Identity and Access Management for Complex Research Data Workflows

Richard Zahoransky, Saher Semaan, Klaus Rechert

Department of Computer Science
University of Freiburg
Hermann-Herder-Str. 10
79104 Freiburg
richard.zahoransky@rz.uni-freiburg.de
semaan@uni-freiburg.de
klaus.rechert@rz.uni-freiburg.de

Abstract: Identity and Access Management (IAM) infrastructures already provide a crucial and established technology, enabling researchers and students to access services like computing facilities and electronic resources. However, the rise of complex and fully digitalized scientific workflows, world-wide research co-operations, and the reliance on external services and data sources poses new challenges to IAM architectures and their federations. Due to the non-uniform structure of such services each service provider is implementing its own access- and security-policy. As a result of license restrictions or privacy concerns, a user has to be authenticated and authorized by different entities in different contexts and roles to access complex research data, i.e. requesting a digital object as well as appropriate processing tools and a rendering environment. In order to enable seamless scientific workflows, an efficient federated IAM architecture is required. In this paper we discuss the use-case of functional research data preservation and the requirements for a common authentication and authorization scheme. The goal is to develop a security architecture allowing the user to login only once, e.g. at his or her university library and the Identity Management (IdM) system should be able to delegate the user's request to the related service providers. All these entities need to interact with and on behalf of the user without the user having to enter his credentials at every point. The results of this work are particularly useful when facing upcoming challenges to securing and managing access to non-uniform and inhomogenous cloud services and external data sources as a basis for today's scientific workflows and electronic business processes.

1 Introduction

Universities and research institutions in general offer simple and convenient access for their staff to journals and scientific papers, using Identity and Access Management (IAM) systems either based on LDAP authentication, Microsoft Active Directory (MSAD), or Single Sign-On (SSO) techniques [GSvS08]. With today's research and development tasks and requirements shifting to pure digital workflows, involving world-wide cooperations and heavy reliance on external data-sources and computational services, traditional IAM

systems face new challenges providing seamless access to efficiently support today's scientific workflows. While enhancements on today's IAM-systems are inevitable, building on already established identity infrastructure is preferable, since this would reduce migration and future management burden by eliminating duplicate user-ids.

Usually, a research institution negotiates a contract with a publisher or service provider and acts as identity provider (IdP) for its staff, using a federated protocol such as Secure Assertion Markup Language (SAML) [CKPM05] to assert a user's identity. For instance, a library user may access a publisher's digital content as long as he or she is affiliated and has a valid university account. These permissions, however, lack granularity and are too simple for today's decentralized settings, with access workflows not constrained to a single service provider or single object. For instance, instead of only assuring the user's identity, service providers may need to interact with other services on behalf of the user. Additionally, service providers may need more detailed information about a person's identity, due to more complex access rights and roles, for instance when requesting sensitive data. Hence, a more sophisticated and granular identity management solution is required to support complex e-Science workflows, also taking potential privacy issues of research data into account [SBSCB06].

In this paper we present requirements and architecture of a distributed architecture to providing access to scientific data together with a solution for a federated IAM.

2 Related Work

Federated IdM-Systems are already successfully integrated in today's research institutions and universities. For instance, Germany's universities commonly use SAML-based Identity Provider systems. Interoperability between these organizations is co-ordinated by the "Deutsches Forschungsnetzwerk" (DFN) [1].

Furthermore, sub-groups can be formed, with members agreeing on specific details to extend the scope of interoperability and cooperation. Such as the bwIdM-project [2] of Baden-Württemberg defines a set of user attributes, which every participant must agree on and guarantee that individual IdM-systems are able to deliver all required attributes.

Other countries have similar organizations relying on SAML as backend infrastructure, e.g. the federation Éducation-Recherche [3] in France or Switzerland's SWITCH [4]. To enhance user privacy and un-traceability, certain setups of IdM-Systems allow a separation between IdP, key provider and service provider (SP). For instance, New Zealand's government implemented an identity management solution with strong emphasis on privacy and security [MCW08].

