Identity in research infrastructure and scientific communication

Report from the 1st IRISC workshop, Helsinki Sep 12-13 2011

Authors: Gudmundur A. Thorisson, Mikael Linden, Anthony J. Brookes, Myles Byrne, Juha Muilu and Tommi Nyronen, with input from workshop participants (listed in Appendix I)

Date published: November 16, 2011

This report was published online at:
http://irisc-workshop.org/irisc2011-helsinki/workshop-report

An executive summary derived from this report will also be published shortly.

The workshop schedule with links to speaker profiles, slides and video recordings is available at:
http://irisc-workshop.org/irisc2011-helsinki/schedule

Table of Contents

1. Introduction ... 3
1.1. Unique identifiers and identity on the Internet ... 3
1.2. Identity federations in Higher Education and Research ... 4
1.2.1. Trust ... 5
1.2.2. Access management ... 5
1.3. Identifying authors and other knowledge contributors ... 6
1.3.1. Identity and attribution ... 6
1.3.2. Author identifier systems ... 7
1.4. Common challenges and opportunities ... 8
1.4.1. Institutional validation of ORCID profile data ... 8
1.4.2. Bridging identity federations with persistent, global identifiers ... 8
1.4.3. Managing access to online biomedical services ... 9
1.5. Motivation for the IRISC workshop ... 9
2. Workshop participation and proceedings ... 11
2.1. The programme ... 11
2.1.1. Plenary sessions ... 11
2.1.2. Parallel interactive breakout sessions ... 12
2.1.3. Reporting from breakouts, discussion and wrapping up ... 12
3. Workshop results ... 13
3.1. Motivating ORCID adoption in the researcher community ... 13
3.1.1. "Killer apps" for end users to accelerate adoption ... 13
3.1.2. An expanded role for ORCID ... 14
3.1.3. Overcoming political blockage ... 14
3.2. Identity federation requirements from service providers ... 15
3.2.1. Retrieval of user information, data protection and scalability ... 15
3.2.2. Reliability of Identity and authentication ... 15
3.2.3. Attribute semantics ... 16
3.2.4. Funding model ... 16
3.2.5. Usability and user awareness ... 16
3.3. Common challenges/solutions and opportunity for collaboration ... 17
3.3.1. Biomedical data services: use cases for IDF integration ... 17
3.3.2. Interoperability between identity federations and ORCID ... 18
4. Final summary and conclusions ... 19
5. Acknowledgements ... 19
Appendix I: Workshop participants ... 20
Appendix II: Workshop session proceedings ... 22
Appendix III: Workshop schedule ... 28

1. Introduction

Reliable digital identification and authentication of researchers is increasingly relevant in all aspects of contemporary scientific research. Digital media and the Internet are used near-ubiquitously for rapid dissemination of scientific knowledge. Scientific research itself is increasingly undertaken, debated and communicated online in highly collaborative and interactive fashion.

This digital and highly networked environment is raising difficult questions concerning identity of the growing number of individuals who use and contribute to an expanding range of electronic publications and online resources as part of their scientific activities. Pragmatically, there is a need for a unique digital identifier scheme or schemes for persons engaged in those activities. This Introduction provides background on unique identifiers, identity and authentication, followed by two contrasting perspectives on those topics which motivated and provided a starting point for the IRISC workshop discussion.

1.1. Unique identifiers and identity on the Internet

In computer security, *identity* refers to a set of attributes describing an entity, often a natural person. The attributes can be descriptions of a person (such as his/her name or date of birth), his/her role in an organization (such as his/her job title or the institution or laboratory he is affiliated to) or contact details (such as email, postal mail or instant messaging address, phone number). Of particular interest are attributes which uniquely identify the end user in a particular context or scope. National identification numbers (e.g. social security numbers) are unique in one country (or a sector in a country, such as taxation). Usernames are unique identifiers in one IT system only or across several systems within an organization.

