Year 1 - Activities

# 1. Summary

The project has three main thrusts, divided over three work packages: infrastructural interoperability (WP1), technical interoperability (WP2) and semantic interoperability (WP3). The goals for these three work packages are as follows:

* WP1: develop a plan for a solution to allow cross-framework access between LAPPS Grid and Weblicht based on the necessary trust measures their respective AAI subsystems.
* WP2: develop a solution that provides access to both LAPPS Grid and WebLicht tools to users of both frameworks.
* WP3: compare the vocabulary for linguistic concepts that provide the means to formallyspecify the linguistic objects that a given service or tool requires as input and produces as output.

# 2. Progress

## WP1 - Infrastructural Interoperability

The authentication and authorization models used in LAPPS Grid framework and CLARIN are based on different technologies. Both frameworks have also different requirements on the users. Because of this, users from one framework cannot use the services and resources from the other automatically. The goal of WP1 (tasks T1.1 and T1.2) is to review all the procedures related to authentication and authorization including assessing the required level of trust of users within each framework and specify changes required in both frameworks to enable cross-framework access based on the necessary trust measures.

We created a detailed review of both frameworks and identified crucial differences in the authentication and authorization process and the level of trust of users. We also analysed an alternative solution based on the eduGAIN federation network but that proved less advantageous than the solution based on the CLARIN infrastructure. Finally, we created a specification of subtasks that when finished, LAPPS Grid users will be able to use CLARIN services by accessing them from both web frontends – LAPPS Grid’s Galaxy or CLARIN’s Weblicht. Correspondingly, CLARIN user will have access to LAPPS Grid services from both web applications too. Based on the analysis of all the possible scenarios in Figure  1, we identified the following different scenarios that require changes in the frameworks:

1. LAPPS Grid user executing public Weblicht services in Galaxy;
2. LAPPS Grid user executing academic[[1]](#footnote-0) Weblicht services in Galaxy;
3. LAPPS Grid user executing public Weblicht services in Weblicht;
4. LAPPS Grid user executing academic Weblicht services in Weblicht;
5. CLARIN user executing LAPPS grid services in Galaxy;
6. CLARIN user executing LAPPS grid services in Weblicht.

