XSIM Final Report:
Modelling the Past and Future of Identity Management for Scientific Collaborations

US Department of Energy Next-Generation Networks for Science (NGNS) program
Grant No. DE-FG02-12ER26111

DOE Award Number: DE-SC0008874
Grantee name: Indiana University
Period of Performance: 09/01/2012 – 08/31/2016

Robert Cowles, Craig Jackson, Von Welch (PI)
Center for Applied Cybersecurity Research
Indiana University

Abstract

The eXtreme Science Identity Management (XSIM[1]) research project: collected and analyzed real world data on virtual organization (VO) identity management (IdM) representing the last 15+ years of collaborative DOE science; constructed a descriptive VO IdM model based on that data; used the model and existing trends to project the direction for IdM in the 2020 timeframe; and provided guidance to scientific collaborations and resource providers that are implementing or seeking to improve IdM functionality. XSIM conducted over 20 semi-structured interviews of representatives from scientific collaborations and resource providers, both in the US and Europe; the interviewees supported diverse set of scientific collaborations and disciplines. We developed a definition of "trust," a key concept in IdM, to understand how varying trust models affect where IdM functions are performed. The model identifies how key IdM data elements are utilized in collaborative scientific workflows, and it has the flexibility to describe past, present and future trust relationships and IdM implementations. During the funding period, we gave more than two dozen presentations to socialize our work, encourage feedback, and improve the model; we also published four refereed papers. Additionally, we developed, presented, and received favorable feedback on three white papers providing practical advice to collaborations and/or resource providers.

1 Introduction

Identity management (IdM) is the practice of creating and maintaining digital identities (composed of an identifier and attributes) and conveying those identities in a trustworthy manner, such that relying entities have some assurance about with whom (or what) they are communicating and providing access. IdM processes allow relying entities to make informed, confident decisions regarding, for example, how to service requests, log activities, and respond to security incidents.

In the early days of scientific computing, resource providers (RPs) had an unmediated relationship with their user communities, and therefore handled all aspects of identity management. We refer to this as the classic model of IdM. As scientific collaborations increased in both number of people and magnitude of computing requirements, they needed to obtain resources from multiple RPs. The concept of a virtual organization (VO) emerged to coordinate the scientific collaboration and its relationship to the multiple RPs serving it. The distributed and heterogeneous nature of the scientific computing resources, and the unique position of the VO in negotiating and managing community relationships resulted in new opportunities and challenges for IdM. In the DOE sciences community's two decades of experience implementing VOs, a number of IdM approaches have been used. An initial objective of the eXtreme Scale Identity Management for Scientific Collaborations

[1] http://cacr.iu.edu/collab-idm

(XSIM) project was to develop an evidence-based, descriptive IdM model that could describe this variety of implementation, and provide insight and heuristics into the factors favoring one implementation over another. Iterating on this model, XSIM worked towards its goal of providing practical advice to VOs and RPs on designing and optimizing IdM implementations fit for their particular needs.

The goal of the XSIM project was to develop a model for IdM to apply generally to scientific collaborations where the scientists are potentially distributed among universities, DOE National Laboratories, and research institutes around the world; the model had to be sufficiently flexible to address how IdM for scientific collaborations could interoperate with existing national (*e.g.,* US Federal PKI[2]) and international (*e.g.,* Interoperable Global Trust Federation[3]) IdM standards.

Foundational to our model is the idea that RPs can and do delegate IdM responsibilities to VOs classically carried out by the RP. VOs play an important role in brokering relationships between scientific communities and RPs, and in the context of IdM can (and often do) play some role in setting up mediated trust relationships between RPs and users. The VO is in a position particularly well suited to take on the user interface role since it is more likely to understand their needs than the RPs. By delegating IdM functions, RPs can also reduce their administrative overhead and, for users, the delegation means they only have to interact with the VO and not with the possibly large number of RPs.

2 Methodology and Project Timeline

2.1 Research Methods

We conducted interviews with the goal to obtain both subjective and objective information regarding identity management implementations across a broad range of VOs[4] and RPs[5] on which to form our VO IdM Model. We developed an semi-structured interview process [eSci], and focused our questions about the following topics:

- governance and stakeholders - what parties had influence over the IdM choices;
- assets, risks, and threats - what were the biggest concerns of the parties involved;
- user management - what were the processes of vetting, enrolling, authenticating and authorizing users;
- incident handling - how were exceptional cases handed when users needed to be contacted;
- lessons learned - what worked well and what would be changed if it could be

[2] http://www.idmanagement.gov/federal-public-key-infrastructure
[3] https://www.igtf.net/
[4] ATLAS, BaBar, Belle-II, CMS, Darkside, Engage, Earth System Grid, Fermi Space Telescope, Fusion Collaboratory, LIGO, LSST/DESC
[5] ATLAS Great Lakes T2, Blue Waters, CERN, FermiGrid, GRIF/LAL, Jefferson Lab, LCLS, LLNL, U. Nebraska (CMS), NERSC, NIKHEF, ORNL, Rutherford-Appleton Lab

re-implemented?

