The Language Application Grid and Galaxy

N. Ide*, J. Pustejovsky, K. Suderman*, M. Verhagen**, C. Cieri†, E. Nyberg‡**
*Vassar College, **Brandeis University, †Linguistic Data Consortium, ‡Carnegie-Mellon University
*Poughkeepsie, NY USA, **Waltham, Mass. USA, †Philadelphia, PA USA, ‡Pittsburgh, PA USA
{ide,suderman}@cs.vassar.edu, ccieri@ldc.upenn.edu, ehn@cs.cmu.edu

Abstract

The NSF-SI^2-funded LAPPS Grid project is a collaborative effort among Brandeis University, Vassar College, Carnegie-Mellon University (CMU), and the Linguistic Data Consortium (LDC) at the University of Pennsylvania, which has developed an open, web-based infrastructure through which massive and distributed resources can be easily accessed, in whole or in part, and within which tailored language services can be efficiently composed, evaluated, disseminated and consumed by researchers, developers, and students across a wide variety of disciplines. The LAPPS Grid project recently adopted Galaxy (Giardine et al., 2005), a robust, well-developed, and well-supported front-end for workflow configuration, management, and persistence. Galaxy allows data inputs and processing steps to be selected from graphical menus, and results are displayed in intuitive plots and summaries that encourage interactive workflows and the exploration of hypotheses. Galaxy provides significant advantages for deploying pipelines of LAPPS Grid web services, including not only means to create and deploy locally-run and even customized versions of the LAPPS Grid as well as running the LAPPS Grid in the cloud, but also access to a huge array of statistical and visualization tools that have been developed for use in genomics research.

Keywords: Web services, Interoperability, Workflow management

1. Overview

The NSF-SI^2-funded LAPPS Grid project[1] is a collaborative effort among Brandeis University, Vassar College, Carnegie-Mellon University (CMU), and the Linguistic Data Consortium (LDC) at the University of Pennsylvania, which has developed an open, web-based infrastructure through which massive and distributed resources can be easily accessed, in whole or in part, and within which tailored language services can be efficiently composed, evaluated, disseminated and consumed by researchers, developers, and students across a wide variety of disciplines (Ide et al., 2014). The LAPPS Grid is part of a larger multi-way international collaboration including key individuals and projects from the U.S., Europe, Australia, and Asia involved with language resource development and distribution and standards-making, who are creating the "The Federated Grid of Language Services" (FGLS) federation (Ishida et al., 2014), a multi-lingual, international network of web service grids and providers. We have also recently entered into a formal partnership with WebLicht/Tübingen[2] and LINDAT/CLARIN (Prague)[3] to create a "trust network" among our sites in order to provide mutual access to all from any one of the three portals. The key to the success of these partnerships is the *interoperability* among tools and services that is accomplished via the service-oriented architecture and the development of common vocabularies and multi-way mappings that has involved key researchers from around the world for over a decade[4].

The LAPPS Grid currently includes a wide range of NLP component web services and provides facilities for service discovery, service composition (including automatic format conversion between tools where necessary), performance evaluation (via provision of component-level measures for standard evaluation metrics for component-level and end-to-end measurement), and resource delivery for a range of language resources, including holdings of the Linguistic Data Consortium (LDC).[5]

The LAPPS Grid project recently adopted Galaxy (Giardine et al., 2005), a robust, well-developed, and well-supported front-end for workflow configuration, management, and persistence.[6] Galaxy allows data inputs and processing steps to be selected from graphical menus, and results are displayed in intuitive plots and summaries that encourage interactive workflows and the exploration of hypotheses. Galaxy provides significant advantages for deploying pipelines of LAPPS Grid web services, including not only means to create and deploy locally-run and even customized versions of the LAPPS Grid as well as running the LAPPS Grid in the cloud, but also access to a huge array of statistical and visualization tools that have been developed for use in genomics research.

2. Galaxy

The Galaxy project[7] started in 2005 to create a system enabling biologists without informatics expertise to perform computational analysis through the web (Giardine et al., 2005). It has since been widely adopted within the life sciences community.

Galaxy is an open-source application[8] that includes tool

[1] http://www.lappsgrid.org

[2] http://weblicht.sfs.uni-tuebingen.de/

[3] https://lindat.mff.cuni.cz/

[4] E.g., the NSF-funded Sustainable Interoperability for Language Technology (SILT) project (NSF-INTEROP 0753069) (Ide et al., 2009), the EU-funded Fostering Language Resources Network (FLaReNet) project (Calzolari et al., 2009), the International Standards Organization (ISO) committee for Language Resource Management (ISO TC37 SC4), and parallel efforts in Asia and Australia, together with the LAPPS project and international collaborators.

