The LAPPS Grid: Current State and Next Steps

Nancy Ide, Keith Suderman
Department of Computer Science
Vassar College
Poughkeepsie, New York USA
{ide,suderman}@cs.vassar.edu

Eric Nyberg
Language Technology Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania USA
ehn@cs.cmu.edu

James Pustejovsky, Marc Verhagen
Department of Computer Science
Brandeis University
Waltham, Massachusetts USA
{jamesp,marc}@cs.brandeis.edu

Abstract

The US National Science Foundation (NSF) SI^2-funded LAPPS Grid project has developed an open-source platform for enabling complex analyses while hiding complexities associated with underlying infrastructure, that can be accessed through a web interface, deployed on any Unix system, or run from the cloud. It provides sophisticated tool integration and history capabilities, a workflow system for building automated multi-step analyses, state-of-the-art evaluation capabilities, and facilities for sharing and publishing analyses. This paper describes the current facilities available in the LAPPS Grid and outlines the project's ongoing activities to enhance the framework.

1 Introduction

The US National Science Foundation (NSF) SI^2-funded LAPPS Grid project is a collaborative effort among Vassar College, Brandeis University, Carnegie-Mellon University (CMU), and the Linguistic Data Consortium (LDC) at the University of Pennsylvania. The LAPPS Grid is an open-source platform for enabling complex analyses while hiding complexities associated with underlying infrastructure. The platform can be accessed through a web interface, deployed on any Unix system, or run from the cloud. It provides sophisticated tool integration and history capabilities, a workflow system for building automated multi-step analyses, state-of-the-art evaluation capabilities, and facilities for sharing and publishing analyses. The LAPPS Grid is highly customizable and integrates with a wide variety of computing environments, ranging from laptop computers to clusters to compute clouds. The LAPPS Grid is part of the Federated Grid of Language Services (FGLS) (Ishida et al., 2014), a multi-lingual, international network of web service grids and providers in Asia and Europe, and is currently federating with two major frameworks in the pan-European CLARIN project. Through these federations, LAPPS Grid users have interoperable access to all the tools and modules in the other platforms, thus creating the largest network of interoperable components for language-related analysis and activity in the world.

This paper describes the current facilities available in the LAPPS Grid, which have been continually expanded over the past year, and provides an overview of the ways in which the LAPPS Grid compares to similar frameworks. We also outline the project's ongoing activities to enhance the framework, in particular its collaboration with the developers of the Galaxy framework and creation of educational materials.

Proceedings of WLSI/OIAF4HLT,
pages 11–18, Osaka, Japan, December 12 2016.

2 Current facilities

The LAPPS Grid currently provides a broad suite of interoperable Human Language Technologies (HLT) tools and data, 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).[1] The list of HLT processing tools and resources currently available in the LAPPS Grid can be found at http://www.lappsgrid.org/language-services and includes dozens of the most commonly used HLT toolkits. In addition, through our federation with international partners, LAPPS Grid users have (or will soon have) access to hundreds of multi-lingual and multi-modal tools, applications, evaluation tools, lexicons, and data sources[2].

The key feature of the LAPPS Grid that differentiates it from most other platforms is the *interoperability* among tools and services, which is achieved through adoption of standard protocols and formats together with the development of additional standards for interchange. Syntactic interoperability among LAPPS Grid services is ensured by a JSON-LD format called the LAPPS Interchange Format (LIF) that groups annotations in views, where each view contains metadata that spells out the information contained in that view, including information necessary to determine compatibility with other tools and data. Semantic interoperability is achieved via references to definitions in the Web Services Exchange Vocabulary (WSEV). The WSEV has been built bottom up, driven by the needs of components in the LAPPS Grid and closely following standard practice in the field as well as adopting, where possible, existing terminology and type systems. Both LIF and the WSEV are described in detail elsewhere (Verhagen et al., 2016; Ide et al., 2014; Ide et al., 2016).

Another distinctive feature of the LAPPS Grid is its Open Advancement (OA) Evaluation system, a sophisticated environment that was used to develop IBM's Jeopardy-winning Watson. OA can be simultaneously applied to multiple variant workflows involving alternative tools for a given task, and the results are evaluated and displayed so that the best possible configuration is readily apparent. Similarly, the weak links in a chain are easily detected, which can lead to module-specific improvements that affect the entire process. The inputs, tools, parameters and settings used for each step in an analysis are recorded, thereby ensuring that each result can be exactly reproduced and reviewed later, and any tool configuration can be repeatedly applied to different data.

