Enhancing Access to Digital Media

The Language Application Grid in the HTRC Data Capsule

James Pustejovsky
Brandeis University
Waltham, Massachusetts
jamesp@cs.brandeis.edu

Marc Verhagen
Brandeis University
Waltham, Massachusetts
marc@cs.brandeis.edu

Keongmin Rim
Brandeis University
Waltham, Massachusetts
krim@cs.brandeis.edu

Yu Ma
Indiana University
Bloomington, Indiana
yuma@iu.edu

Liang Ran
Indiana University
Bloomington, Indiana
lran@umail.iu.edu

Samitha Liyanage
Indiana University
Bloomington, Indiana
shliyana@indiana.edu

Jaimie Murdock
Indiana University
Bloomington, Indiana
jammurdo@indiana.edu

Robert H. McDonald
Indiana University
Bloomington, Indiana
rhmcdona@indiana.edu

Beth Plale
Indiana University
Bloomington, Indiana
plale@indiana.edu

ABSTRACT

The project "Workset Creation for Scholarly Analysis and Data Capsules" is building an infrastructure where researchers have access to text processing tools that can then be used on a copyrighted set of digital data. The infrastructure is built on (1) the HathiTrust Research Center (HTRC) Data Capsule services that can be used to access the HathiTrust Digital Library and (2) the language processing services of the Language Application (LAPPS Grid). The main thrust of the work presented here is the integration of the LAPPS Grid workflow infrastructure with the secure data access computing environment provided by the Data Capsules.

KEYWORDS

Digital Library, Natural Language Processing, Processing Workflows, Jetstream

ACM Reference format:
James Pustejovsky, Marc Verhagen, Keongmin Rim, Yu Ma, Liang Ran, Samitha Liyanage, Jaimie Murdock, Robert H. McDonald, and Beth Plale. 2017. Enhancing Access to Digital Media. In *Proceedings of Practice & Experience in Advanced Research Computing, New Orleans, Louisiana USA, July 2017 (PEARC'17),* 3 pages.
https://doi.org/10.475/123_4

1 INTRODUCTION

The main motivation for this ongoing work is to make text processing tools available to users of a restricted set of digital data. In order to achieve this goal we need to pay attention to the following items:

- Deployment of tools that enhance search and discovery across the library by complementing traditional volume-level bibliographic metadata with new metadata, using specially developed LAPPS/Galaxy-based CL applications;
- 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;
- Creation of a set of exemplar pre-built Data Capsules that incorporate tools commonly used by both the Digital Humanities and the Computational Linguistics communities that scholars can then customize to address their specific needs.

We use the secure HTRC Data Capsule services provided by the HTRC and integrate into this secure environment language processing services from the Language Application Grid. In the following sections we will first describe the HathiTrust Digital Library and the Data Capsule in section 2 and the Language Application Grid in section 3. We then lay out our approach and the architecture of the integration in section 4.

2 THE HATHITRUST DIGITAL LIBRARY

The HathiTrust (HT) is a consortium of members that steward the over 15 million volumes of digitized content from research libraries across the world. It provides long-term preservation and access services for public domain and in copyright content from a variety of sources, including Google, the Internet Archive, Microsoft, and in-house partner institution initiatives.

HathiTrust ensures the reliability and efficiency of the digital library by relying on community standards and best practices, developing policies and procedures to manage content and services at scale, and maintaining a modular, open infrastructure.

The HathiTrust Research Center (HTRC) [1] was created in 2011 by HathiTrust to pioneer models and infrastructure for computational analysis to the HathiTrust Digital Library. This collaboration between Indiana University and University of Illinois providing means for researchers to analyze large swaths of the 15+ million

PEARC'17, July 2017, New Orleans, Louisiana USA

ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.475/123_4

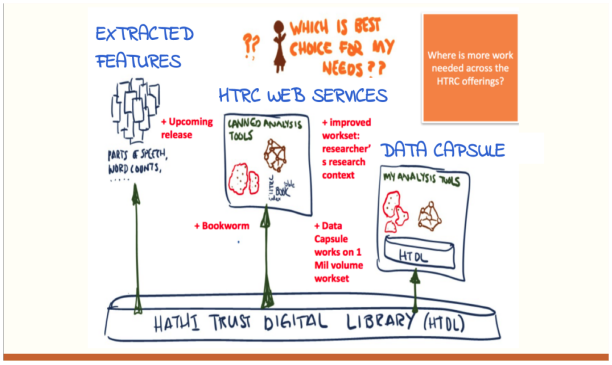

Figure 1: HTRC Architecture

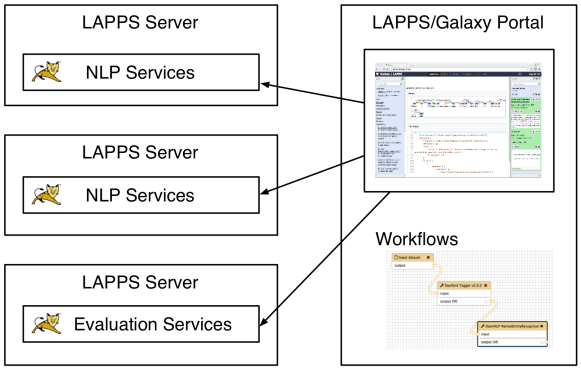

Figure 2: LAPPS Grid Architecture

volumes of HathiTrust drawing on computational resources and tools in house, and externally, such as using Jetstream.