Different technologies exist for providing federated identity management in the Cloud [HLK+11]. OpenID for example is commonly used on websites [Fre08] and works in a

[1] Deutsches Forschungsnetzwerk, http://www.dfn.de, (20/1/2013)
[2] bwIDM Project, http://www.bwidm.uni-freiburg.de, (20/1/2013)
[3] Éducation-Recherche, https://services.renater.fr/federation/en/index, (20/1/2012)
[4] SWITCH, http://www.switch.ch (20/1/2012)

similar fashion as SAML by separating service providers and identity provider. However, OpenID does not support delegation of rights natively, which is a key feature required for distributed data management and scientific workflows in general. A complementary service for OpenID designed especially for delegation of rights is OAuth, for instance enabling convenient access to distributed scientific sites [BG11]. However, oAuth is rather a framework than a standardized protocol, thus different implementations may not be interoperable. Further, due to design and implementation decisions of real-world deployments, critical vulnerabilities were discovered recently [SB12].

This paper will focus on SAML as federative access technology as it natively supports the delegation of user rights and endorses a distributed access model by providing user attributes.

3 Access to Scientific Data and Workflows

Management of research data is becoming a crucial service of memory institutions and university (library) facilities in particular. In order to foster scientific innovation and simultaneously reduce redundant spending on data generation, efficient access to research results as well as their fundamental data are indispensable. Furthermore, with the rise of networked functional services, e.g. Cloud offerings such as software-as-a-service, and data service (Big Data), a process-oriented, holistic approach to research data management becomes more important. For instance, the reproduction of research results requires a complex setup of data(-source) paired with a suitable software environment containing a multi-step software tool-chain to process and to render data.

In most cases the best way to re-enact a digital scientific process and its data is using its original environment, since this covers all original aspects of the process's significant properties, hence providing an authentic and possibly an interactive user experience. Emulation and virtualization are a key strategy to provide a digital object's native environment [VvdH06].

Emulation has evolved into a mature digital preservation strategy providing authentic functional access to a wide range of digital objects using their original creation environments [RvSW10]. In contrast to format migration strategies, a functional, emulation-based approach requires a number of additional components, i.e. the full software-stack required to render a digital object but also its configuration. The bwFLA project [5] provides necessary workflows and an implementation of a distributed framework for emulation-based services to capture a scientific process's environment and configuration and to re-enact the environment at some point in the future in a predictable and authentic way. These technologies try to address Baden-Württemberg state's and higher education libraries and archives, new challenges in digital preservation, and research data management [RVvL12].

[5] Baden-Württemberg Functional Long-Term Archiving and Access (bwFLA) Project Website, http://www.bw-fla.uni-freiburg.de, (20/1/2013)

3.1 Scalable and Distributed Architecture

Workflows and tools developed by the bwFLA project are designed to be used in a distributed, labor- and cost-sharing setting. While the project delivers technical solutions and a distributed service-model, preservation of individual digital objects and accompanying measures are left to individual memory institutions. The goal of the bwFLA framework is to enable these institutions to use tools and perform workflows on certain types of digital objects, both for ingest and access workflows. In addition, the distributed setup enables various memory institutions to specialize on specific installation, e.g., covering tool-chains and environments for CAD models or embedded software development.

Local Memory Institution Memory institutions act as a locally available service facility. The memory institution (e.g. a university library) is able to accept research data for preservation purposes as well as providing access to stored digital objects. While it is not required that all objects are actually stored on-site, basic archival meta-data records are kept in a searchable catalogue pointing to the appropriate storage and computational services. Equally important, local institutions are also able to authenticate local users. Thus, local memory institutions are the main gatekeeper to a complex distributed research data management infrastructure.

