Identity in the Federal Learning Registry

James Klo and Marie Bienkowski
SRI International [1]
{jim.klo, marie.bienkowski}@sri.com

Introduction

Open and commercial education resources are migrating to many digital platforms as computing takes on new forms and as learning becomes more self-directed and reliant on digital and internet-based systems. Users, we believe, will expect to interact not only cognitively with these resources but also to interact socially around these resources using the Web. Acknowledging this trend, development is underway on a service called the Federal Learning Registry, a simple and inexpensive system for distributing information about learning resources and their use (learningregistry.org/). The Learning Registry helps alleviate the problem of disparate standards for describing resources by changing the business model for suppliers from hand-curation of descriptive data (the "library model") to tapping data streams from social networks and learning management systems (among others) to locate and identify resources (the "recommender model"). The Learning Registry will expose many resources, sharing information about their usage and enabling filtering to locate relevant ones.

The Learning Registry concept is predicated on a model of sharing that assumes varying levels of trust and privacy, and we believe we can thereby contribute a unique perspective on the requirements for trustworthy and reputable digital identity management. Information placed in the Learning Registry must come from reputable sources for it to be of value. Information about usage, ratings and the like may be more honestly expressed within small communities of educators yet placed in the Learning Registry in a way that is both trusted and ensures the privacy of the educators. In the remainder of this position paper, we expand upon our problem and possible solutions, responding to the workshop organizer's call for papers on "…use-cases and requirements from enterprise, online banking, government, health, business, regulatory bodies, and activist groups."

The Federal Learning Registry Requirements for Identity

The Federal Learning Registry, under development as an open-source project, is intended to make digital learning resources easier to find and access regardless of where they are stored. Envisioned as a learning-resource distribution network for simple notifications or assertions about resources (roughly, metadata) and how they are used (roughly, paradata), the Learning Registry provides three basic capabilities:

1. a lightweight mechanism to publish metadata or paradata into a distribution network, independent of the standard, format or data type,
2. a simple means to consume the published data and then, in turn, to publish additional feedback about the resources' use into the network as paradata, adding to the overall knowledge about resources, and
3. a high-latency, loosely connected network of nodes for distributing data throughout the network.

The Learning Registry is providing a set of open APIs upon which more complex services can be built. In the core distribution network (built on the CouchDB document-oriented replication system) there is no central control and no central registries or repositories. As in other CouchDB systems, published data can eventually flow to all nodes in the network. By design, the network will be self-assembling. APIs provide connections to nodes to find out what resources are in the network, what has changed, what is being used, and so on.

[1] Our interest in the topic of Identity stems from work we are doing in conjunction with the US Department of Education and the Department of Defense's Advanced Distributed Learning Initiative.

Organizations may build consumer-facing, value-added services upon the Learning Registry nodes to enable using, finding, sharing, and amplifying resources, metadata and paradata for user communities. The Learning Registry provides *social networking for metadata* (trusted social collaboration around learning resources), enabling a *learning layer* on the social web.

Given this general design, identity is important to build trust and reputation in the Learning Registry data for consumers. We need to support the following with an identity solution:

- Assert a claim using a given identity (where the claim is about a resource, metadata or paradata)
- Validate that a claim was made by a given identity

The Learning Registry persistently stores published data for later use, which leads to a requirement unique (we believe) to the Learning Registry: to work with data involving identity long after publication. Over time, as users and organizations publish data—perhaps even years later—consumers may need to validate claims involving identity. Publishing and consuming data associated with an identity will be optional, but we suspect will be necessary to keep the data trustworthy and to permit building a reputation based on a history of behavior.

Trustworthiness encourages consumption of data, while reputation, we believe, may encourage publication. To motivate publishers (organizations and individuals) to participate in the Learning Registry, exposing their work as authors, commentators, curators, and the like can build their digital reputations. As we are building the infrastructure of the Learning Registry, we are linking up publishers of content: large repositories of digital content such as the National Science Digital Library and the National Archives. Over time, to keep the material fresh and up to date, we need to encourage all users, not just digital libraries, to participate with metadata or paradata.

As described above, the Learning Registry is decentralized, and a design goal is to build in no single point of failure and (ultimately) flexible levels of trust. The implementation of identity should be supportable on an architecture that is effectively a peer-to-peer content distribution network, and the solution should be as lightweight as possible, and easy for users to adopt. The Learning Registry is a permanent document store, and because documents are replicated over possibly unsecured connections, and "rogue" nodes could join the network (at present, the policies for network assembly are being specified), methods for determining that a document has not been tampered with are also important in our thinking about a solution. In practice, this means that documents need to be *signed* before entry into the Learning Registry and the digital signature must be able to be checked at any other node and at any future time.