We supplemented the interview results with published papers, presentations, and articles about the interviewed VO or RP. In order to promote honest discussion, the interviews were not recorded and the detailed notes were taken by the authors in confidence.[6]

Following the interviews, we undertook an iterative process to form a descriptive model that best fit the our data. Our goal was to find a model that allowed for both the easy and clear expression of data from all interviews, and could be leveraged to provide guidance in designing a VO IdM implementation. Based on our collective experience, we initially selected, , an initial set of parameters and possible values. We then attempted to match our interview data to those parameters, and then iteratively refined the parameters and values to improve the quality of the matches. Our initial publication (see Timeline below for references), presented at the 9th IEEE International Conference on eScience in 2013, established a simple VO identity model that expressed the VO-RP relationship in terms of the amount of delegation of responsibility for IdM from the RP to the VO. In subsequent work, presented at the 20th International Conference on Computing in High Energy and Nuclear Physics (CHEP2013), we began exploring the motivations that VOs and RPs have for these delegations. It identified the following factors: the need to provide isolation among users; persistence of user data at the RP; complexity of VO roles; cultural and historical inertia; scaling in terms of the size of the VO and number of RPs; and the RP's incentive to support the VO. In the paper presented for the International Symposium on Grids and Clouds (ISGC 2014), we describe our additional interviews, refinements to our VO IdM model, influential factors in applying a transitive trust approach, and conclude with a NERSC use case illustrating and applying our refined model. For the 2015 Workshop on Changing Landscapes in HPC Security (CLHS'15), we presented how, particularly for US DOE Labs, existing policies allow the delegation of IdM functions to collaboratories within the context of acceptable risk management. We suggested strategies that allow for the incremental increase of trust and delegation of IdM functionality.

In all, a total of four iterations of model development were required. The quality was determined subjectively by the authors based on our combined 60+ years of experience in distributed computational science, cybersecurity and identity management.

2.2 Timeline

2.2.1 Describe the approach, obtain Interviews (2013)

We made presentations at various meetings and conferences to explain the goals of XSIM, the methodology we were using, obtain feedback on our approach, and foster adoption of our work. At the same time, we used to opportunity to approach knowledgeable people and request interviews as part of our data collection. Meetings included: the Open Science

[6] While no sensitive or confidential information was exchanged in the interviews, this seemed to relax the interviewees as it lessened the feeling of the need to be guarded in their responses.

Grid (OSG) All-Hands meeting (VO and resource providers in the US); HEPiX (representatives of resource providers in the US and Europe); EUGridPMA (identity and resource providers in Europe); Vo Architecture and Middleware Planning (VAMP) (virtual organizations in Europe); NGNS-PI (NSF-funded researchers).

2.2.2 Presentations of initial results (2013)

In October 2013, we presented our initial interview results and had refereed papers accepted for publication by eScience 2013 [eSci] and CHEP 2013 [CHEP].

2.2.3 Additional interviews and test early model (2014)

In the latter part of 2013 and early 2014, we developed an initial model and performed additional interviews. A number of presentations were made to obtain feedback on the results obtained to date. Meetings included: HEPiX (representatives of resource providers in the US and Europe); EUGridPMA (identity and resource providers in Europe); TAGPMA (identity and resource providers in North and South America); LBNL and NERSC; OWASP (application security community); National Labs Information Technology exchange (NLIT) 2014 (technology experts from the DOE national Labs); NGNS-PI (NSF-funded researchers). We had additional discussions at other conferences and meetings such as XSEDE, SC14, Federated Identity Management for Research (FIM4R), Security for Collaborating Infrastructures (SCI), and Terena Networking Conference (TNC14).

2.2.4 Fully developed model (2014-2015)

Through several more iterations of the model, we provided updated presentations at OSG and EUGridPMA. Presentations were made at CERN, PNNL, The Networking Conference (TNC15), and MAGIC. We developed a whitepaper to address the issues often confronted at DOE Labs [FSC1] when adopting a transitive trust model. A high-level discussion of those issues were presented at a meeting of the National Labs CIOs (NLCIO). Both presentations and refereed papers were developed for the International Symposium on Grids and Clouds (ISGC) 2014 [ISGC], and Changing Landscape in HPC Security (CLHS) 2015 [CLHS].

