The Language Application Grid

Nancy Ide[1(✉)], James Pustejovsky[2], Christopher Cieri[3], Eric Nyberg[4], Denise DiPersio[3], Chunqi Shi[2], Keith Suderman[1], Marc Verhagen[2], Di Wang[4], and Jonathan Wright[3]

[1] Vassar College, Poughkeepsie, NY, USA
{ide,suderman}@cs.vassar.edu
[2] Brandeis University, Waltham, MA, USA
{jamesp,shicq,marc}@cs.brandeis.edu
[3] Linguistic Data Consortium, Philadelphia, PA, USA
{ccieri,dipersio,jdwright}@ldc.upenn.edu
[4] Carnegie-Mellon University, Pittsburgh, PA, USA
{ehn,diwang}@cs.cmu.edu

Abstract. The Language Application (LAPPS) Grid project is establishing a framework that enables language service discovery, composition, and reuse and promotes sustainability, manageability, usability, and interoperability of natural language Processing (NLP) components. It is based on the *service-oriented architecture* (SOA), a more recent, web-oriented version of the "pipeline" architecture that has long been used in NLP for sequencing loosely-coupled linguistic analyses. The LAPPS Grid provides access to basic NLP processing tools and resources and enables pipelining such tools to create custom NLP applications, as well as composite services such as question answering and machine translation together with language resources such as mono- and multi-lingual corpora and lexicons that support NLP. The transformative aspect of the LAPPS Grid is that it orchestrates access to and deployment of language resources and processing functions available from servers around the globe and enables users to add their own language resources, services, and even service grids to satisfy their particular needs.

Keywords: NLP frameworks · Web services · Service grids · Open advancement · Resource licensing

1 Introduction

The need for robust language processing capabilities across academic disciplines, education, and industry is without question of vital importance to national security, infrastructure development, and the competitiveness of American business. However, while the past two decades have produced reliable and accurate tools for the various linguistic analyses required by Natural Language Processing (NLP) applications, component interoperability–and hence, reusability–has remained a serious problem for the field. A few application frameworks have

Y. Murakami and D. Lin (Eds.): WLSI 2015, LNAI 9442, pp. 51–70, 2016.
DOI: 10.1007/978-3-319-31468-6_4

been recently developed for the integration and delivery of end-to-end language software (e.g., UIMA, GATE), but these frameworks provide for interoperability among tools and components only within the frameworks themselves. Additionally, while such frameworks provide for *syntactic interoperability* via internally-defined physical formats, *semantic interoperability* [11], even within a given framework, is still problematic because users must define their own type systems and ontologies, which vary widely. As a result, the field has remained relatively 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 NLP 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. This capability, coupled with access to a repository of interoperable NLP processing components and test data, will enable a major leap in productivity for researchers and developers alike.

To meet this need, the Language Application (LAPPS) Grid project is establishing a framework that enables language service discovery, composition, and reuse and promotes sustainability, manageability, usability, and interoperability of natural language Processing (NLP) components. It is based on the *service-oriented architecture* (SOA), a more recent, web-oriented version of the "pipeline" architecture that has long been used in NLP for sequencing loosely-coupled linguistic analyses. The LAPPS Grid provides a critical missing layer of functionality for NLP: although existing frameworks such as UIMA and GATE provide the capability to wrap, integrate, and deploy language services, they do not provide general support for service discovery, composition, and reuse.

The LAPPS Grid is a collaborative effort among US partners Brandeis University, Vassar College, Carnegie-Mellon University, and the Linguistic Data Consortium at the University of Pennsylvania, and is funded by the US National Science Foundation. The project 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 "Open Language Grid" federation [14], a multi-lingual, international network of web service grids and providers that integrates large-scale computing, high-speed networks, and massive data archives across the world to support the development and testing of integrated natural language applications. The key to the success of this federation 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 have involved key researchers from around the world for over a decade, including members of the LAPPS Grid project[1].

[1] E.g., in the NSF-funded Sustainable Interoperability for Language Technology (SILT) project (NSF-INTEROP 0753069) [12], the EU-funded Fostering Language Resources Network (FLaReNet) project [1], 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.

These efforts laid the groundwork in terms of standards development, raising community awareness and buy-in, and proof-of-concept implementation upon which the creation of a comprehensive, international infrastructure supporting discovery and deployment of web services for language resources and processing components is now being built.

The development and deployment of the LAPPS Grid and its integration in the Open Language Grid has already demonstrated its potential to significantly transform the way language data is accessed, analyzed, and exploited across disciplines for diverse research and development needs, and to ultimately enable a major leap in language processing capabilities that can impact the way people use and interact with computers. The LAPPS Grid offers the following benefits:

- 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;
- substantially increased access to resources for members of the NLP community as well as researchers in sociology, psychology, economics, education, linguistics, digital media, etc., including mono- and multi-lingual lexical, semantic, and ontological resources that provide information relevant to a wide range of sub-domains (e.g., speech, machine translation, information retrieval);
- means to address the current lack of interoperability among NLP components and data by negotiating across formats and categories;
- access to a state-of-the-art, sophisticated evaluation environment that facilitates assessment of component contribution to overall performance and iterative application development;
- capabilities for rapid development of resources for less-resourced and endangered languages, for which automatic language processing capabilities are only beginning or have yet to be developed;
- enhanced capability for state-of-the-art, "on-the-fly" stream processing of language by enabling NLP applications to call services and extract information from service resources;
- enhancement of research, development, and teaching of NLP by providing controlled access to resources that are otherwise too costly to acquire or restricted by intellectual property rights, as well as access to large-scale computing required to process massive language resources.

