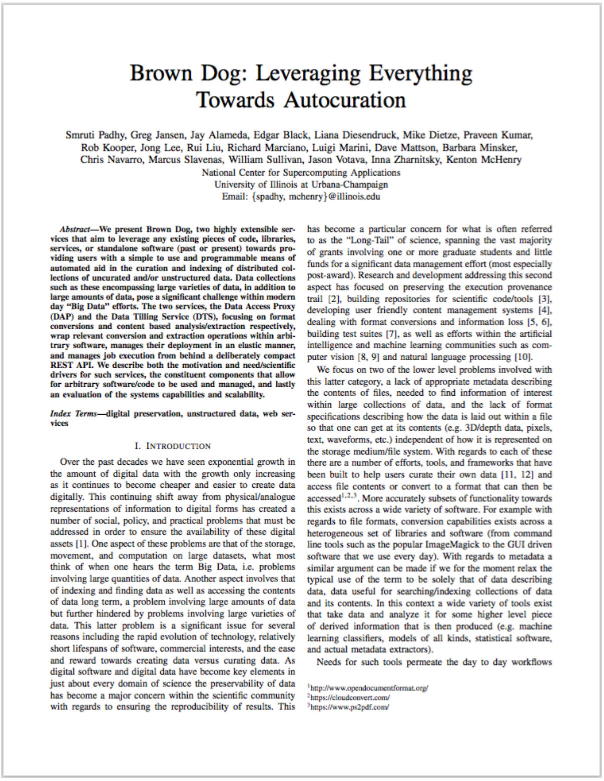
IEEE Big Data

We had a paper accepted into IEEE Big Data this year. The paper goes over the architecture as well as the various components that make up Brown Dog. If you need to cite Brown Dog this is the paper to use:

S. Padhy, G. Jansen, J. Alameda, E. Black, L. Diesendruck, M. Dietze, P. Kumar, R. Kooper, J. Lee, R. Liu, R. Marciano, L. Marini, D. Mattson, B. Minsker, C. Navarro, M. Slavenas, W. Sullivan, J. Votava, K. McHenry, "Brown Dog: Leveraging Everything Towards Autocuration", *IEEE Big Data*, 2015



Brown Dog: Leveraging Everything Towards Autocuration

Smruti Padhy, Greg Jansen, Jay Alameda, Edgar Black, Liana Diesendruck, Mike Dietze, Praveen Kumar,
Rob Kooper, Jong Lee, Rui Liu, Richard Marciano, Luigi Marini, Dave Mattson, Barbara Minsker,
Chris Navarro, Marcus Slavenas, William Sullivan, Jason Votava, Inna Zharnitsky, Kenton McHenry
National Center for Supercomputing Applications
University of Illinois at Urbana-Champaign
Email: {spadhy, mchenry}@illinois.edu

***Abstract*—We present Brown Dog, two highly extensible services that aim to leverage any existing pieces of code, libraries, services, or standalone software (past or present) towards providing users with a simple to use and programmable means of automated aid in the curation and indexing of distributed collections of uncurated and/or unstructured data. Data collections such as these encompassing large varieties of data, in addition to large amounts of data, pose a significant challenge within modern day "Big Data" efforts. The two services, the Data Access Proxy (DAP) and the Data Tilling Service (DTS), focusing on format conversions and content based analysis/extraction respectively, wrap relevant conversion and extraction operations within arbitrary software, manages their deployment in an elastic manner, and manages job execution from behind a deliberately compact REST API. We describe both the motivation and need/scientific drivers for such services, the constituent components that allow for arbitrary software/code to be used and managed, and lastly an evaluation of the systems capabilities and scalability.**