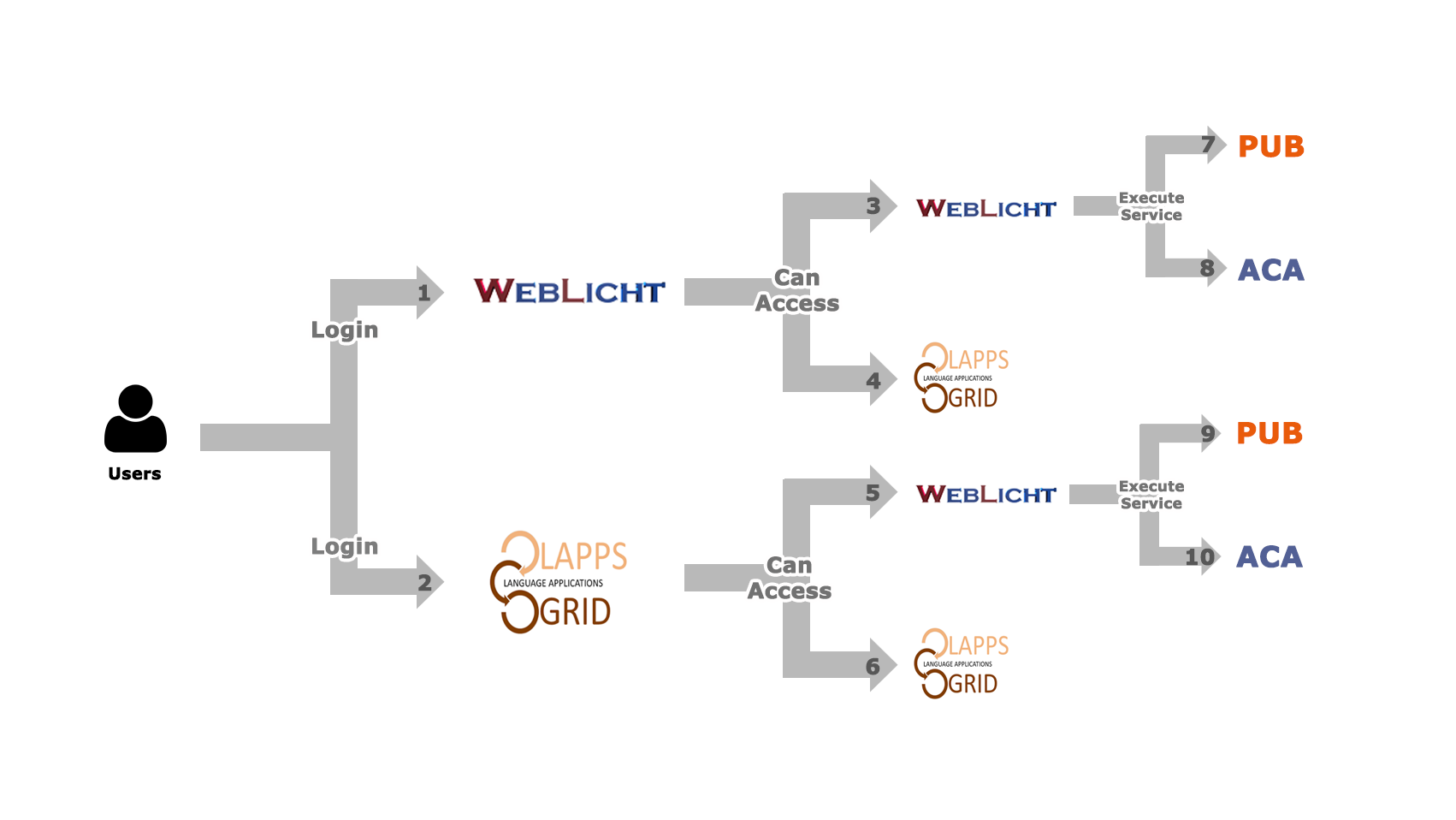


Figure 1: Different user flows while executing services from LAPPS Grid and Weblicht frameworks.

The solution is comprised of the following subtasks: make Galaxy LAPPS Grid a service provider in the CLARIN SPF federation and the InCommon U.S. education and research identity federation, creating a LAPPS Grid Identity provider that will be explicitly trusted by the Weblicht service provider, enhance Weblicht with the notion of public services and users with two different level of trusts (minimal and academic), allow automated elevation of level of trust for LAPPS Grid users as illustrated in Figure 2.