2.1 The HTRC Data Capsule

The HTRC Data Capsule (DC) service [4] provides a secure computing environment for analysis of restricted content. As with Jetstream, users are provisioned with a virtual machine (Capsule) through which they interact with HT volumes. Data Capsule service has restrictions on its use, particularly in limiting how and when the products created by analysis tools leave a Capsule: data must undergo results review prior to release to ensure they meet the HTRC's policy for non-consumptive use research policy.[1] DC represents HTRC's solution to analysis that is closest to the HTDL. Other solutions, HTRC web services, and its Extracted Features dataset[2] guarantee non-consumptive exports.

3 THE LANGUAGE APPLICATION GRID

The Language Application (LAPPS) Grid project [2] has established a framework that enables language service discovery, composition and reuse and promotes sustainability, manageability and interoperability of natural language processing (NLP) components. The LAPPS Grid is built upon the service-oriented architecture (SOA), a more recent web-oriented version of the pipeline architecture that has long been used in NLP for sequencing loosely-coupled linguistic analyses.

At its core, the LAPPS Grid provides language processing tools and allows users to run these tools in a pipeline, thereby combining the results of several types of processing. The LAPPS Grid uses a particular syntactic format called the LAPPS Interchange Format (LIF) [6] which makes it easy to integrate and re-use new tools in pipelines. In addition, the LAPPS Grid uses a small vocabulary called the Web Services Exchange Vocabulary (WSEV) [3]. The WSEV contains annotation categories like Token, NamedEntity and SemanticRole and is used to define the meanings of those categories as produced by processing services.[3]

The somewhat simplified basic LAPPS Grid architecture is shown in Figure 2. Processing services are typically on a different server than the portal that gives access to all services, but this is not necessarily the case. We have used Tomcat web servers as the main way to deploy NLP services. All services are essentially existing popular NLP modules that were wrapped as web services that consume and produce the LIF format and use annotation categories as defined in WSEV. Of special interest are the evaluation services that are based on the approach used in the open-source OAQA project.[4]

LAPPS has adopted the GALAXY[5] workflow engine as its front end to LAPPS services and created the LAPPS/Galaxy server.[6] Galaxy provides the functionality to add arbitrary services to the front end by writing simple XML definition files and then allows the user to chain these services into workflows.

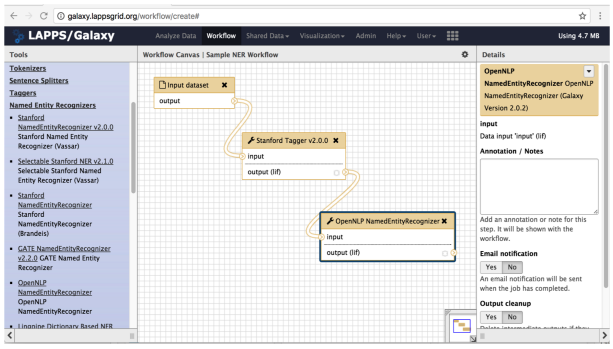

Figure 3: Galaxy Workflow

Workflows can be composed from a variety of components, including components that use different formats, as long as converters are provided. The workflow editor allows us to quickly create workflows ranging from simple ones like the example in Figure 4 to complicated workflows that contain forks and evaluation components that can compare to alternative pipelines.

[1] https://www.hathitrust.org/htrc_ncup
[2] https://analytics.hathitrust.org/datasets.
[3] The vocabulary is available on line at http://vocab.lappsgrid.org/.
[4] Open Advancement of Question Answering Systems, see https://oaqa.github.io/.
[5] http://galaxyproject.org
[6] http://galaxy.lappsgrid.org

4 APPROACH AND ARCHITECTURE

The initial work involved an integrated effort of studying the needs and requirements of the HTRC Data Capsule users: that is, identifying those NLP web services that have already been wrapped and integrated into the LAPPS Grid, as well as modules that are not yet available. Several components were selected such as Named Entity Recognizers that find named entities such as cities, countries, people, etcetera, as well as components performing various levels of constituent- and dependency-based parsing at the sentence level. For these components we (1) assess the overall performance of each component service within the HTRC Data Capsule; and (2) examine the possible workflow configurations of the different services as configured in distinct pipelines to determine the optimal configuration in terms of performance. For evaluation purposes we have annotated a small set of about 100 pages with named entities and further annotation will be performed for relations between entities.

The ability to apply a cyclic process of iterative testing, evaluation, and re-configuration is particularly important for rapid development of workflows to suit specific user needs, and is one of the benefits offered by adopting the LAPPS Grid framework.