Several HLT software developers and projects have contributed their components and systems for inclusion in the LAPPS Grid, in order to provide exposure of these tools to a wide user audience and, crucially, to render them interoperable with other tools in the LAPPS Grid, thus allowing for their immediate inclusion in workflows supporting sophisticated applications as well as evaluation of their performance side-by-side with comparable components. Although many contributors host their own contributed services (which are called from within the LAPPS Grid), where necessary the LAPPS Grid provides hosting to ensure that software remains available to the community. Recently contributed tools include all core tools from University of Darmstadt's DKPro[3], the AIFdb services for Argumentation analysis[4] (Lawrence et al., 2012), the SWIFT Aligner for cross-lingual transfer (Gilmanov et al., 2014), the EDISON feature extraction framework[5] (Sammons et al., 2016) and other tools available from the University of Illinois (e.g., semantic role labelers, entity extractors), among others. In addition, several of the basic components produced by the ARIEL team working within DARPA's Low Resource Languages for Emergent Incidents (LORELEI) program have been integrated into the LAPPS Grid, which include tools and data to support a wide array of under-resourced languages.

The LAPPS Grid 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 the HathiTrust Research

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

[2] For example, see http://langrid.org/service_manager/language-services for a partial list of services available though the FGLS, and https://vlo.clarin.eu/ for tools and resources available through CLARIN.

[3] https://dkpro.github.io

[4] http://www.aifdb.org

[5] https://cogcomp.cs.illinois.edu/page/software_view/Edison

Center's (HTRC) massive digital library)[6], in order enhance search and discovery across the library by complementing traditional volume-level bibliographic metadata with new metadata, using specially-developed LAPPS Grid-based CL applications. Finally, we are currently working with IBM to integrate its Watson services into the LAPPS Grid so that they will be available for community use for the first time, and interoperate with the very wide range of HLT modules the LAPPS Grid provides. This will open the door to rapid evaluation of alternative component pipelines using state-of-the-art metrics and procedures.

The LAPPS Grid uses Docker[7], a recently developed and increasingly popular virtualization platform, as a way to distribute and deploy the complete LAPPS Grid or parts thereof. Docker images are completely self-contained: users can download docker images and run them on their own servers or laptops or in the cloud without the need to install supporting components. We currently have several docker virtual servers running versions of the LAPPS Grid on Amazon Web Services and JetStream (http://jetstream-cloud.org).

To provide an intuitive, easy-to-use interface and management system, the LAPPS Grid project adopted the Galaxy framework (Giardine et al., 2005), a robust, well-developed platform that includes tool 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). In addition to providing an intuitive user interface for workflow composition appropriate for non-technical users, Galaxy also provides support for replication of experiments and sharing of methods and results via automatic recording of all inputs, tools, parameters and settings in an experiment and provisions for sharing datasets, histories, and workflows via web links.[8] Replication capabilities are a vital need in the field of HLT, which has been plagued by a chronic lack of potential for replicability and reuse described in several recent publications (Pedersen, 2008; Fokkens et al., 2013), blogs[9], and workshops[10]. Facilities for reproducibility also enable users to develop organized catalogs of reusable workflows, rather than starting from scratch each time or trying to navigate a collection of *ad hoc* analysis scripts, and/or apply a command history to different data.

We have contributed Galaxy wrappers to call all LAPPS Grid web services to the Galaxy ToolShed[11]. The enables user to create a LAPPS/Galaxy instance locally or in the cloud that includes only the Galaxy "NLP Flavor" (in Galaxy terminology), which comprises all and only LAPPS Grid services and resources, if so desired.

The LAPPS Grid benefits from the Galaxy team's integration of interactive analysis environments, including Jupyter (Perez and Granger, 2007) and RStudio (Gandrud, 2013). Jupyter in particular is of interest to LAPPS Grid users; the LAPPS Grid team has developed a LAPPS Grid Services DSL[12] (LSD) kernel that can be used to interact with the LAPPS Grid services.[13] Jupyter Notebooks can contain executable code and documentation in one location, thus allowing fast templating and prototyping of services. There is no need to compile Java/Groovy code and deploy services to web servers for evaluation and testing, thus making it easy for students and non-technical users to develop sophisticated workflows and/or add their own components with no programming effort. Jupyter also provides "human in the loop" functionality, by allowing one to run a pipeline in the LAPPS Grid, manipulate the Galaxy history items in Jupyter, and finally upload the results to Galaxy for further processing. This facility is especially important for HLT development that involves iterative enhancement of training data on the basis of error analysis, etc.

---

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

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

[8] See (Goecks et al., 2010) for a comprehensive overview of Galaxy's sharing and publication capabilities.

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

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

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

[12] Groovy Domain Specific Language.

