D2.12

Robust SPF 2: Identity Provider and Discovery Service

Document information

Title	Robust SPF 2: Identity Provider and Discovery Service
ID	CLARINPLUS-D2.12 (CE-2017-1068)
Author(s)	Jozef Mišutka
Responsible WP leader	Dieter Van Uytvanck
Contractual Delivery Date	2017-07-31
Actual Delivery Date	2017-07-30
Distribution	Public
Document status in workplan	Deliverable

Project information

Project name	CLARIN-PLUS
Project number	676529
Call	H2020-INFRADEV-1-2015-1
Duration	2015-09-01 – 2017-08-31
Website	www.clarin.eu
Contact address	contact-clarinplus@clarin.eu

Table of contents

1 Executive Summary

This document describes the final set of implementations to make the CLARIN Service Provider Federation (SPF) more robust in terms of quality, reliability and performance. The work reported in this deliverable builds upon the work done in the first phase of Task T2.1 of WP2 related to SPF and that is described in [CLARINPLUS-D2.2].

The first goal was to offer CLARIN users a secure, user and administrator-friendly identity management solution that supports different authentication and authorisation protocols. The second goal was to create a robust and sustainable central service that helps CLARIN members perform federated login.

2 Introduction

CLARIN stands for Common Language Resources and Technology Infrastructure. The primary aim of CLARIN is to provide easy and sustainable access to digital language data for scholars in the Humanities and the Social Sciences. One of the infrastructural pillars that makes this possible is the CLARIN Service Provider Federation (SPF). It connects CLARIN Service Providers to national federations inside the European Union. Users from institutes that are members of a national federation in countries that have joined the CLARIN SPF can automatically access CLARIN's protected resources and services in a secure way. Other users can register with the CLARIN Identity Provider to get access to protected resources and services once their membership is approved.

The first part of Task T2.1 of WP2 was focussed on a technical re-engineering of the CLARIN Identity Provider (IdP) as described in [CLARINPLUS-D2.2]. The work included complex testing of all software, its deployment and the actual migration of the user base to the new service. This is described in Section 3.

The next part of Task 2.1 was meant to ensure that the central *Discovery Service* that is responsible for the up-to-date listing of supported institutions in federated login is deployed in a highly available mode recovering from reasonable failures dynamically. This also included an evaluation of alternatives to the current discovery solution and a setup of processes to passively ensure that the supported IdPs have a reasonable level of trust. This is described in Section 4.

3 Re-engineering of the CLARIN Identity Provider

The deliverable [CLARINPLUS-D2.2] describes the evaluation procedure and the reasons for choosing Unity IDM[1] software as the identity management solution in CLARIN. Unity IDM will also be used as the technological solution for the new CLARIN Identity Provider (IdP). Here, we describe the work that builds upon the work described in [CLARINPLUS-D2.2].[2]

The migration process from the old CLARIN IdP to the new one based on Unity IdM resulted in technical and administrative challenges, which we describe next. Once the challenges were tackled, and after a period of running both the old and the new CLARIN IdP in parallel without any incidents, the old IdP was shut down on the 13th of July, 2017.

3.1 User Base Migration

After a successful test phase, the new CLARIN IdP was ready to be moved from the test to the production environment. To accomplish this task, users from the old CLARIN IdP had to be migrated to the new CLARIN IdP. This process could not be fully transparent and automated because of security implications and the different algorithms used to store passwords (more precisely, password hashes).

The first step in the migration process was to fill the new IdP with the accounts from the old IdP. Old accounts were exported in the ldiff[3] format and then processed and imported to Unity IdM using a REST-based API[4].
CLARIN developers then used the REST-based API of Unity IdM to send an email to every user. The email described how to activate the migrated account at the new CLARIN IdP; for this, users had to click on a provided link. Users were asked to respond to the email within a month after which the activation link was invalidated.