It is important to note that the transformative aspect of the LAPPS Grid is not the provision of a suite of web services and composite workflows, but rather that it orchestrates access to and deployment of language resources and processing functions available from servers around the globe and enables users to add their own language resources, services, and even service grids to satisfy their particular needs. As such, the LAPPS Grid is ultimately a community-based project, to which services will be contributed by members of the community and existing service repositories and grids can be federated to enable universal access.

In this paper we provide an overview of the LAPPS Grid and the technologies we are developing to support its use. Section 2 describes the overall architecture of the LAPPS Grid. In Sect. 3, the development of the LAPPS Web Service Exchange Vocabulary, which enables interoperability among services in the Grid, is described. Section 4 introduces the LAPPS/Galaxy interface for accessing and constructing atomic and composite web services, and in Sect. 5 we overview the open advancement evaluation capabilities that are being provided in the Grid. Section 6 discusses our approach to handling potentially divergent licensing constraints in web service pipelines. Finally, Sects. 7 and 8 discuss user-provided evaluation of the LAPPS Grid and the relation of this project to similar projects in Asia, Australia, and the European Union.

2 LAPPS Grid Design

The fundamental system architecture of the LAPPS Grid is based on the Open Service Grid Initiative's Service Grid Server Software developed by the National Institute of Information and Communications Technology (NICT) in Japan and used to implement Kyoto University's Language Grid, a service grid that supports multilingual communication and collaboration. Like the Language Grid, the LAPPS Grid provides three main functions: language service registration and deployment, language service search, and language service composition and execution. From the perspective of application developers, one of the intended audiences for the LAPPS Grid, several aspects of service deployment are important:

1. *Service Discovery.* An application designer can query for existing components and services that provide some desired functionality, and quickly identify elements in the repository that are suited to the task.
2. *Service Adaptation.* The LAPPS Grid supports straightforward customization and adaptation of each component or service (e.g., by exposing parameters, options, etc.).
3. *Service Composition.* New applications can be built from existing elements and tested on client data with a minimum amount of programming.
4. *Metrics and Measurement.* The LAPPS Grid is instrumented to provide relevant component-level measures for standard metrics, given gold-standard test data. New applications automatically include instrumentation for component-level and end-to-end measurement; intermediate (component-level) I/O is logged to support effective error analysis.

By opting to begin with the software supporting the Japanese grid, we have been able to deploy a new service grid hosted entirely 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 LAPPS Grid extends the core functionality of the Service Grid Software by (1) further enabling composition of tool and resource chains as well as providing sophisticated evaluation services; (2) implementing a *dynamic licensing* system (see

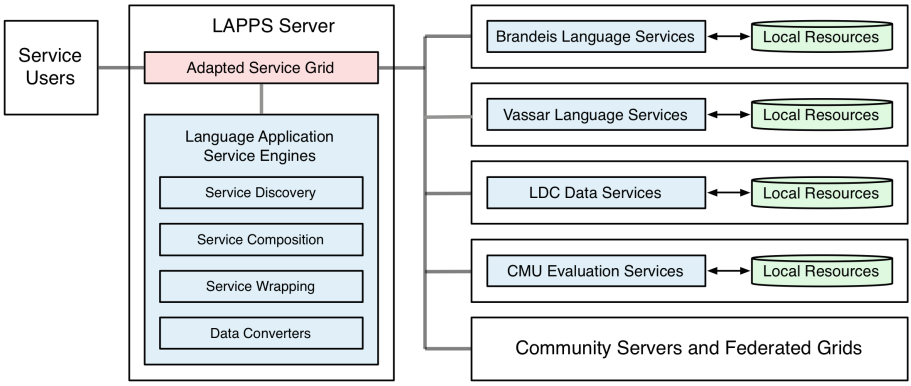

Fig. 1. LAPPS grid architecture

Sect. 6 for handling license agreements on the fly; (3) providing the option to run services locally, with high-security technology to protect sensitive information where required, improve data delivery services; and (4) enabling access to grids other than those based on the Service Grid technology. Also, because the LAPPS Grid is a community-based resource to which members of the community will increasingly contribute as well as use, we provide user-friendly, transparent facilities for wrapping user-provided services.

The basic components of the LAPPS Grid are presented in Fig. 1. The main LAPPS server maintains a workflow repository for composite linguistic services and is equipped with a workflow engine to enable users to develop their own composite (pipelined) services. It also contains modules for discovery, wrapping and conversion. LAPPS Grid nodes housed at Brandeis University and Vassar College maintain repositories of known atomic linguistic services and provide service discovery functionality to users and applications. The LDC node houses various data services, and the node at CMU provides services for automatic measurement and analysis of workflow components, including error analysis at the component and end-to-end application level.

3 Interoperability

Differing specifications of linguistic categories and typologies from application to application have posed a well-known obstacle to interoperability. We have worked with researchers, projects and standards-making bodies from around the world to develop common vocabularies and multi-way mappings, using as a basis the output of various international efforts undertaken over the previous decade[2]. Our developments address both *syntactic interoperability* among web services by providing an implementation of a well-established physical format for web service exchange, as well as *semantic interoperability* to enable services to mutually understand the "meaning" of exchanged objects.

[2] E.g., SILT [12], FLaReNet [1], ISO TC37 SC4, etc.

3.1 LAPPS Interchange Format

Syntactic interoperability among services is enabled via JSON-based serialization for Linked Data (JSON-LD)[3], a widely accepted format that allows data represented in the international standard JSON format[4] to interoperate at Web-scale. The JavaScript Object Notation (JSON)[5] is a lightweight, text-based, language-independent data interchange format that defines a small set of formatting rules for the portable representation of structured data. Because it is based on the W3C Resource Definition Framework (RDF), JSON-LD is trivially mappable to and from other graph-based formats such as ISO LAF/GrAF [13,15] and UIMA CAS[6], as well as a growing number of formats implementing the same data model. JSON-LD enables services to reference categories and definitions in web-based repositories and ontologies (e.g., ISOCat[7], GOLD[8], Dublin Core[9], OLiA[10]) or any suitably defined concept at a given URI.

