The Language Application Grid

Nancy Ide*, **James Pustejovsky****, **Keith Suderman***
Marc Verhagen**, **Chris Cieri†**, **Eric 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, {jamesp,marc}@cs.brandeis.edu, ccieri@ldc.upenn.edu, ehn@cs.cmu.edu

Abstract

We describe the LAPPS Grid and its Galaxy front-end, focusing on its ability to interoperate between a variety of NLP platforms. The LAPPS Grid project has been a leading force in the development of specifications for web service interoperability on syntactic and semantic levels. *Syntactic interoperability* among services is enabled through LIF, the LAPPS Interchange Format, which is expressed using the JSON-LD exchange format. JSON-LD is a widely accepted format that allows data represented in the international standard JSON format to interoperate at Web-scale. *Semantic interoperability* is achieved through the LAPPS Web Service Exchange Vocabulary, which has been developed by closely with interested and invested groups to develop a lightweight, web-accessible, and readily mappable hierarchy of concepts in a bottom-up, "as needed" basis.

Keywords: web services, NLP pipelines, interoperability

1. Overview

The NSF-SI^2-funded Language Applications (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. It has developed an open, web-based infrastructure through which massive and distributed resources can be accessed to support Natural Language Processing (NLP) research and teaching. In the LAPPS Grid, 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., 2014a).

The LAPPS Grid project is not developing new NLP analysis tools, but rather is building the infrastructure to make existing tools and resources easily discoverable, enable their rapid and easy configuration into pipelines and composite services, and most importantly, make them transparently interoperable. The Grid currently provides access to a large suite of commonly used NLP modules[2], together with 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)[3], negotiating licenses where necessary (Cieri et al., 2014). Means to add services and create and save composite workflows are fully in place, and we are adding to the LAPPS Grid Repository routinely while also providing means to enable easy addition of tools and modules to the LAPPS library.

The LAPPS Grid is based upon a deployment and extension of the service grid software[4] used to create the NICT/Kyoto Language Grid[5]. By opting to begin with the software supporting the Japanese grid, we have been able to deploy a new service grid hosted within the United States, without incurring the very significant cost of an entirely new software development effort, although differences in local reality and implementation made it necessary to augment the service grid software in a number of ways. The advantages of a grid supporting development of pipelines of web services include: ability to combine and experiment with individual services from multiple/alternative sources, rather than being confined to those provided in a particular platform such as NLTK or GATE; and reduction of demands on developers by removing the necessity to license the included libraries for distribution, create installation kits (for all relevant OSes/environments), document installation process, and provide technical support to those struggling to install. Perhaps most importantly, it allows for federation with other grids and service platforms in order to provide access to an increasingly large number of resources and tools.

2. Interoperability

Differing specifications of linguistic categories and typologies from application to application have posed a well-known obstacle to interoperability. One of the most important contributions of the LAPPS Grid project is its work in the area of *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 researchers from around the world for over a decade.[6] These efforts laid the groundwork in terms of standards development, raising community

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

[2] For example, Stanford NLP modules, OpenNLP tools, GATE's ANNIE tools, NLTK, BRAT annotation tool, etc., which can now be arbitrarily interchanged as needed by a given task or application. See http://www.lappsgrid.org/language-services for a full list of currently available tools.

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

[4] http://servicegrid.net

[5] http://langrid.org/en/index.html

[6] E.g., the NSF-funded Sustainable Interoperability for Language Technology (SILT) project (NSF-INTEROP 0753069) (Ide

awareness and buy-in, and proof-of-concept implementation upon which a comprehensive, international infrastructure supporting discovery and deployment of web services that deliver language resources and processing components can be built. We have worked with researchers, projects and standards-making bodies from around the world to develop specifications to enable NLP tools and services from diverse sources to seamlessly interoperate and promoted their adoption.

