INFORMATION STANDARDS QUARTERLY
FALL 2014 | VOL 26 | ISSUE 3 | ISSN 1041-0031

ISQ

TOPIC

IDENTITY MANAGEMENT

PRIVACY BY DESIGN
AND THE ONLINE LIBRARY

FROM THE LIBRARY OF CONGRESS
TO THE LIBRARY OF ME

THE INTENTION PUBLISHING ECONOMY:
WHEN PATRONS TAKE CHARGE

A JSON-BASED IDENTITY
PROTOCOL SUITE

NISO
How the information world
CONNECTS

DAN BLUM

PRIVACY BY DESIGN

AND THE ONLINE LIBRARY ENVIRONMENT

This paper focuses on ways that libraries can incorporate advanced identity management concepts within the Privacy By Design framework to meet their needs as they continue their transition from the brick, mortar, and paper era to an era of mixed physical and digital content. In order to add value over and above what researchers can find with search engines and freely available content on the Internet, libraries must excel at supporting both ordinary knowledge seekers and academic researchers in fulfilling their content- and collaboration-related needs. Increasingly, libraries must support a seamless, personalized, and collaborative experience for diverse audiences across the full lifecycle from content discovery to content delivery while at the same time protecting patrons' privacy and intellectual property prerogatives.

Changing Business Trends

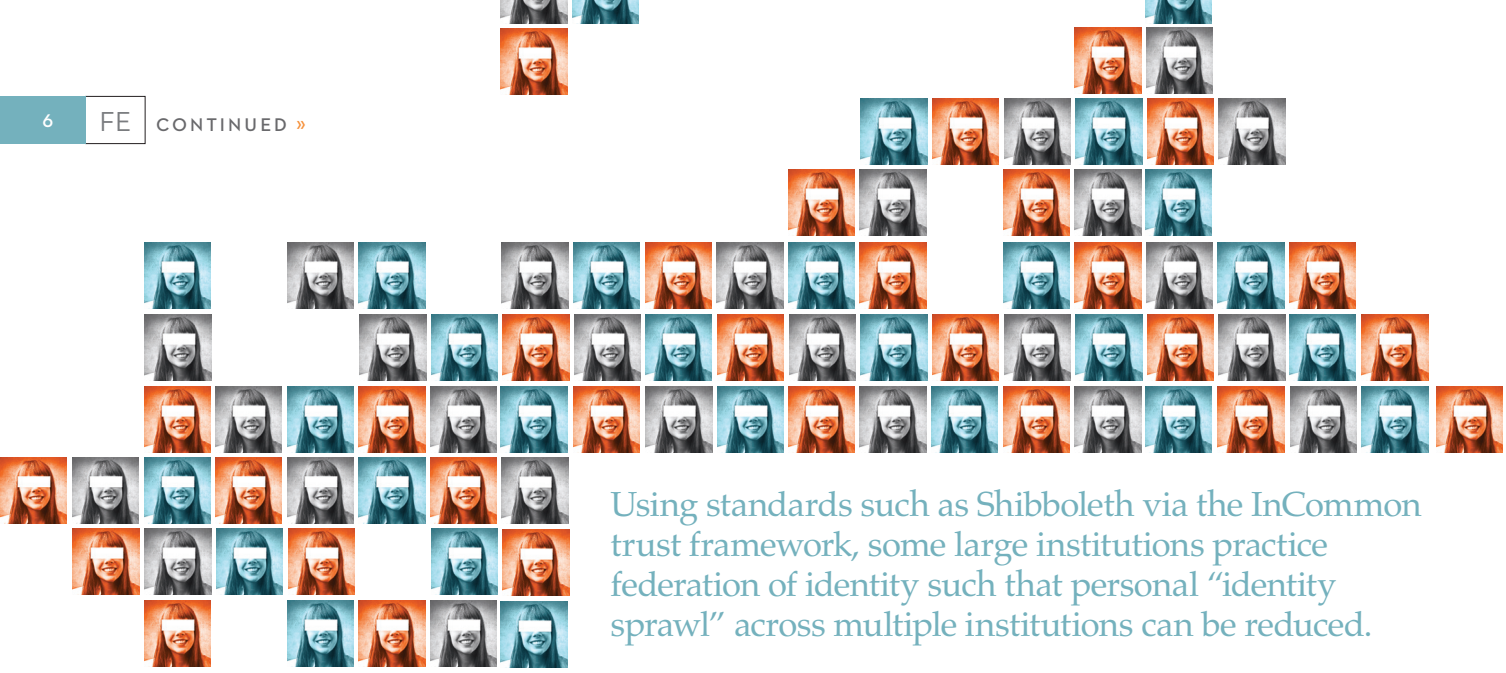
Like many industry segments, higher education and public libraries face a business imperative to support more complex online use cases for patrons and partners. Each library patron has a unique constellation of needs and relationships. Faculty, staff, students, alumni, and even "walk-ins" (or visitors) may be associated with multiple borrowing or authorizing institutions. Each partner library, research institution, business, or content provider may also have different entitlements and licensing or other business practices that must be respected.