We have designed the LAPPS Interchange Format (LIF) to represent linguistically annotated data in JSON-LD. Services that implement a linguistic application (or wrap an existing application) must consume LIF objects and are responsible for creating LIF objects. Each web service in the LAPPS Grid publishes metadata describing what it requires for input and what it produces as output. A process that is constructing a service pipeline can then query each service to determine compatibility. Where necessary, data converters included in the Language Application Service Engines (see Fig. 1) are automatically invoked map from commonly used formats to the JSON-LD interchange format. For a fuller description of LIF, see Verhagen *et al.*, "The LAPPS Interchange Format", in this volume.

3.2 Exchange Vocabulary

Semantic interoperability among web services is a far greater challenge. Although the pipeline architecture has been implemented in several NLP frameworks over the past two decades, including self-contained (non-service) frameworks such as GATE and UIMA, no accepted standard for module description or input/output interchange to support service discovery, composition, and reuse in the language application domain exists. To address this, we have worked closely with interested and invested groups and members of ISO TC 37 SC4 to develop a lightweight, web-accessible, and readily mappable hierarchy of concepts called the Web Service Exchange Vocabulary (WS-EV) that specifies a terminology for a

[3] http://json-ld.org.
[4] http://www.ecma-international.org/publications/files/ECMA-ST/ECMA-404.pdf.
[5] http://www.json.org and http://www.ietf.org/rfc/rfc4627.txt.
[6] The *Common Analysis Structure* (CAS) is the internal format for exchange among modules in the UIMA framework.
[7] http://www.isocat.org.
[8] http://linguistics-ontology.org.
[9] http://dublincore.org.
[10] http://nachhalt.sfb632.uni-potsdam.de/owl/.

core of linguistic objects and features exchanged among NLP tools that consume and produce linguistically annotated data. Development is further guided by collaboration with projects such as the CLARIN Data Concept Registry[11] and ISOcat[12], and integration with existing web service ontologies such as the Language Grid's Language Service Ontology [10]. The WS-EV addresses a need within the community to not only identify a readily usable set of terms, but also specify the relations among them. However, it is crucial to note that the goal of the WS-EV is not to provide a definitive set of terms and relations that will serve every purpose and satisfy every user, but rather to provide a base set of terms, trivially mappable from a substantial number of widely-used schemes, that can be used for exchanging linguistic data among web services. A fuller description of the WS-EV and the philosophy behind it are provided elsewhere in this volume.[13]

Our approach to development of the WS-EV is "bottom-up", in order to avoid *a priori* development of a comprehensive linguistic type system. To that end, we have adopted a "minimalist" strategy of providing a simple core set of objects and features. Where possible, the core is drawn from existing repositories such as ISOCat; however, because many categories and objects relevant for web service exchange are not included in such repositories, we have attempted to identify a set of (more or less) "universal" concepts by surveying existing type systems and schemas–for example, the Julie Lab and DARPA GALE UIMA type systems and the GATE schemas for linguistic phenomena–together with the I/O requirements of commonly used NLP software (e.g., the Stanford NLP tools, OpenNLP, etc.).[14]

We have established an Exchange Vocabulary Repository[15] similar to schema.org, in order to provide web-addressable terms and definitions for reference from annotations exchanged among web services for NLP tools and processes. Wherever possible, terms in the vocabulary are mapped to categories defined in other repositories, ontologies, registries, etc. (including mapping to multiple repositories when appropriate). For this purpose we utilize the taxonomy of relation types defined in RELcat [21], which accommodates multiple vocabularies for relation predicates including those from the Web Ontology Language (OWL) [19] and the Simple Knowledge Organization System (SKOS) [20].

Terms in the repository are organized in a shallow hierarchy, with inheritance of properties, as shown in Fig. 2. WS-EV development is undertaken in collaboration with a Working Group within ISO TC37 SC4, to guarantee substantial

[11] https://openskos.meertens.knaw.nl/ccr/browser/.

[12] http://www.isocat.org.

[13] See Ide *et al.*, "The Language Application Grid Web Service Exchange Vocabulary", in this volume.

[14] The survey of basic linguistic objects was undertaken within a Working Group of ISO TC37 SC4. A working draft and an inventory of type systems are available at http://vocab.lappsgrid.org/EV/ev-draft.pdf and http://vocab.lappsgrid.org/EV/materials/.

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

community involvement and so that our results may ultimately become a part of the larger set of ISO standards for language resource management.

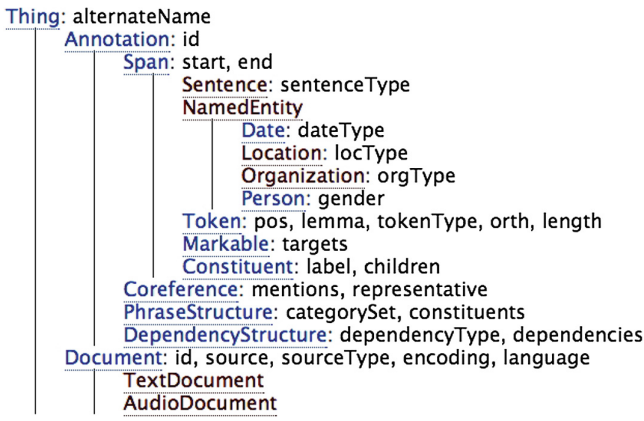

Fig. 2. Fragment of the WS-EV type hierarchy (associated properties in gray)

References in the LAPPS JSON-LD representation exchanged between web services point not only to definitions for specific linguistic categories, but also to documentation for processing software and "rules" for processes such as tokenization, entity recognition, etc. used to produce a set of annotations, which are often left unspecified in annotated resources, thus inhibiting replication of results (see for example [5]). While not required for web service exchange in the LAPPS Grid, the inclusion of such references can contribute to the better replication and evaluation of results in the field.