[13] The kernel is available from https://github.com/lappsgrid-incubator/jupyter-lsd-kernel; see also http://wiki.lappsgrid.org/technical/jupyter.html. A Docker image is also available including Jupyter and the LSD kernel, for installation-free usage.

3 Comparison with other frameworks

The LAPPS Grid differs from existing frameworks primarily because of the standards for interoperability it implements to enable components from different sources to be seamlessly interfaced, and the resulting ease and transparency with which components can be added and manipulated. Frameworks such as UIMA and GATE, which also provide multiple tools that can be pipelined together, require a relatively steep learning curve to use and considerable programming effort to add or modify components. Similarly, the Natural Language Toolkit (NLTK) requires Python programming and effectively limits the user to the tools that are built-in to the system. In contrast, modules can be easily added to the LAPPS Grid by wrapping them as a service, using provided templates; and, more importantly, no programming experience or technical expertise is required, since workflows are constructed using the Galaxy project's workflow management framework. This makes the LAPPS Grid ideal for instructional use.

The recently introduced Kathaa system (Mohanty et al., 2016) provides functionality similar to the LAPPS Grid, but allows modules to be interfaced only if compatible with one another–i.e., there is no attempt to standardize inputs and outputs among modules, so that mixing and matching of different tools that perform the same function is limited. The LAPPS Grid's Open Advancement evaluation modules, which exploit the ability to construct alternative pipelines in order produce statistics identifying the most effective tool sequence and/or components accounting for the largest proportion of error, are also unique; Kathaa in contrast has only basic evaluation facilities.

Another similar framework is the Alveo system (Cassidy et al., 2014), which also uses Galaxy as its workflow engine and renders tools interoperable using representations that are the same as, or isomorphic to, the LAPPS Grid's. Alveo is dedicated primarily to multi-modal data and applications and therefore includes a very different suite of modules and datasets; in the future, federating with Alveo could enable access to its facilities from within the LAPPS Grid.

4 Collaboration with Galaxy

Since the LAPPS Grid project adopted Galaxy in late 2014, we have had multiple interactions with the Galaxy development team to extend and/or modify Galaxy to meet some of our needs. These interactions have been so successful and mutually beneficial that our two projects are establishing a collaboration that will contribute to the continued development of both infrastructures. We believe that the synergistic development of capabilities supporting both HLT and genomic analysis within the Galaxy framework could have a significant impact on both fields, not only by enhancing Galaxy functionality overall, but also by integrating data, tools, as well as workflows and methods from previously distinct scientific communities. For example, HLT researchers will benefit from access to sophisticated visualization software for display and analysis of results common to research in the life sciences, but rarely used in HLT research, and biologists will be able to take advantage of bio-oriented HLT web services for mining of bio-entities and relations from textual sources and (via capabilities already present in Galaxy) integrate them into existing bio-data resources and analysis tools.

There are several key enhancements to Galaxy that are necessary to scale the LAPPS Grid to handle sophisticated HLT development and education that are synergistic with enhancements that the Galaxy team is already proposing to benefit life sciences research. These include the ability to efficiently handle large data collections (i.e., a large number of small datasets, vs. the single large dataset common in genomics research); and enhanced ability to investigate data using Interactive Environments (IEs), which is intended to assist life sciences research by allowing for injection of visual analytics in a workflow, but for HLT would provide "human-in-the-loop" capabilities to support iterative improvement in machine learning. At the same time, the Galaxy developers have proposed enhancements to their framework that are already in development for LAPPS Grid tools, for example, automatic management of software dependencies (both syntactic and semantic), and systematic workflow exploration and optimization, as provided in our OA evaluation suite of tools. Thus the collaboration can provide significant benefits to both the HLT and life sciences communities.

The two projects are also working together to create automatic workflow creation capabilities, such as a "Pipeline Wizard" to provide a wizard-like interface to guide the user through the logical steps in creating instances of known workflow(s); and/or a "Pipeline-via-Dialog" that allows the user to specify the desired analytic goal in simple English, where the system constructs the desired workflow automatically.

The LAPPS Grid will also benefit from Galaxy's efforts to expand their access to the Extreme Science and Engineering Discovery Environment (XSEDE) (Towns et al., 2014) and JetStream, to which the LAPPS Grid has recently deployed several cloud-based instances for use in courses and specific projects. The Galaxy team is implementing means to take full advantage of High Performance Computing resources, including parallelization and scaling support, as well as the latest dependency management, code versioning, and virtualization techniques, all of which will serve LAPPS Grid needs. Our collaboration will ensure that implementation within Galaxy of these and other capabilities that can serve both disciplines are flexible and scalable to other disciplines' needs in the future. It will also enable gathering user feedback from the perspectives of both communities to feed development of both platforms.