With the requirement to differentiate from, or add value to, the ocean of free Internet content, libraries must support value-added services or content that are not provided freely to anonymous users. As research and collaboration enablers, they must support these services from discovery to delivery, in some cases providing a level of full-text search without "giving away the farm" to subscribing institutions, customers, or partners.

At the far end of the spectrum for business value and disruption, many businesses, and even individuals, may simultaneously become both consumers and providers of premium or restricted content in the growing "bring your own cloud" and "bring your own identity" environments. Thus, the library of the future could intermediate research and collaboration exchanges between complex fabrics of lenders, personal clouds, content providers, and businesses. To do so will depend on meeting requirements for richer identity and entitlement information interchange between actors in various use cases.

Each library patron has a unique constellation of needs and relationships.

CONTINUED »



Using standards such as Shibboleth via the InCommon trust framework, some large institutions practice federation of identity such that personal "identity sprawl" across multiple institutions can be reduced.

An increasingly diverse and inter-dependent library environment will bring new challenges as well. Often, information providers sharing their premium content will expect to get personal information on patrons, or to deliver advertising to patrons as a quid pro quo. These advertising endeavors could in turn ensnare libraries in a web of dubious relationships, as the author described happening to other online services in his article *Dark Lords of the Internet*.

Regulatory Risk

Juxtaposed against the growing business need for rich identity and entitlement interchange is the continuing movement for privacy regulation. This trend is creating tremendous tension between the advertising technology model ("ad-tech model") for online service delivery and the law. Libraries are already governed by the Family Educational Rights and Privacy Act (FERPA) and similar regulations. To the extent they operate internationally, engage non-U.S. patrons, and store personal records of students or patrons, libraries may also fall under a growing wave of international regulations.

In 2014, revelations of pervasive public and private surveillance by Edward Snowden, the CBS show *60 Minutes – The Data Brokers*, and other sources outraged public opinion, pouring gasoline on the regulatory fire. Even in the U.S., the Federal Trade Commission (FTC) and privacy consumer activist groups now actively hunt for privacy terms abusers. Libraries that try to expand identity data interchange and retention without a strong leavening of Privacy By Design will do so at increased risk.

Other Risks

Libraries face more than just regulatory risk as both their public and academic industry sub-segments frequently come under cyber-surveillance or cyber-attack. Even libraries that don't deliberately abuse privacy may be held liable for negligence if they:

- allow patrons to be hacked from infected library networks or computers;
- fail to assure the confidentiality and integrity of licensed content against the efforts of malicious patrons, fraudsters, and hackers;
- leak too much personal information on patrons to unscrupulous private data brokers in a harmful manner or on a large scale; or
- cooperate with or allow unwarranted law enforcement or other government searches of patron data and activity.

The endless inventiveness of cybercriminals and scammers is already taking its toll on the industry as seen in reports of Russian websites trafficking in user ids and passwords granting access to library proxy servers.

Identity and Privacy Issues for Libraries

Libraries have multiple issues with operational inefficiency, fraud, and regulatory risk arising from shortfalls in existing identity and privacy-related practices. Some issues—such as resale of proxy user ids, or of an entire patron database and subsequent release of passwords by cybercriminals—can arise for a single institution. Other issues occur in the context of multi-library interactions and the over-sharing of patron information.

In theory under the inter-library loan protocols, lending institutions should not have to obtain patron information—dealing with the patron should be the responsibility of the borrowing institution holding the patron relationship. Using standards such as Shibboleth via the InCommon trust framework, some large institutions practice federation of identity such that personal "identity sprawl" across multiple institutions can be reduced. Often, however, practice lags theory. Proxies may not be well integrated with identity systems providing a single campus id. Many institutions don't participate in the InCommon federation or have use cases—such as the need to support direct end user interaction with non-library content providers—not readily supported by the standards.