As anticipated, the account migration proved to be a non-trivial task. First, sending more than 1600 emails in a strictly secure environment at a commercial provider trying to actively prevent spamming required formally justifying the reasons and also required administration changes in the mail infrastructure configuration. In the end, it was possible to accomplish this task with sending 180 emails per hour. Even after this, the sending procedure had to be done in batches with a considerable delay between them. Second, the one-month time-frame turned out to be too short. Less than 25% of users clicked on the activation before it expired. The version of Unity IdM used at that time did not allow the simple recreation and resending of the activation token. This functionality had to be implemented manually This was done[5] and the implementation was used to send new tokens.

[1] http://www.unity-idm.eu
[2] The source code for all work reported in this deliverable is available at https://app.assembla.com/spaces/unity-public/git/source/ldapEndpoint?type=branch.
[3] https://tools.ietf.org/html/rfc2849
[4] The code is available at https://github.com/kosarko/unity-rest.
[5] https://app.assembla.com/spaces/unity-public/tickets/realtime_list?ticket=590

By July 2017, almost 40% of accounts had been migrated. The reasons for the relatively low percentage have not been inspected in detail yet, but there is one possible explanation: the CLARIN SPF federation has grown significantly so that many users do not need a CLARIN account because they can use their institutional login. There was also an unexpectedly high number of migration emails (11%) that could not reach the recipient, although the emails specified in the old CLARIN account had to be working at some time.

The communication with users was time-consuming. It was not reasonable, for instance, to resend migration emails to all users who have not migrated yet after the one-month period expired because they might have not have migrated on purpose. In fact, many users migrated their accounts to the new CLARIN IdP only after they were not able to login to CLARIN services with their old account. Here, the actual date when old accounts stopped working could differ for each SP because it was dependent when the SP changed the configuration to support only the new CLARIN IdP.

3.2 Service Provider Migration

Support for the new CLARIN IdP did not require any special configuration changes at the SP side because it was added to the standard CLARIN IdP feed published by CLARIN[6]. Following a user friendly approach meant offering both the old CLARIN IdP and the new IdP in parallel for some time. There were several hidden challenges, in part because SPs can choose different software to support federated logins. In fact, several bugs identified in Unity IdM and non-Shibboleth software in handling SAML (Security Assertion Markup Language) were identified, and subsequently, reported and fixed[7]. Deploying the necessary code fixes to more than twenty SPs required substantial effort.

SPs that store user-specific information (*e.g.,* repositories with user submissions or performed queries) had to deal with the new accounts in a more complex way. For instance, the Language Resource Inventory hosted at LINDAT/CLARIN[8] is based on *clarin-dspace*[9]; here, administrators had to reassign accounts for security reasons.

[6] https://infra.clarin.eu/aai/prod_md_about_clarin_erics_idp.xml
[7] E.g., https://app.assembla.com/spaces/unity-public/tickets/realtime_list?ticket=598
[8] https://lindat.mff.cuni.cz/repository/xmlui/?locale-attribute=en
[9] https://github.com/ufal/clarin-dspace

3.3 Security Audit

During consultations with other projects using Unity IdM, it was found that the Polish Grid Infrastructure[10] conducted a security assessment of Unity IdM. In this assessment, no high risk issues were identified. Moreover, all other issues found have been already fixed. As a consequence, we concluded that even with our changes there is no immediate need for another security audit.

3.4 High Availability

During the testing phase of the migration, Unity IdM was deployed in a redundant way behind a reverse proxy: here, all traffic is directed toward the proxy, which then decides which component(s) should handle the request. A high availability can be achieved by running multiple instances of the components; the reverse proxy chooses which instance to forward the request to while monitoring instances to avoid those that are not responding. Also, the reverse proxy performs load balancing by forwarding all requests fairly across the instances.
For the deployment of Unity IdM, the docker[11] platform was used. Docker supports the easy creation of new application instances - in our case of Unity IdM - across multiple data centres in a secure and almost instantaneous manner[12]. This setup has been successfully tested.