The LAPPS Grid project has been a leading force in the development of specifications for web service interoperability on syntactic and semantic levels. *Syntactic interoperability* among services is enabled through LIF, the LAPPS Interchange Format (Verhagen et al., 2015), which is expressed using the JSON-LD exchange format. JSON-LD[7] is a widely accepted format that allows data represented in the international standard JSON format[8] to interoperate at Web-scale. LIF uses the Linked Data aspect of JSON-LD to connect elements used in the JSON format to a vocabulary of semantic categories.

Semantic interoperability is a far greater challenge; we have addressed it by developing a lightweight, web-accessible, and readily mappable hierarchy of concepts in a bottom-up, "as needed" basis, called the LAPPS Grid Web Service Exchange Vocabulary (WSEV) (Ide et al., 2014b). The goal is not to define a new set of terms, but rather to to provide a basic, common terminology that can handle the basic types that are exchanged among LAPPS Grid services, regardless of the internal representations they use, with the intention that where possible, commonly used linguistic types (whatever their names, and whether they are objects or properties in the the original scheme) are mapped to terms in the WSEV. A second goal is to define relations among the terms that can be used when linguistic data are exchanged. The fundamental design principle of the WSEV is atomicity, to enable easy mapping between existing formats and the exchange vocabulary, together with ease of access and web-based addressing. Therefore, rather than a heavy interface, the vocabulary is accessible as a set of web pages[9], and reference is via a standard URI. Vocabulary items are defined and accompanied by a "sameAs" link to known web-based definitions that correspond to them[10].

WSEV development is guided by collaboration with interested and invested groups, including members of ISO TC 37 SC4 and projects such as the Technische Universität Darmstadt DKPro project[11], the Alveo Virtual Laboratory (Cassidy et al., 2014) project, WebLicht/Tübingen[12]and LINDAT/CLARIN (Prague)[13], as well as integration with existing web service ontologies such as the Language Grid's Language Service Ontology (Hayashi et al., 2011). Working closely with relevant groups and projects can help to ensure community input, buy-in, and, ultimately, widespread adoption.

It is important to note that the interoperability solutions implemented in the LAPPS Grid are not intended to provide an ultimate solution to the problem, but rather represent our best effort to carefully develop means to achieve, especially, semantic interoperability for NLP tools. They cannot readily address more fundamental sources of tool input/output incompatibility such as differences in tokenization and wildly different conceptual approaches to linguistic category definition; at present, the WSEV requires each service to publish input and output specifications in the form of a reference to rules (e.g., tokenization rules) and linguistic categories used by the tool in question, in order to provide means to check for compatibility. Obviously, more work in this area is greatly needed and must involve the entire community, if eventual success is to be achieved.

3. Federation with other grids and platforms

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" (Ishida et al., 2014), a multilingual, international network of web service grids and providers. Members currently include the Language Grid (NICT and Kyoto University, Japan)[14], grids operated by NECTEC (Thailand)[15] and the University of Indonesia[16], and the European Language Resources Association (ELRA) grid currently under development. We have recently entered into a formal partnership with WebLicht/Tübingen and LINDAT/CLARIN (Prague) to create a "trust network" among our sites in order to provide mutual access to all from any one of the three portals. We are also collaborating closely with the Australian Alveo Virtual Laboratory and the DKPro projects, with the intention to eventually federate with these platforms as well.

The federation of the LAPPS Grid with grids and platforms in Asia and Europe represents a landmark international collaboration that is unprecedented in the language processing field, which has the potential to lead to a paradigm shift in NLP development and research as well as work in the digital humanities, sciences, and social sciences. The key to the success of these partnerships is the *interoperability* among tools and services that is accomplished via the

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.