Figure 3 shows the information for *Token*, which defines the concept, identifies application types that produce objects of this type, cross-references a similar concept in ISOCat, and provides the URI for use in the JSON-LD representation. It also specifies the common properties that can be specified for a set of Token objects, and the individual properties that can be associated with a Token object.

The LAPPS WS-EV is intended to support URI-based references to basic concepts used in the description and processing of linguistically annotated corpora from JSON-LD and other linked data representations such as W3C RDF, or any linguistically annotated resource. There is no requirement to use any or all of the specified properties, and we foresee that many web services will require definition of objects and properties not included in the WS-EV or elsewhere. We therefore provide mechanisms for (principled) definition of objects and features beyond the WS-EV. Two options exist: users can provide a URI where a new term or other documentation is defined, or users may add a definition to the WS-EV. In the latter case, service providers use the name space automatically assigned to them at the time of registration, thereby avoiding name clashes and providing a distinction between general categories used across services and more idiosyncratic categories.

Thing > Annotation > Span > Token

Definition	A string of one or more characters that serves as an indivisible unit for the purposes of morpho-syntactic labeling (part of speech tagging).
Similar to	http://www.isocat.org/datcat/DC-1403
URI	http//vocab.lappsgrid.org/Token

Metadata

Metadata from Annotation

Properties	Type	Description
producer	List of URI	The software that produced the annotations.
rules	List of URI	The documentation (if any) for the rules that were used to identify the annotations.

Properties

Properties	Type	Description
pos	String or URI	Part-of-speech tag associated with the token.
lemma	String or URI	The root (base) form associated with the token. URI may point to a lexicon entry.
tokenType	String or URI	Sub-type such as word, punctuation, abbreviation, number, symbol, etc. Ideally a URI referencing a pre-defined descriptor.
orth	String or URI	Orthographic properties of the token such as LowerCase, UpperCase, UpperInitial, etc. Ideally a URI referencing a pre-defined descriptor.
length	Integer	The length of the token

Properties from Span

Properties	Type	Description
start	Integer	The starting offset (0-based) in the primary data.
end	Integer	The ending offset (0-based) in the primary data.

Properties from Annotation

Properties	Type	Description
id	String	A unique identifier associated with the annotation.

Properties from Thing

Properties	Type	Description
alternateName	String	An alias for the item.

Fig. 3. Token definition in the LAPPS WS-EV

4 LAPPS/Galaxy Workflow Engine

The Galaxy project[16] started in 2005 to create a system enabling biologists without informatics expertise to perform computational analysis through the web [7]. Galaxy is an open-source application[17] 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 [8]. 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.

Rather than duplicate the extensive work of the Galaxy project, we recently adopted it as the primary workflow management system for the LAPPS Grid.[18] We have worked with the Galaxy development team in order to adapt the system

[16] http://galaxyproject.org.

[17] Distributed under the terms of permissive Academic Free License; http://getgalaxy.org.

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

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 provide Galaxy wrappers to call all LAPPS web services to the Galaxy ToolShed[19]. 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. 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[20] and passing 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. 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-data resources and analysis tools. 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.

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, and (2) incorporation of authentication procedures for protected data using the open standard OAuth[21], which specifies a process for resource owners to authorize third-party access to their server resources without sharing their credentials. We also have contributed a "Galaxy Flavor" for LAPPS, 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 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. In addition, if the images are downloaded ahead of time, no network connection is required.

Figures 4 and 5 show a simple workflow configuration and a visualization of named entity annotation over a document.

[19] https://toolshed.g2.bx.psu.edu.
[20] http://www.ncbi.nlm.nih.gov/pubmed.
[21] http://oauth.net.