[5] http://www.ldc.upenn.edu

[6] http://galaxy.lappsgrid.org

[7] http://galaxyproject.org

[8] Distributed under the terms of permissive Academic Free Li-

integration and history capabilities together with a workflow system for building automated multi-step analyses, a visualization framework including visual analysis capabilities, and facilities for sharing and publishing analyses (Goecks et al., 2012). It is accessed through a graphical interface where data inputs and computational steps are selected from dynamic menus, and results are displayed in plots and summaries that encourage interactive workflows and the exploration of hypotheses.

The main Galaxy site at http://usegalaxy.org is an installation of the Galaxy software combined with many common tools and visualizations that is available to anyone to analyze their data free of charge. The Galaxy ToolShed provides a central location where tool developers can upload both their tool configurations and"recipes" describing how to install necessary dependencies.

Rather than duplicate the extensive work of the Galaxy project, we recently adopted it as the primary workflow management system for the LAPPS Grid.[9] We have worked with the Galaxy development team in order to adapt the system to our domain, and continue this collaboration to both enhance the capabilities we require as well as contribute to the expansion of Galaxy to domains outside the life sciences, which is a current goal of the Galaxy project.

We have contributed Galaxy wrappers to call all LAPPS web services to the Galaxy ToolShed[10]. This enables the creation of complex workflows involving standard NLP components and composite services from a wide range of sources from within an easy-to-use, intuitive workflow engine with capabilities to persist experiments and results. In addition to access to LAPPS Grid tools and data, we have developed and contributed the following capabilities of the LAPPS Grid for use in Galaxy in order to support NLP research and development within that platform, including (1) exploitation of our web service metadata to allow for automatic detection of input/output formats and requirements for modules in a workflow and subsequent automatic invocation of converters to make interoperability seamless and invisible to the user; (2) incorporation of authentication procedures for protected data using the open standard OAuth[11], which specifies a process for resource owners to authorize third-party access to their server resources without sharing their credentials; and (3) addition of a visualization plugin that recognizes the kind of input (coreference, phrase structure) and then uses appropriate off-the-shelf components like BRAT and Graphviz to generate a visualization. Figures 1 and 2 show a simple workflow configuration and a visualization of named entity annotation over a document.

We have adopted and, as necessary, adapted Galaxy strategies for the following:

1. Replication of experiments, pervasive sharing of methods and results. Reproducing experimental results is an essential part of scientific inquiry, providing the foundation for understanding, integrating, and extending results toward new discoveries. However, the field of NLP research and development has been plagued by a chronic lack of potential for replicability of results, as discussed in several recent publications (Pedersen, 2008; Fokkens et al., 2013), blogs[12], and workshops[13]. As a result, there is not only a great deal of re-inventing of the wheel and wasted effort, but also serious inhibition to progress that can be made possible by tapping into the collective intelligence of the community. Evaluation of results is also seriously hampered when details of an experiment (including versions and parameters for data, software) are not included in papers, which is all too often the case. Our adaptation of the Galaxy workflow system enables us to foster replicability and reuse for NLP by providing the following capabilities (see (Goecks et al., 2010) for a comprehensive overview of Galaxy's sharing and publication capabilities):

- automatic recording of inputs, tools, parameters and settings used for each step in an analysis in a publicly viewable history, thereby ensuring that each result can be exactly reproduced and reviewed later;
- provisions for sharing datasets, histories, and workflows via web links, with progressive levels of sharing including the ability to publish in a public repository;
- ability to create custom web-based documents to communicate about an entire experiment, which represent a step towards the next generation of online publication or publication supplement.

In addition to enabling other users to replicate an experiment, the individual user can develop a rich, organized catalog of reusable workflows rather than starting from scratch each time or trying to navigate a collection of *ad hoc* analysis scripts. Similarly, it is possible to repeatedly apply a command history on different data. Once an analysis is done, the record eliminates ambiguity as to which result used which settings provide critical information for follow-up analysis.

2. Transparency. Research publications involving computationally intensive analysis can be difficult to understand (Nekrutenko and Taylor, 2012; Sandve et al., 2013). Galaxy provides means for researchers to make their analyses available to others in ways that are easy to understand, primarily via Galaxy histories that can be shared or pointed to in papers to demonstrate exactly what has been done. In addition, Galaxy Pages and free-form annotations provide ways to add context to analysis to describe the reasoning behind an analysis and parameter settings.