A closely related concept is authentication: that is, the process by which the service verifies the user's identity and creates a binding between the identity and the related entity in real life. Somehow, the individual convinces the authenticator that he/she is indeed the same entity whose identity is registered to the system. Passwords, tokens or biometric systems can be used for this purpose. Authentication can take place either locally against a user account on the same system or in a federated manner involving an external authentication service where the user account is hosted (the identity provider, or IdP).

Hierarchy, notably the DNS system which underpins the Internet itself, is often used to make local person identifiers globally unique (e.g. an email address *firstname.lastname@example.org*). This eases the management of the global uniqueness, because different branches of the hierarchy can be managed by different organizations independently. Hierarchies form the basis of the identity concept in the identity federations further discussed below. However, hierarchical organization introduces challenges when a person moves from one organization to another, effectively losing his/her previous identity in the process and gaining a new one. This makes organization-issued, distributed identifiers problematic for use in persistently identifying the nomadic scientists whose career path traverses several organizations.

The incompatibility of organizational identifiers and nomadic users as led to the development of "flat" models of centralized assignment and management of person identifiers which are unique across the entire Web. Early attempts were not successful; for example, Microsoft failed in their attempt a decade ago to make the Passport service (now rebranded as Live ID) into a global Web-wide identity service, in

part due to the proprietary, closed nature of the technology and security concerns[1] but not the least because of concerns over any one entity - especially a for-profit company - storing and controlling personal information for all Internet users.

More recently, “lightweight” solutions based on open, non-proprietary standards such as OpenID (federated authentication) and OAuth (delegated authorization) have been widely adopted in Web 2.0 space to create decentralized identity systems based on assigning globally-unique Universal Resource Identifiers (URIs) to persons[2]. A major driver in these developments is *user-centric identity*[3], which refers to the empowerment of Internet users to link together their accounts across the various online services they use and - critically - control which personal identifier(s) and potentially-identifiable information represents them online and how this information is shared with other parties.

The net effect of the above is that the majority of Internet users now have one or more globally-unique identifiers, assigned to them by providers of the services they use (notably Facebook and Google). Increasingly, these IDs are federated: in other words, instead of creating new user accounts, users can - in principle at least - use their existing ID or IDs to sign into a wide range of 3rd party services. Although now quite widely used in this way in primarily social networking contexts, there are numerous obstacles to using such “social IDs” in a scholarly research setting. Chief amongst these are persistence, privacy concerns, and authority of the associated information. These issues are further elaborated in the next two subsections.

1.2. Identity federations in Higher Education and Research

Proliferation of IT systems in the beginning of the last decade made the number of information systems that need authenticated end users grow also in the academic community. End users, such as researchers, teachers and students, needed to get registered and receive a new username and password from various systems in their home organization and also in services they need in their daily work/studies outside their home organizations. This started to cause usability, efficiency and security problems to the users and the organizations responsible for protecting their services.

Athens (in the UK higher education), PAPI (in Spain) and Moria (in Norway) were early adopters of federated identity management in the higher education and research. The Shibboleth project of Internet2 and the commercial sector’s Liberty alliance project started the development of standards-based federated identity management, and lead to introduction of the SAML 2.0 standard by OASIS in March, 2005. In higher education and research (HER), the first SAML-based identity federations (IDFs) were introduced in the same year by the national research and education networks, which became the advocates of federated identity in the academic sector.

In the beginning, IDFs focused mostly on serving academic publishers whose content was licensed by the libraries. Learning management systems, portals and collaboration services such as wikis were the

[1] Oppliger, R. Microsoft .NET Passport and identity management. *Information Security Technical Report* **9**, 26-34 (2004). http://dx.doi.org/10.1016/S1363-4127(04)00013-5

[2] Weitzner, D.J. In Search of Manageable Identity Systems. *IEEE Internet Computing* **10**, 84-86 (2006). http://dx.doi.org/10.1109/MIC.2007.95

[3] Maler, E. The design of everyday identity. *Online Information Review* **33**, 443-457 (2009). http://dx.doi.org/10.1108/14684520910969899

next to follow. Recently, also researcher services and infrastructures have started to show interest in IDFs[4].

