, over the entire project lifecycle caused some delay in work plan, which necessitated a no-cost extension request, submitted and approved in 2017.

Lessons Learned:

Workset Builder 2.0

A key lesson we have learned about workset building by HTRC users has been that the idea of a “standard” use case for HTRC services is difficult to define. Initial HTRC engagement had been with worksets of hundreds, perhaps thousands, of volumes. As services have evolved, and data has grown and access increased, HTRC has seen a shift toward thousands/tens of thousands of volumes in most worksets, an interesting, but sometimes challenging revelation. Collaboration with the Oxford team has indicated that the fundamentals of established information-seeking models have held up in the context of HathiTrust, but, without extension, are not suited for characterizing and mediating between the new classes of user needs and the engineering requirements that fulfill them (i.e., traditional search and workset building are related but quite different operations). With the completion of WCSA+DC, HTRC will be targeting a new series of user studies to better gauge and evaluate how users are most likely to interact with HTRC services to better inform current information seeking models, and the evolution and development of services, both through software, hardware and human support.

The another major lesson learned is that data at the scale of the HT collection is inherently inconsistent in ways that that can be quite surprising and sometimes frustrating. For example, inconsistent date data in volume metadata has proven a constant challenge for both WSB 2.0 developers and end-users. Many users have asked for time slices of the data which are very difficult to provide accurately at present. Variance in author names, titles, publication places, etc., have also shown to be problematic and hinder both retrieval and duplication detection. Unigrams text data has helped get EF 1.5 and WSB 2.0 to be very helpful, but users are asking for phrase detection operations that we do not currently support.

Data Capsule

We have enhanced DC computation environment with pre-installed sample data, popular analysis tools and home-brewed software packages, both the subject of enthusiastic requests from users. It remains highly desirable to develop a standard and easy process to share custom-built tools among users across Capsules, which is likely to become a focus of the HTRC cyberinfrastructure team in the coming years. Additionally, the interim report discussed a potential use case of analyzing a half-million volumes inside a Capsule. In pursuit of this use case, we discovered that the total number volumes in a workset is not a reliable indicator of the analysis capacity for data capsules. Individual volume sizes vary greatly within HT corpus, and analysis tools vary greatly on their resource demands. As we encounter huge worksets, we encourage researchers onto traditional High Performance Computing platforms, which are still more preferable in super large analysis tasks.

Key Research Partners

The source of HT volumes, mostly historical text that has been digitized using optical character recognition (OCR), informed the work of the Brandeis team, as many of the machine learning or

NLP tools that are publicly available and used have been tuned for modern public data, and not historical text. HathiTrust text data often retains errors created during the OCR pipeline, which further muddies the analysis pipeline. The Waikato team also found issues with the OCR data, and eventually uncovered that using curated metadata was often more reliable than the OCR full text. It was also found that comparison of tagging approaches differed depending on the domain in which the tags would be applied, with different types of source material often presenting different seeding and tagging methods as most successful.

Goals Following Grant End:

Workset Builder 2.0

WSB 2.0 search tool is planned for deployment as part of the HTRC Analytics production services in the first half of 2019. After deployment of WSB 2.0, continued development and improvements will continue, especially to address issues of scale and continued evolution of domain-standard software tools. A new release of an enhanced EF 2.0 Dataset is also planned for the first half of 2019. EF 2.0 will cover more of the ever-growing HT collection and use the Stanford NLP tools which should provides use with better part of speech tagging. Based on user feedback, we plan on extracting and publishing new classes of entities from the HT (and then incorporating in the the WSB 2.0 framework). We hope to explore phrase detection operations that would identify real phrases such as “New York” or “peace treaty.” This approach would meet a real-world need without publishing security-compromising bigrams and trigrams. We would also like to explore methods of dealing with the data variance issues and if solutions are found, also incorporate them into the WSB 2.0 framework. Data reduction techniques to help speed up such things as duplicate or quality detection are also being considered. We are eager to find a way to socialize the creation, use and curation of worksets as scholarly products. In a similar vein, we are keen to explore how new but standardized “extracted feature” types can be generated from both the HathiTrust and other collections and then shared in an open but federated manner (e.g., LOD).

Data Capsule

This award has allowed us to progress the DC framework in increased availability, stability, scalability, and usability. There remain two chief challenges. The DC framework is limited to running on physical servers, but today most resources that are abundantly available and inexpensive provide resources in the form of VMs. To move beyond the current restriction of physical machines would take a rewrite of the low-level virtual machine management software. This rewrite is important to embrace full scalability. The second objective is to support cross-Capsule analysis where a researcher utilizes more than one Capsule to accomplish a larger task. In order to do this, the DC design would have to allow certain cross-Capsule communication while dis-allowing other forms. Finally, the workset can stand as a publishable result of one’s research. Exploration of how the DC can contribute to reproducibility of a result with infrastructure provenance is also another area of potential inquiry.

Key Research Partners

The Brandeis team is targeting potential user identification and authentication integration to improve user experience when interacting with both LAPPS Grid / Galaxy and the DC. Further integration of HTRC tools such as the Workset Toolkit is also needed to better support large worksets and demanding text analysis tools. Lastly, continuing development of existing tools and integration of new tools will continue to be an area of exploration.

The Waikato team hopes to continue to refine and improve the current approaches to seeding and tagging concepts in HT data, and also sees improved Capisco semantic analysis language support, including more languages and cross-language corpora. Additionally, there is an opportunity to further improve the CiC knowledge base through scholar adaptation and annotation.

The Oxford team plans to continue to collaborate with HTRC on leveraging LOD for workset building and refinement, including joining bibliographic metadata with feature metadata, modeling feature relationships, exploring options for interoperability across heterogeneous RDF semantic sets, and documenting and evaluating cost of cross-corpora workset building versus ingestion/merging of the datasets. In addition, evaluating the information seeking needs and strategies in the context of large-scale digital libraries is a key piece of future work, and would necessitate further development of both simulated approaches as well as conducting user interviews and trials. Once done, this could yield a handbook of best practices which could be an impactful output that would better shape and evolve current and future HTRC services.