Enhancing Access to Media Collections and Archives Using Computational Linguistic Tools

James Pustejovsky, Marc Verhagen
Department of Computer Science
Brandeis University
E-mail: {jamesp,marc}@cs.brandeis.edu

Nancy Ide, Keith Suderman
Department of Computer Science
Vassar College
E-mail: {ide,suderman}@cs.vassar.edu

Abstract

In this paper, we outline the strategies, methodology, and infrastructure needed to bring advanced computational linguistic tools to researchers and archivists in the humanities. We discuss three use cases involving the application of the Language Application Grid (LAPPS), an open, web-based infrastructure providing interoperable access to hundreds of computational linguistic (CL) component web services, together with facilities for multi-step analyses via tools pipelining, performance evaluation, and resource delivery. These include: CL analysis of corpora restricted under copyright; the challenge posed by radio and television media collections; and the use of LAPPS for assisting archivists in their collection and cataloguing efforts. We believe that the adoption and use of CL platforms such as LAPPS by the digital humanities (DH) will help foster better communication, sharing, and research between the two communities.

1 Introduction

In the 1960s, the fields that are now called “computational linguistics" and “digital humanities" were not recognized as distinct [9, 19, 20]. In the 1970s, when computational linguistics (CL) began to be heavily influenced by advances in the field of Artificial Intelligence and adopted logic- and rule-based, symbolic methods, “Humanities Computing" retained the fundamentally statistical approach prevalent in the previous decade. Over the ensuing 40 years, the two fields have evolved in relative isolation. Some efforts were made in the early 1990s to reunite the two when CL once again took up statistical methods, using the argument that statistical methods adopted and adapted in CL had much to offer the field of digital humanities, and also that digital humanities, with its vast store of creative language data, provided a challenge to current methods that could yield fresh insights into the ways language conveys meaning. These efforts failed, and as a result, the two fields continued to evolve along their own paths. CL pushed empirical methods forward into machine learning and, most recently, “deep learning" involving neural networks and similar structures, while humanities computing evolved along a very different path, encompassing the creation, maintenance, and use of massive libraries of digitized data, including not only literary, historical, and similar texts but also images, audio, and video, representing artifacts relevant to the arts and humanities. Thus the term “digital humanities" was coined.

Within the past few years, Digital Humanities (DH) has looked to CL for methods to enable richer analysis of literary, historical, and other kinds of documents, recognizing that CL methods and procedures can in fact enhance the kinds and amount of information that can be automatically extracted from language data [22]. However, re-marrying the two disciplines has proven non-trivial [18, 20]. The difficulty is invariably attributed to a lack of accommodation in CL tools for users who are not technically inclined, and indeed, this is largely true. However, the roots of the problem go far deeper, stemming from two complementary factors: differences in the goals for which the same analyses are applied in each area, and differences in the methodological norms and perspective of the researcher.

In this paper, we outline a general methodology towards accomplishing the goal of re-integrating the two fields, and list the requirements on what tools are needed by humanities researchers and archivists. We first review the platform of the LAPPS Grid, and then examine three case studies of how the platform can help the humanities scholar and archivist in their research.

2 The Language Applications (LAPPS) Grid

Over the past ten years, there has been increased activity in efforts to integrate Human Language Technologies (HLT) applications, corpora, as well as development platforms. This stems from an obvious and growing demand for robust language processing capabilities across academic disciplines, education, and industry. However, one of the major problems in this area has been and remains component interoperability, reusability, and integration. This has resulted in much of the field of HLT being fragmented, characterized by a lack of standard practices, few widely usable and reusable tools and resources, and much redundancy of effort. Rapid development and deployment of HLT applications has also been hindered by the lack of ready-made, standardized evaluation mechanisms, especially those which enable evaluation of component performance in applications consisting of a pipeline of processing tools.

To address these problems, we have developed an open, web-based infrastructure, called the LAPPS Grid[1][10, 21], that provides interoperable access to hun-

[1] Funded by the NSF-SI^2 initiative.

dreds of HLT component web services, together with facilities for multi-step analyses via tools pipelining, performance evaluation, and resource delivery for a wide range of language resources [11, 13]. As an easy-to-use interface and management system, the LAPPS Grid has adopted the Galaxy framework[2] [6], a robust, well-developed platform for workflow configuration and management, and persistence of results. The LAPPS Grid affords the possibility of creating ready-made workflows to perform specific analytic tasks that can be used off-the-shelf or customized to accommodate specific projects, as well as means to compose and evaluate workflows from atomic NLP components [12]. The LAPPS/Galaxy platform can be accessed through a web interface (http://www.lappsgrid.org), deployed locally on any Unix system, or run from the cloud. Figure 1 provides an overview of the LAPPS Grid architecture.