3 Accomplishments

3.1 Technical Accomplishments

3.1.1 A functional definition of trust incorporating the role of risk

Trust is a foundational concept in IdM, and poses particular challenges in multi-party trust relationships like the ones we studied. As we proceeded through the interview process and saw increasing levels of delegation, it was clear the that trust was linked to and hence a critical factor for successful delegation; therefore, to ground our work, we needed a working definition of trust based on prior research and suitable for the DOE science community. Trust is a complex concept, and is subject to myriad definitions informed by work in fields including philosophy, sociology, law, psychology, and information theory. Even within the field of information security trust has a variety of definitions.

In studying this prior work, we based our definition for both terms on the recent risk theory work of Philip J. Nickel and Krist Vaesen on the relationship between trust and risk [Risk]:

"Trust is a disposition to willingly accept the risk of reliance on a person, entity, or system to act in ways that benefit, protect, or respect one's interests in a given domain."

This definition captures the relationship we see between RP and VO with regard to IdM; that is, the more the RP chooses to delegate to the VO in terms of IdM, the more trust it has in the VO and the more risk it accepts in VO relationship (risk may be reduced in other areas).

3.1.2 An evidence-based, descriptive VO-IdM model

We now present our refined VO IdM model. Our final model introduced two significant refinements based on the interview data and our desire to maximize the model's clarity: (a) We decomposed the data into three basic information types common to VOs, and (b) we introduced the notion of *producers* and *consumers* of the information. The model is based on the *types* identity information, the *entities* and *functions* that produce and/or consume that information, and the information *flows*.

3.1.2.1 VO identity information types: data for supporting scientific workflows

Our initial model considered a single flow of user-centric identity information. In analyzing the results from our interviews, we noted there actually exist three different types of user information that are commonly produced and consumed in the context of VO-IdM:

- ***Digital Identifier:*** That is, an identifier of the scientist/VO member issued by an IdP. Examples of this information type include an X.509 distinguished name, an eduPerson Principal Name (ePPN) in a SAML assertion and a username.
- ***VO Membership & Role:*** Minimally, each VO tracks information about who is a member of the VO. For example, attribute data as captured in Virtual Organization Management Registration Service (VOMRS) system[7]. Some VOs have richer expressions of membership that include a scientist's role(s) and privileges in the VO.
- ***Contact Information:*** Often, but not always, contact information (e.g., email address, phone number, postal address) is collected from a scientist.

The latter two types of information can be, and often are, referred to as "user attributes"; and certain attributes are included in the VOMS Attribute Certificate associated with some grid requests. However, we distinguish them because, as we describe subsequently, they are often generated by different parties and utilized for different purposes.

[7] http://www.fnal.gov/docs/products/vomrs/

A reader familiar with IdM also will notice we do not include other types of attributes (*e.g.* a scientist's institution and their role and department at that institution). Our interviews have not revealed evidence of these attributes being in common use in the VO context. There are, at least, two possible reasons for this: (1) There is a lack of demand for this information, that is, it is not useful to VOs or RPs; and/or (2) Sufficient information is available by using clues from the email address or the authentication domain associated with the Identity Provider (IdP).

3.1.2.2 Identity production and consumption: functions enabled by IdM

Early iterations of our model focused on transmissions of identity information at stages of a VO user's lifecycle, and did not account for the multiple purpose-driven flows of specific types of identity information in VO-RP relationships. What that earlier iteration gained from simplicity it lost in utility as we began working with VOs to address the mechanics of designing or evolving their IdM implementations. We found it necessary to evolve our model to account for this complexity. For example, consider the multi-user pilot job factory: VO membership information is used to authorize a request, the user's identity information is recorded for audit purposes, and contact information is collected and retained for incident response (e.g., user support or security investigations).

Our refined model reflects our observation that identity data is, similar to any data, produced, stored, transformed, transmitted and consumed. As such we turned to the concept of Data Flow Diagrams (DFD) [DFD1] to model the flow of the three types of identity data in the context of VOs. While DFD offers a rich framework, we borrow its simple concepts of entities being *sources* and *terminations* of identity data, though we use the terms *producer* and *consumer* as we believe they more clearly convey the process in the IdM context.