Basic business practices may be only marginally compliant with FERPA. Although FERPA provides a substantial loophole where institutions can designate large amounts of personal information as "directory information" to support over-sharing and over-storing arrangements, they often don't provide sufficient transparency to patrons or the ability for patrons to opt out of integration with third-party services that could result in information leakage to data brokers, advertisers, or worse. Should the regulatory climate tighten, even large institutions such as Harvard University could come under pressure to narrow their definition of "directory common elements" and provide greater permissioning granularity to patrons.

Technology Trends

Shibboleth and the Security Assertion Markup Language (SAML) on which it is based are showing their age. While some provisions exist for handling attribute assertions as well as authentication, a new crop of "claims-based identity standards" are emerging. Implementing these standards to provide claims-based access control may help libraries reduce their privacy compliance risks from identity sprawl and over-sharing. For example, a library could reply with a "U.S. citizen" token or "age over 18" token rather than personally-identifying information about the patron to enable certain authorization use cases.

While necessary, current claims-based identity standards won't be sufficient. Unfortunately, the OAuth 1.0 and 2.0 specifications on which most of the standards are based have numerous security weaknesses, and when used in practice by providers such as Facebook, Google, and Microsoft, tend towards the over-sharing and overly-permissive registration practices characteristic of the model rather than a Privacy By Design-based approach.

Standards groups in the Internet Engineering Task Force (IETF) are working to remedy some of these flaws by adding proof of possession, JavaScript Object Notation (JSON) cryptographic tokens, and new dynamic registration specifications, but it may take years before major online providers driving the identity technology space implement them. Thus, although emerging pre-standards such as OpenID Connect and User Managed Access (UMA) may provide some basic claims-based plumbing, more assurance is needed on the security robustness and trustworthiness of the underlying OAuth protocol they currently rely on.

Some in the industry, such as members of the FIDO Alliance, envision that ubiquitous mobile devices never far from the users' hands may provide better identity assurance. They hope to leverage native device capabilities such as Apple's iTouch to use the mobile device as a strong identity token for online interactions. But skepticism abounds that interoperability will be universally attained, or that sub-$500 commodity devices floating around in users' purses and pockets can gain the hoped-for assurance.

Bring your own identity (BYOI) is emerging not only from the FIDO Alliance, but from a new category of personal information management (PIM) products and services. PIM product categories, such as personal data stores and user-centric personal clouds, are often premised on the individual, rather than some centralized cloud service, controlling both storage and sharing of personal data in keeping with strict privacy principles.

CONTINUED »

Some in the industry, such as members of the FIDO Alliance, envision that ubiquitous mobile devices never far from the users' hands may provide better identity assurance. They hope to leverage native device capabilities such as Apple's iTouch to use the mobile device as a strong identity token for online interactions.

BYOI solutions are sometimes criticized for only providing self-asserted identity, as if organization-asserted identity was always much more trustworthy. This misses the larger point that, whatever the original source of identity information, the risks of impersonation and fraud will always be with us, especially as the information drifts through chains of intermediaries that take it further and further from the source. Protocols alone cannot solve this problem of assurance; what's needed are trust frameworks and/or reputation systems that operate at the legal and social layer of the relationships of online communities relying on them.

Having trust frameworks (or agreements that enable participants who share or accept identity credentials—and identity, authorization or reputation claims—to operate under well-defined policies) is especially important when a strong requirement for privacy is added to traditional security objectives such as confidentiality and integrity. Some providers in the personal information management category are banding together around user-centric trust frameworks such as Respect Network. In these frameworks, privacy is the default setting, informed consent is required for all permissions, pseudonymity is an option, and the right to be forgotten is also specified.

How Online Libraries Can Apply the Seven Principles of Privacy By Design

Privacy by Design is an approach to IT systems development that takes privacy into account throughout the whole engineering process. The concept is analogous to "safety by design," i.e., to take human safety into account in a well defined manner. The concept is believed to have originated in a 1995 report by Canada's Information and Privacy Commissioner and Netherlands' Registratiekamer on *Privacy-Enhancing Technologies*. Dr. Ann Cavoukian, the former Information and Privacy Commissioner, Ontario, Canada, has promoted the concept of Privacy by Design since the late 1990s and manages a website with the name.