Federated Software Archive In a distributed archival model the costs of archiving secondary digital objects, for instance standard software components, can be shared. Through mutual specialization, niches and specific areas can be covered without giving up generality. Similar to emulators, the number of available system-environments is limited and changes rather slowly. Therefore, the number of software components to be collected is limited and almost fixed for a certain time span, while the number of digital objects produced by these environments is usually not bound. For efficient archival and retrieval of standard software components various individual software archives are accessible through a common API as web-service. However, each archive is able to produce its own access policy and may neither be organizationally nor legally in the same realm as the memory institutions. Thus, the archive may have different access policies for each memory institutions and may additionally have specific rules for individual users, e.g. based on subscription, pay-per-use or similar models.

Emulation-as-a-Service (EaaS) Emulation and virtualization technologies are able to resemble a complete computer system. While the technical challenges developing emulators are not considered in this paper, usability and accessibility of emulators for non-technical users are crucial. Since the number of different ancient and current computer systems (i.e. hardware architectures) is limited, the number of required emulator-setups is limited, too. Hence, providing access to emulation is suitable for standardized services. In order to allow a large, non-technical user-group to interact with virtual computer systems an abstract emulation component has been developed to standardize usage and hide individual system complexity. Each emulation component provides a uniform API as web-

service and an abstract interfaces for graphical user interaction. Currently, the user is able to interact with emulation components through a website using either a Java plug-in or a HTML5 implementation. Furthermore, standard machine interaction is available, such as attaching/detaching removable drives (e.g. floppies, CD/DVDs) and attaching hard-drives to an emulator. The components are designed as atomic units suitable to be run in a distributed setup and are especially suitable for computing grids or clusters. bwFLA currently provides emulation components for a wide range of emulators and virtualization solutions, covering all major current and past computer platforms and their operating systems.

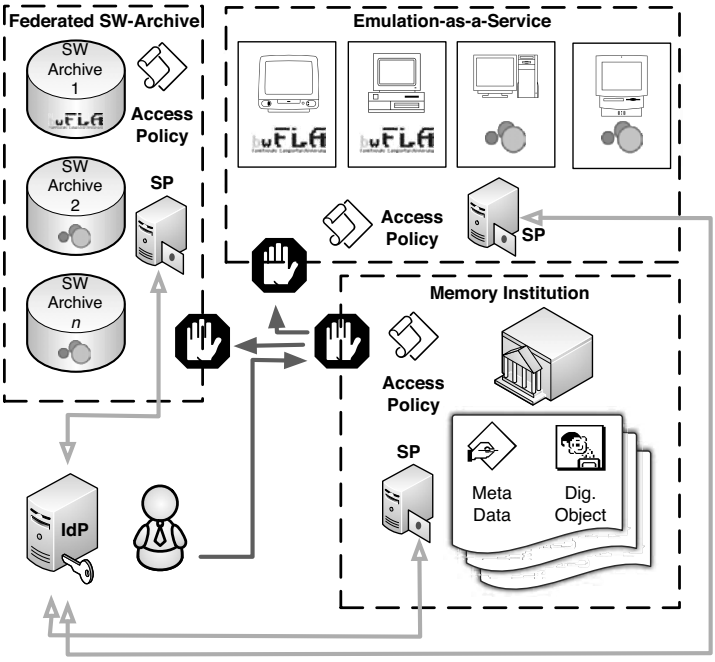


Figure 1: Distributed Architecture for Scientific Data Access Workflows

3.2 Identity and Access Requirements

In order to reenact digital data as well as the required software environment all three aforementioned entities have to cooperate. Usually the user consults his or her local memory institution's catalogue for digital objects of interest. If we assume that suitable metadata [DA12] is available for these objects the user is able to start the bwFLA access workflow. A detailed technical description of the bwFLA-workflows can be found in earlier work [RVvL12].