Identity information is produced by administrative action by an entity. In the VO context, producers may include the VO, the RP, or (introducing a new, but well recognized party to our model) an IdP. Each of the three types of identity information can be produced by any one of these three parties. Production may entail generation of previously non-existent information (creating a username) or conversion/translation of existing information into digital form (e.g., recording contact information). For example, some common patterns in the VO context are:

- Identity is produced by an identity provider when a credential is generated for the user.
- VO membership information is produced by a VO when the user successfully applies for membership.
- Contact information is collected by the VO when VO membership is granted to the user.

Flowing from a producer, identity data may arrive at one or more consumers who use the data for the purpose of providing some service. We have identified seven common *functions* supported by identity information in the VO context (see the figures below):

- ***Authentication***. Consumes externally provided identity information and produces an internally trusted identity/attribute "bundle" for use by other functions.
- ***Authorization***. Consumes identity information (identity, VO membership/role) to implement access controls on resources.
- ***Allocation / Scheduling of resources.*** Consumes identity information (identity, VO membership/role) to make decisions regarding how to allocate or schedule resources to service a request.
- ***Accounting.*** Consumes identity information to account for resource consumption.
- ***Auditing.*** Consumes and records identity information to allow for the proper decision making regarding a request and to provide information in case user support or incident response is necessary.
- ***User Support.*** A typically manual process that consumes identity information in order to communicate directly with the user initiating a computing workflow in order to resolve some apparent malfunction.
- ***Incident Response.*** A process (typically manual) that consumes identity information in order to communicate directly with the user initiating a computing workflow in order to resolve a possible security violation.

All consumption may take place at the RP, or, for a function that has been delegated, at the VO. The location of the production and consumption is an indicator of whether responsibilities have been delegated to the VO; information flows between the producers and consumers serves to show what identity information is used for a particular function.

We note that the Data Flow Diagramming methodology can scale to much more complex system descriptions than we set out here, and the method includes a number of concepts that may prove useful in to setting out the fine details of an IdM system design. For example, it defines *stores* (roughly equivalent to databases and credential stores), and processes which can transform data (which seem equivalent to security token services such as CILogon [ref]). This assures us that the Data Flow Diagram can be used to reflect highly-complex identity flows if needed, but we resist incorporating them into our model until proven necessary in order to keep it simpler.

3.1.2.3 Example applications of the model

In Figure 1, we show the simplest possible example: a classic implementation with the RP handling all identity management. In this case, two types of identity information are produced and consumed by the RP (and, there is no VO membership information since there is no VO).