The Problem

In contrast to many other identity efforts, the Learning Registry identity problem is not to verify identity in the strictest sense of complete assurance that a claimed identity is correct, but instead to create a "lower level of assurance" that is agile and workable. We assume that users will respond most favorably to the need for a simple yet persistent validation of their identity: thus, we believe we have exposed a new requirement for the Identity topic area with this application. In our search for a solution to match the needs of the Learning Registry (described in detail below), we believe that in the browser and web arenas, there is no public, unrestricted, irrevocable means to validate claims involving identity,

We are interested, as are others, in leveraging the broad set of items that collectively constitute the digital, virtual web identity for an individual or organization. We wish to leverage identities as established in social media sites, reputations built up from participation in public events and published materials, accomplishments, and web site material that establish reputations. The simplest and most similar related idea is WebFinger. WebFinger defines a method for using an account-based URI scheme as an identifier to

discover additional information regarding an individual or entity without requiring the owner of the identity to authenticate or authorize such action.

The Learning Registry requires the capability to:

- validate the integrity of data published into the network from within any node of the network,
- link separately published data based on the identity of the publisher,
- validate the existence of an external identity, and
- validate an identity using a variety of methods that can adapt over time.

As we have described above, the solution should be lightweight, decentralized, and flexible in design, and, ideally, the Learning Registry should not require negotiation of a pre-established trust with an Identity Provider to verify an external identity, i.e., a node should not have to manage or configure a relationship with *n* identity providers.

Possible Solutions

The Learning Registry, as an open source project should, to the extent possible, leverage existing and emerging standards for identity. In this section, we review possible solutions.

One common and "low assurance" form of validating claims involves the use of an email address that is validated by user action: clicking on a link in an email to validate that a real user owning an email address is behind the action taken or assertion made with that email address. This works to establish a login/password-based account or an initial connection between service and provider, but does not provide a solution that has persistence so that claims can be validated after a user has disconnected or otherwise left a session.

A second form of validation we considered is an emerging standard such as OAuth[2]. OAuth is designed to allow a user of one Web service to authorize use of all or part of a resource owned by the user available at a partnered Web service. As part of this service-to-service exchange, validation of the user (via an account login with an existing account) takes place. This validation could be viewed as a stand-in for an identity validation process[3]. The main obstacle to using OAuth for the Learning Registry, we have learned, is that it assumes the exchange of a "shared secret" between the identity provider and the service that is authorized to use that identity (or a portion of it). The Learning Registry is an ad-hoc network in which the lifetime of a node is not guaranteed. What goes in one node can be replicated to any other node, and the receiving node may choose to perform validation, but will not have access to the "secret" that was exchanged between the identity provider and the original node to validate the signature.

A third consideration, again looking for standards that could solve our problem, was OpenID[4]. OpenID has similar implementation features to OAuth and email validation, as it also assumes a centralized store and an active, user-in-the-loop validation process. Ideally, we would like the Learning Registry to interoperate with these identity solutions so that our users have a low barrier to entry for publishing in the Learning Registry, and to leverage the efforts of larger institutions for the benefit of improving educational resource location and utilization.

[2] http://oauth.net/documentation/spec/

[3] http://dev.twitter.com/pages/sign_in_with_twitter

[4] http://openid.net/developers/specs/

Finally, WebFinger[5] or WebID[6] both have the advantage of being an open method for consolidating multiple identities on the Web. However, neither are widely adopted and in use, and, like the other approaches, do not guarantee a method to confirm the integrity or validity of a document once it has been published and replicated throughout the registry nodes. Specifically, both utilize a discovered XRD document[7] to describe how to locate additional identity information tied to specifications such as OpenID and OAuth. WebID can be used to support signature validation, however this support is limited to X.509 RSA Public Key support. Given this limitation, our concern is that X.509 limits certificate signing to a single "Trusted Introducer" or "Certificate Authority", which creates a central authority for identity verification. WebID fails to satisfy the needs of the Learning Registry: X.509 as a guaranteed method of signature validation does not support decentralized identity management. WebFinger and WebID could be utilized to link a profile to other forms of Public Key verification, however neither specification defines a vocabulary for linking public keys. The Learning Registry needs these fields to support our solution so that (1) only links which may not be persistent, (2) optional in that neither format requires one to declare linked keys, and (3) the vocabulary to link keys are limited to descriptive comments which would require our users to define a community convention when defining links to keys within their profile. Ideally a solution for storing public keys should be encapsulated into the solution in a manner that WebID provides, but with more flexible Public Key options.

Our Approach

After reviewing the above approaches, we are experimenting with an ad-hoc implementation of PGP (and Gnu Privacy Guard) and leveraging existing public keystores as public key storage and retrieval (see Figure 1). This approach requires the publisher to submit their public key into a public keystore prior to publishing into the Learning Registry (1 in Figure 1). Once an assertion is signed and published (2,3), any node may retrieve the public key (4) and use it for checking the integrity of the published contents via signature validation based upon the policies of the node (5). We have discussed adding a requirement for publishers to utilize keystores that perform email validation to validate the existence of a real entity (robot, agent, person). This implies a need for certification and enforcement of "trusted introducers"[8] which is currently viewed as a value-added (not core) service by a node in the Learning Registry.