3. Enhancement of the user base and community involvement. The Galaxy project has had notable success in community building and outreach, comparable to what we hope to achieve for the LAPPS Grid. Inspired by their success, we are adopting the Galaxy project's outreach strategies in order to most effectively reach, teach, and involve the community in the LAPPS Grid, as well as promote community engagement in LAPPS development via

cense: http://getgalaxy.org

[9] http://galaxy.lappsgrid.org

[10] https://toolshed.g2.bx.psu.edu

[11] http://oauth.net

[12] E.g., http://nlpers.blogspot.com/2006/11/reproducible-results.html

[13] E.g., Replicability and Reusability in Natural Language Processing: from Data to Software Sharing: http://nl.ijs.si/rrnlp2015/

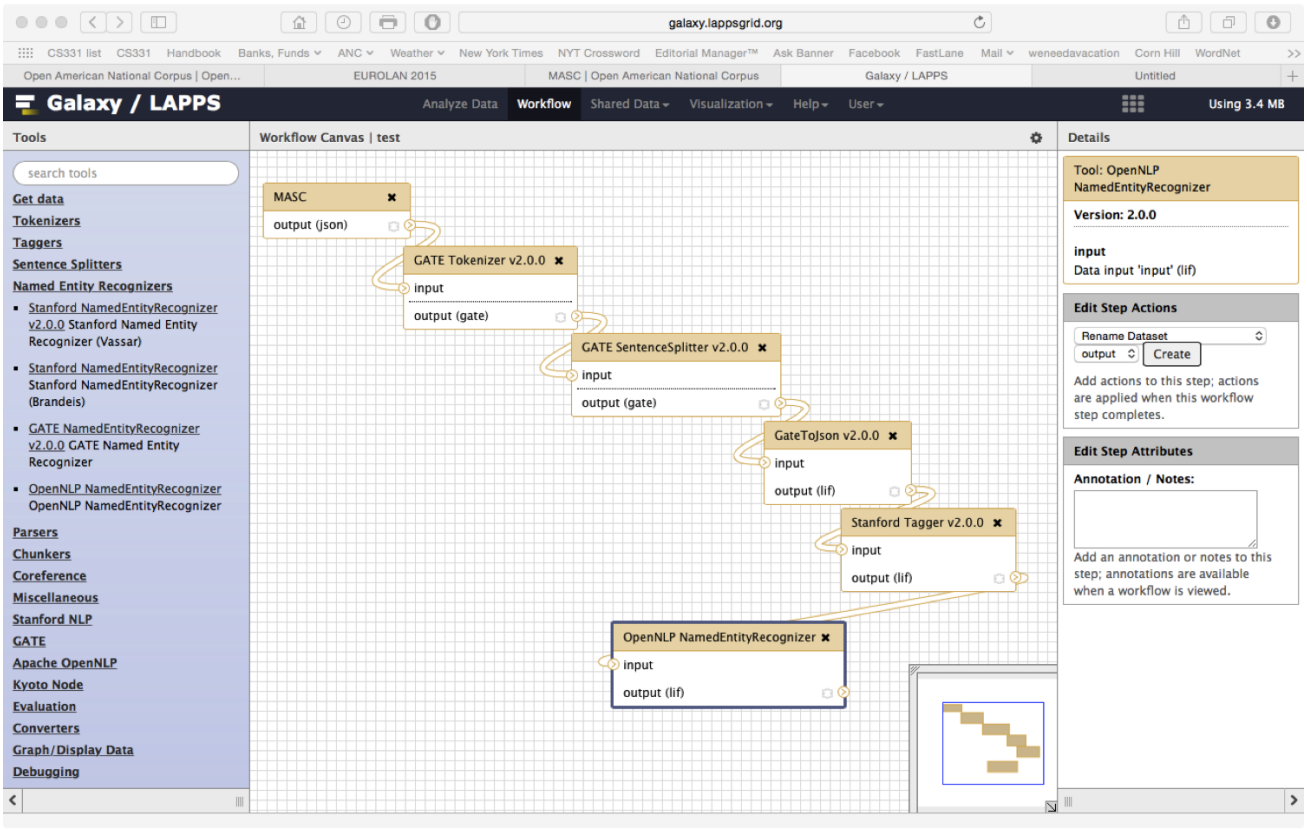

Figure 1: The LAPPS/Galaxy Interface: Workflow configuration

sharing of tools, data, and (especially) workflows and results. The community consists of two main constituencies: the user community, who is supported through outreach and training activities including both carrying out training directly and developing training materials, the developer community, consisting both of LAPPS tool developers and contributors to the LAPPS Grid development itself, whom we support through reviewing and merging pull requests (additions and modifications to the source code submitted to us through the version control system). We have also set up direct support for both of these constituencies by providing a question/answer environment on the LAPPS wiki and dedicated mailing lists, which are routinely monitored by the LAPPS Grid development team.