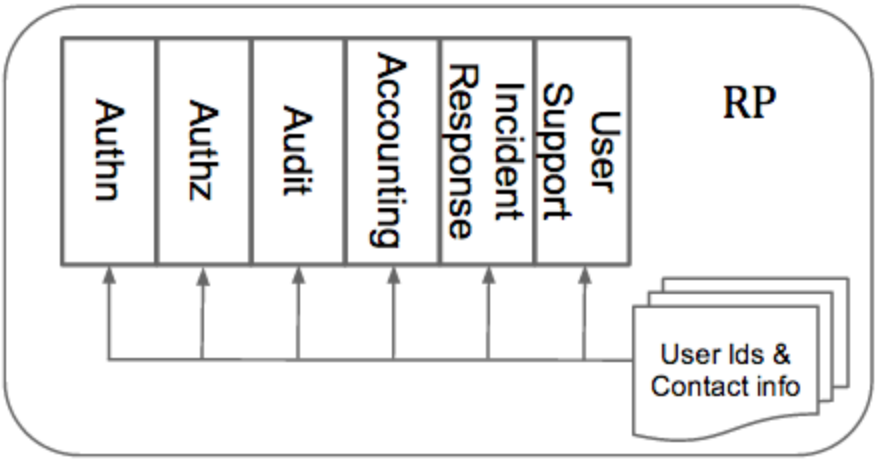

***Figure 1**: The classic model with RP handling all identity management.*

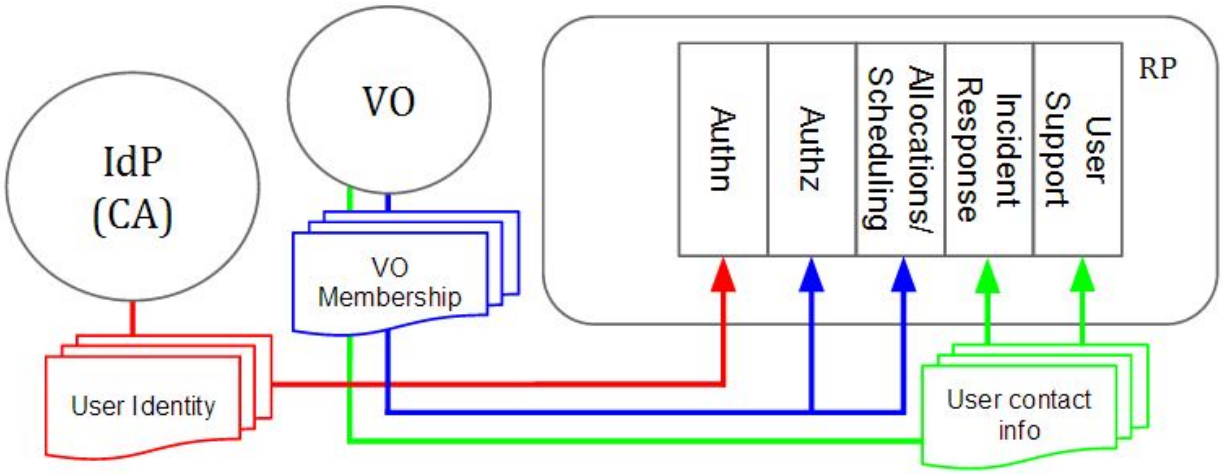

***Figure 2**: An example of a multi-user a job factory expressed in our VO IdM Model. User identities come from an identity provider (a certificate authority) are used to authenticate the user's compute job, VO membership information is used by the RP to authorized and allocate resources for the job, and the user's contact information from the VO is used by the RP in the event there is an incident or user support needs to be undertaken.*

Figure 2 shows a more complex model where the RP has retained responsibility for identity management consumption and associated functionality, but has delegated production of user identity to a certificate authority, and determination of VO membership and collection of user contact information to the VO. We believe our model supports the clear expression a complex implementation. It not only conveys the flow of identity information, but also allows for ready inference of the trust relationships and delegations involved.

3.1.3 Factors impacting RP-to-VO IdM delegation

We now turn to the factors we found commonly influence whether and what identity management responsibilities are delegated from RPs to VOs and IdPs. We felt it was important to identify these factors for two reasons: (1) They tended to come up in the

interviews; (2) they provide a richer factual picture of the relationships and decision-making process; and (3) they provide a lens through which we can offer informed, actionable advice to new or evolving RP/VO trust relationships.

3.1.3.1 Motivations for delegation

Scaling and Dynamicity of the VO. Scale can affect the VO/RP relationship in two main ways: The number of RPs involved and the number of users (both total and in terms of turnover) involved in the VO may motivate the parties to delegate production of IdM identity to the VO or an IdP, where it will be centralized instead of replicated at multiple RPs. Aggregate identity management effort is roughly O(#RPs x #Users) if all the IdM for a service is done at the RP. The more control is centralized to the VO, the more the number of RPs drops out of this equation bringing the effort down to O(#Users). The amount of effort based on #RPs is initially very steep, but once it exceeds a handful of RPs, the mechanisms are typically in place to support a much larger number. We note that with the inclusion of identity providers in our model, this factor is the main factor influencing the use of a third party identity provider.

Complex VO Member Roles and Privileges. The more heterogeneous the privileges of different VO users, the more complex the access control policies will be and, if RPs are responsible for enforcing those policies, the more complex the communication between the VO and RP will need to be to communicate the policy and necessary information to enforce it. Hence greater complexity of VO roles tends to push authorization functionality to the VO.

VO-wide Collaboration Services. Many VOs have the need to have provide services that support collaboration to their communities: e.g., forums for communication, source code repositories for development, means for sharing and collaboratively analyzing data. Since operating these services requires both effort and identity information (to authenticate and authorize users), this encourages RPs to delegate identity information consumption to the VO so that it can take on this effort.

Alignment with RP's Mission. RPs have their own missions, often heavily influenced by the missions of their funding agencies. In the context of scientific VOs, typically RPs are generally motivated to help VOs achieve science results, though they may be more strongly motivated by VOs tightly aligning with their missions or when specifically funded to help a particular VO. Commercial RPs (e.g., cloud providers) are primarily motivated by payment.

3.1.3.2 Enablers of delegation

This set of factors serve to reduce the barriers to the delegation (i.e., reduce the amount of motivation needed from the first set of factors), but do not themselves motivate delegation.

Established Trust Relationships. When the RP has an established trusting relationship with the VO, this reduces the barriers to the delegation. Examples include a history of prior collaboration, the VO being closely associated with the RP organizationally, and a reputational history of trustworthy VO behavior with other RPs.

Available VO IT/IdM Effort and Expertise. A VO's available IT staff time and expertise in running services (IdM services in particular) is a straightforward, but critical enabler of delegation. A VO that is highly capable, or at least on par with the RP, makes delegation easier. VOs without members with IT expertise, or interest in operating IT services, naturally dissuade delegations of IdM to them.