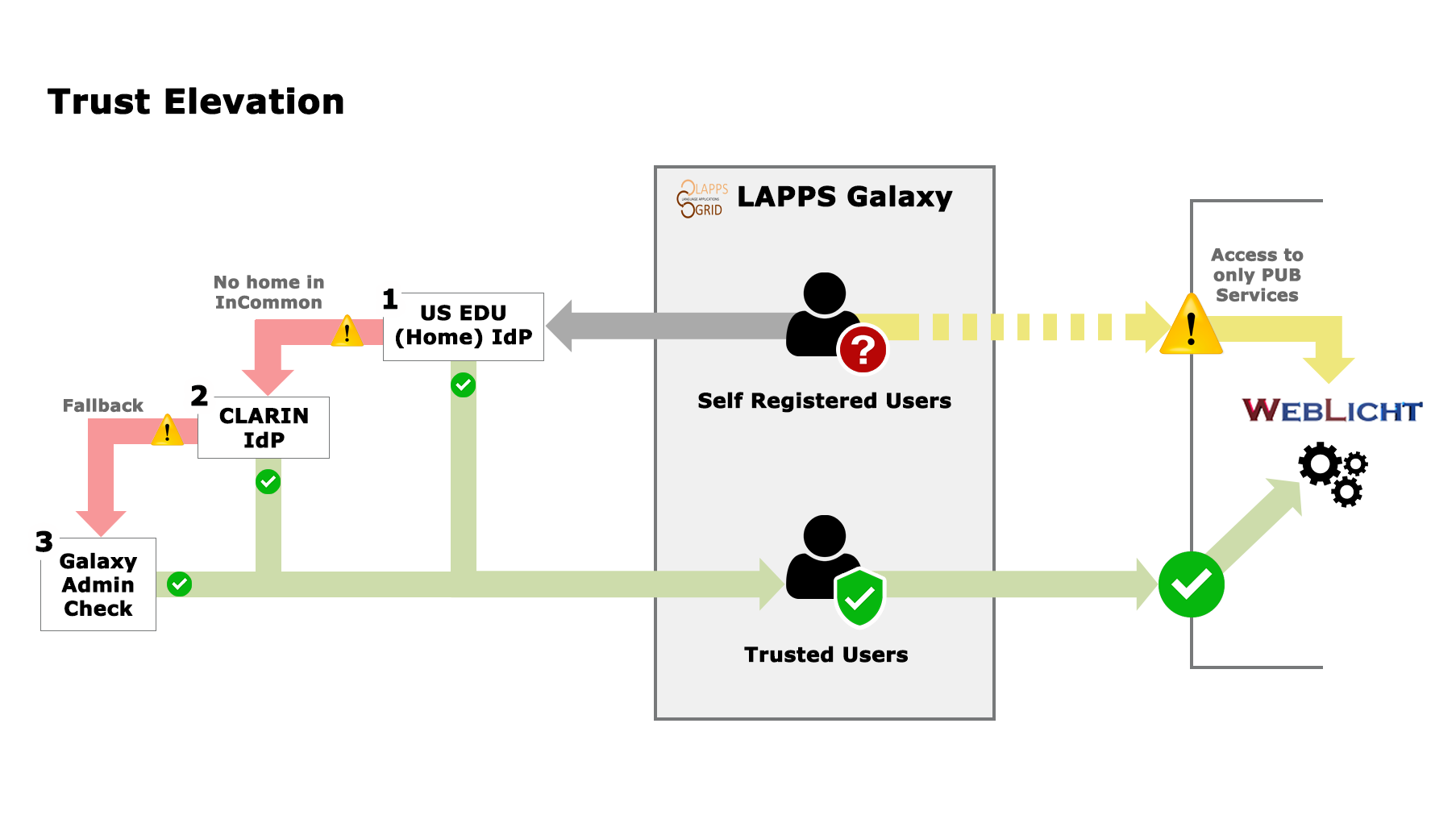


Figure 2: Automation of Level of Trust Elevation.

## WP2 - Technical Interoperability

LAPPS Grid and WebLicht have different underlying architectures and internal data exchange formats. The goals in this reporting phase were to implement prototype conversion software between the data formats used by the two systems, and to make services in each infrastructure visible to the other infrastructure, thereby enabling invocation of services from both sides. In this section, we briefly discuss on the development carried out on each of the two tasks.

Data Converters:

In the LAPPS Grid, language resources and NLP tools are made available from the Galaxy workflow engine, as well as programmatic access through the LAPPS Grid API[[2]](#footnote-1).The tools consume and produce data in the LAPPS Interchange Format (LIF),[[3]](#footnote-2) a JSON-LD-based format designed to serve as an internal interchange format for linguistically annotated data. WebLicht[[4]](#footnote-3) is an environment for building, executing, and visualizing the results of NLP pipelines, which is integrated into the CLARIN infrastructure. WebLicht NLP tools consume and produce the Text Corpus Format (TCF)[[5]](#footnote-4) data, an XML format designed for use as an internal data exchange format for WebLicht processing tools. The structural differences and granularity of annotation data in LIF and TCF impose several challenges to the conversion task.

* One challenge faced in converting token-level annotations from TCF to LIF stems from the fact that LIF requires start and end character offsets into the primary text to be part of the tokens-level annotation, but TCF token annotations normally do not contain this character offset information.
* Another challenge was encountered in converting LIF to TCF when the structures of the two document types were considered. LIF allows multiple occurrences of the same annotation type within a document, whereas TCF allows only one occurrence of an annotation type per document. For example, the LIF representation contains multiple part-of-speech annotations in one document. Only one of these can be chosen for conversion into TCF.

These challenges, along with others, and technical implementation details were discussed on the milestone report in months 1-6. Software for converting documents in both directions between the two formats was implemented and made available as web service.

Service Metadata:

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. The main challenges were to convert the service metadata (to make services visible to the other infrastructure) and to develop a proxy service to convert between the differing communication protocols used in each infrastructure (to enable invocation of services). The following steps were taken to achieve the goals in this phase:

* LAPPS Grid uses the Simple Object Access Protocol (SOAP), a messaging protocol in which each web service provides its own metadata. The WebLicht services are stored using CMDI[[6]](#footnote-5), a framework developed within CLARIN that provides a way to describe and reuse metadata components. WebLicht service metadata was mapped to the LAPPS metadata format automatically. Mapping LAPPS service metadata to WebLicht metadata was done with the help of the CMDI Orchestration Metadata Editing Tool (COMET[[7]](#footnote-6)), which is a tool for creating and validating WebLicht service metadata.
* 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 services. 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.