[10] http://www.plgrid.pl/en
[11] https://gitlab.com/CLARIN-ERIC/docker-unity-idm
[12] To be precise, this does not cover having the (file based) user database deployed redundantly but there exist standard paradigms that can be followed.

4 Discovery Service in SPF

The CLARIN Service Provider Federation (SPF) connects CLARIN Service Providers to most of the national federations inside the European Union. Users from institutes that are members of a national federation in countries that have joined CLARIN SPF can automatically access CLARIN's protected resources and services in a secure way. Other users can register with the CLARIN Identity Provider to access the protected resources and services once their membership is approved.

More precisely, when accessing a protected resource or service the user should be presented with a list of supported institutes, see Figure 1.[13]

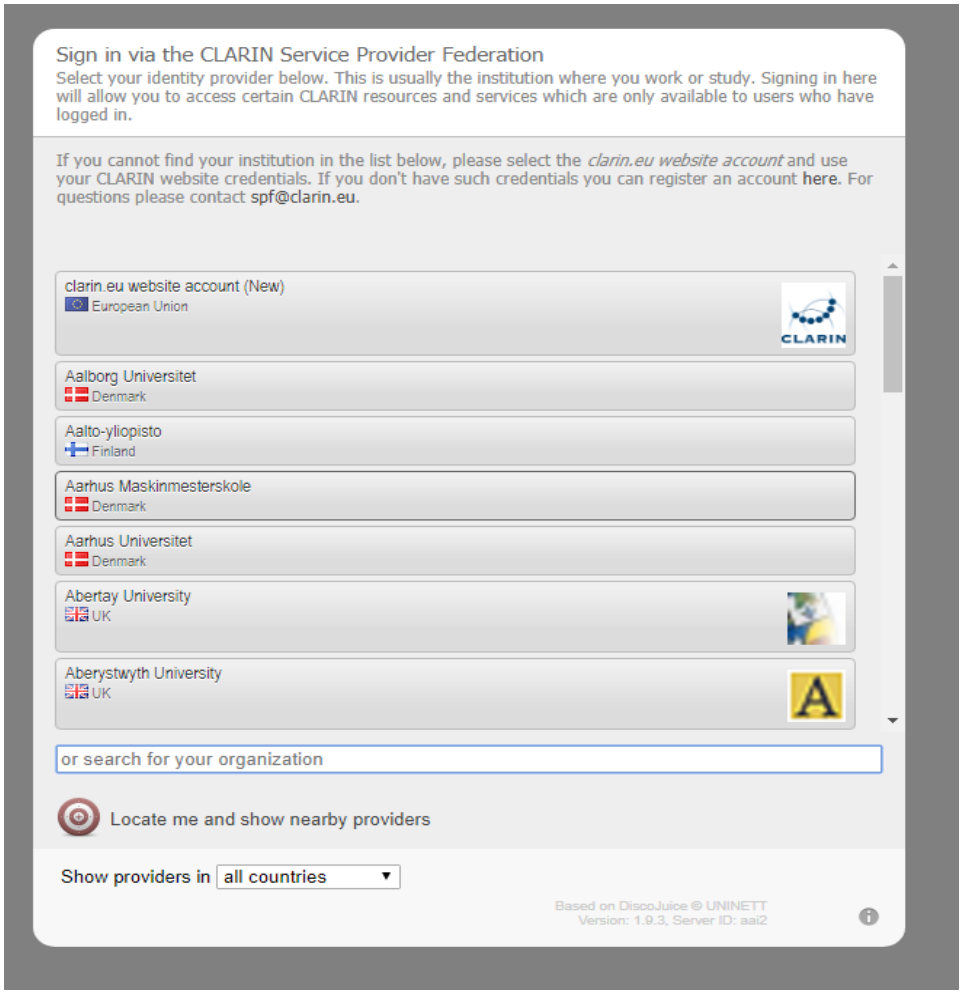

Figure 1: Presenting a list of supported institutes

The graphical visualisation should be created from a set of institutes (more precisely IdPs) that the SP trusts. The trust must be mutual and this is the main goal of the SPF federation. The implementation of such an IdP selection can be done either on the SP directly (which induces more work on the SP side) or outsourced to a central discovery service. The latter means that the SP can simply configure their federation software to use this service. Moreover, for all the services across different CLARIN centres that use the central discovery service the user experience is consistent. The selection of IdP from