Availability of Traceability Mechanisms. Increasingly, traceability [EndU] -- i.e., the ability to trace events back their initiator on an as-needed basis to facilitate user support and security incident response -- is a viable and in-demand mitigation against the reduced RP real-time visibility into user identity that comes along with increased IdM delegation.

3.1.3.3 Barriers to delegation

Historical Inertia and Introduction of Risk. For RPs with a history of doing their own identity management, the delegation of identity management will often require some time for acclimatization. This may also be true for RPs' funding agencies or other stakeholders who set policy for them. These entities frequently have formal policies, informal cultures, and respected reputations around information security and risk which have evolved over time. Delegating even a portion of the information security domain means a change in risk profile, as "decisions to establish trust relationships are expressions of acceptable risk." [NIST] Our recent interviews with supercomputing centers have reinforced the validity and importance of this factor. We observe these RPs taking more conservative steps, and beginning to delegate IdM to VOs once implementations have proven viable and benign in other settings.

Compliance and Assurance Requirements. IdM-related compliance and/or assurance may present barriers to delegation. Strength of authentication, traceability, auditing, and accounting may be critical responsibilities, and usually lie with the RP by default. Note that external stakeholders of the RP and VO must often be considered here. Stakeholders of RPs, in particular, tend to introduce higher requirements for IdM. There has been some recent relaxation of requirements in recognition, in some cases, that the identity requirements were for persistence and/or valid contact information rather than traceability to a legal identity.

Technology Limitations. The technologies (e.g., software stacks) to be used in the VO/RP context must be considered. Many contemporary tools require identity to function, but allow only for authenticated individual access (e.g., an individual logging in with a username and password or certificate), access by an undifferentiated group to an individual user account, or anonymous access (e.g., a public website, a read-only data server). Some

technologies have been extended to allow access by a group to an individual user account while carrying information about the individual user to the RP. For example, this is what VOMS does by embedding a VO credential in a batch job request. The less sophisticated the technologies in terms of their IdM support, the more effort is required to distribute IdM functionality between two parties and hence encourages IdM to be concentrated at one party or the other (typically the one that is more resourced).

3.1.4 The historical trend toward transitive trust and potential for incremental implementation

Our interviews revealed a historical trend of resource providers increasingly delegating more and more IdM identity production and consumption activities to VOs. Figure 3 depicts an increasingly common and often desirable approach for VO identity management (IdM): transitive trust. In this approach, the VO manages its community and RPs trust the VO to do so with little-to-no cognizance of the individuals, seeing them only as a members of the VO community. As such, transitive trust implementations represent the most extensive feasible delegation of IdM.

This approach has become desirable because it produces a clear separation of responsibilities between the VO and RP, establishes a simpler workflow, and reduces administrative overhead in relatively low risk environments. The VO does not need to communicate information about all of its members to the RPs and can utilize any mechanism for managing their identities it desires (i.e., a mechanism that meets any agreed-to assurance level with the resource providers). Transitive trust relationships can provide significant benefits in international collaborations where user privacy and personal data location requirements would otherwise constitute barriers to science.

Figure 3: Transitive trust approach with VO managing its members and RPs trusting the VO to do so.
No direct trust relationship between members and RP.

However, the transitive trust approach does introduce issues that should be considered:

- ***Lack of persistent personal data storage at resource provider.*** Data that is either shared by the VO or temporary to a specific compute job can be stored by an RP readily because the lifetime of the data corresponds to the lifetime of the VO or compute job respectively. However, storing persistent data that is private to individual VO members is a challenge because the RP isn't aware of individual VO

members from an IdM perspective. Addressing this challenge typically entails the VO arranging for any member-specific data to be migrated between VO and RP storage before and after a computational job (often referred to *staging* of the data).

- ***VO-hosted collaborative services.*** Many collaborative services (e.g., source code repositories, discussion forums, data storage) expect user identities to function. Because resource providers are not participating in individual user management, these services need to be hosted by the VO or a party (perhaps a individual resource provider) acting on the VO's behalf. Where the VO provides a portal as its primary user interface, this is a common place to host such services.
- ***User support and incident response coordination.*** During the course of normal operation, unexpected or adverse events will happen at the resource provider in servicing requests from the VO. Because the resource provider has no ability to contact individual VO members, RP and VO should have an agreement in place to handle these events -- e.g., an expectation that the resource provider can report events to the VO, who will handle them in some reasonable amount of time. OSG Document 1149 [OSG1] explains the security requirements to execute user jobs submitted without an end user certificate.

