Transatlantic Collaboration between LAPPS and CLARIN
M6 Milestone Report

WP1, Tasks T1.1 and T1.2

User Authentication and Identification

Introduction

This document reports on the progress in WP1 in the first 6 months of the project. The goal was to review authentication and authorization infrastructure (AAI) of both LAPPS Grid and CLARIN. Services and applications in LAPPS Grid and CLARIN require different trust levels and are using different technologies. The result of this report is a unified summary that will be further used to prepare the implementation plan to support users from both communities.

Background

The Language Application (LAPPS)[1] Grid framework is based on the Langrid Service Manager[2] platform (Ide et al., 2014) and provides access to basic natural language processing (NLP) tools and resources. The Galaxy[3] platform enables pipelining these tools in a workflow engine. Even though most of the Galaxy workflow engine functionality can be used without any registration there are use cases when the service must identify the user such as when the user wants to have a persistent history of his actions or when data and services are not publicly available.

European Common Language Resources and Technology Infrastructure (CLARIN)[4] Research Infrastructure is based on a distributed network of national organizations (centres). The centres offer unified access to language data and provide advanced tools and services. WebLicht[5] is an environment for building and executing NLP pipelines integrated into the CLARIN infrastructure (Dima et al., 2012) that provides easy access to selected CLARIN services. CLARIN is focusing primarily on research and Weblicht enforces academic usage by requiring authentication via CLARIN Service Provider Federation (SPF)[6].

---

[1] https://www.lappsgrid.org/
[2] http://langrid.org/oss-project/en/index.html
[3] https://galaxyproject.org
[4] https://www.clarin.eu/
[5] https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main_Page
[6] https://www.clarin.eu/content/service-provider-federation

LAPPS Grid registration, authentication and authorization procedures

In this section, we firstly describe the workflow when using LAPPS Grid workflow engine. Then, we describe the details and requirements of the stakeholders in this process and visualize it in a summary table.

When a user invokes a pipeline in LAPPS Grid several components have to communicate together to handle the request. Everything starts in the Galaxy *user interface*[7] where user connects data and services he wants to invoke. The user interface must know which services are available. The list of available services is obtained from *service managers.* Service managers are web applications that can be also used to manage individual services. Service managers then invoke individual services feeding them with input data or intermediate results.

The stakeholders in LAPPS Grid AAI workflow are:

- **graphical user interface (GUI)** - e.g., Galaxy;
- **command line interface (CLI)** - not publicly supported (even though possible either via service manager or urls pointing directly to services);
- **backend** - service manager;
- **individual services** - specific tools;
- **underlying data** - e.g., 3rd party rights holders.

In a standard use case, user registers in a web form filling out a username (in the form of an email), password and a public name. There is currently only basic validation on the input fields themselves (e.g., that username must be at least three characters long). Right afterwards, user can start using most of the functionality. If data that are not public are required, user has to authenticate to a 3rd party system that verifies authorization to that data (e.g., if a fee has been paid). The system returns secret token that can be used to delegate access to LAPPS Grid.

Summary Table

In the table below, each row represents one stakeholder in the standard workflow of LAPPS Grid. If all standard users have the same set of functionality available after logging in it means that the authorization is equal to authentication. Relations that are not possible are marked as N/A (not applicable).

	authentication	**authorization**
GUI	• optional via self registration without validation	equals to authentication
CLI	N/A	N/A
backend	• username and password specific to service manager	equals to authentication

[7] https://galaxy.lappsgrid.org/

	• GUI uses special user to communicate with service manager	
services	none required	none required
data	• OAUTH[8] protocol • 3rd party credentials used to get OAUTH token	• transparent 3rd party validation based on OAUTH token filled in by the user ○ valid for specified amount of time e.g., six hours

CLARIN registration, authentication and authorization procedures

Similar to the above, when a user invokes a pipeline in Weblicht several components have to communicate together to handle the request. Everything starts in a *graphical user interface* where user connects data and services he wants to invoke. Only example data (text) is offered by Weblicht on the contrary to LAPPS Grid. The user interface must know which services are available. The list of available services is available directly in Weblicht by harvesting configured OAI-PMH[9] repository endpoints exporting specific metadata profiles. Weblicht then invokes individual services often located remotely feeding them with input data or intermediate results. There is a possibility to download the workflow from graphical user interface and and execute it via *command line interface*.

One of the requirements for a standard use case for Weblicht is that the user is related to an academic institute or university. Verifying this requirement can be achieved by joining the CLARIN SPF as a service provider. This means that users from any identity provider from CLARIN SPF are able to authenticate to Weblicht. In 2017, there are 18 European states directly connected to CLARIN SPF. However, CLARIN SPF service providers are exported to eduGAIN[10] inter-federation which means they are accessible also by organizations from the InCommon[11] federation in the United States. The vast majority of the organizations in these federations are universities or research organizations and users from these organizations are affiliated with research.
In case when the user is not affiliated with an organization from CLARIN SPF, (s)he can use CLARIN Identity Provider[12] and undergo a verification procedure to ensure the user is a real person and that the intended use does not violate any licenses. In order to guarantee a level of trust also on the international scale, CLARIN SPF maintains a list of blacklisted identity

[8] OAuth is an open standard for access delegation used to grant applications access to their information but without giving them the passwords.
[9] Open Archives Initiative Protocol for Metadata Harvesting.
[10] https://www.geant.org/Services/Trust_identity_and_security/eduGAIN
[11] https://www.incommon.org/
[12] https://www.clarin.eu/content/clarin-identity-provider

providers that do not meet basic requirements e.g., users have to be identifiable to real person[13].

The stakeholders in Weblicht AAI work flow are:

- **graphical user interface (GUI)** - Weblicht frontend;
- **command line interface (CLI)** - WebLicht as a Service;
- **backend** - Weblicht;
- **individual services** - specific tools;
- **underlying data** - no data available other than simple example ones.

Summary Table

	authentication	**authorization**
GUI	● via CLARIN SPF or CLARIN Identity provider	equal to authentication
CLI	● CLARIN SPF or CLARIN Identity Provider GUI login used to get API key ○ generated API key used valid for specified amount of time e.g., three months	equal to authentication
backend	none required	none required
services	none required	none required
data	N/A	N/A

Comparison

In the table below, we summarize the advantages, disadvantages and the overall trust level of different authentication methods described in the previous sections. Each row summarises one aspect of the trust level provided by different methods. In this context it means, what pieces of information are available to the system and who is responsible for the accuracy of the information. For example, self asserted (provided by the user itself) information about employment is not as much trustworthy as if it is provided by an organization or university that is the user's employer.

Local accounts	**Identity federation account**	**CLARIN Identity Provider**
● self asserted ● not validated	● validated at the university or organization	● validated manually by CLARIN ERIC staff on

[13] General description of how this works is at https://lindat.mff.cuni.cz/en/how-do-i-sign-up

	• very strict procedures (often at least identity documents are required) because of employment or the possibility of getting academic degrees • assertions based on the obtained information	case-by-case basis • proof required in form of organizational email address or similar
user can be anyone/anything	user is a real person and can be identified	user is a real person and can be identified
user can be anyone/anything	user is related to academic organization or university	user is related to research

Table 1. Comparison of different authentication methods and their trust level.

References

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

E. Dima, E. Hinrichs, M. Hinrichs, A. Kislev, T. Trippel, and T. Zastrow. 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.