To access a specific digital object the user has to be authenticated and authorized by the local memory institution w.r.t. to local access policies. In case this object requires further

rendering or processing to be useful, local meta-data is retrieved and interpreted by the bwFLA framework. In a second step, appropriate software archive web-service ports are discovered and bound. On behalf of the user, the memory institution requests access to the required software components. The local IdM system delegates the user's rights to the memory institution which than will act as the user and can access the specific resource on the software archive in the user's name. If the user or his or her home institutions has been grated access to the requested software in a final (authentication and authorization) step, an appropriate EaaS site is discovered and bound. Due to the distributed setup, access to computing is restricted and usually bound to some kind of cost model. Access to the environment may also be restricted depending on the user's identity. The required computation time may also be limited and thus only be available to a specific user group. Once again, the local memory institution must relay the user's identity to render the requested environment and associated digital object. Figure 1 provides an overview of the general architecture and the data paths of an access workflow.

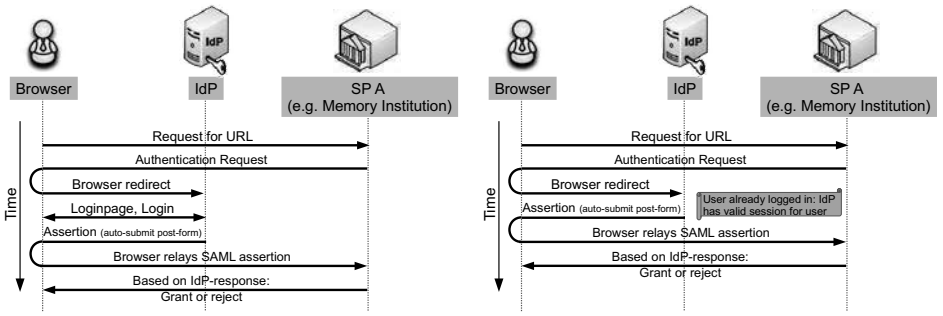
To improve the user experience, the delegation of rights should remain unperceived by the user [AHS03] but also he or she should be informed and have the facility to approve or to deny that a trusted entity (e.g. the memory institution) acts on behalf of him/herself, i.e. using his or her identity to perform certain tasks for the user.

4 SAML and Hybrid Authorization in Distributed Systems

To support federative access, SAML builds a trust relation between service providers and identity providers through the exchange of meta-data. SAML meta-data contains digitally signed information about the identity of IdP and SP and their corresponding digital certificates. This trust relation enables service providers to delegate their user management to dedicated identity providers. Thus, a service provider no longer has to manage individual user accounts. SAML further provides single-sing-on solutions: instead of providing user credentials to each single service individually, the user logs in only once at his home IdP.

Normally, SSO authentication is performed when the user initially accesses a SAML-enabled service with his web browser. Instead of providing the service with username and password, the user is redirected to his home IdP and will be presented with a login page of his home institution. Once the user is identified and authenticated, his browser is redirected to the requested service. This time the browser holds a secure SAML assertions that the service provider can consume and verify. The service may then decide to grant or deny access based on the user's attributes transported by the SAML assertions. The complete sequence is depicted in Figure 2(a).

Additionally, after authentication by the IdP, the browser holds a cookie containing a session ID, enabling single-sign-on for cooperating sites. If the browser requests another service, it is redirected to the home IdP. However, this time the IdP can identify the user based on his browser's cookie. Immediately the browser is redirected back to the service, holding a new set of SAML assertions tailored for the specific service provider. Single-sing-on, as depicted in Figure 2(b) is designed to work in absent of user interaction. As no



(a) First visit.

(b) If the user is already authenticated by the IdP he is immediately redirected back to the requested service.

Figure 2: SSO authentication with SAML.

information like entering the password is requested, the user most likely won't notice the cascade of redirects and will only experience seamless and personalized services.