While transitive trust is increasingly common and has substantial benefits, our research reveals that IdM delegation from RP to VO is not an all-or-nothing affair. Partial delegation can establish a simpler workflow with greatly reduced administrative overhead (for both RPs and users) and provide greater assurance by placing the responsibility with the entity best suited to address it.

3.2 Actionable Guidance for the DOE Community

After we felt confident in the model, we developed white papers aimed at giving advice on identity management to (1) virtual organizations joining OSG [OSG2]; (2) the Dark Energy Science Collaboration (DESC) (associated with LSST) [DESC]; and (3) DOE Labs [FSC1], the latter with particular emphasis on removing perceived barriers to delegating IdM functions -- and promoted that white paper by making presentation to a meeting of the National Labs CIO organization (NLCIO).

4 Lessons Learned at the Project Level

4.1 Successful Innovations

4.1.1 Composition of the team

The project team had varied backgrounds that both overlapped and complemented each other in a manner that contributed greatly to its success. Areas of expertise included Open Science Grid and supercomputer facilities; DOE cyber security environment; project management; law; philosophy; social science research methods; LHC Grid Security; academic paper writing, and Global PKI requirements for science. The team had strong

connections to resources in both the US and Europe; to educational institutions, NSF funded facilities and DOE Labs.

4.1.2 Emphasis on knowledge rather than code

With a significant amount of technical development in progress around IdM and years of applied experimentation, producing one additional code product did not seem useful. Rather than produce a final product consisting of code that was likely to disappear once the resources ran out to further develop or maintain it, we decided to produce an evidence-based "knowledge product" that could be used to alter the way a broad spectrum of developers think about IdM.

4.1.3 Comprehensive and comprehensible model

Our early descriptive models either were inadequate to describe the observed variations in our research data or were too complicated. We searched for a model that was comprehensive, but still simple enough to explain and use in novel situations. It was important that the model be comprehensible to both researchers and IT/Cyber security experts to support a dialog between stakeholder groups with different lexicons.

4.1.4 Evidence-based research

Rather than developing an idealized model only looking at the future, we mapped the course of IdM as it has developed in some of the major science collaborations, taking into account the direction they were moving in the last fifteen years. Using that data, we produced a framework that fit past, present and the projected future of identity management. Grounding our model in real world historical and present data from our interviews dramatically increases our confidence in our projections.

4.2 Challenges

4.2.1 Collaboration engagement at the right time

We learned is it very difficult to engage with a collaboration at the right point in their lifecycle to create a measurable impact on that project. Too early and projects are typically still awaiting to hear about their funding and not yet engaged in technical design. Once design is underway to the point design decisions around IdM have been made, it is very difficult to motivate the revisiting those decisions.

We would have liked to see a collaboration utilize our model and guidance for its IdM implementation from start to finish of its own design and implementation during the course of our funding, however the timing for such turned out to be difficult. We attempted an engagement with DESC at SLAC, we found that we were both too early and too late for a profitable engagement. We were too late in the sense that they already had tentative plans for how their IdM was likely to work using designs and technical staff from previous scientific experiments. We were too early in the sense that various construction and funding delays meant that the first real data collection for the project was not scheduled

for 6-7 years; so, there was not a feeling of urgency to design the IdM during the timeframe of the XSIM project.

In retrospect, we would have liked to have identified a scientific collaboration early on in the project, that would have been entering the window of IdM design towards the end of our project. This would have allowed us to work in parallel to build the relationship as we build our model and guidance.

4.2.2 Reaching the target audience

The target audience for our work is technical leadership in scientific collaboration. We were challenged to find natural meetings where such congregated (as opposed to their counterparts in the National Laboratories, from whom we could identify numerous meetings to address them as a group). As such we had to engage with them individually, which required significant time and travel. DOE may want to encourage a gathering of technical leadership in its scientific collaborations to exchange experiences and hear from projects in programs such as NGNS they may benefit from.

5 Next Steps

5.1 Exploring the promise of and systemic barriers to transitive trust

Our interviews revealed a clear historical trend to greater delegation of IdM to VO's, and transitive or near-transitive trust implementations are taking hold in some corners of the community. Feedback from presentations of the model centered around two issues. First: Use of a transitive trust model where the collaborators provided contact information to the VO and the VO was then responsible for any subsequent contact, including incident response, structurally reduced or eliminated the need for the release of user attributes by identity providers and the logging of personally identifiable information by resource providers, reducing privacy and data protection concerns those provided may have had. Second: Attendees expressed significant concern that only a relatively small number of VO were actually competent to manage user registration and to follow-up on incidents. In our experience, for the case of the "long tail of science" a large number of scientific collaborations do not have the expertise to perform these functions.