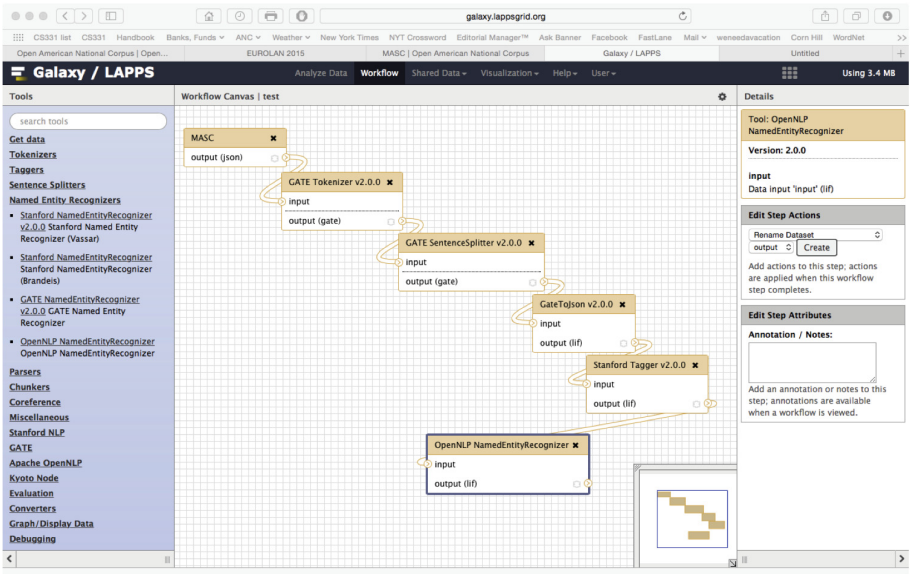

Fig. 4. The LAPPS/Galaxy interface: workflow configuration

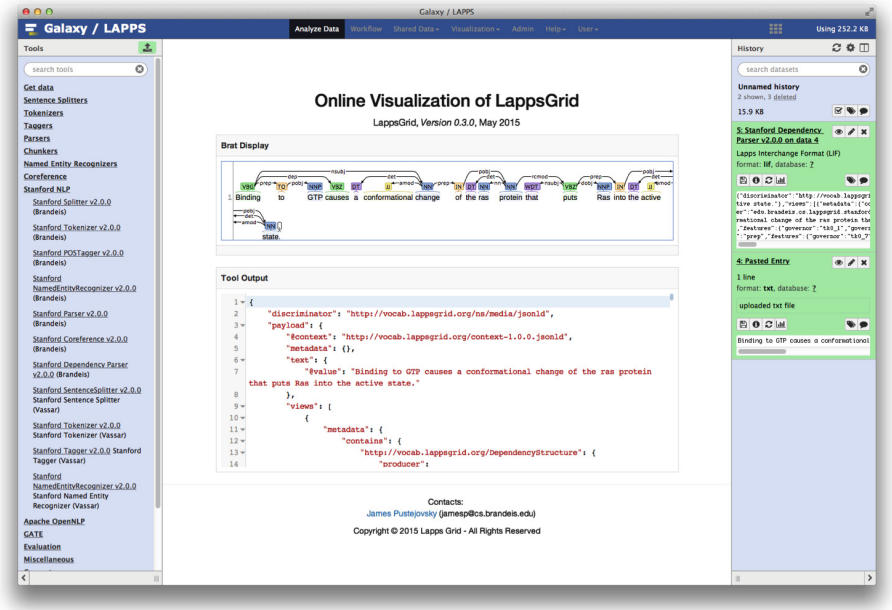

Fig. 5. Visualization of a named entity annotation using LAPPS/Galaxy

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 [5,17]), blogs[22], and workshops[23]. 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 [9] 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. 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 will adopt 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 sharing of tools, data, and (especially) workflows and results.

[22] E.g., http://nlpers.blogspot.com/2006/11/reproducible-results.html.
[23] E.g., Replicability and Reusability in Natural Language Processing: from Data to Software Sharing: http://nl.ijs.si/rrnlp2015/.

5 Open Advancement

CMU has provided the tooling and infrastructure for two major services, based in part on the existing OAQA framework developed at CMU and deployed on a service node housed at CMU. The availability of this type of evaluation service, which implements state-of-the-art Open Advancement techniques, provides an unprecedented tool for NLP development that could, in itself, take the field to a new level of productivity. The open advancement (OA) approach for component- and application-based evaluation has been successful in enabling rapid identification of frequent error categories within modules and documents, together with an indication of which module(s) and error type(s) have the greatest impact on overall performance, thus contributing to more effective investment of resources in both research and application assembly [3,22]. The OA approach was used in the development of IBM's Watson to achieve steady performance gains over the four years of its development [4]. More recently, the open-source OAQA project has released software frameworks which provide general support for open advancement of information systems [6,22]; the OAQA software has been used to rapidly develop information retrieval and question answering systems for bioinformatics [16,22].

A fundamental element of open advancement involves evaluating multiple possible solutions to a given problem, to find the optimal solution available using given components, resources and evaluation data. The output of the optimal solution is then subjected to error analysis, to identify the most frequent errors with the highest impact on system output quality. Possible enhancements to the system are then considered, with an eye toward achieving the largest possible reduction in error rate by addressing the most frequent error types. The performance of each new configuration is evaluated to determine whether a significant improvement has been achieved in comparison with prior baselines or best known configurations. When multiple teams collaborate to implement this process across several sites, types of components, etc. it is possible to make rapid progress in improving solution quality, as measured by the chosen metrics and evaluation dataset [3,22]. To support rapid, open advancement, a developer can add new components to the system and test them in the context of existing pipelines by "plugging them in" to existing solutions. We also provide capabilities for parallel exploration of alternative workflows, evaluation of module-by-module results, and "best path" analysis to determine the optimal workflow.

The LAPPS/Galaxy workflow engine described in the previous section provides easy configuration and re-configuration of pipelines, and represents the first step in supporting open advancement by allowing users to rapidly configure and evaluate a new, single pipeline on a chosen dataset and metrics. In addition, the user can specify an entire range of pipeline configurations for comparative evaluation; the system evaluates each possible pipeline configuration and generate metrics measurements, plus variance and statistical significance calculations. We are working to extend the LAPPS/Galaxy interface to allow easy specification of configuration descriptors (ECD); [22] that define a space of possible pipelines, where each step in the pipeline might be achieved by multiple components or