[7] http://json-ld.org

[8] http://www.ecma-international.org/publications/files/ECMA-ST/ECMA-404.pdf

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

[10] E.g., in existing repositories, type systems, and ontologies such as the CLARIN Data Concept Registry (https://openskos.meertens.knaw.nl/ccr/browser/), OLiA (http://nachhalt.sfb632.uni-potsdam.de/owl/), GOLD (http://linguistics-ontology.org), the NIF Core Ontology urlhttp://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core/nif-core), and general repositories such as Dublin Core (http://dublincore.org), schema.org, and the Friend of a Friend project (http://www.foaf-project.org).

[11] http://dkpro.github.io/info/

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

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

[14] http://langrid.org/en/index.html

[15] http://langrid.servicegrid-bangkok.org/en/overview.php

[16] http://langrid.portal.cs.ui.ac.id/langrid/

service-oriented architecture as well as collaborative development common vocabularies and multi-way mappings among tools and resources.

4. Galaxy workflow interface

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.[17] 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.

We provide Galaxy wrappers to call all LAPPS web services to the Galaxy ToolShed[18]. 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 several 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[19], 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.

Galaxy recently added support for running tools from the Galaxy ToolShed within Docker containers. Docker[20] 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.

Figure 1 shows a simple workflow configuration in LAPPS/Galaxy that invokes a chain of processors from different sources (in this example, GATE, Stanford NLP tools, and OpenNLP tools) to perform named entity recognition.

Our adaptation of the Galaxy workflow system also enables us to foster replicability and reuse for NLP by providing the following capabilities[21]: (1) 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; (20 provisions for sharing datasets, histories, and workflows via web links, with progressive levels of sharing including the ability to publish in a public repository; and (3) 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. Individual users 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 and repeatedly apply a command history on different data. Galaxy also 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; and Galaxy Pages and free-form annotations, which provide ways to add context to analysis to describe the reasoning behind an analysis and parameter settings.

5. Evaluation services

The Open Advancement (OA) Evaluation system implemented in the LAPPS Grid provides access to a sophisticated evaluation environment for NLP development. OA can be simultaneously applied to multiple variant workflows involving alternative tools for a given sub-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 and can lead to improvements that will affect the entire process. In addition, 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.

Until its incorporation into the LAPPS Grid, OA capabilities, which contributed significantly to the success of IBM’s Jeopardy-winning Watson, were not available for general use within the community. In addition, the federation of the multiple grids described above will make it possible to evaluate the performance of vast arrays of alternative