The seven foundational principles of Privacy by Design, which have been translated into over 35 languages, are:

1. *Proactive* not Reactive; *Preventative* not Remedial
2. Privacy as the *Default Setting*
3. Privacy *Embedded* into Design
4. Full Functionality – *Positive-Sum*, not Zero-Sum
5. End-to-End Security – *Full Lifecycle Protection*
6. *Visibility and Transparency* – Keep it Open
7. *Respect* for User Privacy – Keep it *User-Centric*

The following sections consider each foundation principle from the library industry perspective, building on Ann Cavoukian's and Drummond Reed's paper *Big Privacy: Bridging Big Data and the Personal Data Ecosystem Through Privacy by Design*.

PROACTIVE NOT REACTIVE; PREVENTATIVE NOT REMEDIAL

"The *Privacy by Design* approach is characterized by proactive rather than reactive measures. It anticipates and prevents privacy-invasive events before they happen. PbD does not wait for privacy risks to materialize, nor does it offer remedies for resolving privacy infractions once they have occurred—it aims to prevent them from occurring. In short, *Privacy by Design* comes before the fact, not after."

Libraries, and online businesses in general, have many opportunities to deploy proactive Privacy by Design solutions. That is because both advertising-based and non advertising-based systems have tended greatly towards centralized models of personal information storage. In this model, the organization with access to all their users' personal information sets all the terms and conditions of use. Centralized systems create, in effect, a single information silo that cuts individuals off from meaningfully participating in a market "based on a resource that they themselves (mostly) produce, namely their personal information." Privacy risks abound under such a model, not only because the data controllers have incentives to exploit personal information without regard to the subjects' preferences, but also because risk aggregates in the large centralized systems, and the more of them there are, the more identity information sprawls.

Solutions that decentralize control of personal information either to the individuals themselves (e.g., personal clouds) or at least to the organizations that have a closer relationship to the individual (e.g., borrowing libraries rather than lending libraries) may prevent many privacy risks from arising by putting people more in control of their information. They can also improve operational efficiency and assurance by moving the authoritative source for data closer to the individual, thus improving its quality and accuracy.

Every library, patron, and partner has its own unique constellation of relationships and entitlements. Library use cases are becoming more complex to address enhanced research and collaboration functionality enabling everything from discovery to delivery of a mixed universe of free and restricted content. Thus, the library community will need a mix of centralized, decentralized, and hybrid

identity topologies. Different topologies will favor different technologies falling broadly into the federated identity, claims-based access control, and BYOI technology categories deployed in a proactive, Privacy by Design manner. As the breadth of the communities grows and the use cases and privacy challenges become more advanced, trust frameworks and semantic authorization standards will also be required.

PRIVACY AS THE DEFAULT SETTING

"*Privacy by Design* seeks to deliver the maximum degree of privacy by ensuring that personal data is automatically protected in any given IT system or business practice. If an individual does nothing, their privacy still remains intact. No action is required on the part of the individual to protect their privacy—it is built into the system, by default."

Today, large online content and service providers replicate the personal information they've collected in duplicative, centralized databases. They then seek to monetize this information through data sharing arrangements for advertising. Privacy is not the default; instead it is obscured under the cover of complex privacy policies. While the requirement for privacy policies was intended by regulators to promote openness and greater transparency of an organization's processing of personal information, most in fact do precisely the opposite with long, difficult to understand legalese, which the user is required to accept as is or not use the service.

As libraries seek to expand research and collaboration services to patrons, they run the risk of being drawn into relationships with content providers that participate more heavily in the ad-tech economy and become tainted by association. To avoid such situations from occurring, cooperative library industry trust frameworks that are user-centric should be developed to control the web of relationships underpinning services. Such frameworks should ensure that privacy is the default setting and that all sharing of personal information is by permission only.

A trust framework legally binds *all* members of a trust community—both individuals and organizations—to a set of business, legal, or operational policies, as a condition of membership. For example, the Respect Trust Framework is a user-centric trust framework that sets down global terms and conditions for interacting with personal information in a manner that respects the privacy of individuals, with strong assurances of security. Libraries could participate in this trust framework or develop something similar for themselves. They could also strengthen privacy provisions for themselves as a sub-community of the InCommon trust framework.