4.1 ECP and Delegation of Rights

Accessing complex research data within a federated and distributed system requires services to interact with each other using the context of an individual's identity. For services to act on behalf of a user's identity, the IdM-System has to support the delegation of rights and identities. The so-called Enhanced Client or Proxy (ECP) profile [Can09] as part of the SAML standard facilitates the required functionality [6]. This profile allows for an SP to relay an IdP-Statement regarding a user's identification to a second SP. For this, the user's password is not shared between the entities involved. Instead the user's identity is transmitted through secured assertions. The process is depicted in Figure 3.

The complete cycle of accessing and rendering a digital object starts from the user's perspective by accessing the memory institution catalogue system which holds a reference to the intended object. The user logs on with his home IdP using the aforementioned SAML web SSO login. Additionally to the standard login process, the memory institution already requests a secure token that allows to authenticate back to the IdP as the current user. The IdP provider checks if the memory institution is allowed to delegate the user's right and answers with the secure token as additional SAML assertions. Steps 1-6 in figure 3 show the delegation of rights in addition to the common user login procedure.

Furthermore, depending on the digital object, the memory institution discovers the required runtime environment to display its content and requests access to it on the software-archive. The archive's request for authentication is relayed by the memory institution using its secure token to the IdP. Based on the secure token, the IdP can recognize the memory institution's right to act on behalf of the user and signs the archive's authentication request.

[6] SAML Specifications, http://saml.xml.org/saml-specifications, (20/1/2012)

The memory institution forwards the IdP's answer to the software archive and thus gets access to the requested software components. Steps 7-12 in Figure 3 illustrate this process. The same procedure is repeated for accessing and presenting the emulation service (depicted as steps 13-18 in Figure 3).

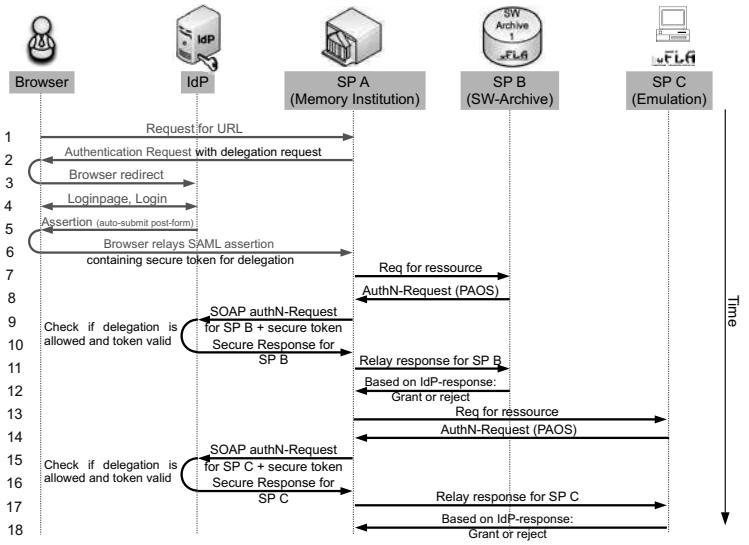


Figure 3: Delegation of rights within the SAML workflow. Standard user login through web SSO system is shown in gray color. The memory institution asks for the right to act on behalf of the user (delegation request).

As a proof of concept implementation a Pluggable Authentication Module (PAM) was implemented within the bwIdM projects [SZ12], [SWS+12] to demonstrate the capabilities and functionalities of the ECP specifications. This PA-Module was installed and used for non-web-based services, e.g. bwGRiD project[7], a nationwide federated high performance computing grid. Even in such a kind of uniform and homogeneous environment it was not obvious and unproblematic to define a set of user attributes to give the SP the ability to make exact and well-defined authorization decisions for each user. For more complex and federated e-Science workflows there is need of "hybrid" approaches.

4.2 Hybrid Authorization in Distributed Systems

With ECP a suitable technique for distributed user authentication has been identified. However, authorization is needed after authentication. This step is usually performed by each service provider itself. Each service provider may decide to grant or reject access based on the attributes provided by the IdP. Unfortunately, in a federated system, the understanding of user attributes may differ, thus making it hard to derive access rules.

