Consensus - Identifiers

- Identifier Requirements
- Proposals
  - Define a Pair of Identifier Attributes
  - Scoping
  - Attribute or NameID
  - Implications for Entity Categories
- Next Steps
  - Implications for SAML Persistent NameID and eduPersonTargetedID
  - OASIS Spec for Attributes and Signaling of Capabilities and Requirements
  - Response to Questions about Pursuing the OASIS Route Rather than eduPersonUniqueId

The working group has studied and weighed a number of problems with current approaches to user identification in R&E SAML federations. Recently discovered security concerns, long-standing deployment challenges, and emerging approaches in alternate technologies such as OIDC have been studied, together with the other requirements communicated to the group by its participants, representing various communities.

The group's emerging consensus is to propose a pair of SAML Attributes to be used for long-term identification of subjects. These may be entirely new Attributes, or may (in part) reuse existing work, depending on decisions yet to be made.

The following presents a summary of the requirements gathering and motivations for this proposal and the next steps needed.

Identifier Requirements

The requirements for these attributes have been shaped by studying a number of well-understood properties of identifiers and discussing them in the context of the use cases for which the working group believes its output can best be applied. This also serves as a natural scoping exercise; applications for which the requirements cannot hold are necessarily not in scope for the profile. As will be seen below, this rules out (as one example) applications that can support only a single identifier that must be a recognizable e-mail address.

To that end:

1. Persistent: **YES**
   - This is non-controversial; identifiers must persist for a reasonable period of time to be useful to most applications; for most use cases, the longer an identifier persists, the better.
2. Non-reassignable: **YES**
   - The reassignability of both "eduPersonPrincipalName" and "mail" have been a source of long-standing frustration and continue to lead to complex workarounds and out-of-band agreements to maintain their usability for many applications. Any new identifier clearly needs to fully and unequivocally disallow reassignment in all cases.
3. Human-friendly: **NO**
   - The overlapping of identification with search and selection has been a major source of frustration, and the group believes the other requirements cannot be reasonably met by continuing to allow that overlap. We explicitly believe it should not be necessary for identifiers to be displayable or known by users themselves. By extension, it is not necessary that identifiers be derived from a person's name.
4. Size-Bounded: **YES**
   - Applications frequently require, and should be able to expect, reasonable constraints on the maximum size of an identifier.
5. Simple Syntax: **YES**
   - Applications must be able to directly consume and store a simple string value, or at most a pair of values. More complex encodings are barriers to use.
6. Case-Sensitive: **NO**
   - The realization that many if not most applications do not support case-sensitive identifier comparison leads to the conclusion that we need to ensure any identifiers are safe to use in spite of that fact. The cost of an inaccurate match is potentially serious.
7. Portable: **NO**
   - While there are use cases for globally-managed identifiers that follow users as they move between organizations, the working group feels this use case is, for the time being, better left as an account linking exercise by services and a subject of future study.
8. Correlatable: **YES**
   - A significant proportion of important applications require an identifier that can be used to correlate activity with other applications. Thus, it must be possible for identifiers to be omni-directional, having the same value in all cases.
9. Targeted / Non-correlatable: **YES**
   - Another significant proportion of important applications, and their users, believe it is important to prevent the correlation of activity between different applications. Thus, it must be possible for identifiers to be uni-directional, having a different value for some sets of different use cases.

Clearly requirements 8 and 9 are mutually exclusive. It follows that either two identifier types are required, or a single identifier must implicitly vary in its semantics along that axis. OpenID Connect apparently chose the latter approach, allowing its 'sub' claim to potentially behave in either fashion, depending on the manner in which the OP operates and the options used by the client when it registers.

The working group's consensus is that SAML is better served with two explicitly separate attributes. This separation allows applications to deliberately support one or both attributes, and to signal its requirements using existing methods, avoiding the need for new work in this area or on out of band signaling.

Proposals

Define a Pair of Identifier Attributes

We believe two attributes are required, notionally referred to here as "uniqueID" and "directedID" (the names are merely placeholders for discussion). Both attributes would carry stable/persistent, non-reassignable, case-insensitive values, unique within the scope of the issuing/controlling organization. The former would be a correlatable, globally uniform value and the latter would be a value that may (and should) vary across different services in a manner appropriate to deployment of the SAML protocol.

