FEDERATED IDENTITY MANAGEMENT FOR LIBRARIES (FIM4L): DRAFT GUIDELINES & RECOMMENDATIONS

Access to online library resources can be quite complex. Patrons normally have easy access when signed on to a campus network but when working from other locations — as modern work patterns often demand — the same patrons are increasingly asked to 'log in to their institution'. This process can release identifying information.

Known as federated authentication, delivering Single Sign On (SSO), this process, if not configured correctly, is at odds with the responsibility of libraries to protect their patrons' privacy.

In order to preserve patron privacy, while also making the configuration and management of federated SSO connections easier for both libraries and publishers, LIBER's FIM4L Working Group has drafted 10 Implementation Principles for SSO. The principles drafted by the group are now open for public comment until 30 April 2020.

Your comments will influence a final set of recommendations which libraries can use to give patrons seamless access while preserving privacy as much as possible. If you prefer to provide feedback online, please comment on our online discussion. You can also share feedback by emailing liber@kb.nl.

Introduction

Publishers and suppliers of licensed online resources want to provide authorized users of institutions for higher education and research with access to their services in a controlled way. The commonly used access method based on IP address has limits when users want access from anywhere and any device at any time. With the new solutions, based on federated authentication and Single Sign-On (SSO), it depends on how you configure the connection whether and which parties can identify individual users. As always, libraries want to protect the privacy of their patrons, and give them control over that privacy.

In order to make configuration and management of federated authentication easier for both libraries as well as publishers, a number of scholarly libraries from around the world have agreed upon the following guidelines to control access to services based on licensed content.

To understand the rest of the text, the following terms are important:

- Publishers are Service Providers (SP)
- Institutions/libraries are Identity Providers (IdP)

SSO Implementation Principles

Principle 1

The configuration and solution has to be in line with data protection regulations, in particular the General Data Protection Regulation (EU GDPR).[1]

Principle 2

For access to services based on licensed content, next to the option of access based on IP addresses, it is recommended to use the SAML 2.0 protocol (or its follow-up technology OIDC/OAuth2 if the involved IdPs are able to handle it) to connect and control access.

Principle 3

eduGAIN has been established as a proper means to interfederate between identity federations, and thus enables service providers to greatly expand their user base. Thus scholarly libraries should prefer publishers who are connected to eduGAIN. Libraries should encourage publishers to make use of eduGAIN.

Principle 4

The following lists the recommended options for authentication attributes, ordered by degree of privacy control, with a. being better privacy preserving than b. and so on:

A - The publisher only requires a transient identifier - "privacy star". During a session the user is identified by a transient identifier (NameID) containing a unique string (for example: bd09168cf0c2e675b2def0ade6f50b7d4bb4aae) for this Service Provider (SP). If the user logs in again, a new transient identifier will be generated. This allows for maximum privacy. However, it doesn't allow the publisher to recognize a returning customer, which makes it impossible for instance to know what resource is downloaded by the same user. A profile page for the user thus also doesn't make sense with this option. It also doesn't allow the library to translate the transient ID to a patron in case of misconduct (e.g. excessive downloads). (In exceptional cases it could be done however by some IdPs and federations by thoroughly investigation of log files).

B - The publisher requires a persistent but targeted identifier - "personalisation and subject tracking possible". A persistent identifier (ID) contains a unique string, like the transient one, identifying the user for a specific SP, but persisting over multiple sessions: on every authentication, for the same user the same ID is used. This is an option for services that have a need to recognize returning customers, for instance so it can present you your files, your orders etc. In SAML the Pairwise Subject Identifier is preferred over eduPersonTargetedID (deprecated) and SAML 2.0 persistent NameID.[2] When opting for a persistent ID, consider the following:

- A persistent ID allows the library (not the publisher) to translate the ID to a patron in case of misconduct.
- It is possible to lock down access for a particular user in case of misconduct.
- A persistent ID (like the Pairwise Subject Identifier, pairwise-id) is sufficient for the SP to provide personalization features. Sometimes an SP requests more information, like

[1] Regulation (EU) 2016/679 (General Data Protection Regulation) in the current version of the OJ L 119, 04.05.2016; cor. OJ L 127, 23.5.2018, https://gdpr-info.eu
[2] This is in line with this argumentation: https://wiki.univie.ac.at/display/federation/eduPersonTargetedID

a name and email address. Adding personal information like Name and Email to enrich the user profile should be optional (not mandatory) for the user. Libraries/ institutions are advised not to transfer that information during authentication, but have the SP offer the user a profile page in their service, where users provide consent and can voluntarily provide name, email or other information. Minimize the attribute set provided to the service during the authentication-flow.

