AAI: systematically addressing the attribute release problem

Jozef Mišutka
Institute of Formal and Applied Linguistics
Charles University in Prague
misutka@ufal.mff.cuni.cz

Ondřej Košarko
Institute of Formal and Applied Linguistics
Charles University in Prague
kosarko@ufal.mff.cuni.cz

Amir Kamran
Institute of Formal and Applied Linguistics
Charles University in Prague
kamran@ufal.mff.cuni.cz

Abstract

CLARIN Service Provider Federation is not the only inter-federation but it is the first one to systematically address the attributes release issue; when Identity Providers do not release mandatory attributes to a service. For instance, if a data set is licensed under a restrictive license the user must be uniquely identifiable over time. However, if the Identity Provider does not release such information, the service cannot let the user download the data.

1 Introduction

CLARIN is the Common Language Resources and Technology Infrastructure project and provides easy and sustainable access for scholars in the Humanities and Social Sciences to digital language data. One of the pillars that make this possible is the CLARIN Service Provider Federation (SPF) that connects CLARIN Service Providers to the majority of national federations inside the European Union.

There are many different services based on many different platforms in the CLARIN SPF. These services require specific information about authenticated users in order to operate. For instance, if a data set is licensed under a restrictive license the user must be uniquely identifiable over time or the data set might be offered only to users with a specific entitlement. If the Identity Provider does not release such information, the service cannot fully operate.

To address this issue on a project level, in a uniform way, it is required to know the attributes that have not been released properly. Furthermore, it is often that application developers neither have access to the machines where the federation enabling software resides, nor they have the proper knowledge to handle such situations. The CLARIN Attribute Name Aggregator framework has been developed to collect information about the authentication attempts to participating Service Providers. Most importantly, the framework aggregates **names of attributes** that Identity Providers release to Service Providers. Together with the metadata about every Service Provider and every Identity Provider, it is possible to decide if the attribute release has been successful or not. The framework rely on the fact that Service Providers in CLARIN SPF have to meet strict requirements about their metadata and policies enforced by the CLARIN assessment process (https://www.clarin.eu/content/assessment-procedure) and by the requirement to implement the Data Protection Code of Conduct (http://hdl.handle.net/11346/GAIU).

The aggregator provides a user interface that simplifies the notification of affected entities in a unified way and tracks the progress of on-going issues. From the underlying data, statistics can be compiled about every Service Provider and Identity Providers from federations and the report can be used to improve problematic federations.

For users of a SP, the CLARIN Attribute Name Aggregator can be used to transparently improve the quality of academic federating. For SP admins, the aggregator offers a "one-click send email" feature that sends a description of an authentication problem together with more detailed information to emails parsed from the Identity Provider and Service Provider metadata. This rather simple feature uses information from multiple different sources and makes managing and improving the quality of a feed with more than thousand Identity Providers feasible. Federation operators can use the statistics to understand how their Identity Providers are configured in the context of attribute release.

The Attribute Name Aggregator consists of two components. The first one must be installed on the Service Provider itself and the second one is the central component where the statistics are collected and are displayed.

The software for the Service Provider side of the aggregator project is publicly available at https://github.com/ufal/clarin-sp-aaggregator. The Readme.md file contains a detailed description of how to deploy it to both the Shibboleth federation solution or as a web application dependency. The central aggregator application is publicly available at https://github.com/ufal/lindat-aai-attribute-aggregator.

2 Attribute Name Aggregator on the Service Provider

There are several software solutions that allow for federating. The most used one is Shibboleth[1]. We developed a script that can be integrated with Shibboleth by using the *sessionHook*[2] feature. The flow of the authentication is redirected to the sessionHook script before finally getting to the application operated by the service provider. This effectively means that we can read the released **attribute names** in this script and inform the central service in the background. The workflow is shown in Figure 1.