Thus:

1. subject-id
    a. **public/correlatable**
    b. non-reassignable
    c. persistent
2. pairwise-id
    a. **targeted/non-correlatable**
    b. non-reassignable
    c. persistent

Consistent with the earlier requirements summary, neither type of identifier is expected to be human-friendly, suitable for display, or appropriate for searching or user selection. Applications requiring such features are expected to rely on the "mail" attribute for that purpose.

⚠ **Persistent Identifier**

A persistent identifier is **not** necessarily a permanent identifier.

Scoping

Short of requiring cryptographically random values, any identifier scheme must contain some notion of scope or namepace to prevent unintentional collisions. Most federation-safe applications already apply some notion of namespace separation in order to support commonly used identifiers like student or employee numbers, and some organizations may well find those values suitable as a "uniqueID", depending on their IDM practices. It has also been observed that applications without this notion of scope tend to be vulnerable to identifier collision/spoofing attacks. Thus, it behooves us to support collision-avoidance in whatever manner seems best.

The working group has yet to reach a consensus on the best approach to this problem.

Today there are two common scoping schemes:

- domain-based scoping (e.g., "mail", "eduPersonPrincipalName", "eduPersonUniqueId")
- explicit or implicit scoping by issuer/context (e.g., "eduPersonTargetedID", SAML Persistent NameIDs, OIDC 'sub' claim)

Anecdotally, the domain syntax has proven easier for developers to understand and apply, and is flexible, but at a cost of a tendency by some to treat any domain-scoped value as an email address. It also requires a policy layer to map protocol-specific issuers to scopes, and gaps in implementing or applying that policy layer can lead to risk. In addition, the decision to internationalize domain names has rendered processing them far more complex in theory than in actual practice, creating opportunities to exploit implementations that we likely have not begun to understand.

On the other hand, both SAML and OIDC rely on URIs to identifier "issuers", and storing/managing identifier/issuer pairs is less well-understood by developers and, when only a single field is available for storage, leads to unexpected syntax such as "identifier@https://idp.example.org" that look odd and sometimes break software. It also muddies the issue of case-sensitivity, because in both SAML and OIDC, the issuer name is explicitly case-sensitive.

On the subject of relying party scoping, SAML made this necessary by introducing the notion of affiliations of services that can collectively receive a common identifier. Thus, the receiver of an identifier isn't always the "target scope" for a SAML "persistent" NameID. If we dispense with this notion, then the need for explicit identification of the receiver as a qualifier goes away, which is also consistent with OIDC, and avoids the challenges associated with handling identifiers made up of "triples" of data.

Attribute or NameID

The question of how the identifier data should be communicated in a SAML Assertion breaks down to essentially one of two choices, either as an Attribute or the NameID in the Subject. The use of the NameID element, while popular with one-off SAML deployments seeking to minimize generality, tends to be used in a sloppy, often incorrect fashion that breaks the uniformity of relying on Attributes for other information. Also, the NameID is more difficult to configure in at least one of the popular software solutions used in the community (Shibboleth). It also forks the signaling that services use to express data requirements since there are separate mechanisms defined in SAML for expressing NameID and Attribute requirements.

Also, perhaps more seriously, the NameID is the unique key used in SAML LogoutRequest messages. When the most common binding for such messages (HTTP-Redirect) is used, XML Encryption is the only way to prevent the NameID value from appearing unencrypted (though obfuscated) in web server logs. This can be a significant privacy concern for both omni- and uni-directional identifiers. XML Encryption is not commonly deployed today in messages produced by SPs, and potentially requires deployment and management of a new set of IdP keys, and a potentially long process of convincing SPs to encrypt the data.

Of course, logout itself is not (and may never be) widely deployed and is significantly limited in practice, but the working group also received strong feedback that many SPs feel it is a customer requirement they must try to meet.

Taken together, along with existing concerns around the "persistent" NameID Format and its suitability to meet the "directedID" requirements (discussed below) the working group believes the profile should course-correct the community in the direction of absolute avoidance of the NameID element apart from the "transient" Format suitable for logout purposes, and that identifiers should be exclusively defined as SAML Attributes.

