Bridging the LAPPS Grid and CLARIN

Erhard Hinrichs*, Nancy Ide, James Pustejovsky†, Jan Hajič‡, Marie Hinrichs*, Mohammad Fazleh Elahi*, Keith Suderman**, Marc Verhagen†, Kyeongmin Rim†, Pavel Straňák‡, Jozef Mišutka‡**

*University of Tübingen, **Vassar College, †Brandeis University, ‡Charles University

{erhard.hinrichs, marie.hinrichs, mohammad-fazleh.elahi}@uni-tuebingen.de, {ide, suderman}@cs.vassar.edu, {jamesp, marc}@cs.brandeis.edu, krim@brandeis.edu,{hajic, stranak, misutka}@ufal.mff.cuni.cz

Abstract

The Andrew K. Mellon Foundation has funded a project to create a "trust network" between the Language Applications (LAPPS) Grid, a major framework for composing pipelines of natural language processing (NLP) tools, and the WebLicht workflow engine hosted by the CLARIN-D Center in Tübingen. The project also includes integration of NLP services available from the LINDAT/CLARIN Center in Prague. The goal is to allow users on one side of the bridge to gain appropriately authenticated access to the other and enable seamless communication among tools and resources in both frameworks. The resulting "meta-framework" provides users across the globe with access to an unprecedented array of language processing facilities that cover multiple languages, tasks, and applications, all of which are fully interoperable.

Keywords: Language Applications Grid, WebLicht, Text Corpus Format (TCF), LAPPS Grid Interchange Format (LIF), syntactic interoperability, semantic interoperability, user identification and authentication

1. Introduction

The Andrew K. Mellon Foundation has funded a project to create a "trust network" between the Language Applications (LAPPS) Grid (Ide et al., 2014), a major framework for composing pipelines of natural language processing (NLP) tools, and the WebLicht workflow engine (Dima et al., 2012) hosted by the CLARIN-D Center in Tübingen. The project also includes integration of NLP services available from the LINDAT/CLARIN Center in Prague[1]. The goal is to allow users on one side of the bridge to gain appropriately authenticated access to the other and enable seamless communication among tools and resources in both frameworks. The resulting "meta-framework" provides users across the globe with access to an unprecedented array of language processing facilities that cover multiple languages, tasks, and applications, all of which are fully interoperable.

The LAPPS Grid/CLARIN Mellon project involves two major tasks: (1) establishing a joint single sign-on user authentication and authorization mechanism; and (2) enabling seamless interoperability at both the syntactic and semantic levels among tools available from both the LAPPS Grid and WebLicht, so that users can mix and match these tools regardless of provenance and without concern for differing I/O requirements. In this paper we describe the work required to accomplish these tasks.

2. Overview

In the LAPPS Grid, language resources and NLP tools are made available from the Galaxy workflow engine (Giardine et al., 2005), as well as programmatic access through the LAPPS Grid API[2]. LAPPS Grid tools consume and produce data in the LAPPS Interchange Format (LIF) (Verhagen et al., 2015), a JSON-LD-based format designed to serve as an internal interchange format for linguistically annotated data. Semantic interoperability among services is accomplished via URI references to the LAPPS Grid Web Service Exchange Vocabulary (WSEV) (Ide et al., 2016). NLP tools are accessed as web services that deliver metadata about the content at a standardized URI and are at present invoked using the SOAP protocol.

WebLicht is an environment for building, executing, and visualizing the results of NLP pipelines, which is integrated into the CLARIN infrastructure (Hinrichs and Krauwer, 2014). WebLicht NLP tools are implemented as web services that consume and produce the Text Corpus Format (TCF)[3] data, an XML format designed for use as an internal data exchange format for WebLicht processing tools. The TCF also ensures semantic interoperability among all WebLicht tools and resources by defining a common vocabulary for linguistic concepts. Metadata descriptions of WebLicht tools are stored in repositories located at the CLARIN center hosting the service. WebLicht web services are invoked using the RESTful protocol.

LINDAT/CLARIN provides various NLP services[4] based on single-purpose tools or pre-configured chains of tools, often for multiple languages, most notably the UDPipe service (Straka et al., 2016). UDPipe produces CoNLL-U, the UD annotation format (Nivre et al., 2016), which is based on the column-based "BIO" format used in the Conference on Natural Language Learning (CoNLL) exercises. Subchains of UDPipe are being exposed in WebLicht, and will be made interoperable with WebLicht's other TCF-based tools.

