# State of the art at LINDAT/CLARIN

LINDAT/CLARIN hosts approx. 20 services related to linguistics listed at <https://lindat.mff.cuni.cz/en/services> and a couple more related to the technical infrastructure.

Majority of LINDAT/CLARIN services related to linguistics do not need explicit AAI (authentication and authorization) features because they are for free.

Several linguistic services do require authentication (and authorization) because they either directly contain data with less permissive licenses, or offer features connected to individual users like their processing history, or relax restrictions for logged in academic users, or are used to identify users that want to add add/modify new submissions.

Infrastructure related services do require authentication because the underlying content is available only to particular users.

In many cases, authentication is equal to authorization which means that no additional information is needed to decide if the user has or has not access for particular resources than successful authentication.

## Overview

If a linguistic services requires AAI, it has to support Shibboleth (SAML2) authentication. Shibboleth is a single sign-on (log-in) system. It allows people to sign in using just one identity to various systems run by federations of different organizations or institutions.

Federations consists of Identity Providers (IdPs) and Service Providers (SPs). An IdP is often an institute or university that knows identities (users) and stores their login information. A SP is typically a service that uses information released by IdP about particular user. The information are assertions released as attributes (e.g., givenname:Jozef). Note, that a user always authenticates to an IdP and only then the IdP decides which information is released to the SP (e.g., password is not sent to the SP).

As mentioned above, IdPs and SPs form federations and federations are commonly formed on a national level but can extend to inter-federations.

LINDAT/CLARIN is a SP in:

* Czech national federation eduID.cz (our “home” federation);
* EduGAIN inter federation;
* Clarin SPF federation.

There are number of advantages with this approach:

* single sign on - one identity (username/password) to many services;
* needs trust only in one place - your IdP login screen;
* simplifying services - they do not need to store passwords;
* simplifying services - login screen, error messages can be centralised and services do not need to know about them.

There are also a couple of disadvantages:

* Users that are not affiliated with any of the federated institutions. The solution for this problem is a “home-less” IdP often created for particular projects that can accept any users that meet the requirements. In our case, it is the CLARIN IdP (<https://www.clarin.eu/content/clarin-identity-provider>). As a last resort, services can provide local accounts which we also do for several of our services; however, these local accounts are specific for each service.
* Some IdPs do not release enough information to SPs to be able to identify a user (e.g., they do not release attributes designed for authentication). LINDAT/CLARIN created a tool that monitors released attributes to several SPs and notifies IdPs in case not all required information has been released. See also #Authorization considerations.

There are a few other very specific problems (and solutions) that I consider out of the scope of this light overview.

## Federated login case study (slightly simplified)

1. User visits <https://lindat.mff.cuni.cz/repository>
2. Clicks Login button ()
3. Local selection of federated IdPs is presented, search for “clarin”, click on \*clarin.eu website account\*
4. Redirected to [https://idp.clarin.eu/profile/SAML2/Redirect/…](https://idp.clarin.eu/profile/SAML2/Redirect/%E2%80%A6)
5. Enter credentials for CLARIN IdP
6. Redirected back to <https://lindat.mff.cuni.cz/repository/xmlui/?locale-attribute=en> but already logged in

CLARIN IdP released the following attributes

* edupersonprincipalname: misutka\_ufal.mff.cuni.cz@clarin.eu
* edupersonscopedaffiliation: member@clarin.eu
* entitlement: http://www.clarin.eu/entitlement/academic
* cn: misutka
* o: CLARIN
* mail: misutka@ufal.mff.cuni.cz
* sn: misutka

In this case, mail with edupersonprincipalname was used as the identifier of the user.

## Authorization considerations

Authorization can be either done locally for each authenticated user or based on the attributes released by IdPs e.g., entitlement or as a mix of the two approaches. However, relying on harmonised attributes from thousands of IdPs in the whole world is problematic.

## Services Case studies

### Digital Repository (https://lindat.mff.cuni.cz/repository)

AAI is required to:

* submit data;
* sign licenses for datasets that require license signing before they can be downloaded;
* role management (users/editors/curators/administrators).

To identify a user, attributes released by federated login are used to create the unique identification.

Authorization information is stored “locally” in a database.

Supports both federated login and local accounts.

### PML-TQ Treebank Query (<https://lindat.mff.cuni.cz/services/pmltq/>)

AAI is required to:

* licensed datasets;
* administration;
* history of queries.

To identify a user, attributes released by federated login are used to create the unique identification.

Authorization information is stored “locally” in a database.

Supports both federated login and local accounts.

### Treex::Web (https://lindat.mff.cuni.cz/services/treex-web/)

AAI is required to:

* storage of processing results and history.

Supports both federated login and local accounts.

### Cesilko (<http://lindat.mff.cuni.cz/services/cesilko/>)

AAI is required to:

* lifts the limit of input text length.

Also available in weblicht as service (under development).

### Dialogy.org (<http://lindat.mff.cuni.cz/services/dialogy.org/>)

AAI is required to:

* get access to sensitive data.

### Kontext (<https://lindat.mff.cuni.cz/services/kontext/>)

AAI is required to:

* corpus search with authentication used to store customisations and settings;
* possible to also use for authorisation to datasets, but currently only Open Access datasets on the production system.

### CMS joomla system / Monitoring services

AAI is required to:

* identify administrators/editors.

Supports both federated login and local accounts.