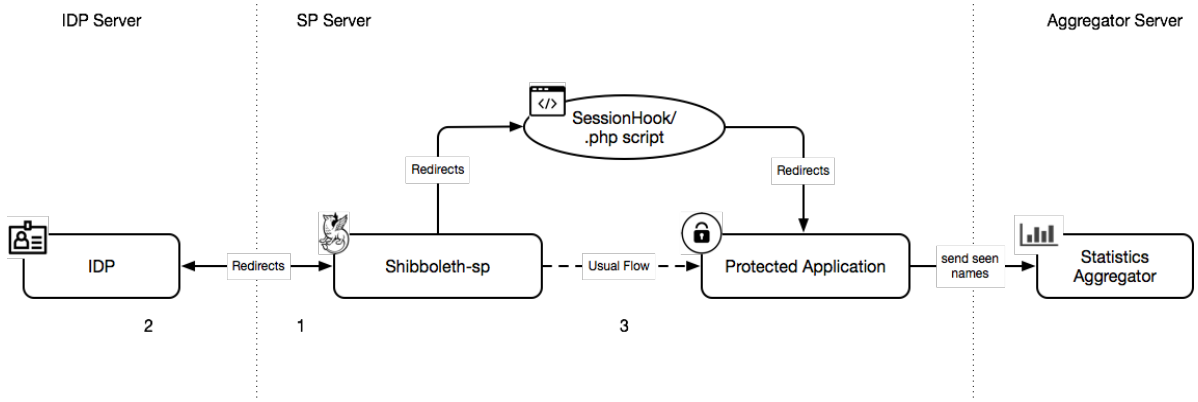


Figure 1. Authentication flow when using the sessionHook functionality

The sessionHook functionality can be used to intercept the authentication flow and execute custom code that inspects the session using exposed Shibboleth handlers. We require that the assertion export handler defined via *exportLocation* is available but only locally; otherwise, it would introduce security risks.

The attribute name aggregator script is written in php and is executed on the server side. The script receives two parameters: target and return. The return parameter contains the location that would be used if no sessionHook were present. The target is the resource destination URL from the authentication point of view.

In order to not break the authentication execution flow, the first action of the script is to redirect the browser to the URL specified in the return parameter. This is done by sending a HTTP Location header. The script then reads the *Shib-Assertion-Count* environment variable that contains the number of exported assertions and reads environment variables from *Shib-Assertion-NN* (for all NN between 01 and Shib-Assertion-Count). The values contain URLs from which we can get the assertions. Only

[1] https://shibboleth.net/
[2] http://hdl.handle.net/11346/SN4A

the name of the released attribute is stored from each assertion. Finally, an external program is executed to send the attribute names to the central service. The script does not wait for the external program to finish, minimising the performance overhead. The final step is to inform the central service using the provided REST API. Technically it ends up with a HTTPS GET request similar to

```
GET /aaggreg/v1/got?idp=https://idp.clarin.eu
&sp=https://ufal-point.mff.cuni.cz/shibboleth/eduid/sp
&timestamp=2016-06-14T11:32:21.165Z
&attributes[]=urn%3Aoid%3A2.5.4.10
&attributes[]=urn%3Aoid%3A1.3.6.1.4.1.5923.1.1.1.9
&attributes[]=urn%3Aoid%3A0.9.2342.19200300.100.1.3
&attributes[]=urn%3Aoid%3A2.16.840.1.113730.3.1.241
&attributes[]=urn%3Aoid%3A1.3.6.1.4.1.5923.1.1.1.6
&attributes[]=urn%3Aoid%3A1.3.6.1.4.1.5923.1.1.1.7
&attributes[]=urn%3Aoid%3A2.5.4.3
```

The request can be interpreted that a user tried to authenticate to LINDAT/CLARIN service provider using the CLARIN Identity Provider and that several attributes have been released including e.g., urn:oid:0.9.2342.19200300.100.1.3 which is the identifier for mail. Please note, that no private information is being processed only names of the released attributes.

In case the direct integration of the aggregator is not possible, the fallback plan is to include a JavaScript script in the web application itself that will send the relevant information to the central service by parsing the default Shibboleth handler. The workflow of this solution is depicted in Figure 2.

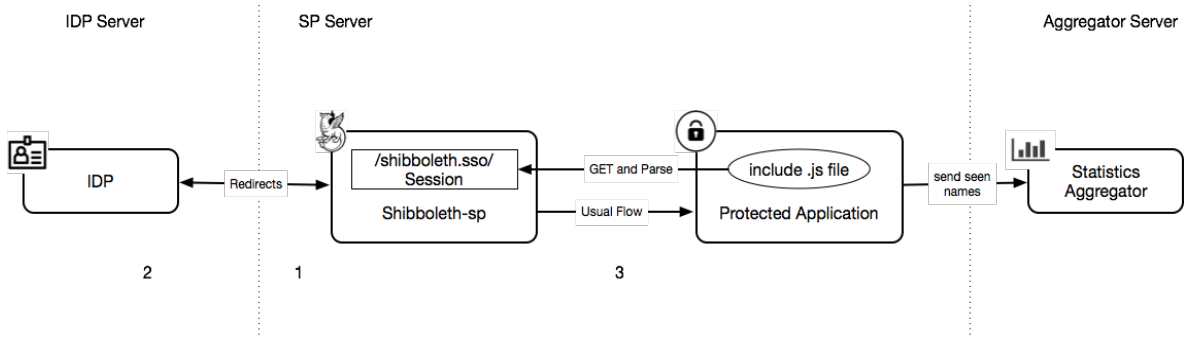


Figure 2. JavaScript fallback solution workflow

3 Attribute Name Aggregator as a central service

When the user is in the authenticating process and the aggregator script on the Service Provider side is executed, it sends the information to a central server. The central server is a Node.js application offering REST API to both collect and expose the attribute name aggregated information.