The challenges to bridging The LAPPS Grid and WebLicht frameworks arise from differences in the architectures of the two systems, in particular the differences in data exchange formats, access to and format of metadata, and the protocols used to invoke web services. In addition, it is necessary to provide support for authentication and authorization mechanisms that allow users to access resources and services provided by each framework as easily and seam-

[1] https://lindat.mff.cuni.cz/en

[2] http://wiki.lappsgrid.org/Developing.html

[3] https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/The_TCF_Format

[4] http://lindat.mff.cuni.cz/services/

```
"text":{
      "@value": "Karen flew to New York."
},
"views": [
    { "metadata": {}
        "annotations": [
            {
                "id": "tok0","start": 0,"end": 5,
                "features": {"word": "Karen"
                }},...]},
    { "metadata": {},
        "annotations": [
            {
                "id": "pos_0","start": 0,"end": 5,
                "features": {"pos": "NNP"
                }},...]},
    { "metadata": {},
        "annotations": [
            {
                "id": "tok0","start": 0,"end": 5,
                "features": {"pos": "NNP","word": "Karen"}},...]}]
```

```
<text>Karen flew to New York. </text>
<tokens>
  <token ID="t_0">Karen</token>
  <token ID="t_1">flew</token>
  <token ID="t_2">to</token>
  <token ID="t_3">New</token>
  <token ID="t_4">York</token>
  <token ID="t_5">.</token>
</tokens>
<POStags tagset="pennTB">
  <tag tokenIDs="t_0">NNP</tag>
  <tag tokenIDs="t_1">VBD</tag>
  <tag tokenIDs="t_2">TO</tag>
  <tag tokenIDs="t_3">NNP</tag>
  <tag tokenIDs="t_4">NNP</tag>
  <tag tokenIDs="t_5">.</tag>
</POStags>
```

Table 1: Example LIF (top) and TCF (bottom) formats

lessly as those within the framework they typically use. Each of these tasks is described in the following sections.

3. Data Format Conversion

Conversion between the LAPPS Grid and WebLicht formats is addressed at the level of service protocols, and at the syntactic (data representation) and semantic (linguistic objects and their relations) levels.

Communication Protocols. The LAPPS Grid and CLARIN services use different communication protocols. The LAPPS Grid uses the Simple Object Access Protocol (SOAP), whereas the CLARIN tools are implemented as RESTful[5] services. To invoke LAPPS services registered in CLARIN, it is currently necessary to convert CLARIN's RESTful requests to LAPPS Grid SOAP requests. A SOAP-PROXY service has been implemented to take a REST service request as input, convert it to a SOAP message, invoke the service with the SOAP request, and return the response from the service. A similar mechanism enables access to CLARIN RESTful services from the LAPPS Grid.

Syntactic Interoperability. The problem of differing data exchange formats has been addressed at the syntactic level by implementing data converters between LIF and TCF as web services and registering them in both frameworks. Structural differences and the granularity of annotation data in LIF and TCF imposed several challenges:

- LIF is a stand-off annotation format and therefore requires character offset anchors to the primary text as part of an annotation. TCF token annotations, on which all further WebLicht annotations rely, do not contain character offsets; therefore, conversion from LIF to TCF requires mapping character offset anchors from/to each token.
- TCF allows only one occurrence of an annotation type per document, whereas LIF allows multiple occurrences of the same annotation type within a single output document (contained in different "views"). Therefore, only one of multiple LIF annotations of the same type can be chosen for conversion into TCF. For example, in Table 1, the LIF representation contains multiple part-of-speech annotations. Therefore, it is necessary to identify the optimal alternative for conversion into TCF. This remains an open problem at this time; currently, the last view for any given annotation type is included in the TCF representation.

Semantic Interoperability. The three frameworks in this project reference different sets of linguistic objects (with some overlaps), using differing terminology and expressing relations among these objects in differing configurations. To enable semantic interoperability among the services in the three frameworks, we provide means to specify the linguistic objects that a given service or tool requires as input and produces as output, so that that other producers and consumers (i.e., other services and/or tools) can determine if its requirements are satisfied. In a pipeline of tools or web services, this information is provided as metadata that must be checked automatically for compatibility. This in turn demands that identical concepts can be identified as such, either by direct match or by reference to a common web-addressable entity. Internally, a given tool may use different terminology; the only necessity is that the tool is wrapped to map the exchange vocabulary into the internal terminology and vice versa. Thus a single mapping of a tool's specific terminology into and out of the common exchange vocabulary is sufficient to enable information exchange with all others.