Upon closer examination of this approach, we have found that there is still no standard for interfacing with public keystores to retrieve keys, so a custom solution must be developed. Some documentation and APIs for accessing keystores attempt to define standards as extensions of HTTP, SMTP, LDAP, DNS and other proprietary or undocumented protocols, yet none appear to be implemented much further than a stable reference implementation of the protocol. Considering our effort to support decentralized identity, having to focus or limit publishers to one solution of key distribution breaks our ability to support decentralized identity solutions. While we believe we can potentially work with what's available through the construction of a tool to unify key retrieval for the LR needs. We claim that there currently exists no common set of interface methods across implementations that can easily be used to construct a robust, persistent, and universal means for accessing public keys to confirm digital signatures.

We are left to challenge the identity community to aid us in defining an easily adoptable standard for validating assertions made by an identity that can be used across the Web, inside and outside of the browser.

[5] http://code.google.com/p/webfinger/wiki/WebFingerProtocol

[6] http://www.w3.org/wiki/WebID

[7] http://docs.oasis-open.org/xri/xrd/v1.0/xrd-1.0.html

[8] http://www.pgpi.org/doc/pgpintro/#p19

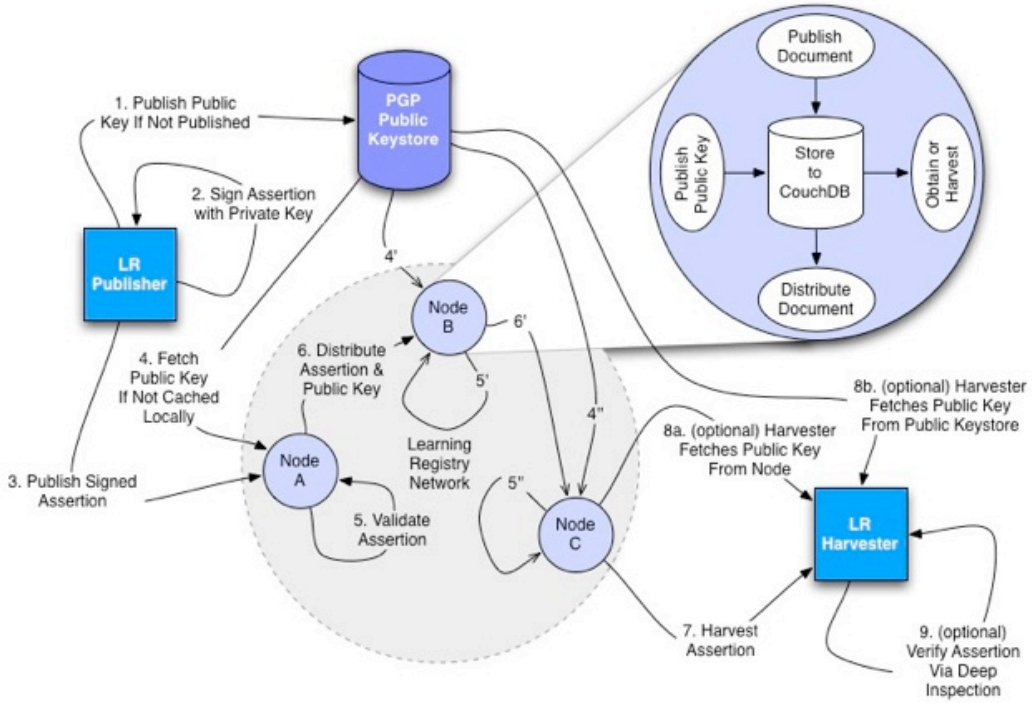

Figure 1: Data Flow for Proposed Learning Registry Identity Solution

Additional Notes on the Learning Registry

Community Activity

The Federal Learning Registry was announced in July 2010 by Arne Duncan, Secretary of Education, G. Wayne Clough, Secretary of the Smithsonian Institution, and Federal Communications Commission Chairman Julius Genachowski. The project has engaged many other agencies and organizations, both inside and outside the federal government.

Federal organizations participating include the Office of Education Technology of Department of Education, Advanced Distributed Learning (ADL) Initiative of the Department of Defense, Office of Science and Technology Policy at the White House, National Institute of Standards and Technology, National Archives and Records Administration, the data.gov team, and the Federal CIO and CTO. The technical team has also had partnership discussions with NASA, the Smithsonian, Library of Congress, National Science Foundation, and the Department of Energy.

Representatives from PBS, Creative Commons, CapStone Digital Publishing, international organizations and other educational partners are members of the project technical working group. Technical working group members are interacting with all interested parties via a public e-mail list. A preliminary implementation of the Learning Registry was used at the 2011 OER Hackdays in Manchester, UK (April).

Public comment has been solicited via an IdeaScale site with a corresponding open Google discussion group. The IdeaScale site allows comment and voting for ideas. learningregistry.org (maintained by ADL) also contains news items, as does the @learningreg twitter feed.

Acknowledgements

Many of the ideas in this paper are due to thoughtful work by Steve Midgley of the Department of Education and Daniel Rehak of the Advanced Distributed Learning Initiative (Department of Defense), and participants on the learningregistry.org mailing list. Mingyu Feng of SRI did background research, Eve Maler of Forrester Research gave us pointers, and Phil Zimmermann contributed thoughtful input and critiques.