[7] bwGRiD Standort Freiburg, http://www.bwgrid.uni-freiburg.de, (20/01/2013)

Common cultural, organizational and legal understandings do not apply especially for services, which may be spread over the world. The meaning of attributes may be subject to to regional differences as well as cultural, organizational and legal specialties. Even simple things as name and surname of a user may be ambiguous in countries where naming schemes differ. When it comes to more complex attributes like affiliation or employee status, common concepts are hard to describe.

Generalization of attributes is required, at least to a certain extent, to be able to cope with this problem set[8]. With only a small subset of possible attributes a large user-base is already covered. A broad attribute set would however be needed to map only the "tip of the iceberg" of all possible user rights, as Figure 4 depicts. Still, a service can grant authorization based on attributes but, as not all possible user roles can be covered by a commonly shared set of attributes, SPs may want to employ a local set of authorization rules. This is best expressed by a hybrid authorization model which utilizes different levels of generalization until individuals can be identified a service locally (pseudo anonymous), if needed. One SP may allow all users with a valid e-Mail address to access its service while another service allows only users with "student" in their entitlement attribute set. Yet another provider may, for example, only allow users that payed for a specific service identified by their userID and institution's name.

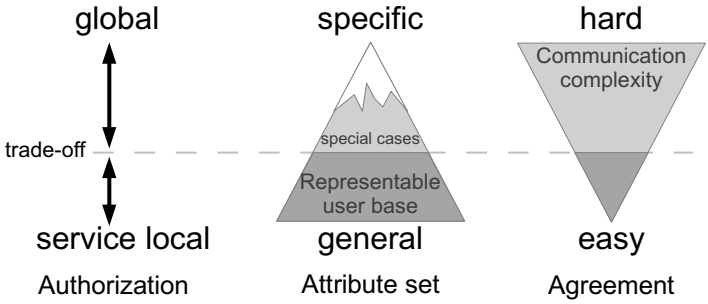


Figure 4: Trade-off between easy to implement authorization rules but a hard to manage set of user attributes and service local set of authorization rules but easy to manage set of user attributes.

5 Conclusion and Outlook

We identified current challenges for identity management systems coping with distributed services outside of the scope of the user's identity provider and proposed a solution using existing infrastructure based on SAML. As an example of a distributed data access model, we presented a workflow to access and render complex scientific data. In this example, the various service providers involved must interact on behalf of the requesting user in order to present seamless and efficient data proliferation and its processing.

[8]The eduPerson and eduOrg schemata should be considered as a generalization for widely-used person and organizational attributes in higher education: http://middleware.internet2.edu/eduperson, (20/1/2013)

This delegation of rights between heterogeneous services is a new key challenge for identity management systems. Today's infrastructures, like SAML-based implementations with its ECP-profile, already bring along the required functionality – however, this functionality seems not to be fully utilized. Furhtermore, with distributed services and identity management the challenges of distributed access control arise. We have discussed an authorization model that partly depends and generalizes on user attributes but also allows service local sets of authorization entries that can identify and manage access rules down to individual users. We see this as necessity as it will be hard to communicate and to plan a set of user attributes which can map all possible user roles.

References

[AHS03] Michael Amberg, Markus Hirschmeier, and Deniz Schobert. DART - Ein Ansatz zur Analyse und Evaluierung der Benutzerakzeptanz. In Wirtschaftinformatik Proceedings, 2003.

[BG11] Jim Basney and Jeff Gaynor. An OAuth service for issuing certificates to science gateways for TeraGrid users. In Proceedings of the 2011 TeraGrid Conference: Extreme Digital Discovery, TG '11, pages 32:1–32:6, New York, NY, USA, 2011. ACM.

[Can09] Scott Cantor. SAML V2.0 Condition for Delegation Restriction Version 1.0. Technical report, OASIS Security Services TC, 2009.