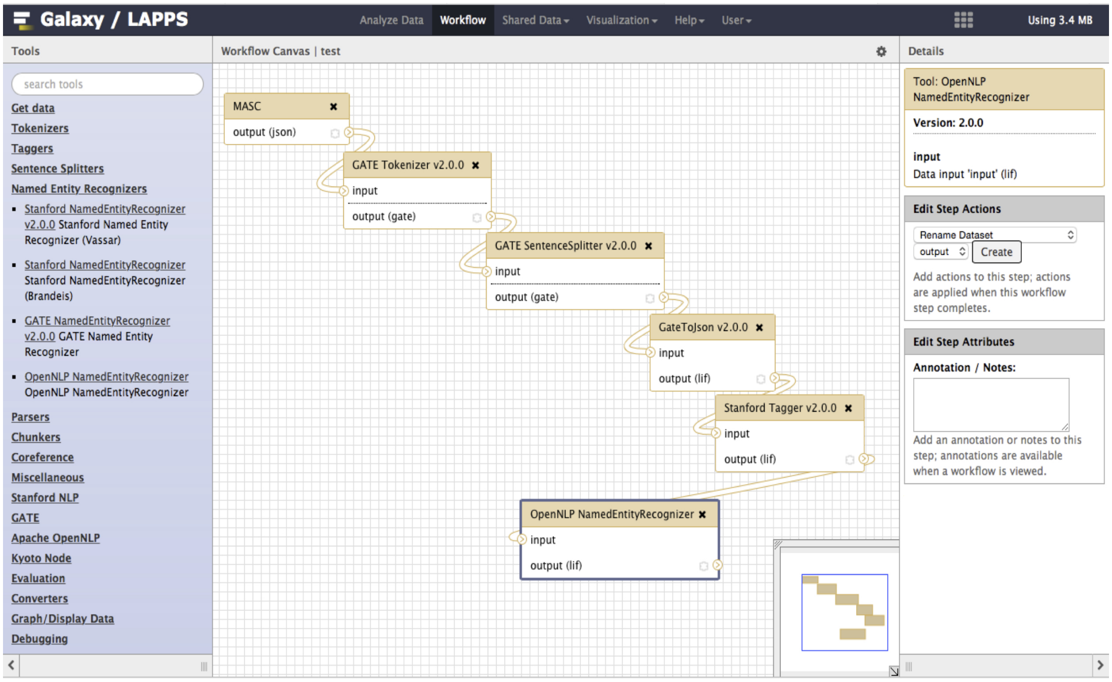
[17] http://galaxy.lappsgrid.org

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

[19] http://oauth.net

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

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



Add actions to this step; actions are applied when this workflow step completes.

Add an annotation or notes to this step; annotations are available when a workflow is viewed.

Figure 1: The LAPPS/Galaxy Interface: Workflow configuration

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 tool performance on data in a huge set of different languages. We are currently extending the evaluation capabilities in the LAPPS Grid to support parallel evaluation and broader adoption by end-user communities, including (1) the ability to assess the performance of an individual component in a pipeline; (2) parallel exploration of alternative pipelines; and (3) 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). The ability to combine processing modules from different sources becomes especially valuable when used in combination with the Open Advancement (OA) evaluation services in the LAPPS Grid, which provides performance statistics for each component in the pipeline as well as statistics reflecting the cumulative performance. This facility enables users to explore parallel workflows and evaluate module-by-module results in order to ultimately identify the optimal workflow configuration. Figure 2 shows a screenshot of the use of the OA evaluation service in a (simplified) workflow.

6. License navigation capabilities

The LAPPS Grid project is committed to open data and software; however, we would do the community a disservice if we did not allow access to licensed data and software, which in fact accounts for the vast majority of the language data available over the web. Within the LAPPS Grid, service providers and grid node hosts are not necessarily the owners of the software that drives their services, contrary to what seems to have been a core assumption of the Japanese grid. This has required us to build more sophisticated license management components, including “click through” licenses that can be accepted in real time (the LAPPS Grid retrieves any agreements from the service nodes and requires the user to agree to them before processing continues), as well as handling permissions that must be acquired in advance (Cieri and DiPersio, 2014). For this second type, the grid passes a request to the licensing entity, which then prompts for user credentials; if confirmed, the entity passes a timed token back to the grid allowing access to the resource.

7. Conclusion

The LAPPS Grid project’s efforts to make tools and data interoperable among platforms, tools, and services has enabled access to high-performance computing NLP facilities for members of the research and education communities who would otherwise have no such access, or who have little background in NLP, while reducing the often prohibitive overhead now required to adapt or develop new components. It is important to note that our goal is not to develop a monolithic grid nor a prescriptive set of standards that may never be widely adopted outside the LAPPS Grid, but rather to foster interoperability among existing grids, platforms, and frameworks so that the thousands of tools and resources available from sites everywhere in the world can be transparently shared, reused, and combined to create sophisticated NLP applications. We recognize that this cannot be accomplished within one or even a few projects, but rather must rely on the input and collaboration among projects around the globe to work toward means to achieve this interoperability at both the syntactic and semantic levels. Technology has evolved to the point where syntactic interoperability is less problematic, but for semantic interoperability, continued effort is required. We therefore solicit