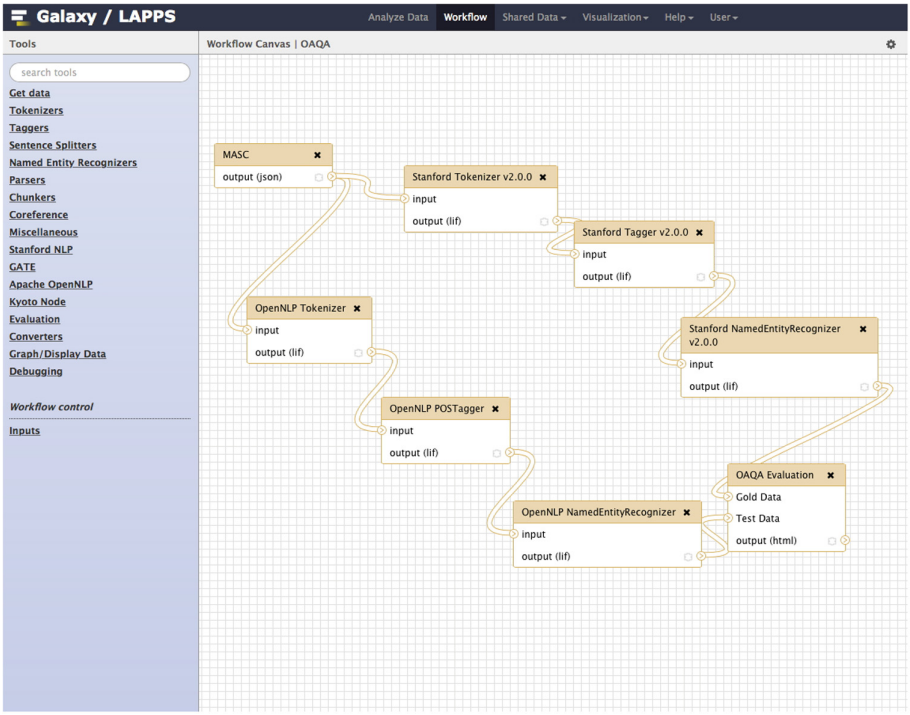

Fig. 6. The LAPPS/Galaxy interface: evaluation configuration for two workflows

services; each component or service may also have configuration parameters with more than one possible value to be tested. We are also extending the system to support automatic evaluation of each configuration so specified, by implementing a service-oriented version of the Configuration Space Exploration (CSE) algorithm [22].

Figure 6 shows a simple evaluation configuration in LAPPS/Galaxy, which compares evaluative statistics for two parallel pipelines performing named entity identification.

6 Resource Access

LDC's contributions to the multi-site LAPPS Grid focus naturally on data. LDC is creating services that provide grid access to the contents of its LDC Online service: multilingual newswire and transcribed conversational telephone speech in English, as well as to lexical databases. The challenges of this work lie in developing useful and efficient service interfaces to these data. In each case, we envision the interface as containing a number of simple operations: requests to retrieve the features of the supplied data, queries into the data using those features that return identifiers and requests to fetch data elements by identifier,

via iteration or randomly. LDC already deploys data services, both internal and external, so our Grid work emphasizes enclosing those services in a thin wrapper within a Grid node that we host. Using the data source API developed by the LAPPS project, we pass on Grid requests to LDC services. Some LDC services, including the Grid node, run on virtual machines, allowing us to easily adjust system resources to match changing demand. LDC's infrastructure also includes a Solr[24] server for searching text, including some of the content available to the Grid.

Along with the flexibility the LAPPS Grid offers to users seeking to create service pipelines comes an increase in the complexity of intellectual property arrangements. We anticipate two major pipeline types. In the first, users request language resources from a given source (or supply their own) and route them through a workflow of multiple grid services with the final result returned to the user. In the second type, language resources are routed through a single service and then back to the user before being routed along to the next service. The difference between these user case types has implications for licensing and constraints imposed on grid users, services and operators. Moreover, within those cases, one must consider constraints imposed by the language resources, data and software enabling the web services.

At each point in either pipeline above, constraints depend upon the language resources or resulting services, processing and user. Resources may be constrained or unconstrained. Constraints may be imposed by legal principles such as copyright or by contract. Constraints may prohibit commercial use, derivative works or re-distribution or insist upon attribution or in-kind sharing of the user's intellectual products. Resources may be constrained as to user, typically forbidding use by commercial organizations, or as to use, whether for education, basic research, applied research, technology development, evaluation and deployment or resale. Processing may also be constrained, for example, ruling out derivative works and only permitting so-called transformative works. Users may be licensed or not. Their licensing may be defined by enumeration or by user features, for example whether they work in an academic, non-academic, not-for-profit, government, pre-commercial or commercial environments.

We manage this complexity by identifying the licenses associated with each Grid service and analyzing them into their component constraints. Those constraints are accumulated as the service pipeline is constructed, and users are notified about them before the pipeline is executed. Constraints are of two types, requirement and notification. Required constraints block the pipeline until the constraint is removed. Examples include cases where users must pay a fee or sign a specific agreement in order to access the desired resource or service. Other constraints, such as redistribution, commercial/non-commercial use, use of derivatives and so on are presented as conditions which users must acknowledge before the pipeline will be executed. Figure 7 summarizes that process.

Variation in license terms notwithstanding, the human language technology community has for some time envisioned open source-based models for language

[24] https://lucene.apache.org/solr/.

Constraint	Action
Redistribution	Notify
Use	Notify
Derivatives Use	Notify
Attribution	Notify
Share Alike	Notify
Fee	Require
Other Specific License	Require
Other Specific Constraint	?

Fig. 7. LAPPS grid license constraint enforcement

resource development and distribution. Most recently, META-SHARE proposes a network of distributed repositories that license resources from a single platform via open source agreements (META-SHARE Commons licenses) as well as more restrictive arrangements [18]. Although all levels of licensing complexity are acknowledged in the LAPPS Grid, the LAPPS license scheme depends on the utilization of open source software and resource licenses to the greatest extent possible. By limiting distribution and processing constraints, we aim to promote the project goal of community engagement through sharing, federation and other means. By developing a comprehensive model for addressing constraints on the intellectual property used in the Grid we hope to create a resource that is maximally open to users ranging from open source developers to commercial users of languages services.

7 User Evaluation

To a large extent, the measure of success for LAPPS is a matter of the ease with which the user community–both NLP researchers and developers and those with little knowledge of the field–can use the infrastructure to serve their needs. The project therefore includes an on-going user-evaluation component involving a range of user types, including those whose computational expertise may be limited, who provide periodic feedback concerning Grid access, adding applications to the Grid, using external applications or services in combination with the Grid, etc. In the spirit of open advancement, we measure the total time and effort required to determine the optimal configuration of existing components for a given problem and use these measures to improve the system's design.