***Index Terms*—digital preservation, unstructured data, web services**

I. Introduction

Over the past decades we have seen exponential growth in the amount of digital data with the growth only increasing as it continues to become cheaper and easier to create data digitally. This continuing shift away from physical/analogue representations of information to digital forms has created a number of social, policy, and practical problems that must be addressed in order to ensure the availability of these digital assets [1]. One aspect of these problems are that of the storage, movement, and computation on large datasets, what most think of when one hears the term Big Data, i.e. problems involving large quantities of data. Another aspect involves that of indexing and finding data as well as accessing the contents of data long term, a problem involving large amounts of data but further hindered by problems involving large varieties of data. This latter problem is a significant issue for several reasons including the rapid evolution of technology, relatively short lifespans of software, commercial interests, and the ease and reward towards creating data versus curating data. As digital software and digital data have become key elements in just about every domain of science the preservability of data has become a major concern within the scientific community with regards to ensuring the reproducibility of results. This has become a particular concern for what is often referred to as the "Long-Tail" of science, spanning the vast majority of grants involving one or more graduate students and little funds for a significant data management effort (most especially post-award). Research and development addressing this second aspect has focused on preserving the execution provenance trail [2], building repositories for scientific code/tools [3], developing user friendly content management systems [4], dealing with format conversions and information loss [5, 6], building test suites [7], as well as efforts within the artificial intelligence and machine learning communities such as computer vision [8, 9] and natural language processing [10].

We focus on two of the lower level problems involved with this latter category, a lack of appropriate metadata describing the contents of files, needed to find information of interest within large collections of data, and the lack of format specifications describing how the data is laid out within a file so that one can get at its contents (e.g. 3D/depth data, pixels, text, waveforms, etc.) independent of how it is represented on the storage medium/file system. With regards to each of these there are a number of efforts, tools, and frameworks that have been built to help users curate their own data [11, 12] and access file contents or convert to a format that can then be accessed[1,2,3]. More accurately subsets of functionality towards this exists across a wide variety of software. For example with regards to file formats, conversion capabilities exists across a heterogeneous set of libraries and software (from command line tools such as the popular ImageMagick to the GUI driven software that we use every day). With regards to metadata a similar argument can be made if we for the moment relax the typical use of the term to be solely that of data describing data, data useful for searching/indexing collections of data and its contents. In this context a wide variety of tools exist that take data and analyze it for some higher level piece of derived information that is then produced (e.g. machine learning classifiers, models of all kinds, statistical software, and actual metadata extractors).

Needs for such tools permeate the day to day workflows

[1] http://www.opendocumentformat.org/
[2] https://cloudconvert.com/
[3] https://www.ps2pdf.com/

Figure 6: **Top:** Conversion simulation results for document conversions to PDF. **Bottom:** Conversion simulation results for image conversions to TIFF.

some archives, we do not assume that TIFF and PDF are the most desirable formats for all archives. Instead, the TIFF and PDF conversions ensure that the content is not trapped in the original format, that Brown Dog can open the original file and get content out of it. Thus far we have identified numerous samples that are trapped in WordPerfect (WP and WPD), Photoshop (PSD), and Windows MetaFile (WMF) formats. A more robust simulation is also being developed through a policy-based format migration which will sample files randomly from the CI-BER collection and consult a lookup table to obtain the preferred preservation format. The simulation will report missing migration paths, as well as missing migration policies, i.e. data files or formats for which no preservation format has been recommended.

V. Conclusion

We have deployed an alpha release of the two services and begun incorporating tools in support of a number of our use cases (e.g. supporting ecological model conversion via PEcAn, supporting Lidar analysis for our hydrology use case, supporting human preference modeling for our green infrastructure use case, as well as other capabilities suited for more general usage). Further, towards supporting the wide range of users across our use cases we have begun developing a number of client interfaces[16] to leverage the DAP and DTS. These include language specific libraries, a bookmarklet interface that can be used to call the services on arbitrary web pages, a Google Chrome extension, a command line interface, incorporation into a scientific workflow system [21], and incorporation back into Clowder as an example within a content management system. Efforts moving forward aim to add additional cloud infrastructure support to the elasticity module, refine the level of granularity considered during scaling by deploying Docker instances of converters/extractors within a single VM, exploring how we might optimize data movement so as to as efficiently as possible handle large data collections, and incorporating/using additional information relevant to provenance such as the estimates of information loss incurred during specific conversions described in [5].