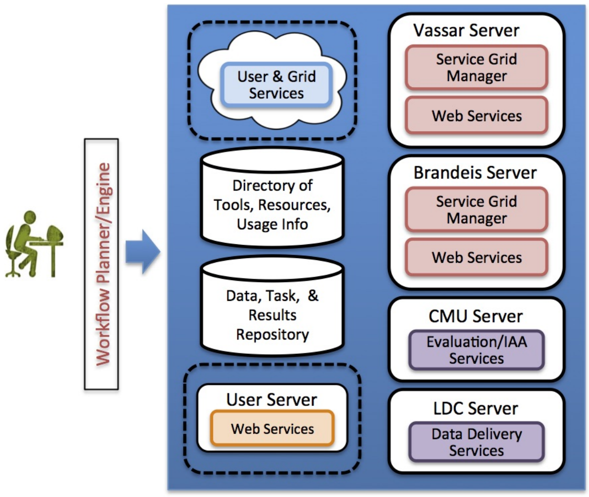

Figure 1: The LAPPS Grid supports discovery, adaptation and composition of language technologies.

3 Language Analysis within the HathiTrust Data Capsule

In our first case study, we examine the role that the LAPPS Grid can play in ensuring access for CL tools over copyright-restricted content, specifically, over the HathiTrust Library [17, 5]. The HathiTrust Digital Library comprises the digitized representations of 13.68 million volumes, 6.84 million book titles, 359,528 serial titles, and 4.79 billion pages. Approximately 39% of the items in the HathiTrust corpus are digital representations of print volumes in the public domain. The remaining 61% are works under copyright. Because of copyright restrictions, scholars have come to see this 61% of the HathiTrust collection of volumes as sitting

[2] Funded by NSF awards 0543285 and 0850103.

behind a ‘copyright wall’ that makes it next to impossible for them to have meaningful access to their content.

The HathiTrust Research Center (HTRC) develops software infrastructure, models, and tools to help digital humanities (DH) scholars conduct new computational analyses of works of the HathiTrust corpus, with a focus on analysis of larger datasets than can be done today (what they call “analysis at scale”). One of the key infrastructure components of HTRC is the Data Capsule (DC). Recently, LAPPS/Galaxy has been adopted by a Mellon-funded project at the University of Illinois, which is utilizing the platform to apply sophisticated HLT text mining methods to HTRC’s massive digital library[3]. The HTRC’s DC Project involves a collaboration between Illinois, Indiana, Brandeis, Oxford, and Waikato Universities [7]. Working with our Illinois and Indiana collaborators, the project is focused on implementing specific LAPPS tools that are most needed by the digital humanities scholar within the HTRC user community.

The HTRC Data Capsule [23], shown below in Figure 2, is a solution to provisioning secure researcher access directly to the raw data objects of the HathTrust.

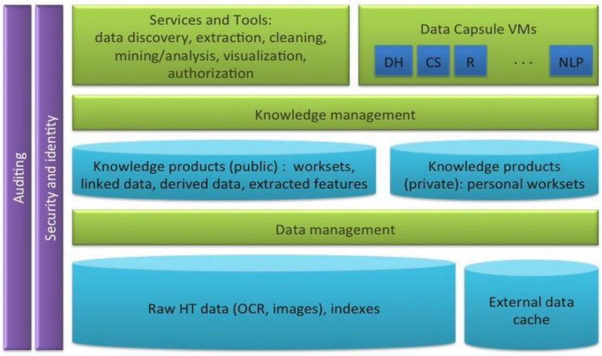
Auditing
Security and identity
Services and Tools:
data discovery, extraction, cleaning,
mining/analysis, visualization,
authorization
Data Capsule VMs
DH
CS
R
. . .
NLP
Knowledge management
Knowledge products (public) : worksets,
linked data, derived data, extracted features
Knowledge products
(private): personal worksets
Data management
Raw HT data (OCR, images), indexes
External data
cache

Figure 2: HTRC Secure Commons architectural components.