2.1. NLP Galaxy "Flavor"

Galaxy recently added support for running tools from the Galaxy ToolShed within Docker containers. Docker[14] allows users to package an application with all of its dependencies into a standardized unit into a Docker image, which is an easily distributable full-fledged installation that can be used for testing, teaching, and presenting new tools and features. Within Galaxy, Docker support can be used to create a *Galaxy Flavor*, which is a Galaxy image configured with a tool suite for a particular task or application.

We have contributed a "Galaxy Flavor" including all LAPPS Grid services and resources, which is effectively a pre-configured virtual machine (VM) that can be run in any of several VMs (e.g., VirtualBox, AmazonEC2, Google, Microsoft Azure, VMWare, OpenStack, etc.). This enables users to access only the NLP subset of tools if desired, as well as to download a Galaxy-stable image and run it locally. This capability is ideal for class work, workshops, and presentations as it allows full-blown installations to be easily shared and run. This also provides the capability to run the LAPPS Grid in environments where there is no internet access, or where security requires a completely local environment.

2.2. LAPPS/Galaxy Synergy

An additional, and potentially hugely significant, outcome of the LAPPS/Galaxy collaboration is that it enables the use of LAPPS Grid NLP services to extract information from repositories of biomedical publications such as PubMed[15] and pass it on to biomedical analysis and visualization tools available in Galaxy. The synergistic development of capabilities supporting both NLP and genomic analysis within the Galaxy framework can have a significant impact on work in both fields, providing each with access to methods for exploration and analysis of data that were previously unfamiliar. For example, NLP researchers will benefit enormously from access to sophisticated visualization software for display and analysis of results common to research in the life sciences, but rarely used in NLP research. Similarly, biologists will be able to take advantage of bio-oriented NLP web services for text mining of bio-entities and relations from textual sources, and via capabilities already present in Galaxy, integrate them into existing bio-