4.1 Federation with CLARIN

The Mellon Foundation recently funded a collaborative effort make the LAPPS Grid interoperable with the pan-European CLARIN project's WebLicht/Tübingen[14] and LINDAT/CLARIN (Prague)[15] frameworks. The effort will create a "trust network" among the LAPPS Grid and CLARIN sites, in order to to make the services we currently provide, as well as future services we will develop, transparently interoperable and mutually accessible from our respective infrastructures. A focus of activity will be to adapt the LAPPS Grid to accommodate the Shibboleth (SAML 2) protocol in order to allow access to login-protected (but otherwise generally accessible) content and services by anyone authenticated through the CLARIN authentication and identification mechanism.

Collaborative access between the LAPPS Grid and CLARIN is achievable largely due to standards and best practices for interoperability that have emerged over the past decade and that were applied internally to both the LAPPS Grid and CLARIN. The federation with CLARIN complements the LAPPS Grid's membership in the Federated Grid of Language Services (FGLS), which provides interoperable access to the University of Kyoto's Language Grid[16] and four other Asian grids, plus a grid under development at the European Language Resources Association (ELRA)[17]. Through these two federations we have vastly increased the availability of multi-lingual and multi-modal resources and tools in the LAPPS Grid and expanded the range of users beyond the HLT community we originally intended to serve, including users involved in inter-cultural communication and the digital humanities research.

4.2 Course development

The LAPPS Grid team recently developed and delivered a four-day short course on Intelligent Information Systems for Analysts, using a purpose-built instance of the LAPPS Grid software deployed in the cloud. The LAPPS Grid "Discovery" instance supports access to selected corpora from the LDC Gigaword data source, along with a variety of open-source HLT tools (sentence splitters, named entity recognizers, passage extractors, passage rankers), and was the focus of hands-on exercises for comparative evaluation of HLT components, composition of analytic pipelines from individual components, and analytic approaches that can learn from user knowledge and feedback.

Based on this experience, we recognized that the LAPPS Grid is an ideal framework to help students interact with and design complex HLT systems without the need for sophisticated technical skills. To that end, we are focusing on the development of curriculum materials to enable students to learn the skills required to rapidly analyze large bodies of language data in practical contexts, in collaboration with several partners at universities in the US, Canada, and Europe (Brandeis University, Carnegie Mellon University, Vassar College, University of Dundee (Scotland), George Washington University, and University of British Columbia). The course materials are being designed to effectively teach to a first-year computer science or computational linguistics student the skills required to create new analytic modules, compose pipelines integrating those modules, and compare them to pipelines which integrate existing services–all

[14]http://weblicht.sfs.uni-tuebingen.de/
[15]https://lindat.mff.cuni.cz/
[16]http://langrid.org
[17]http://www.elra.info.

without the often prohibitive need to acquire and manage data sets and software suites on an individual basis–within the context of on-line, self-driven, hands-on exercises supported by the LAPPS Grid.

A secondary goal is to effectively teach to a first-year computer science graduate student the skills required to compose, evaluate and optimize state-of-the-art software solutions using standard datasets, evaluation metrics and corpora for multi-language, multi-media information systems that process text, audio, images and video. Providing such a capability for graduate instruction would unlock the potential of the vast store of data present in repositories such as the Linguistic Data Consortium and the multilingual resources available from federated grids. The end result will be to train the next generation of HLT researchers, developers, and language analysts, who will use advanced technologies such as the LAPPS Grid on a regular basis to augment simpler analyses available via web search and the use of standalone tools.

All course materials will be freely available from the Open Learning Initiative (OLI)[18] and a dedicated repository maintained by the LAPPS Grid project. The materials will be accompanied with ready-made docker images that can be used as is or easily customized to suit specific pedagogical goals.

5 Enhancement of Open Advancement Evaluation Capabilities

A final focus of current activity is extension of the component and pipeline evaluation capabilities in the LAPPS Grid to support parallel evaluation and broader adoption by end-user communities. These include

- parallel exploration of alternative pipelines, drawing on recent work on Configuration Space Exploration (CSE) and Phased Ranking Models (Garduno et al., 2013; Yang et al., 2013; Liu and Nyberg, 2013). We will provide the capability to specify an abstract pipeline with multiple possible components per phase, and the corresponding capability to explore the different alternatives automatically in parallel using the CSE technique.

- accessible display of results, with support for different visualizations for pipeline results (both the data objects produced by the pipeline as well as the evaluation metrics measured for each pipeline test).