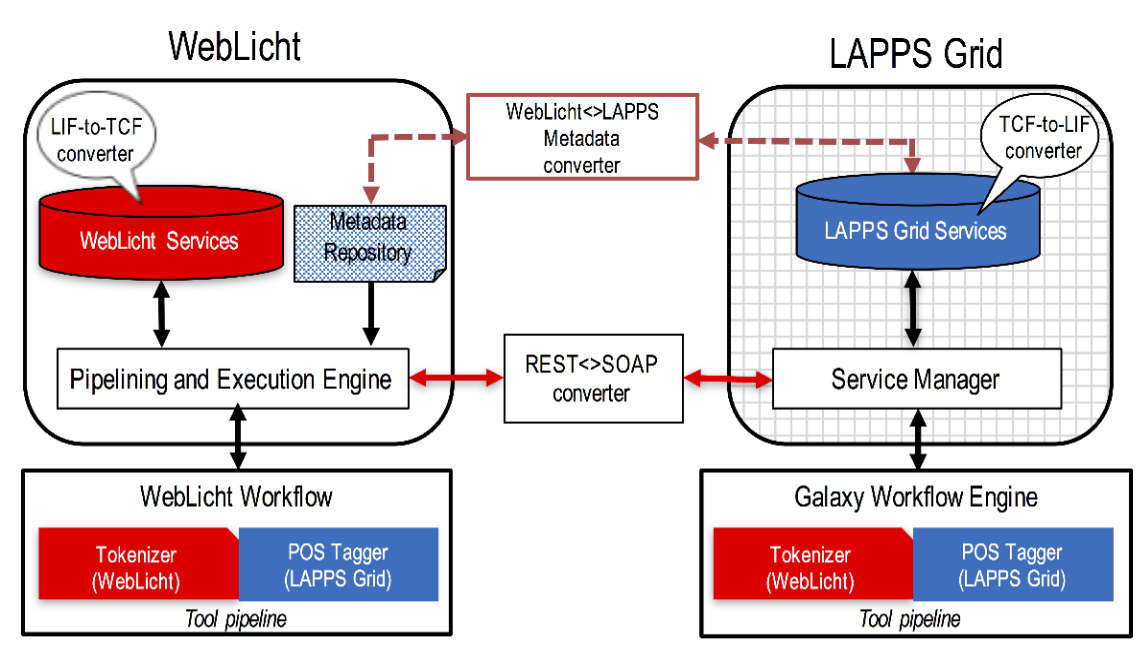


Figure 1: Integration framework

State of the Integration:

LAPPS and WebLicht are partially integrated and now it is possible to create and execute processing pipelines cross-platform. 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. The technical details of constructing processing pipeline cross-platform are discussed on the milestone report in months 7-12.

## WP3 - Semantic interoperability

The goal for this work package is to compare the vocabulary for linguistic concepts that provide the means to formallyspecify the linguistic objects that a given service or tool requires as input and produces as output. We compared the CLARIN WebLicht TCF format and the LAPPS Web Services Exchange Vocabulary (WSEV), at first just for the linguistic objects as used by tokenizers, sentence splitters, part-of-speech tagger, parsers and other common processing tools in WebLicht and the LAPPS Grid and wrote an interim report that provides an inventory of all objects and their features, as well as the mappings between them.

In general, there are fairly easy mappings between linguistic categories in TCF and the annotation types in WSEV. There are a couple of differences though that are worth mentioning:

1. In a few cases either TCF or WSEV lacks a category that is used by its counterpart. For example, at the onset of this project TCF had no category for noun chunks or verb chunks, this category has been added.
2. Features in WSEV are sometimes expressed as objects in TCF. For example, in LAPPS the Token annotation type contains features for part-of-speech and lemma, whereas in TCF part-of-speech and lemma are their own objects. As a result, some feature needed to be added to the LAPPS Token annotation type in order to express TCF features on part-of-speech and lemma.
3. We intend to extend both TCF and LAPPS so that they can incorporate tools that generate Universal Dependencies. This boils down to adding vehicles to express multi-word tokens.

In the process of comparing the two vocabularies we have created a first draft of a simple formal mapping language. The language maps elements from the TCF schema to the LAPPS vocabulary. An example is

constituent#cat ⟺ Constituent#label

This states that the TCF *constituent* element has an attribute *cat* and this attribute maps to the label attribute of the WSEV *Constituent* annotation type.

# 3. Board, management, or staff changes;

No significant changes to report.

# 4. Plans and goals

In the next period, the goal in WP1 is to formalise the solution into a detailed and actionable list which will lead to a successful cross-framework access between LAPPS Grid and CLARIN. For WP2 we intend to further extend CLARIN and LAPPS Grid platform, allowing users to access more services cross-platform. And for WP3, we intend to expand the number of linguistic types in our analysis and further expand the mapping language.

1. Weblicht services can have academic usage/affiliation condition in terms of usage. [↑](#footnote-ref-0)
2. http://wiki.lappsgrid.org/Developing.html [↑](#footnote-ref-1)
3. http://wiki.lappsgrid.org/interchange/ [↑](#footnote-ref-2)
4. https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main\_Page [↑](#footnote-ref-3)
5. https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php /The TCF Format [↑](#footnote-ref-4)
6. http://www.clarin.eu/cmdi [↑](#footnote-ref-5)
7. http://weblicht.sfs.uni-tuebingen.de/comet/ [↑](#footnote-ref-6)