- Before a service that receives a persistent identifier creates a profile for the user, the user should be asked for permission to store and process personal data, for instance via a button "personalize account" or at least be informed by a message on data privacy.[3] In no way should the permission request be mandatory or seemingly mandatory for the user; the user must be free whether or not to have a personal profile.

C - In addition to 4A or 4B the SP can require extra ('non-identifiable') information. If more information is needed to allow for billing, access control etc. identity providers can supply one or more of the following attributes (from most to least preferred):

- eduPersonEntitlement, with the specific value urn:mace:dir:entitlement:common-lib-terms
- eduPersonScopedAffiliation
- eduPersonEntitlement, with other values, representing group or role memberships in alignment with AARC Guidelines on expressing group membership and role information
- Usage of schacLocalReportingCode attribute is recommended for statistics purposes once it is well defined.[4]

Principle 5

SP software should be able to handle more attributes, but not require more attributes. Some publishers state "I need an email address, as my software can't function without it". Publishers with (older) systems that require more attributes for authentication to function should adapt their systems ASAP. Libraries are recommended to stop or don't start using services that require more personally identifiable information (PII) than a transient or persistent ID during authentication.

[3] E.g., "By connecting to this service, I agree that the service provider stores my person related data (ID, affiliation, entitlements sent by my IdP, my IP address sent by my client, and my actions on this platform). Only if I want to receive emails from the service or if I want to be addressed by my name, I will add my email address and name respectively, but this is not needed for any other personalisation features like 'point me to the last document and its last page I read', 'my last searches', <include your personalisation feature here>, etc. Whenever I wish to do so, I may request to see and to have deleted all data stored about me."

[4] Please note that this attribute is not available in many federations and IdP's, so if the SP would like to receive that attribute, it will take specific communication between SP and IdP and possible the federation.

Principle 6

Apart from generally working according to the GDPR, when requesting information from users, for instance in a profile page, publishers have to adhere to the most recent EU "Guidelines on Consent"[5] to make sure that free consent is given in compliance with the GDPR.

Principle 7

When providing PII to a SP, whether based on consent[6] or not, a respective data processing agreement (DPA) may be needed.

Principle 8

Publishers are encouraged to declare compliance with the GÉANT Data Protection Code of Conduct.

Principle 9

Publishers are encouraged to declare compliance with the assertions of the REFEDS Sirtfi framework (Research and Education FEDerations group, Security Incident Response Trust Framework for Federated Identity).

Principle 10

Publishers are encouraged to follow the guidelines from the SeamlessAccess.org coalition (formerly known as RA21).

Risks & Concerns

The above recommendations do impact some risks, that we want to make explicit in this section:

- Deanonymization: If you provide a targeted ID, as recommended in Principle 4, Part B above, you have to be aware that other data, already collected by the SP, could be linked to this ID.

- Apart from the fact that for GDPR pseudonym IDs (and even IP-addresses) are PII, normally users would see a consent or information screen when accessing an SP for the first time and would see what attribute release policy the IdP has opted for. There might be cases where everybody is fine with releasing certain PII. But if possible, give users a choice, for instance by not releasing information during authentication, but by offering a profile page within a service, where an individual can voluntarily share more information.

[5] Guidelines on Consent under Regulation 2016/679, https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=623051
[6] We know of and are tracking the internet2 CAR-initiative about consent for optional release of attributes.

Terms & Definitions

AARC - Authentication and Authorization for Research Collaborations, Project funded by the European Union's Horizon 2020 research and innovation programme under Grant Agreements 653965 and 730941. AARC was successful in establishing a Blue Print Architecture for the deployment of FIM technologies in research infrastructures, as well as in establishing guidelines on respective technical and policy matters .