To support community use, we regularly offer tutorials and training workshops on LAPPS Grid use at major conferences in the field[25], including venues associated with other disciplines, with the goal of introducing scientists and

[25] E.g., *Web Services for Effective NLP Application Development and Evaluation: Using and Contributing to the Language Application (LAPPS) Grid*, offered at LREC 2014.

engineers from diverse disciplines to a broad-based and integrated set of NLP services that has the potential to impact their research and development needs. We envision that research from sociology, psychology, economics, education, linguistics, digital media, as well as engineering, can be impacted by the ability to manipulate and process diverse data sources in multiple languages.

Another major effort aimed toward both development of the LAPPS Grid and user evaluation is inclusion of LAPPS use in courses offered at Carnegie-Mellon University and Brandeis University. At Carnegie-Mellon, two courses will use the LAPPS framework: a master's level seminar course including a project on "automatically building customized search engines with LAPPS", and a Question Answering course including development of a world history question-answering pipeline. At Brandeis, the LAPPS Grid will be deployed as the development, testing, and evaluation platform for several projects in a course on Fundamentals in Computational Linguistics course. We are also pursuing the development of courses relying on the LAPPS Grid for use in US Government agencies. Feedback from these courses on all aspects of the LAPPS Grid–configuration, availability of relevant services, usability of interfaces, etc.–will provide valuable input to iterative development of the LAPPS Grid.

8 Relation to Other Projects

The LAPPS Grid effort builds on the foundation laid in several recent U.S., European, and Asian projects, including the NSF-funded Sustainable Interoperability for Language Technology (SILT) project [12] and the EU-funded Fostering Language Resources Network (FLaReNet) project [1]. At the same time, the International Standards Organization (ISO) committee for Language Resource Management (ISO TC37 SC4)[26] has addressed the need for standards for linguistic data. Through these and other projects and parallel efforts in Asia and Australia, substantial groundwork—in terms of standards development, raising community awareness and buy-in, and proof-of-concept implementation—has been laid to turn existing, fragmented NLP technologies and data into web-accessible, stable, and interoperable resources that can be readily reused across several fields. As a result, existing and potential projects across the globe are beginning to converge on common data models, best practices, and standards, and the vision of a comprehensive infrastructure supporting discovery and deployment of web services that deliver language resources and processing components is an increasingly achievable goal.

Our vision is therefore not for a monolithic grid, but rather a heterogeneous configuration of federated grids that implement a set of best practices for managing and interchanging linguistic information, so that services on all of these grids are mutually accessible. To that end, the LAPPS Grid project has entered into a multi-way international collaboration among the US partners and institutions in Asia and Europe. The basis of the collaboration is the federation of the LAPPS Grid, the Language Grid (Kyoto University, Japan), NECTEC

[26] ISO/TC 37/SC4, Language Resources Management, http://www.tc37sc4.org.

(Thailand)[27], grids operated by the University of Indonesia[28] and Xinjiang University (China)[29], and LinguaGrid[30], to be formally as the "Open Language Grid" announced in January 2016.[31] The connection of these six grids into a single federated entity will enable access to all services and resources on any of these grids by users of any one of them and, perhaps most importantly, facilitate adding additional grids and service platforms to the federation in the future. Currently, the European MetaNet/Meta-Share[32] initiative is committed to joining the federation in the near future, which will provide access to the substantial resource holding of the European Language Resources Association (ELRA) as well as web services developed in the EU project PANACEA. We are also working with the EU CLARIN initiative[33], a large-scale pan-European collaborative effort aimed at making language resources and technology readily available for the whole European Humanities (and Social Sciences) communities, as well as the LINDAT-CLARIN Centre for Language Research Infrastructure's open digital repository of tools and data (Charles University, Prague), and the Australian Alveo Virtual Laboratory [2] to similarly share access to services and resources in the near future

One goal of our work is to ensure that all relevant parties can provide input to the development and/or refinement of standards and practices that promote increased interoperability among web service platforms. Therefore, we continue to reach out to other projects to join the collaboration and, where appropriate, grid federation, including EU projects such as KYOTO[34] as well as large projects developing NLP components and data such as the Global WordNet Grid[35] and U-Compare[36], which provides an interface to UIMA-based components primarily for the Biomedical domain. We are also pursuing potentially fruitful uni-directional federations, in which other grids and service nodes are one-way users of the LAPPS Grid; for example, users of an e-Learning Grid could be users of the LAPPS Grid in order to develop e-learning resources, but the LAPPS Grid need not be a user of the e-Learning Grid.

9 Conclusion

The LAPPS Grid project is currently in its third year and has so far provided the basic functionality of the framework. The next steps include expanding the range

[27] http://langrid.servicegrid-bangkok.org/en/overview.php.
[28] http://langrid.portal.cs.ui.ac.id/langrid/.
[29] Under development.
[30] http://www.linguagrid.org/.
[31] Funding for the LAPPS Grid involvement in the federation has awarded as a supplement to the NSF SI^2 grants ACI-1147912 and ACI-1147944.
[32] http://www.meta-net.eu/.
[33] http://eudat.eu/communities/clarin-common-language-resources-and-technology-infrastructure.
[34] http://www.kyoto-project.eu/.
[35] http://www.globalwordnet.org/gwa/gwa_grid.html.
[36] http://u-compare.org/.

of services offered and enhancing the integration with Galaxy. As noted above in Sect. 7, another important activity is the evaluation of current LAPPS Grid capabilities on the basis its use in several graduate-level courses in computational linguistics at major U.S. universities, which we hope will lead to significant enhancements of its usability as well as the range of available services. Another focus of activity will be to adapt the LAPPS Grid in order to empower users to carry out computational analyses without having to be an expert in computer science, so that users can focus on scientific rather than technical questions.