[16] http://browndog.ncsa.illinois.edu/blog.html

Acknowledgment

This research & development has been funded through National Science Foundation Cooperative Agreement ACI-1261582.

References

[1] N. webmaster, "Vision for cyberinfrastructure framework for 21st century science and engineering," 2013. [Online]. Available: http://www.nsf.gov/od/oci/cif21/CIF21Vision2012current.pdf
[2] E. Deelman *et al.*, "Pegasus: A framework for mapping complex scientific workflows onto distributed systems," *Scientific Programming Journal*, 2005.
[3] G. Klimeck *et al.*, "nanohub.org: Advancing education and research in nanotechnology," *Computing in Science and Engineering*, 2008.
[4] L. Marini *et al.*, "Medici: A scalable multimedia environment for research," *The Microsoft e-Science Workshop*, 2010.
[5] K. McHenry, R. Kooper, and P. Bajcsy, "Towards a universal, quantifiable, and scalable file format converter," *The IEEE Conference on e-Science*, 2009.
[6] F. Soper, "The pronom file format registry," *Experts Workgroup on the Preservation of Digital Memory*, 2004.
[7] J. Heard and R. Marciano, "A system for scalable visualization of geographic archival records," *IEEE Symposium on Large Data Analysis and Visualization*, 2011.
[8] L. Diesendruck, R. Kooper, L. Marini, and K. McHenry, "Using lucene to index and search the digitized 1940 us census," *Concurrency and Computation: Practice and Experience*, 2014.
[9] M. Lew, "Content-based multimedia information retrieval: State of the art and challenges," *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2006.
[10] D. Garette and E. Klein, "An extensible toolkit for computational semantics," *International Conference on Computational Semantincs*, 2009.
[11] W. Michener *et al.*, "Participatory design of dataone - enabling cyberinfrastruture for the biological and environmental sciences," *Ecological Informatics*, 2012.
[12] J. Myers *et al.*, "Towards sustainable curation and preservation: The sead project's data services approach," *Interoperable Infrastructures for Interdisciplinary Big Data Sciences Workshop, IEEE eScience*, 2015.
[13] M. Dietze, D. LeBauer, and R. Kooper, "On improving the communication between models and data," *Plant, Cell & Environment*, 2012.
[14] K. McHenry *et al.*, "A mosaic of software," *The IEEE International Conference on eScience*, 2011.
[15] A. Rajasekar, M. Wan, W. Schroeder, and R. Moore, "From srb to irods: Policy virtualization using rule-based data grids," 2005.
[16] L. Marini *et al.*, "Versus: A framework for general content-based comparisons," *IEEE eScience*, 2012.
[17] T. Yeh, T. Chang, and R. Miller, "Sikuli: Using gui screenshots for search and automation," *UIST*, 2009.
[18] M. Beckerle and S. Hanson, "Data format description language (dfdl) v1.0 specification," *Open Grid Forum*, 2014.
[19] G. Galante and L. C. E. d. Bona, "A survey on cloud computing elasticity," in *Proceedings of the 2012 IEEE/ACM Fifth International Conference on Utility and Cloud Computing*, ser. UCC '12, 2012, pp. 263–270.
[20] J. Goecks, A. Nekrutenko, and J. Taylor, "Galaxy: A comprehensive approach for supporting accessible, reproducble, and transparent computation research in the life sciences," *Genome Biology*, 2010.
[21] R. Kooper *et al.*, "Cyberintegrator: A highly interactive scientific process management environment to support earth observations," *Geoinformatics Conference*, 2007.