[13] The service can be tested by accessing https://catalog.clarin.eu/secure/shib_test.pl.

the list is remembered by the service and is presented to users on their next access. Because the central setup introduces a single point of failure, a high availability of the service is a basic requirement.

CLARIN's software stack used behind the discovery service consists of an application periodically downloading metadata of IdPs from CLARIN SPF member federations, a web application that processes and serves the metadata as a list of IdPs and a graphical user interface DiscoJuice[14] that visualises the list as seen in Figure 1.

4.1 Compiling a List of Supported IdPs

The list of IdPs is compiled after processing metadata feeds from CLARIN SPF members[15]. A metadata feed is a web accessible xml file[16] provided by each federation. These files need to be post-processed because they can contain entries that are not relevant, *e.g.*, SPs and IdPs not registered at that particular federation. To automate the process, pyff aggregator[17] and other tools were used.[18] The process is as follows:

1. fetch metadata feeds from CLARIN SPF members;
2. filter out IdPs not registered directly via federation;
3. filter out IdPs not meeting core CLARIN requirements[19];
4. normalise the final xml by filtering out elements not needed, and adding required ones *etc.*;
5. write feeds for the list of trusted IdPs to file; write output feeds for the list of CLARIN IdP to file;
6. sign all files and verify their signatures;
7. publish all files;
8. validate all files using https://github.com/clarin-eric/SAML_metadata_QA_validator to meet the guidelines from https://www.clarin.eu/content/guidelines-saml-metadata-about-your-sp .

A similar approach is used to publish CLARIN SPF SPs; those can be harvested by IdPs with the exception that the input source for the list of SPs is hosted by a version control system with access to CLARIN members only. The website https://centres.clarin.eu/spf lists the current registration status of each SP with each identity federation.

An internal evaluation of the alternatives to pyff tool has been conducted. Because the solution CLARIN uses is working, an alternative would have to be a perfect match that is widely used. No tool has been found that would not require additional nontrivial steps.

---

[14] http://discojuice.org/

[15] https://www.clarin.eu/content/service-provider-federation .

[16] See file "about_identity_federations_md.xrd" in https://github.com/clarin-eric/pyFF_config/ for the complete list of processed federation feeds (in <Subject> element)

[17] https://github.com/leifj/pyFF .

[18] These are publicly available at https://github.com/clarin-eric/pyFF_config.

[19] No official requirements are published but one of the basic requirements is that each user can be identified to a real person. Therefore, IdPs that allow for *anonymous* or not officially verified users are filtered out *e.g.,* those that allow self-registration or are used for testing purposes.

In the current setup, offering custom lists of IdPs means only to change the xml data source and publish it on a different URL. If the need arises, it will be technically easy. However, at the moment there is no other inter-federation that meets basic requirements as mentioned in the previous sections (*e.g.,* traceability to a real person) either actively or passively.

4.2 High Availability

The discovery service is running in a high availability setup. This means that there are two instances of the discovery web application running (https://github.com/clarin-eric/aai-discovery) in a data centre that is different to the one running the reverse proxy that is the public endpoint (https://discovery.clarin.eu/). The reverse proxy uses fair load balancing (*round robin*) to distribute the load to these two instances. Both periodically update the list from https://infra.clarin.eu/aai/prod_md_about_spf_idps.xml [20] and perform additional filtering to hide IdPs that have a specific category set - https://refeds.org/category/hide-from-discovery [21]. The instances are stateless and are automatically kept up-to-date. The spinning of other instances is therefore trivial.