Our federation with multiple grids and frameworks will make it possible to evaluate the performance of vast arrays of alternative tool pipelines that would otherwise be unavailable or prohibitively difficult to use together. It will also provide, for the first time, the capability to study and evaluate tools performance on data in a huge set of different languages.

6 Future Developments and Conclusion

The LAPPS Grid project is developing in two main directions. First, it is expanding not only the range of tools, resources, and components available within the framework itself, but also providing access to hundreds, if not thousands, of resources and tools available from federated partners. Second, it is focusing on development of cloud-deployed, customizable course materials that can dramatically enhance the training of the next generation of HLT researchers, and provide significantly improved materials for non-technical users.

In addition to the above developments, we are also engaging in significant community outreach to encourage tool developers to contribute to the LAPPS Grid repository, through which process they become interoperable with all other the LAPPS Grid tools and components. Ultimately, by this means and federation with other grids and frameworks throughout the world, we hope to develop a massive library of interoperable components for HLT research, development, and education.

Acknowledgements

The work described here is supported by US National Science Foundation awards ACI-1147912, ACI-1147944, ACI-1547621, and ACI-1547270, and a grant from the Mellon Foundation.

[18]http://oli.cmu.edu

References

Steve Cassidy, Dominique Estival, Timothy Jones, Denis Burnham, and Jared Burghold. 2014. 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. European Language Resources Association (ELRA).

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.

C. Gandrud. 2013. *Reproducible Research with R and RStudio*. CRC Press.

Elmer Garduno, Zi Yang, Avner Maiberg, Collin McCormack, Yan Fang, and Eric Nyberg. 2013. CSE Framework: A UIMA-based Distributed System for Configuration Space Exploration Unstructured Information Management Architecture. In Peter Klgl, Richard Eckart de Castilho, and Katrin Tomanek, editors, *UIMA@GSCL*, CEUR Workshop Proceedings, pages 14–17. CEUR-WS.org.

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.

Timur Gilmanov, Olga Scrivner, and Sandra Kbler. 2014. SWIFT Aligner, A Multifunctional Tool for Parallel Corpora: Visualization, Word Alignment, and (Morpho)-Syntactic Cross-Language Transfer. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

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

Nancy Ide, Keith Suderman, Marc Verhagen, and James Pustejovsky. 2016. The Language Application Grid Web Service Exchange Vocabulary. In *Revised Selected Papers of the Second International Workshop on Worldwide Language Service Infrastructure - Volume 9442*, WLSI 2015, pages 18–32, New York, NY, USA. Springer-Verlag New York, Inc.

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.

John Lawrence, Floris Bex, Chris Reed, and Mark Snaith. 2012. AIFdb: Infrastructure for the Argument Web. In *Computational Models of Argument - Proceedings of COMMA 2012, Vienna, Austria, September 10-12, 2012*, pages 515–516.

Rui Liu and Eric Nyberg. 2013. A Phased Ranking Model for Question Answering. In *Proceedings of the CIKM'13*.

Sharada Prasanna Mohanty, Nehal J Wani, Manish Srivastava, and Dipti Misra Sharma. 2016. Kathaa: A visual programming framework for nlp applications. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 92–96, San Diego, California, June. Association for Computational Linguistics.

Ted Pedersen. 2008. Empiricism is Not a Matter of Faith. *Computational Linguistics*, 34(3), September.

Fernando Perez and Brian E. Granger. 2007. IPython: A System for Interactive Scientific Computing. *Computing in Science and Engg.*, 9(3):21–29, May.

Mark Sammons, Christos Christodoulopoulos, Parisa Kordjamshidi, Daniel Khashabi, Vivek Srikumar, Paul Vijayakumar, Mazin Bokhari, Xinbo Wu, and Dan Roth. 2016. EDISON: Feature Extraction for NLP, Simplified. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA).

John Towns, Timothy Cockerill, Maytal Dahan, Ian T. Foster, Kelly P. Gaither, Andrew S. Grimshaw, Victor Hazlewood, Scott Lathrop, David Lifka, Gregory D. Peterson, Ralph Roskies, J. Ray Scott, and Nancy Wilkins-Diehr. 2014. XSEDE: Accelerating Scientific Discovery. *Computing in Science and Engineering*, 16(5):62–74.

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, pages 33–47, New York, NY, USA. Springer-Verlag New York, Inc.

Zi Yang, Elmer Garduno, Yan Fang, Avner Maiberg, Collin McCormack, and Eric Nyberg. 2013. Building Optimal Information Systems Automatically: Configuration Space Exploration for Biomedical Information Systems. In *Proceedings of the CIKM'13*.