We have developed a mapping and linkage among concepts defined in the the vocabularies of WebLicht, LINDAT/CLARIN, and the WSEV to cover entities contained in the others. Where necessary, we have expanded specifications to meet the needs of additional services and data types. In some cases, it was necessary to split and/or merge concepts, revise/refine partially overlapping concepts, allow specification of alternatives, etc. Where necessary, we extended the vocabularies of WebLicht, LINDAT/CLARIN, and the WSEV to cover entities contained in the others. Details are omitted here for lack of space, but a full inventory of the kind and number of modifications required will be provided in the final paper, if accepted.

Figure 1 shows the LAPPS Grid - WebLicht integration framework. When one framework calls a service from the other, metadata from the called service is converted and made available to the other, after which it can be processed with the caller's usual handlers. Similarly, data conversion services allow each platform to consume and produce data in its native format. Service calls are tunneled through a proxy, which invokes services using the required protocol.

[5] https://www.w3.org/2001/sw/wiki/REST

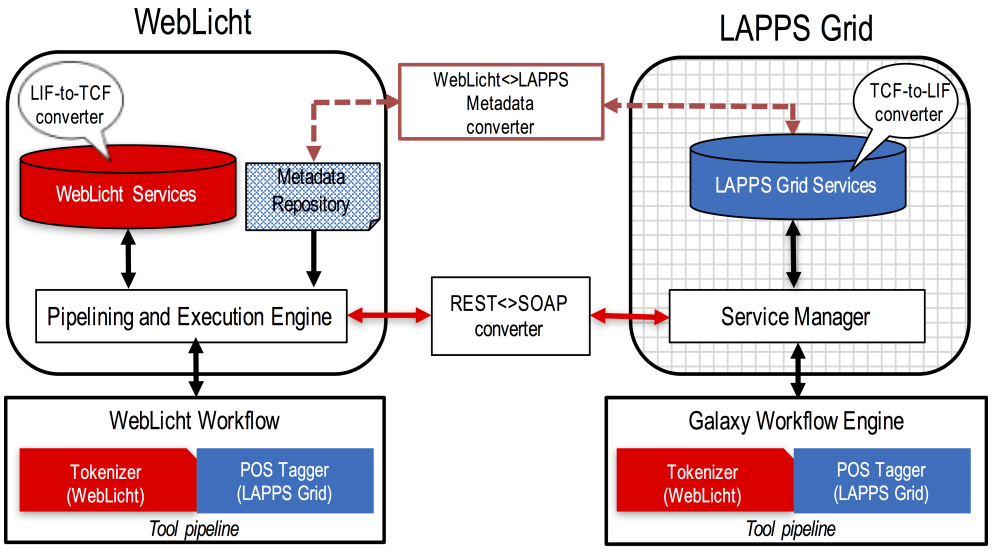

Figure 1: Integration framework

4. Metadata Conformance

Metadata about the web services available in LAPPS Grid and WebLicht contain information needed to invoke the services from their respective frameworks. The two frameworks handle web service metadata differently with respect to content, storage location, and fetching.

In the LAPPS Grid, each web service delivers its own metadata on demand, whereas WebLicht web service metadata is retained in CLARIN Center repositories. WebLicht metadata, which follows the specifications of CLARIN CMDI[6] framework, includes information corresponding LAPPS Grid metadata as well as additional details about the format and contents of a service's input and output. Therefore, WebLicht metadata can be converted to LAPPS Grid metadata automatically, but the reverse is not true. To handle this, LAPPS Grid metadata is stored in the WebLicht repository and augmented manually to include the required additional information.

5. User Authentication and Identification

Using the LAPPS Grid through the Galaxy interface requires a simple registration in order to provide a uniquely-named workspace for each user, but there are no usage restrictions depending on the user type or affiliation. When it is necessary to provide secure access to licensed data and software, the LAPPS Grid uses "click through" licenses that can be accepted in real time as well as verification via timed tokens (Cieri and DiPersio, 2014). In the CLARIN infrastructure, users must be authenticated via identity providers belonging to EU national identity federations that in turn ensure secure authentication services.