As our intention is to provide one piece of what is envisioned to become a global network of federated grids and services for NLP, another important activity is to pursue additional collaborations with similar projects around the world and work to ensure the maximal involvement of the community in the development of exchange mechanisms. We are also seeking means to incorporate individual services and composite service pipelines into the LAPPS Grid (either via direct inclusion or federation with grids that provide these services) for tasks relevant for research in areas such as digital humanities and bioinformatics, and in general to better accommodate the non-technical user.

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

References

1. Calzolari, N., Baroni, P., Bel, N., Budin, G., Choukri, K., Goggi, S., Mariani, J., Monachini, M., Odijk, J., Piperidis, S., Quochi, V., Soria, C., Toral, A. (eds.) Proceedings of The European Language Resources and Technologies Forum: Shaping the Future of the Multilingual Digital Europe. ILC-CNR (2009)
2. Cassidy, S., Estival, D., Jones, T., Burnham, D., Burghold, J.: The alveo virtual laboratory: a web based repository API. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014). European Language Resources Association (ELRA), Reykjavik, May 2014
3. Ferrucci, D., Nyberg, E., Allan, J., Barker, K., Brown, E., Chu-Carroll, J., Ciccolo, A., Duboue, P., Fan, J., Gondek, D., Hovy, E., Katz, B., Lally, A., McCord, M., Morarescu, P., Murdock, B., Porter, B., Prager, J., Strzalkowski, T., Welty, C., Zadrozny, W.: Towards the open advancement of question answering systems. Technical report, IBM Research, Armonk (2009)
4. Ferrucci, D.A., Brown, E.W., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A., Lally, A., Murdock, J.W., Nyberg, E., Prager, J.M., Schlaefer, N., Welty, C.A.: Building Watson: an overview of the DeepQA project. AI Mag. **31**(3), 59–79 (2010)
5. Fokkens, A., van Erp, M., Postma, M., Pedersen, T., Vossen, P., Freire, N.: Offspring from reproduction problems: what replication failure teaches us. In: Proceedings of the Conference of The Association for Computational Linguistics, pp. 1691–1701. The Association for Computational Linguistics (2013)
6. Garduno, E., Yang, Z., Maiberg, A., McCormack, C., Fang, Y., Nyberg, E.: CSE Framework: a UIMA-based distributed system for configuration space exploration unstructured information management architecture. In: Klgl, P., de Castilho, R.E., Tomanek, K. (eds.) UIMA@GSCL, pp. 14–17 (2013). Proceedings of the CEUR Workshop, CEUR-WS.org

7. 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., Nekrutenko, A.: Galaxy: a platform for interactive large-scale genome analysis. Genome Res. **15**(10), 1451–55 (2005)
8. Goecks, J., Coraor, N., Team, T.G., Nekrutenko, A., Taylor, J.: NGS analyses by visualization with trackster. Nat. Biotechnol. **30**(11), 1036–1039 (2012)
9. Goecks, J., Nekrutenko, A., Taylor, J.: Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. **11**, R86 (2010)
10. Hayashi, Y., Declerck, T., Calzolari, N., Monachini, M., Soria, C., Buitelaar, P.: Language service ontology. In: Ishida, T. (ed.) The Language Grid - Service-Oriented Collective Intelligence for Language Resource Interoperability, pp. 85–100. Springer, Heidelberg (2011)
11. Ide, N., Pustejovsky, J.: What does interoperability mean, anyway? toward an operational definition of interoperability. In: Proceedings of the Second International Conference on Global Interoperability for Language Resources (ICGL 2010), Hong Kong, China (2010)
12. Ide, N., Pustejovsky, J., Calzolari, N., Soria, C.: The SILT and FlaReNet international collaboration for interoperability. In: Proceedings of the Third Linguistic Annotation Workshop, ACL-IJCNLP, August 2009
13. Ide, N., Suderman, K.: The linguistic annotation framework: a standard for annotation interchange and merging. Lang. Resour. Eval. **48**, 395–418 (2014)
14. Ishida, T., Murakami, Y., Lin, D., Nakaguchi, T., Otani, M.: Open language grid-towards a global language service infrastructure. In: The Third ASE International Conference on Social Informatics (SocialInformatics 2014). Cambridge, Massachusetts, USA (2014)
15. ISO-24612: Language Resource Management - Linguistic Annotation Framework. ISO 24612 (2012)
16. Patel, A., Yang, Z., Nyberg, E., Mitamura, T.: Building an optimal QA system automatically using configuration space exploration for QA4MRE'13 tasks. In: Proceedings of CLEF 2013 (2013)
17. Pedersen, T.: Empiricism is not a matter of faith. Comput. Linguist. **34**(3), 465–470 (2008)
18. Piperdis, S.: The META-SHARE language resources sharing infrastructure: principles, challenges, solutions. In: Proceedings of the Eighth International Language Resources and Evaluation (LREC12). European Language Resources Association (ELRA), Istanbul (2012)
19. W3C OWL Working Group: OWL 2 Web Ontology Language: Document Overview. W3C Recommendation (2012)
20. W3C SKOS Working Group: SKOS Simple Knowledge Organization System Reference. W3C Recommendation (2009)
21. Windhouwer, M.: RELcat: a Relation Registry for ISOcat data categories. In: Calzolari, N., Choukri, K., Declerck, T., Dogan, M.U., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S. (eds.) LREC 2012, pp. 3661–3664. European Language Resources Association (ELRA), Istanbul (2012)
22. Yang, Z., Garduno, E., Fang, Y., Maiberg, A., McCormack, C., Nyberg, E.: Building optimal information systems automatically: configuration space exploration for biomedical information systems. In: Proceedings of the CIKM 2013 (2013)