Currently, IDFs are tend to be national in scope, with slight technical and policy differences between countries. REFEDS is a global communication and collaboration forum for the IDF community. The EC-funded eduGAIN project is bridging together the various national IDFs into a global authentication and authorisation infrastructure (AAI) for the HER sector.

1.2.1. Trust

The core concept of IDFs involves researchers, teachers, students and other affiliated users receiving usernames and credentials from their home university or research institution (the identity provider), which they can then use to prove their identity (i.e. authenticate) to other organizations (service providers) outside their institution. Service providers are also able to retrieve up-to-date user data (i.e. the user's attributes) from the user's home institution. IDFs have trust frameworks in place to ensure the reliability of the institutional identities and authentication.

Trust in IDF has several directions. The Service Provider must trust the Identity Provider has carried out its own part as agreed; the end user is authenticated according to the IDF standards and his/her attributes are up-to-date. The Identity Provider must trust the Service Provider that it processes and protects any personal data received from the Identity Provider in a way which conforms the data protection laws. Both the Identity and Service Providers need to trust the IDF and the operator of the IDF that is follows the operational guidelines of the IDF.

The foundation of trust in the IDFs are in the "home" organization which have a close relationship to the end user. The home organization is able to carry out a face-to-face identity vetting process at the time the user receives his/her username and passwords (or other means) for authentication. In an IDF, the home organization is able to make sure the user data is up-to-date in its user management systems. If the user has problems with his/her account, they are resolved in the organization's IT service desk. Most importantly, the home organization knows best when the employee departs or student graduates and is able to close his/her accounts swiftly. Due to this close relationship of the end user and his/her home organization, the scheme is often called *organization-centric* identity management.

As a flipside, IDF identity is grounded to a person having an affiliation with an institution and credentials in the institution's identity management service. End user without an affiliation to enabled institution (either, because the end users are freelancers with no home organization, or because their home organization has not joined an IDF) cannot use it. Furthermore, IDFs don't easily support nomadic end users i.e. persons who frequently hop from one institution to another.

1.2.2. Access management

After receiving the authenticated user's identifier, the service decides if access should be grant to him/her. This can be done out-of-band, for example, by maintaining an access control list containing the identifiers of the eligible users. However, in some cases, it is easier to grant access based on the end user's affiliation. Because the home organization knows the end user's affiliation best, this kind of attribute can technically be retrieved via the IDFs (for instance, the eduPersonAffiliation attribute). However, defining attribute semantics becomes a challenge. For example, how is a 'researcher' type role with a given institution distinguished from other non-eligible roles such as 'student' or

[4] See, for instance: Basney, Koranda , Welch: An Analysis of the Benefits and Risks to LIGO When Participating in Identity Federations: https://dcc.ligo.org/public/0070/G1100964/002/LIGOIdentityFederationRiskAnalysis.pdf

'administrator'. The lack of reliable, globally recognized attribute semantics can lead to the attribute being unreliable for authorization use.

There are also formal or informal groups or collaborations of which the user's home organization isn't aware. Those collaborations, often called virtual organizations (VOs), may have members from several home organizations. The virtual organizations may have their own roles, permissions or attributes that they manage for their members by themselves. Instead of pushing these attributes to each user's home organization's IdP server, the virtual organization may rely on the VO management service which takes care of the virtual organization's local attributes and provides them to the services the virtual organization has deployed for its members. In this case, the user is still authenticated by his/her home organization's Identity Provider (so that users can make use of the home organization's authentication credentials), but the virtual organization releases extra attributes to the service for the user. This can be done, for instance, by an Attribute Provider the virtual organization has deployed.

1.3. Identifying authors and other knowledge contributors

Scholarly publishing is another major area where identity and identification of researchers is of key importance. Attributing published works to their creators (that is, the question of "who published what") has been fundamental to communication of scientific ideas and knowledge since the 17th century. Historically, identification and attribution in scholarly publishing has been based on authors' names. Now, in the 21st century, ambiguity resulting from non-uniqueness of person names has become a very acute problem. A 2009 study[5] found that around two thirds of the approximately six million authors in the MEDLINE bibliographic database share a family name and first initial with at least one other author, and that an ambiguous name refers to ~ eight people on average. The sheer number of individuals contributing to the tens of millions of existing scholarly works is a key factor, along with a steady increase in the number of scholarly publications produced each year.