Implications for Entity Categories

The proposed attributes have implications for attribute handling for the existing Research and Scholarship category and the potential to facilitate new categories. R&S today relies on the combination of two different attributes to provide a scheme for non-reassignment, human readability, and e-mail. It is largely designed around the assumption that applications are broken and require a single human-friendly identifier that would be "mail" in most other sectors but has historically been "eduPersonPrincipalName" in R&E.

The consensus is that this is not a good solution for non-broken applications, and a new iteration of this category might take this into account by simply requiring one of the two proposed identifier attributes. Or it might be appropriate to focus R&S on the correlatable identifier case, and define a new category (perhaps termed "privacy-preserving") for applications that can tolerate, or actively encourage, the use of directed identifiers.

Next Steps

The proposal for a pair of SAML Attributes is intentionally abstract at this point to emphasize that it may or may not map directly onto existing work. It is a given that there should be legitimate reasoning to develop new work rather than reuse already-deployed solutions, and so the next steps are to evaluate the requirements against existing identifiers proposed or adopted across the profile's target audiences, such as "eduPersonUniqueId", SCHAC attributes, and of course existing SAML approaches.

There is also the consideration that SAML lacks a good, consistently adopted solution to this problem and that if one were to be drafted and made part of SAML itself or the saml2int deployment profile, both of which are not specific to the R&E sector, there might be at least long-term potential to see more standardized practice across sectors. If this is a legitimate possibility, then reuse of existing R&E attributes may require at least some finesse to be palatable.

Implications for SAML Persistent NameID and eduPersonTargetedID

There are a number of disadvantages to continue to promote the use of SAML's built-in directed identifier (i.e., the Persistent NameID), leading the working group to conclude that it is not a good choice for satisfying the requirements for a "directedID" attribute.

Some of the problems are inherent to any use of the SAML NameID element to carry personally identifiable information (see above). While "eduPersonTargetedID" was defined by eduPerson to allow the same construct to be usable as a SAML Attribute, its complex XML syntax is unsupported by all but a few open source implementations, and its use has been discouraged at least informally for a number of years.

The specific formulation of the SAML Persistent NameID also has a number of ongoing problems, one of which is fatal in the strictest security sense: It was defined to be case-sensitive, which allows issuers to supply identifiers for different users that differ only by case. This literal requirement is not met by a variety of, if not the majority of, common web applications. Even though e-mail address itself is not defined to be case-insensitive, in practice it's treated that way by applications, many of which assume all identifiers should be handled that way. While there are techniques to fix SAML implementations such that even identifiers produced from hashes are not case-sensitive (e.g., using Base32), existing deployments would have to rekey users.

There are also other challenges, such as its large theoretical maximum size and the use of a "triple" containing SAML IdP **and** SP entityIDs to fully qualify the identifier, which is difficult for developers to understand and manage.

OASIS Spec for Attributes and Signaling of Capabilities and Requirements

We will use the OASIS committee process to define two new SAML-level, URN-named identifiers, one for non-targeted ID and one for targeted ID.

Draft is tracked at https://wiki.oasis-open.org/security/SAMLSubjectIDAttr

Signaling of the identifier requirements of the SP will be via the following mechanism:

SPs that are compliant with the profile MUST decorate their entity descriptor metadata with an `<EntityAttribute>` signaling NO LESS THAN ONE of the following:

1. Require pairwise ID
2. Require standard ID
3. Require none
4. Require any ID

Behavior in the presence of both `<RequestedAttribute>` metadata and the `<EntityAttribute>` is undefined by the profile.

Response to Questions about Pursuing the OASIS Route Rather than eduPersonUniqueId

The working group understands that some may not see additional value in the direction we are pursuing. This group feels that there is added value in pursuing the OASIS option and little if any loss in speed. Potential cross-sector acceptance of the identifiers are one such example. Federations are going to be the ones that have to do the push for adoption of this, so any value they get out of doing this via a specific mechanism should be considered. The possibility of applicability to vendors, etc, means this approach has a potentially broader appeal and could mean eventual wider adoption and increased value for federation participants. If it doesn't, nothing significant will have been lost from doing so.