4.3 Alternative Implementations

In order to evaluate alternative implementations, a list of options was compiled, including Switch WAYF[22], Shibboleth Discovery Service[23], CESNET WAYF[24] and Account chooser[25]. The following requirements for the CLARIN central discovery service were identified:

1) handle thousands of IdPs,
2) have minimal dependencies and
3) have active user community.

Switch WAYF and CESNET WAYF seem to meet most requirements but they have not offered any obvious advantage compared to the existing stack.

We have optimised the IdP list in two ways: redundant information such as embedded logos were removed, and compression was enabled. As a result, the list of more than 1500 elements is now served using less than 60 KB[26]. With these results and with the evaluation showing that the alternatives have no clear advantages over the current solution it was decided to stay with the current solution.

[20] https://www.clarin.eu/content/service-provider-federation
[21] https://wiki.refeds.org/display/ENT/Hide+From+Discovery
[22] https://www.switch.ch/aai/guides/discovery/embedded-wayf/
[23] https://wiki.shibboleth.net/confluence/display/SHIB2/DiscoveryService
[24] https://github.com/CESNET/wayf
[25] https://www.accountchooser.com/learnmore.html
[26] 60 KB is considered to be almost nothing in today's network speeds.

4.4 Passive Ensuring Level of Trust

Work in this WP made it possible to formally describe the process of ensuring a basic *level of trust*. Please note that a strict definition of the term *level of trust* in this context is not possible as its meaning varies across national federations and individual institutions.

To ensure a level of trust, the following process is applied with respect to the CLARIN infrastructure[27]:

1. suspicious IdPs are identified – semi-automated for SPs that are members of the Attribute Aggregator as described in [CLARINPLUS-D2.2] because if a new IdP is used to access the CLARIN infrastructure, an email including links to more information is sent to several recipients (*i.e.,* SPF administrators, several CLARIN AAI task force members, Attribute Aggregator administrators);
2. an issue is created in the CLARIN issue tracker[28] with the *AAI IdP Blacklist* tag; CLARIN central office takes care of such issues but also the CLARIN AAI task force is informed where multiple people can verify whether any requirement has not been met;
3. the CLARIN central office closes the ticket and if there is a violation a new filter is added to step 3 (see Section 4.1, *Compiling list of supported IdPs).*

[27] This process is also described at https://trac.clarin.eu/wiki/ServiceProviderFederation/IdpBlacklist (access to CLARIN member s only).

[28] https://trac.clarin.eu

5 Conclusion

Given the complexity of both the technical and the administrative aspects, the re-engineering of the CLARIN Identity Provider and the user migration can eb considered been successful. It should be noted that the use of existing CLARIN workflows and formal processes helped the migration process. One non-obvious outcome is that inactive[29] users are no longer present in the new CLARIN IdP.

There have been updates to the parts of code of Unity IdM that have been developed in this part of WP2[30]. Those code pieces need to be reviewed and merged with the Unity IdM code base. After this, the code will be maintained directly by the Unity developers ensuring sustainability.

Tightly connected to the identity management solution is the ability for users to select from a list of supported institutional logins. This has been offered as a central service in CLARIN from the very beginning. It is worth noting that this can simplify the implementation of a federated login for CLARIN centres. Because this service represents a single point of failure, it must be set up in a reliable and high availability mode.
The discovery service in CLARIN has been implemented as a central service with high availability simplifying the work for member SPs. Its main goal is to enable CLARIN SPF discovery and it has been serving this purpose reliably.

[29] And some less active users too but they can create new accounts.
[30] The current version is available at https://github.com/clarin-eric/unity-ldap

References

[CLARINPLUS-D2.2] Mišutka, J. 2016. *Robust SPF 1: workflow and monitoring.* https://office.clarin.eu/v/CE-2016-0809-CLARINPLUS-D2_2.pdf