Privacy is not the default; instead it is obscured under the cover of complex privacy policies.

PRIVACY EMBEDDED INTO DESIGN

"*Privacy by Design* is embedded into the design and architecture of IT systems and business practices. It is not bolted on as an add-on, after-the-fact. The result is that privacy becomes an essential component of the core functionality being delivered. Privacy is integral to the system, without diminishing functionality."

Libraries can embed privacy into design by:

- Moving to decentralized or federated architectures that minimize the collection of personal information from patrons
- Establishing network-wide trust frameworks so that information is shared only with privacy as the default and standards exist for de-identification of data required for analytics
- Using generalized roles (such as "student", "faculty", "staff", "visitor", "librarian") rather than identifiers or groups for authorization
- Using claims tokens, such as "over 18" or "U.S. citizen", rather than revealing private personal attributes
- Using pseudonymous identifiers for patrons

CONTINUED »

FULL FUNCTIONALITY – POSITIVE-SUM, NOT ZERO-SUM

"*Privacy by Design* seeks to accommodate all legitimate interests and objectives in a positive-sum "win-win" manner, not through a dated, zero-sum approach, where unnecessary trade-offs are made. *Privacy by Design* avoids the pretense of false dichotomies, such as privacy vs. security, demonstrating that it is indeed possible to have both."

A 2012 study by Edelman Digital found that "seven in ten people globally are more concerned about data security and privacy than they were five years ago, and a full 68% believe that consumers have lost control over how online personal information is shared and used by companies."

Privacy by Design advocates have been saying for years that privacy is good for business. When customers are knowledgeable about and fully involved in decisions about sharing of their personal data, they will have more confidence and trust and be more willing to share their personal information with libraries. This information can, in turn, be shared by permission—often in a de-identified manner—to personalize services both from the core library networks and from private sector partners. By providing de-identified patron analytics, libraries can, for example:

- Optimize acquisitions and collections management
- Incentivize content holders to make information more available
- Personalize content for different classes of patrons

By maintaining a strong reputation for integrity and privacy, academic and public libraries can protect or even expand their "market share" versus "freemium" information-based products and services on the Internet.

END-TO-END SECURITY – FULL LIFECYCLE PROTECTION

"*Privacy by Design*, having been embedded into the system prior to the first element of information being collected, extends security throughout the entire lifecycle of the data involved. This ensures that all data is securely retained, and then securely destroyed at the end of the process, in a timely fashion. Thus, Privacy by Design ensures cradle to grave, secure lifecycle management of information, end-to-end."

Through federated identity systems and claims-based access control, libraries can improve identity assurance overall. That is because institutions will put more effort into maintaining accurate information or protecting credentials for an identity that's relied on for single sign-on (such as a campus id) than for a one-off proxy service account. Alternatively, user-centric federations—such as those enabling personal cloud networks or BYOI—apply protection at the interface of the patron or partner. The patron will keep his or her own data accurate, both as the first to know of most changes and for self-protection.

However, such security measures become more complex and harder to manage as more parties are involved, such as multiple libraries and content providers. Federated identity systems through trust frameworks are again a solution to consider when data is shared among multiple stakeholders.

Personal information has a lifecycle, just like records, and must be destroyed on a timely basis in a secure and privacy-protective manner. Personal information should also not be replicated in multiple databases to avoid the existence of excessive copies, which might not get destroyed simultaneously. In the BYOI model, the authoritative source for private information is an individual's personal cloud, and a "subscription" model can be used to provide others with access. The individual retains control over when to delete data or turn off access.

VISIBILITY AND TRANSPARENCY – KEEP IT OPEN

"*Privacy by Design* seeks to assure all stakeholders that whatever the business practice or technology involved, it is in fact operating according to the stated promises and objectives, subject to independent verification. Its component parts and operations remain visible and transparent, to users and providers alike. Remember—trust but verify!"

The only way that users will have a real sense of control over their private information is with full transparency and understanding of how their personal data will be accessed, used, and shared by anyone who is party to it. As previously noted, a user-centric trust framework is the recommended method for such transparency and understanding. Because their terms and conditions are publicly reviewed and published and all members agree to follow them, trust networks can establish a community's best practices for privacy. Inter-library and third-party audits are one method of verifying and enforcing the trust's policies are being followed.