The goals of the present work include: the 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; 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; 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 address their specific needs.

The initial work carried out within the WCSA+DC research involved an integrated effort of studying the needs and requirements of the DC 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. This is being

[3] https://www.hathitrust.org

followed by the integration of document-level and document collection processing (genre and topic identification) modules into the DC, as well as the most basic low-level processing (sentencization, tokenization, and POS tagging). The next level of processing planned includes more the computationally intensive NLP modules, such as finding "Named Entities" such as cities, countries, people, etc., as well as performing various levels of constituent- and dependency-based parsing at the sentence level. This will be followed by a detailed evaluation of the NLP services. This involves: (a) assessing the overall performance of each component service within the Data Capsule; and (b) examining the possible workflow configurations of the different services as configured in distinct pipelines to determine the optimal configuration in terms of performance. 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.

4 CL-Facilitated Access to Media Collections

While the previous use case looked at how HLT tools can assist in the discovery of content in text-based corpora that are subject to limited access due to copyright restrictions, the second use case we consider examines the role that CL tools can play in broadcast media collections. We are currently examining the content of the *American Archive of Public Broadcasting*, a collaboration between WGBH and the Library of Congress[4].

The last sixty years of our shared history and culture has been well documented on broadcast media. Many important events, persons, issues, and conflicts have been recorded and discussed in programs at the national and local levels on public television and radio, but much of this material is currently inaccessible and has yet to become part of the historical record. Scholars have long recognized the value of media for the evidence such material can provide about the past as well as the manner in which the public has experienced the news. Likewise, educators have appreciated the ability of audiovisual materials to make history come alive in the classroom setting. Both scholars and educators have been frustrated by the difficulties associated with accessing these materials [8, 14]. From our discussions with archivists and historians, it seems that making historical public broadcasting programs accessible and searchable would be a great enabler for scholarship.

Broadcast media is especially important because of the era it reflects, such as that contained in the American Archive. Broadcast media, once it is made accessible, will add rich archival material to enhance the historical record. Not only is much of our broadcast media history inaccessible, but it is also in danger of degrading and becoming lost to posterity if it remains much longer on station and archival shelves. Reformatting to a digital format is necessary for long-term preservation, but digitizing is only the first step towards improved access of this material for use

[4] `americanarchive.org`

by scholars and educators. Most materials held in storage by stations contain minimal descriptive information, in many cases only a program title [1]. Once this material is digitized, cataloging becomes an extraordinarily labor intensive endeavor. Using CL applications for language-based analysis can help extract significantly more descriptive information; however, optimizing these tools for humanities research requires a digital history team working closely with the CL team to map the output to historical events, places, people, and themes; iteratively improving the computation by revising the assumptions the tool makes; and provide historical interpretations of the new data. These are some of the tasks we are currently carrying out with WBGH and their affiliates.

5 A Language Application Toolkit for Archivists

As our final case study, we examine the role that an HLT platform such as LAPPS Grid can play in helping digital archivists manage their collections [3]. We have recently begun collaborating with several media and archival organizations to explore the applicability of the LAPPS Grid platform to the specific needs of cataloguing, indexing, and retrieving data from media collections. In particular, we have partnered with WGBH of Boston to determine how the configuration and combination of existing computational linguistic (CL) tools can significantly transform the way archivists and librarians describe their media collections. Using WGBH's corpus of archival video and audio transcripts and metadata as a research data set, we have started to develop a toolkit that will be evaluated on its ability to create and enhance metadata and improve discoverability of large and diverse media collections, allowing for substantial progress in the effort and time spent creating and improving records from their collections. This toolkit will be built on top of the Language Application (LAPPS) Grid and will leverage both the tools and workflow composer environment already present in the LAPPS Grid framework.

Audio and video media are, by definition, not text, and therefore opaque to text search engine capabilities. Finding content relevant to one's research question among thousands of hours of programming is time-consuming, involving watching and/or listening to potentially hours of content in order to zero in on relevant content [15]. The availability of descriptive, structured, textual metadata about the content of these collections and about the items they include radically improves search and browse capabilities for researchers; however, the effort to fully describe and catalog these materials is highly labor-intensive and therefore costly.