[14] https://www.docker.com

[15] http://www.ncbi.nlm.nih.gov/pubmed

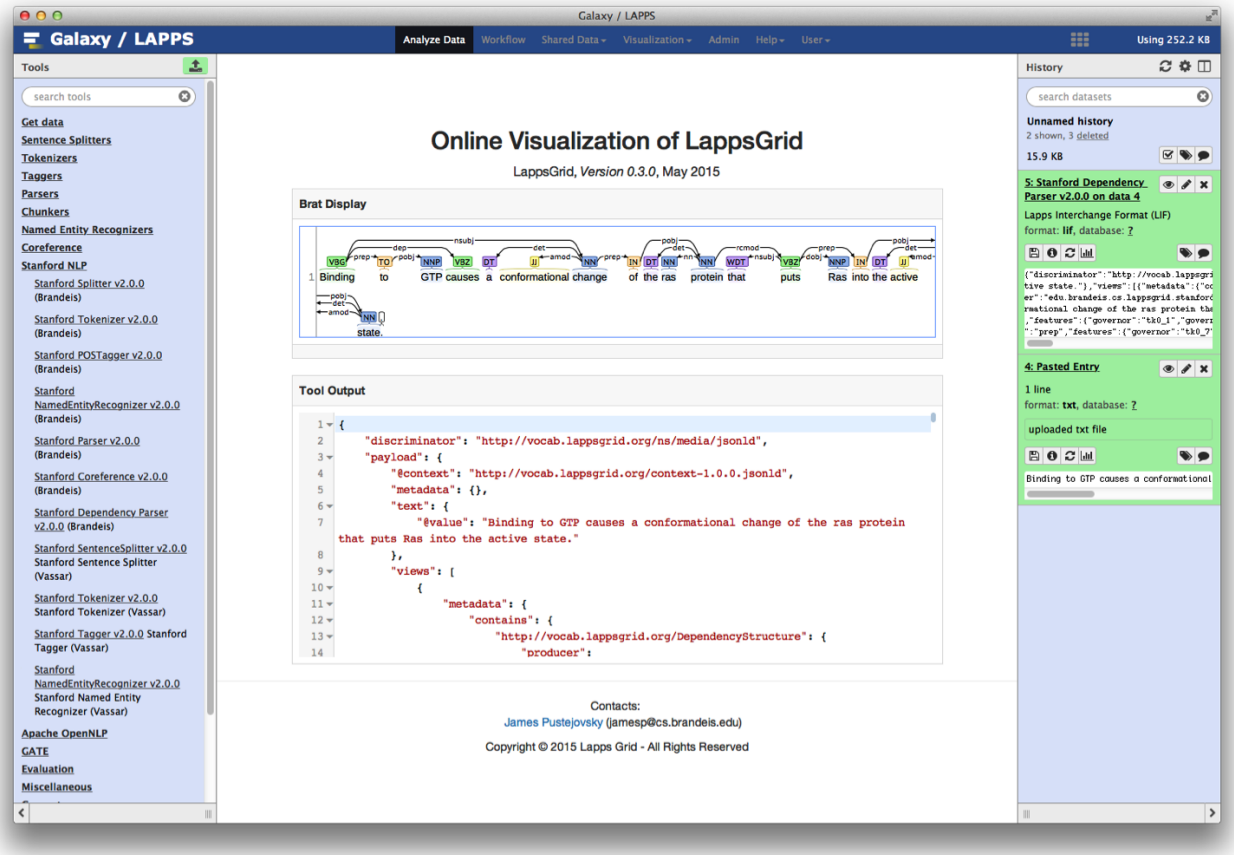

Figure 2: Visualization of a named entity annotation using LAPPS/Galaxy

data resources and analysis tools. Provenance indexing to the article would enable researchers to both verify and contextualize the results. Researchers in biology could also exploit human-in-the-loop capabilities to enhance iterative analysis of genomics and other data.

3. Conclusion

The LAPPS Grid / Galaxy collaboration provides an ideal bridge between the NLP and life science domains. The integration of data, tools, as well as workflows and methods from previously distinct scientific communities can provide unprecedented capabilities for both the emerging field of BioNLP and biomedical and genomic science.

NOTE: We hope to provide a demo of the LAPPS/Galaxy framework at the conference.

Acknowledgments

This work was supported by National Science Foundation grants NSF-ACI 1147944 and NSF-ACI 1147912.

Nicoletta Calzolari, P. Baroni, N. Bel, G. Budin, K. Choukri, S. Goggi, J. Mariani, M. Monachini, J. Odijk, S. Piperidis, V. Quochi, C. Soria, and A Toral, editors. 2009. *Proceedings of "The European Language Resources and Technologies Forum: Shaping the Future of the Multilingual Digital Europe"*. ILC-CNR.

Antske Fokkens, Marieke van Erp, Marten Postma, Ted Pedersen, Piek Vossen, and Nuno Freire. 2013. Offspring from reproduction problems: What replication failure teaches us. In *Proceedings of the Conference of The Association for Computational Linguistics*, pages 1691–1701. The Association for Computational Linguistics.

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. 2005. Galaxy: a platform for interactive large-scale genome analysis. *Genome Research*, 15(10):1451–55.

Jeremy Goecks, Anton Nekrutenko, and James Taylor. 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome biology*, 11:R86.

Jeremy Goecks, Nate Coraor, The Galaxy Team, Anton Nekrutenko, and James Taylor. 2012. NGS Analyses by Visualization with Trackster. *Nature Biotechnology*, 30(11):10361039.

Nancy Ide, James Pustejovsky, Nicoletta Calzolari, and Claudia Soria. 2009. The SILT and FlaReNet international collaboration for interoperability. In *Proceedings of the Third Linguistic Annotation Workshop, ACL-IJCNLP*, August.

Nancy Ide, James Pustejovsky, Christopher Cieri, Eric Nyberg, Di Wang, Keith Suderman, Marc Verhagen, and Jonathan Wright. 2014. The Language Application Grid. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May. European Lan-

guage Resources Association (ELRA).

Toru Ishida, Yohei Murakami, Donghui Lin, Takao Nakaguchi, and Masayuki Otani. 2014. Open Language Grid–Towards a Global Language Service Infrastructure. In *The Third ASE International Conference on Social Informatics (SocialInformatics 2014)*, Cambridge, Massachusetts, USA.

Anton Nekrutenko and James Taylor. 2012. Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nature Reviews Genetics*, 13(9):667–672, September.

Ted Pedersen. 2008. Empiricism is not a matter of faith. *Computational Linguistics*, 34(3), September.

Geir Kjetil Sandve, Anton Nekrutenko, James Taylor, and Eivind Hovig. 2013. Ten simple rules for reproducible computational research. *PLoS Computational Biology*, 9(10):e1003285, 10.