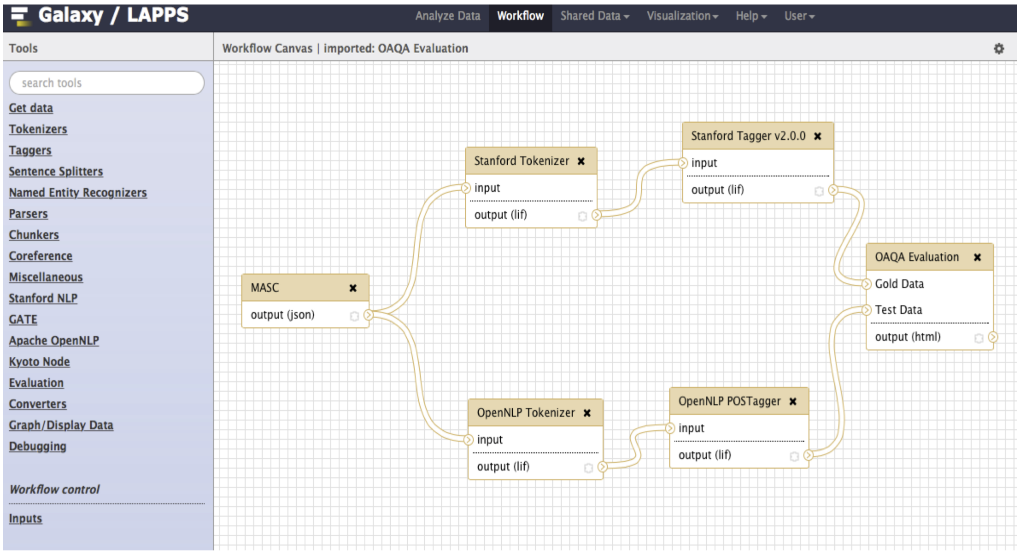


Figure 2: The LAPPS/Galaxy Interface: OA Evaluation on two pipelines

the cooperation of all, in order to achieve what we assume is a common end.

8. Acknowledgments

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

9. Bibliographical References

Nicoletta Calzolari, et al., editors. (2009). *Proceedings of "The European Language Resources and Technologies Forum: Shaping the Future of the Multilingual Digital Europe"*. ILC-CNR.

Cassidy, S., Estival, D., Jones, T., Burnham, D., and Burghold, J. (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).

Cieri, C. and DiPersio, D. (2014). Intellectual Property Rights Management with Web Service Grids. In *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT (OIAF4HLT)*, Dublin, Ireland.

Cieri, C., Dipersio, D., Liberman, M., Mazzucchi, A., Strassel, S., and Wright, J. (2014). New directions for language resource development and distribution. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

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

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

Hayashi, Y., Declerck, T., Calzolari, N., Monachini, M., Soria, C., and Buitelaar, P. (2011). Language service ontology. In *The Language Grid - Service-Oriented Collective Intelligence for Language Resource Interoperability*, pages 85–100. Springer.

Ide, N., Pustejovsky, J., Calzolari, N., and Soria, C. (2009). The SILT and FlaReNet international collaboration for interoperability. In *Proceedings of the Third Linguistic Annotation Workshop, ACL-IJCNLP*, August.

Ide, N., Pustejovsky, J., Cieri, C., Nyberg, E., Wang, D., Suderman, K., Verhagen, M., and Wright, J. (2014a). The Language Application Grid. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Ide, N., Pustejovsky, J., Suderman, K., and Verhagen, M. (2014b). The Language Application Grid Web Service Exchange Vocabulary. In *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT (OIAF4HLT)*, Dublin, Ireland.

Ishida, T., Murakami, Y., Lin, D., Nakaguchi, T., and Otani, M. (2014). Open Language Grid–Towards a Global Language Service Infrastructure. In *The Third ASE International Conference on Social Informatics (SocialInformatics 2014)*, Cambridge, Massachusetts, USA.

Verhagen, M., Suderman, K., Wang, D., Ide, N., Shi, C., Wright, J., and Pustejovsky, J. (2015). The LAPPS Interchange Format. In *Proceedings of the Second International Workshop on Worldwide Language Service Infrastructure (WLSI'15)*, Kyoto, Japan, January.