Such a toolkit is an excellent example of how current CL tools can be configured and combined into a drag-and-drop toolkit and incorporated into archival accessioning and cataloging workflows to significantly ease the work involved in creating rich, descriptive metadata records for each item. The toolkit extends the current capabilities of LAPPS to include tools (already available in projects such as Alveo [4]) for accessing text content in audio and video, as well as access to the publicly available audio and video materials in the WGBH archives. By coupling these tools with sophisticated CL modules for information retrieval, question answering, and text mining, the goal is to be able to create composite workflows to extract and analyze information gleaned from these resources. Through a process of iterative refinement involving both testing of tools and augmentation of supporting resources, we will develop a set of optimal workflows for information extraction from audio and video and evaluate the results on both the collection and individual item levels, to determine the degree to which the annotation process is facilitated. These ready-made workflows will be made available to archivists who can customize them for particular domains and applications, augment supporting resources with additional data, either extracted in earlier steps or derived from other sources. It is our expectation that the project will ultimately produce a set of ready-made (but still customizable) workflow 'packages' that will dramatically reduce the time and cost of metadata production for digital archival materials.

Current archival practice involves the need to dedicate many human hours to create, normalize, and catalog collections; however, cataloging is so time-consuming that it is often the case that collections are put into a queue for cataloging, creating a huge backlog of unprocessed collections. By using such a toolkit, cataloging will be incorporated into the acquisitions workflow and will become a duty of the computer, allowing humans to reallocate their time to work on tasks that still require a human to perform. Instead of catalogers watching hour-long programs and recording descriptive information about the material, the LAPPS toolkit would automate creation of metadata such as creating speech-to-text transcripts, identifying proper names, locations, organizations, and even dates, and would perform metadata clean-up and normalization, and the output of the system could be ingested into the archives' metadata repository automatically. This new workflow could save months, even years, of an archivist's time.

6 What the Digital Humanities Need from CL

The example uses of CL technologies for DH research outlined above demonstrate some of the ways in which the use of CL tools differs between the two fields. In broad terms, the goal of CL is to achieve some level of *automatic* understanding or interpretation of human language data for sophisticated applications such as question answering, machine translation, information retrieval, or summarization [2, 16]. Tool chains for end-to-end performance of this kind of task are developed and tested for their efficacy; the focus is on the final result of applying the tool chain comprising the application, with a relatively high tolerance for error or “noise". In contrast, for DH the focus is more often on finding information that may then be subjected to further human analysis, and may require what in CL are considered to be relatively low-level, enabling tasks, such as tokenization, sentence boundary detection, part-of-speech tagging, named entity recognition, or gross-level dependency analysis. Furthermore, DH deals with data from vastly varying domains and genres, while CL at present tends to focus on a somewhat more limited range of data. Thus for DH the availability of robust, highly accurate tools for low-level tasks that are applicable or configurable to handle vastly different domains is perhaps the highest priority, rather than the overall performance of high-level sophisticated NLP applications. The LAPPS Grid, which provides easy-to-use access to a wide variety of customizable low-level CL tools together with means to evaluate performance on dataset from different domains, already addresses these needs to a large extent. Its adaptation to accommodate the projects described in the previous sections should continue to augment its capabilities to serve the needs of DH research.

References

[1] Howard Besser. The next stage: Moving from isolated digital collections to interoperable digital libraries. *First Monday*, 7(6), 2002.

[2] Julian Brooke, Adam Hammond, and Graeme Hirst. *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, chapter GutenTag: an NLP-driven Tool for Digital Humanities Research in the Project Gutenberg Corpus, pages 42–47. Association for Computational Linguistics, 2015.

[3] Karen Cariani and Casey Davis. Let the computer do the work. In *Presentations from FIAT/IFTA 2016 World Conference*, Warsaw, Poland, http://fiatifta.org/, October 2016.

[4] Steve Cassidy, Dominique Estival, Timothy Jones, Denis Burnham, and Jared Burghold. The alveo virtual laboratory: A web based repository api. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may 2014. European Language Resources Association (ELRA).

[5] J Stephen Downie, Kirstin Dougan, Sayan Bhattacharyya, and Colleen Fallaw. The hathitrust corpus: A digital library for musicology research? In *Proceedings of the 1st International Workshop on Digital Libraries for Musicology*, pages 1–8. ACM, 2014.