Authentication - The process of verifying the the identity of a user, process or deviceability of a user to access an account, often, but by no means exclusively, use of a username and password.

Authorization - The process of verifying against a set of access controls whether an account is authorized to access a given service or resource.

eduGAIN - eduGAIN enables trustworthy exchange of identity information between federations without many bilateral agreements, reduces the costs of developing and operating services, improves the security and end-user experience of services, enables service providers to greatly expand their user base and enables identity providers to increase the number of services available to their users. Speaking about costs of operating services when a resource provider is updating its metadata it easy to send it to just one federation and then propagate it to eduGAIN instead of having to contact many national federations separately. On the federation side getting updated metadata from eduGAIN has no maintenance costs is undoubtedly an advantage. See AARC and eduGAIN: expanding access to online resources for students, teachers and researchers, How to reach global customers with Federated Identity Management and How to Join eduGAIN as Service Provider for more details.

Federated Authentication - The mechanism by which an identity provider, such as a home organization, indicates to one of more service providers that the user has been authenticated and may be authorized by the service provider to access relevant resources.

Federated Identity - A digital identity which is asserted by one system (an identity provider) which may be consumed by other systems (service providers) by means of federated authentication.

Federation - A federation is an association of organizations that agree to exchange information as appropriate about their users and resources in order to enable collaborations and transactions such as user authentication.

Identity Provider (IdP) - An organization that manages digital identities and issues authentication assertions and potentially other attributes to Service Providers.

Identity Provider (IdP) Persistence - The storage and re-use of a previous IdP choice made during an identity provider discovery process.

IP address-based Authorization - A method where a SP and a home organization have agreed that every request coming from a range of network/Internet Protocol (IP) addresses associated with the home organization should be authorized for the services provided by the SP.

Multilateral Identity Federation - A form of identity federation where a trusted third party registers and publishes all entity metadata to all members, preventing the need for bilateral agreements between an IdP and SP.

REFEDS R&S - The REFEDS Research and Scholarship Entity Category (R&S) has been designed as a simple and scalable way for Identity Providers to release minimal amounts of required personal data to Service Providers serving the Research and Scholarship Community. Candidates for the Research and Scholarship (R&S) category are Service Providers that are operated for the purpose of supporting research and scholarship interaction, collaboration or management, at least in part. Example Service Providers may include collaborative tools and services such as wikis, blogs, project and grant management tools that require some personal information about users to work effectively. This entity category should not be used for access to licensed on-line resources as described in the category definition. For more details see REFEDS documentation.

Service Provider (SP) - An organization that makes online resources available to users based in part on information, in particular authentication assertions, from IdPs.

Single Sign On (SSO) - The ability of a user to access multiple discrete systems or sets of resources with a single set of access credentials. This is often achieved by the mechanism of Federated Authentication.

Web Storage[7] **-** Where web applications can store data locally within the user's browser. Before HTML5, application data had to be stored in cookies, included in every server request. Use of browser local storage prevents sending the data to server with each server calls (which is what cookies do).

IP address-based Authorization - A method where an SP and a home organization have agreed that every request coming from a range of network/Internet Protocol (IP) addresses associated with the home organization should be authorized for the for services provided by the SP.

Security Assertion Markup Language (SAML)[8] **-** A standards-based approach to federated or single sign-on (SSO) authentication. Many interoperable open source and commercial implementations of SAML are available.

Background

- Understanding Federated identity, RA21 and other authentication methods
- Stanford Libraries Statement on Patron Privacy and Database Access
- Protecting Patron Privacy in Digital Resources
- Video with considerations related to attribute release
- JISC blog about federated authentication and privacy (GDPR)
- Discussion on Introduction of an Entity Category for Library Services *(see the comments for the latest thinking/argument)*

[7] Web Storage (Second Edition). World Wide Web Consortium. 19 April 2016. https://www.w3.org/TR/2016/REC-webstorage-20160419
[8] https://en.wikipedia.org/wiki/Security_Assertion_Markup_Language