Because of the need for authentication, prior to this project LAPPS Grid users were in general not able to access login-protected CLARIN services. To provide this access, we have devised means for LAPPS Grid users with appropriate credentials to access these services by registering the LAPPS Grid as both a service and identity provider with the relevant organizations.

[6] http://www.clarin.eu/cmdi

6. Example

Figure 2 presents an example use of tools from both the LAPPS Grid and WebLicht frameworks, accessed via the WebLicht user interface. An input text corpus is converted to LIF format, tokenized and sentence-split by LAPPS services, followed by a LIF→TCF format conversion to allow processing to continue using CLARIN services. The lower window in the figure 'Input and Chain Selection'shows the tool chain that was selected for execution. After the LAPPS Grid services (Stanford Tokenizer and Stanford Splitter) are executed, the LIF → TCF converter is used to switch back to using WebLicht; the upper window ('Next Choices') shows the available WebLicht services. If TCF → LIF is chosen, then the chain will switch again from WebLicht to the LAPPS Grid. In this way, a user can alternate between LAPPS Grid services and WebLicht services and vice versa, without ever leaving the WebLicht interface.

Figure 3 shows a portion of the LAPPS Grid Galaxy interface and a workflow in which a tokenizer and part-of-speech tagger from WebLicht are invoked, followed by a named entity recognizer from the LAPPS Grid. In the second step of the workflow, the user specifies the appropriate format for wrapping the input text (in this case, TCF). Output from the part-of-speech tagger is converted to LIF before invoking the LAPPS Grid entity recognizer. Note that while conversion to and from TCF and LIF is done explicitly in this example, in the final implementation this will not be required on the LAPPS Grid side as the Galaxy interface automatically detects and converts formats as needed, without intervention from the user.

7. Broader Impact

The meta-framework providing for mutual access between the LAPPS Grid and the two CLARIN frameworks has the potential to transform scholarship and development across

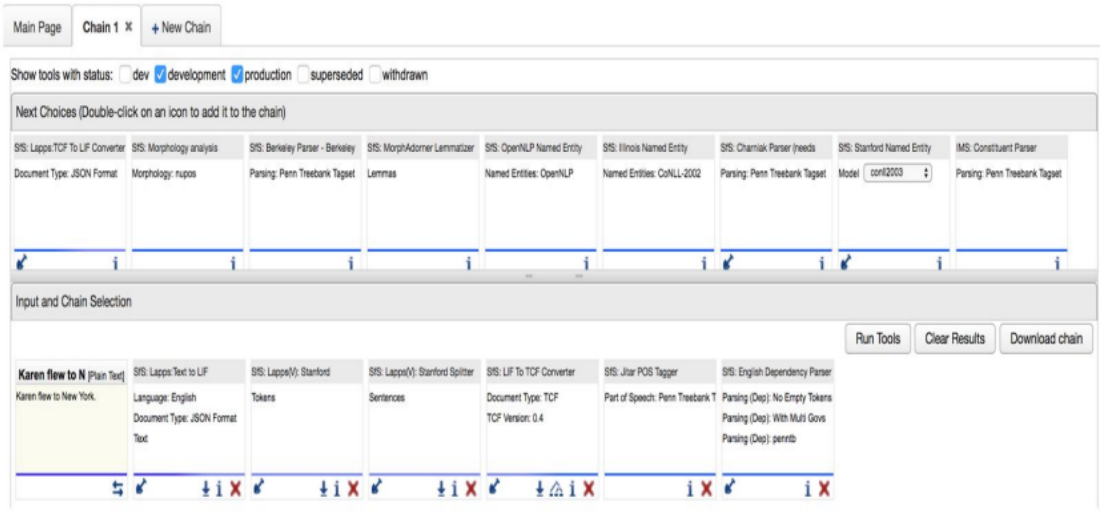

Figure 2: Invoking LAPPS Grid services from WebLicht.