All authentication attempts are stored in Apache Solr[3] search platform and are indexed to enable searching. However, in order to build a usable service we need additional information about the entities. Therefore, a task is being executed regularly that downloads and processes the metadata of the following Identity Providers and Service providers:

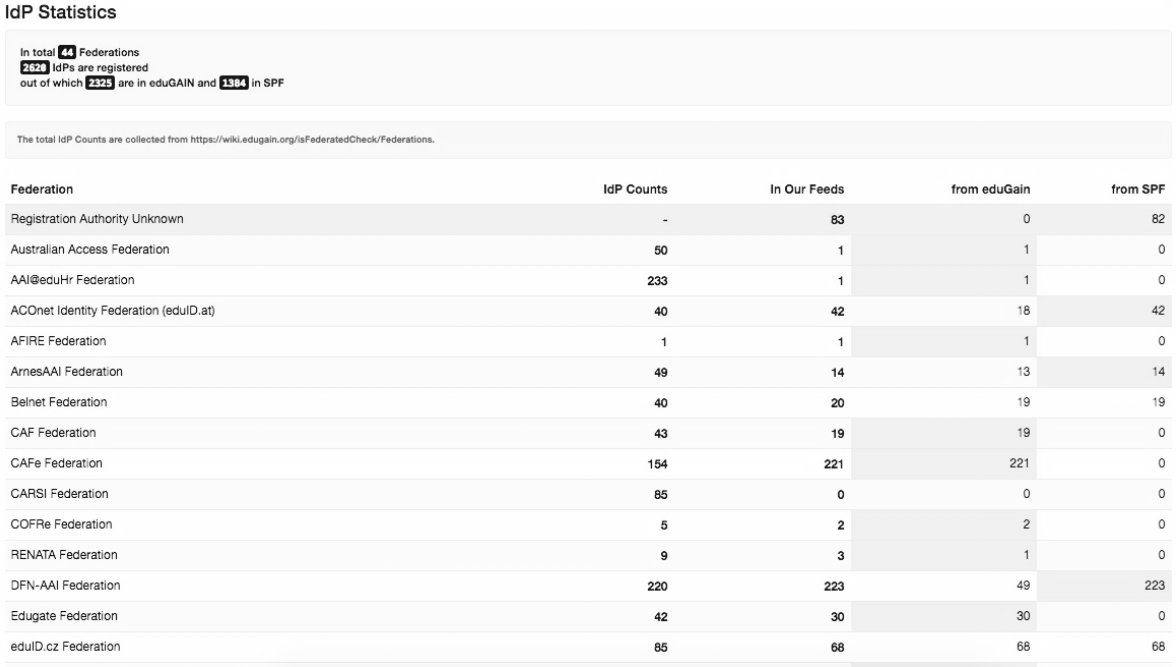
- Service Providers from CLARIN SPF;
- Identity Providers from CLARIN SPF;
- Identity Providers from eduGAIN inter-federation;
- CLARIN Identity Provider.

The metadata are parsed and important information is indexed by Apache Solr. With this information available, we can decide and display if the authentication attempt has met all the requirements or not. Required attributes by the Service Provider together with the actual released

[3] http://lucene.apache.org/solr/

attributes and important entity categories (Research and Scholarship[4], Data Protection Code of Conduct) of both the Service Provider and the Identity Provider that define behaviour when it comes to attribute release are displayed.

Summary and statistics of the attempted logins are also available as illustrated in Figure 3a and 3b.