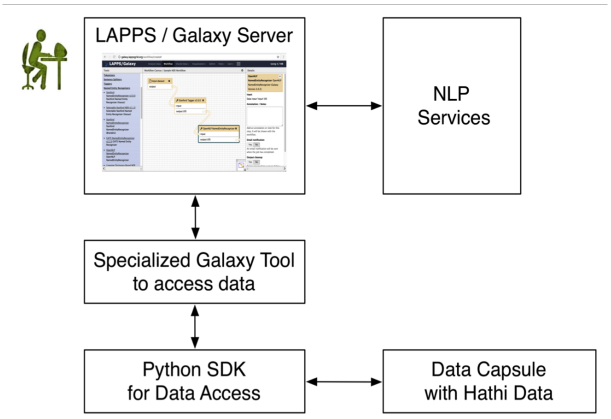

Figure 4: System Architecture

The image in Figure 4 shows the layout of the LAPPS Grid for the HTRC Data Capsules service. The LAPPS Galaxy Server is very similar to the standard LAPPS Server, but it differs in several important respects:

(1) the Galaxy server resides on a secure server with limited access for the outside world,
(2) the processing services do not run on remote servers but locally on a secure server,
(3) the selection of services is tailored to needs of HT Data researchers,
(4) specialized data access protocols were added.

The data access protocol consists of two components and differs from the standard LAPPS practice, which is to create a web service to accesses local data. Instead we have a local Galaxy tool that can be mostly considered to be a wrapper around a Python SDK that accesses data on a HTRC Data Capsule.

4.1 Jetstream as a Test Bed

Experimentation was done at Brandeis University using the standard LAPPS Grid. However, access to the in-copyright content needed to be through a more secure environment like the HTRC Data Capsule service and access to that environment for experimentation with the LAPPS Grid was not seamless.

JetStream[7] [5] provides cloud-based computation and allows researchers to quickly create virtual machines on the remote resource and make these virtual machines look and feel like the researcher's own home machines. We used Jetstream to mimic the security-constrained HTRC Data Capsule service. Jetstream VMs were configured as specialized version of the LAPPS Grid, and access was given to a public domain portion of the HathiTrust. In this way LAPPS interaction with the HT data was tested without security concerns and various configurations could be tested prior to their deployment in the more restricted Capsule environment.

5 CONCLUSION AND FUTURE PLANS

We presented the current state of the continuing work to apply language processing components to restricted digital data. We have set up a prototype with an as of yet restricted set of components and continue experimentation with other NLP components useful for HT Data users. One of the main pain points at the moment is to scale the implementation to larger chunks of data. This involves fixing a bottleneck in Galaxy which tends to slow down when processing documents at volume due to user interface manipulations.

ACKNOWLEDGMENTS

This work was funded in part by the Andrew W. Mellon Foundation under grant number 41500672.

REFERENCES

[1] Mike Furlough Robert H. McDonald Beth Namachchivaya Beth A. Plale Downie, J. Stephen and John Unsworth. 2016. The HathiTrust Research Center: Exploring the Full-Text Frontier. (2016). http://er.educause.edu/articles/2016/5/the-hathitrust-research-center-exploring-the-full-text-frontier EDUCAUSE Review 51, no. 3.
[2] Nancy Ide, James Pustejovsky, Christopher Cieri, Eric Nyberg, Denise DiPersio, Chunqi Shi, Keith Suderman, Marc Verhagen, Di Wang, and Jonathan Wright. 2014. The Language Application Grid.
[3] Nancy Ide, James Pustejovsky, Keith Suderman, and Marc Verhagen. 2014. The Language Application Grid Web Service Exchange Vocabulary.
[4] Alexander Crowell Atul Prakash Jiaan Zeng, Guangchen Ruan and Beth Plale. 2014. Cloud computing data capsules for non-consumptive use of texts. In *Proceedings of the 5th ACM workshop on Scientific cloud computing (ScienceCloud '14)*. ACM, 9–16]. https://doi.org/10.1145/2608029.2608031
[5] Cockerill T.M. Foster I. Hancock D. Merchant N. Skidmore E. Stanzione D. Taylor J. Tuecke S. Turner G. Vaughn M. Stewart, C.A. and N.I. Gaffney. 2015. Jetstream: a self-provisioned, scalable science and engineering cloud environment. In *Proceedings of the 2015 XSEDE Conference: Scientific Advancements Enabled by Enhanced Cyberinfrastructure*.
[6] Marc Verhagen, Keith Suderman, Di Wang, Nancy Ide, Chunqi Shi, Jonathan Wright, and James Pustejovsky. 2016. The LAPPS Interchange Format. In *Revised Selected Papers of the Second International Workshop on Worldwide Language Service Infrastructure - Volume 9442 (WLSI 2015)*. Springer-Verlag New York, Inc., New York, NY, USA, 33–47.

[7] https://jetstream-cloud.org/