[6] B. Giardine, C. Riemer, R. C. Hardison, R. Burhans, L. Elnitski, P. Shah, Y. Zhang, D. Blankenberg, I. Albert, J. Taylor, W. Miller, W. J. Kent, and A. Nekrutenko. Galaxy: a platform for interactive large-scale genome analysis. *Genome Research*, 15(10):1451–55, 2005.

[7] HathiTrust. The workset creation for scholarly analysis + data capsules. https://www.hathitrust.org/2016-spring-update.

[8] Annika Hinze, Craig Taube-Schock, David Bainbridge, Rangi Matamua, and J Stephen Downie. Improving access to large-scale digital libraries through

semantic-enhanced search and disambiguation. In *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 147–156. ACM, 2015.

[9] Susan Hockey. The history of humanities computing. *A companion to digital humanities*, pages 3–19, 2004.

[10] Nancy Ide, James Pustejovsky, Christopher Cieri, Eric Nyberg, Di Wang, Keith Suderman, Marc Verhagen, and Jonathan Wright. The language application grid. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may 2014. European Language Resources Association (ELRA).

[11] Nancy Ide, James Pustejovsky, Keith Suderman, and Marc Verhagen. The Language Application Grid Web Service Exchange Vocabulary. In *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT (OIAF4HLT)*, Dublin, Ireland, 2014.

[12] Nancy Ide, Keith Suderman, James Pustejovsky, Eric Nyberg, Christopher Cieri, and Marc Verhagen. The Language Application Grid and Galaxy. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia, 2016. European Language Resources Association (ELRA).

[13] Nancy Ide, Keith Suderman, Marc Verhagen, and James Pustejovsky. The Language Application Grid Web Service Exchange Vocabulary. In *Revised Selected Papers of the Second International Workshop on Worldwide Language Service Infrastructure - Volume 9442*, pages 18–32. Springer-Verlag New York, Inc., 2016.

[14] Peter Leonard. Mining large datasets for the humanities. *IFLA WLIC*, pages 16–22, 2014.

[15] Johan Oomen, Riste Gligorov, and Michiel Hildebrand. Waisda?: making videos findable through crowdsourced annotations. *Crowdsourcing our Cultural Heritage*, pages 161–184, 2014.

[16] Sandford Bolette Pedersen, Sussi Olsen, and Lars Borin. *Proceedings of the workshop on Semantic resources and semantic annotation for Natural Language Processing and the Digital Humanities at NODALIDA 2015*, chapter Proceedings of the workshop on Semantic resources and semantic annotation for Natural Language Processing and the Digital Humanities at NODALIDA 2015. Northern European Association for Language Technology, 2015.

[17] Beth Plale, Robert McDonald, Yiming Sun, Inna Kouper, Ryan Cobine, J Stephen Downie, Beth Sandore Namachchivaya, and John Unsworth. Hathitrust research center: computational access for digital humanities and

beyond. In *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries*, pages 395–396. ACM, 2013.

[18] Susan Schreibman, Ray Siemens, and John Unsworth. *A companion to digital humanities*. John Wiley & Sons, 2008.

[19] Patrik Svensson. The landscape of digital humanities. *Digital Humanities*, 2010.

[20] Edward Vanhoutte. The gates of hell: History and definition of digitall humanitiesl computing. *Defining Digital Humanities: A Reader*, pages 119–56, 2013.

[21] Marc Verhagen, Keith Suderman, Di Wang, Nancy Ide, Chunqi Shi, Jonathan Wright, and James Pustejovsky. The LAPPS Interchange Format. In *Revised Selected Papers of the Second International Workshop on Worldwide Language Service Infrastructure - Volume 9442*, pages 33–47. Springer-Verlag New York, Inc., 2016.

[22] Christopher Welty and Nacy Ide. Using the right tools: enhancing retrieval from marked-up documents. *Computers and the Humanities*, 33(1-2):59–84, 1999.

[23] Jiaan Zeng, Guangchen Ruan, Alexander Crowell, Atul Prakash, and Beth Plale. Cloud computing data capsules for non-consumptiveuse of texts. In *Proceedings of the 5th ACM workshop on Scientific cloud computing*, pages 9–16. ACM, 2014.