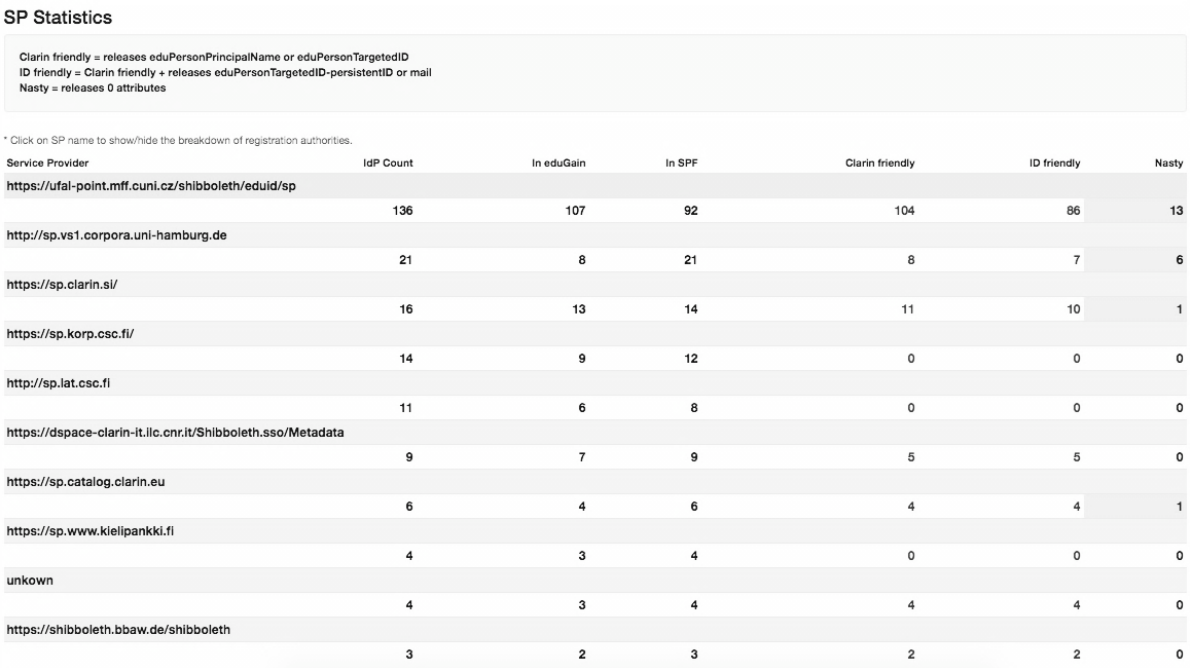
IdP Statistics

In total 44 Federations
2621 IdPs are registered
out of which 2325 are in eduGAIN and 1384 in SPF

The total IdP Counts are collected from https://wiki.edugain.org/isFederatedCheck/Federations.

Federation	IdP Counts	In Our Feeds	from eduGain	from SPF
Registration Authority Unknown	-	83	0	82
Australian Access Federation	50	1	1	0
AAI@eduHr Federation	233	1	1	0
ACOnet Identity Federation (eduID.at)	40	42	18	42
AFIRE Federation	1	1	1	0
ArnesAAI Federation	49	14	13	14
Belnet Federation	40	20	19	19
CAF Federation	43	19	19	0
CAFe Federation	154	221	221	0
CARSI Federation	85	0	0	0
COFRe Federation	5	2	2	0
RENATA Federation	9	3	1	0
DFN-AAI Federation	220	223	49	223
Edugate Federation	42	30	30	0
eduID.cz Federation	85	68	68	68

Figure 3a. Attribute aggregator IdP summary report



SP Statistics

Clarin friendly = releases eduPersonPrincipalName or eduPersonTargetedID
ID friendly = Clarin friendly + releases eduPersonTargetedID-persistentID or mail
Nasty = releases 0 attributes

* Click on SP name to show/hide the breakdown of registration authorities.

Service Provider	IdP Count	In eduGain	In SPF	Clarin friendly	ID friendly	Nasty
https://ufal-point.mff.cuni.cz/shibboleth/eduid/sp						
	136	107	92	104	86	13
http://sp.vs1.corpora.uni-hamburg.de						
	21	8	21	8	7	6
https://sp.clarin.si/						
	16	13	14	11	10	1
https://sp.korp.csc.fi/						
	14	9	12	0	0	0
http://sp.lat.csc.fi						
	11	6	8	0	0	0
https://dspace-clarin-it.ilc.cnr.it/Shibboleth.sso/Metadata						
	9	7	9	5	5	0
https://sp.catalog.clarin.eu						
	6	4	6	4	4	1
https://sp.www.kielipankki.fi						
	4	3	4	0	0	0
unkown						
	4	3	4	4	4	0
https://shibboleth.bbaw.de/shibboleth						
	3	2	3	2	2	0

Figure 3b. Attribute aggregator SP summary report

4 Conclusion

The CLARIN Attribute Name Aggregator addresses specific problems of federated access on an inter-federation level for all member Service Providers in real time. Together with the statistics, it can be used to not only ensure the quality by monitoring but also to improve the quality and usability of federating – in our case the CLARIN SPF.

[4] https://refeds.org/category/research-and-scholarship