[CKPM05] Scott Cantor, John Kemp, Rob Philpott, and Eve Maler. Assertions and Protocols for the OASIS Security Assertion Markup Language (SAML) V2.0. Technical report, OASIS, March 2005.

[DA12] Janet Delve and David Anderson. The Trustworthy Online Technical Environment Metadata Database – TOTEM. Number 4 in Kölner Beiträge zu einer geisteswissenschaftlichen Fachinformatik. Verlag Dr. Kovač, Hamburg, 2012.

[Fre08] Beverly Freeman. OpenID: One Key, Many Doors. Technical report, Yahoo!, July 2008.

[GSvS08] Tarik Gasmi, Gerhard Schneider, and Dirk von Suchodoletz. Von der Accountverwaltung zum erweiterten Identity Management. In Heinz-Gerd Hegering, Axel Lehmann, Hans Jürgen Ohlbach, and Christian Scheideler, editors, GI Jahrestagung (2), INFORMATIK 2008, Beherrschbare Systeme - dank Informatik, Band 2, pages 589–595. GI, 2008.

[HLK+11] Kevin Hamlen, Peng Liu, Murat Kantarcioglu, Bhavani Thuraisingham, and Ting Yu. Identity management for cloud computing: developments and directions. In Proceedings of the Seventh Annual Workshop on Cyber Security and Information Intelligence Research, CSIIRW '11, pages 32:1–32:1, New York, NY, USA, 2011. ACM.

[MCW08] Robin McKenzie, Malcolm Crompton, and Colin Wallis. Use Cases for Identity Management in E-Government. IEEE Security and Privacy, 6(2):51–57, 2008.

[RvSW10] Klaus Rechert, Dirk von Suchodoletz, and Randolph Welte. Emulation based services in digital preservation. In Proceedings of the 10th annual joint conference on Digital libraries, JCDL '10, pages 365–368, New York, NY, USA, 2010. ACM.

[RVvL12] Klaus Rechert, Isgandar Valizada, Dirk von Suchodoletz, and Johann Latocha. bwFLA - A Functional Approach to Digital Preservation. PIK - Praxis der Informationsverarbeitung und Kommunikation, 35(4):259–267, 2012.

[SB12] San-Tsai Sun and Konstantin Beznosov. The devil is in the (implementation) details: an empirical analysis of OAuth SSO systems. In Proceedings of the 2012 ACM conference on Computer and communications security, CCS '12, pages 378–390, New York, NY, USA, 2012. ACM.

[SBSCB06] Anna Squicciarini, Abhilasha Bhargav-Spantzel, Alexei Czeskis, and Elisa Bertino. Traceable and automatic compliance of privacy policies in federated digital identity management. In Proceedings of the 6th international conference on Privacy Enhancing Technologies, PET'06, pages 78–98, Berlin, Heidelberg, 2006. Springer-Verlag.

[SWS+12] Michael Simon, Marcel Waldvogel, Sven Schober, Saher Semaan, and Martin Nussbaumer. bwIDM: Föderieren auch nicht-webbasierter Dienste auf Basis von SAML. In Paul Müller, Bernhard Neumair, Helmut Reiser, and Gabi Dreo Rodosek, editors, DFN-Forum Kommunikationstechnologien, 5. DFN-Forum Kommunikationstechnologien: Verteilte Systeme im Wissenschaftsbereich, pages 119–128. GI, 2012.

[SZ12] Saher Semaan and Richard Zahoransky. bwIDM: Anbindung nicht-webbasierter IT-Infrastrukturen an eine SAML/Shibboleth-Föderation. In 8th Joint BFG/bwGRiD Conference & Workshop, 2012.

[VvdH06] Remco Verdegem and Jeffrey van der Hoeven. Emulation: To be or not to be. In IS&T Conference on Archiving 2006, Ottawa, Canada, May 23-26, pages 55–60, 2006.