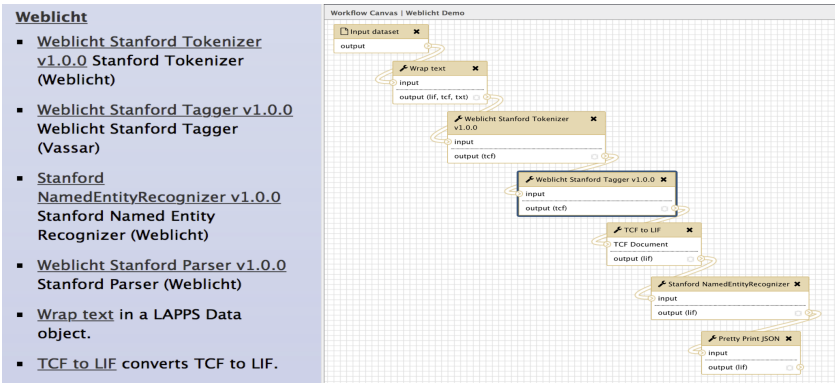

Figure 3: Invoking WebLicht services in a LAPPS Grid/Galaxy pipeline

multiple disciplines in the sciences, language and social sciences, and digital humanities by providing a transparent interface to a massive range of tools and resources at an unprecedented level.

Bridging the LAPPS Grid and WebLicht frameworks will significantly extend the capabilities of each, by providing seamless access to services that are currently unavailable. For example, the LAPPS Grid will benefit from availability of a more extensive suite of tools for output visualization than currently exists in the LAPPS Grid, and WebLicht will gain access to the sophisticated evaluation services the LAPPS Grid provides.

The potential impact extends even farther than the two frameworks involved, as both the LAPPS Grid and WebLicht are federated with other frameworks to which they provide a gateway. WebLicht is a member of the EU CLARIN network and therefore provides access to multilingual tools and resources from CLARIN Centers hosted throughout Europe. The harmonization will also extend to Asia because the LAPPS Grid is federated with seven other grids[7], including the Language Grid housed at Kyoto University[8]. Like the LAPPS Grid-CLARIN bridge, this federation provides interoperability and seamless access among atomic and composite web services available from any of the grids involved.

A more wide-ranging impact of this project may result from its success in providing interoperable access to services in two major frameworks that were developed entirely independently. Although we acknowledge that universal interoperability for NLP tools is far from a solved problem, we believe this project takes an important step towards its achievement. Additionally, our solutions to the problems of authentication, authorization, and access to licensed data and tools can serve as a model for other project facing the same issues. Finally, the work performed takes a major step toward the harmonization of software and data developed across the globe that can vastly ameliorate and eventually eliminate the current lack of reusability of resources and tools that thwarts research and development in the field and hampers collaboration. Ultimately, the LAPPS Grid-CLARIN meta-network may lay the groundwork for the eventual creation of a global network of grids and frameworks to serve researchers, developers, and users of NLP technologies.

[7] Federated Grid of Language Services (FGLS) (Ishida et al., 2014). [8] http://langrid.org

8. Bibliographical References

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*, pages 93–100, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Dima, E., Hinrichs, E., Hinrichs, M., Kislev, A., Trippel, T., and Zastrow, T. (2012). Integration of weblicht into the clarin infrastructure. In *Proceedings of the Joint CLARIN-D/DARIAH Workshop at Digital Humanities Conference 2012: Service-oriented Architectures (SOAs) for the Humanities: Solutions and Impacts*, pages 17–23, Hamburg, Germany.

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.

Hinrichs, E. and Krauwer, S. (2014). The clarin research infrastructure: Resources and tools for ehumanities scholars. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

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

Ide, N., Suderman, K., Verhagen, M., and Pustejovsky, J. (2016). The language applications grid web service exchange vocabulary. In *Revised Selected Papers of the Second International Workshop on Worldwide Language Service Infrastructure WLSI (2015)*, pages 18–32, Kyoto, Japan. Springer-Verlag New York, Inc.

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.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1659–1666, Paris, France. European Language Resources Association.

Straka, M., Hajič, J., and Straková, J. (2016). UDPipe: Trainable pipeline for processing coNLL-u files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4290–4297, Paris, France. European Language Resources Association.

Verhagen, M., Suderman, K., Wang, D., Ide, N., Shi, C., Wright, J., and Pustejovsky, J. (2015). The lapps interchange format. In *Revised Selected Papers of the Second International Workshop on Worldwide Language Service Infrastructure WLSI (2015)*, pages 33–47, Kyoto, Japan. Springer International Publishing.