RESPECT FOR USER PRIVACY – KEEP IT USER-CENTRIC

"*Privacy by Design* requires architects and operators to keep the interests of the individual uppermost by offering such measures as strong privacy defaults, appropriate notice, and empowering user-friendly options....At its core, respecting the user means that, when designing or deploying an information system, the individual's privacy rights and interests are accommodated right from the outset. User-centricity is anticipating and designing in a person's privacy perceptions, needs, requirements, and default settings. It means putting the interests, needs, and expectations of people first, not those of the organization or its staff. Empowering people to play active roles in the management of their personal data helps to mitigate abuses and misuses."

Libraries also can adopt user-centricity as an operating principle. For various use cases they can offer users the convenience and control of BYOI (or a secure institutional identity), the protection of a user-centric trust framework, the option to use either pseudonymous or public identifiers, and the ability to share personal data under contracts that bind relying parties to de-identify the data.

Conclusion

The library industry, in seeking to become a network of scholarship, research, collaboration, and knowledge amidst oceans of uncurated Internet information, should adopt Privacy By Design into its core guidelines. Not only will Privacy by Design improve compliance postures, it can also be good for growing the evolving roles of libraries in information discovery and delivery. By taking a proactive approach to preventing privacy infractions, setting privacy as the default, and maintaining transparent, user-centric identity and privacy policies, libraries can find positive-sum solutions for participating institutions, partners, and patrons.

| FE | doi: 10.3789/isqv26no3.2014.02

DAN BLUM (dan@respectnetwork.net) is Chief Security and Privacy Architect with Respect Network and author of the blog Security Architect (http://security-architect.blogspot.com/). He is dedicated to addressing security and identity management issues from the enterprise, individual, and social perspectives.

RELEVANT LINKS

Blum, Dan. "Dark Lords of the Internet." ***Security Architect*** **[blog], June 9, 2014.**
http://security-architect.blogspot.com/2014/06/dark-lords-of-internet.html

Cavoukian, Ann, and Drummond Reed. ***Big Privacy: Bridging Big Data and the Personal Data Ecosystem Through Privacy by Design*****. Ontario: Information and Privacy Commissioner Ontario, Canada, December 2013.**
http://www.privacybydesign.ca/index.php/paper/big-privacy/

"The Data Brokers." ***60 Minutes*****. CBS, March 10, 2014.**
https://www.youtube.com/watch?v=_Cty7ctycsI

Family Educational Rights and Privacy Act (FERPA)
http://www2.ed.gov/policy/gen/guid/fpco/ferpa/

FIDO Alliance
https://fidoalliance.org/

Harvard University Common FERPA Directory Elements
http://security.harvard.edu/files/it-security-new/files/ferpa_directory_common_elements.pdf

InCommon Identity Assurance Assessment Trust Framework
http://www.incommon.org/docs/assurance/IAAF.pdf

The JavaScript Object Notation (JSON) Data Interchange Format
http://tools.ietf.org/html/rfc7159

The OAuth 2.0 Authorization Framework
http://tools.ietf.org/html/rfc6749

OpenID Connect
http://openid.net/connect/

Privacy & Security: The New Drivers of Brand, Reputation and Action**. Edelman Digital, 2012.**
http://www.edelmandigital.com/2012/04/05/privacy-security-the-new-drivers-of-brand-reputation-and-action/

Privacy By Design
http://www.privacybydesign.ca/

Privacy-Enhancing Technologies: The Path to Anonymity. **Information and Privacy Commissioner, Ontario, Canada and Registratiekamer, The Netherlands, August 1, 1995.**
Vol. 1: http://www.ipc.on.ca/English/Resources/Discussion-Papers/Discussion-Papers-Summary/?id=329
Vol. 2: http://www.ipc.on.ca/English/Resources/Discussion-Papers/Discussion-Papers-Summary/?id=242

Respect Network
https://www.respectnetwork.com/

Respect Trust Framework
http://openidentityexchange.org/trust-frameworks/respect-trust-framework/

Security Assertion Markup Language (SAML) specifications
http://saml.xml.org/saml-specifications

Shibboleth
http://shibboleth.net/

User-Managed Access (UMA) Profile of OAuth 2.0
http://docs.kantarainitiative.org/uma/draft-uma-core.html