5.2 Integration into collaboration supporting infrastructure

There are a number of efforts in the US, Europe, and Australia to provide virtual environments tailored to particular scientific disciplines and these environments provide almost everything in terms of IT infrastructure (including IdM) a collaboration might need to perform its work. Integration of our IdM model would allow these virtual environments to operate coherently with DOE Laboratories and other organizations. There is currently no obvious effort in the U.S., as there is in Europe by EGI and Géant, to deploy and operate such an infrastructure, meaning an entity taking this on could provide leadership in this integration.

5.3 Developing a taxonomy of scientific data and their security requirements

Risk and trust are tightly intertwined. (*See above*, 3.1.1, for our definition of trust.) Nowhere is this more apparent than in the contemporary fields of information security, privacy, and identity management. We encountered uncertainty in the level of risk involved in delegation of IdM for access to scientific collaboration, in part due to the uncertainty around the sensitivity of the data involved. We believe this lack of clarity contributes to conservative decisions around delegation, which in turn may unnecessarily inhibit scientific collaboration and discovery.

Data involved in the science and operation of scientific collaboratories can have varying requirements for confidentiality, integrity, and availability. These requirements come from various sources. Systematic research is needed to develop a comprehensive, comprehensible framework of these requirements that will ease the burden for risk and security decision makers involved in IdM delegation (and other facets of setting up a collaboratory). We are aware of no current work in this area. XSIM's structured interview, analysis, and socialization methods would be well-suited to produce an evidence-based, highly usable, and high impact framework.

References

[CHEP] R. Cowles, C. Jackson, V. Welch. Identity management factors for HEP virtual organizations. 20th International Conference on Computing in High Energy and Nuclear Physics (CHEP2013), 2013, http://www.vonwelch.com/pubs/CHEP2013.

[CLHS] R. Cowles, C. Jackson, and V. Welch. Facilitating Scientific Collaborations by Delegating Identity Management: Reducing Barriers & Roadmap for Incremental Implementation, CLHS '15 Proceedings of the 2015 Workshop on Changing Landscapes in HPC Security, 2015. https://dl.acm.org/citation.cfm?id=2752501

[DESC] B. Cowles, C. Jackson, and V. Welch (PI). DeSC Identity Management: Analysis and Recommendations. Unpublished Technical Report, August, 2014.

[DFD1] P. D. Bruza, and Th. P. Van der Weide, "The Semantics of Data Flow Diagrams", University of Nijmegen, 1993.

[EndU] A. Padmanabhan, M. Altunay, and K. Hill. Assessing Traceability of User Jobs in Absence of End User Certificates in GlideinWMS, ISGC 2014. 2014. http://pos.sissa.it/archive/conferences/210/006/ISGC2014_006.pdf

[eSci] R. Cowles, C. Jackson, V. Welch. Identity Management for Virtual Organizations: A Survey of Implementations and Model, 9th IEEE International Conference on eScience, 2013.
http://www.computer.org/csdl/proceedings/escience/2013/5083/00/5083a278-abs.html

[FSC1] R. Cowles, C. Jackson, and V. Welch. Facilitating Scientific Collaborations by Delegating Identity Management, CACR/XSIM Technical Report, March, 2015.
http://cacr.iu.edu/sites/cacr.iu.edu/files/FSCbyDIM0408.pdf

[ISGC] R. Cowles, C. Jackson, V. Welch, and S. Cholia. A Model for Identity Management in Future Scientific Collaboratories, International Symposium on Grids and Clouds (ISGC) 2014, 2014. http://pos.sissa.it/archive/conferences/210/026/ISGC2014_026.pdf

[NIST] NIST Special Publication 800-39, Managing Information Security Risk, March 2011.

[OSG1] M. Altunay. The OSG Traceability Requirements for VOs Submitting Jobs without End User Certificates, May, 2013.
http://osg-docdb.opensciencegrid.org/cgi-bin/ShowDocument?docid=1149

[OSG2] V. Welch, R. Cowles, and C. Jackson. XSIM OSG IdM Guidance OSG-doc-1199, July 2014. http://osg-docdb.opensciencegrid.org/cgi-bin/ShowDocument?docid=1199

[Risk] P.J. Nickel and K. Vaesen, “Risk and trust,” in Handbook of Risk Theory, S. Roeser, R. Hillerbrand, P. Sandin, M. Peterson, Eds. New York: Springer, 2012, pp. 858-873