1.3.1. Identity and attribution

Giving credit to the ideas and contributions of those that have come before by citing their works remains a fundamental part of scholarly communication. Peer recognition, acknowledgement and reputation are major drivers for scientists as creators of scholarly content. For scientists as content "consumers", knowledge of who authored a given work is an important factor in establishing its trustworthiness and provenance, especially for digital content published online[6]. It follows that researchers as individuals are a key stakeholder group in researcher identification and attribution.

Reliably linking contributors to research output is also a key concern to stakeholders in research evaluation. Universities and other organizations where research takes place are interested in monitoring outputs generated by their researchers for evaluation purposes. Funders are interested in tracking and measuring impact of outputs from the projects they fund, in a drive to maximize efficiency, effectiveness and overall impact (economic, social etc.) of their investment in research.

Reputation and evaluation are currently based almost entirely on i) conventional publications as the primary 'currency' of scientific output and ii) the prestige of journal those publications appear in, as

[5] Torvik, V.I. & Smalheiser, N.R. Author Name Disambiguation in MEDLINE. *Discovery* **3**, 1-29 (2009). http://dx.doi.org/10.1145/1552303.1552304

[6] Bilder, G. Identify This! Identifiers and Trust. *Information Standards Quarterly* **23**, 20 (2011). http://dx.doi.org/10.3789/isqv23n3.2011.05

measured by the controversial ISI Journal Impact Factor (IF). There is growing awareness of and interest in the need to progress beyond the current status quo and take into account:

- a wide range of so-called 'non-conventional'[7] research outputs, including datasets and scientific workflows published online, curation of scientific databases and more.
- a far more diverse types of impact metrics which take into account, for example, online access to full-text articles, PDF-downloads and online social bookmarking.

We will not elaborate on this in detail here, but instead we refer readers to a published report[8] from the recent Beyond Impact workshop (organized by a IRISC2011 speaker and session chair Cameron Neylon). Beyond Impact was convened to tackle many urgent issues in this area, especially the numerous policy-related, political, social non-technical obstacles to progress. Although these largely non-technical obstacles should not be underestimated, it is also evident that major technical challenges also need to be addressed, not the least the current lack of a common infrastructure to enable reliable online identification of research outputs and of the individuals that contribute to those outputs. Other key factors include inadequate access to (or non-existence of) comprehensive information on the use & reuse of scholarly resources, including but not limited to traditional article citation links, full-text article downloads and social bookmarking[9].

1.3.2. Author identifier systems

A global identification infrastructure and open access to data on resource use are the foundation of the 'knowledge discovery' use cases that will underpin scholarly evaluation and incentives/rewards systems of the future. Fortunately, the problem of identifying conventional Science, Technology and Medicine (STM) publications is largely solved, via the establishment and widespread adoption of the Digital Object Identifier (DOI) system in the scholarly publishing sector. More recently, the same underlying infrastructure and principles are being repurposed and extended to enable registration of persistent DOI names for scientific datasets globally, under the umbrella of the newly-formed international DataCite Consortium.

Creating unique identifier systems for researchers has proved a much more difficult task. All attempts made to date have either failed to be widely adopted (Thomson Reuters ResearcherID), or been successful but limited in scope to certain countries (LATTES in Brazil, DAI in the Netherlands) or disciplines (RePEc in economics) or organizations (NIH intramural researchers) (see Table 1). As Martin Fenner notes in his overview article[10], creating unique identifiers for persons is more challenging than creating identifiers for digital objects, and especially considering the requirements for scholarly identifiers, notably long-term persistence, trust and authority.

There is now broad agreement amongst a wide range of stakeholders in this area that these identification challenges must be tackled globally, not the least because research is international and increasingly interdisciplinary, and thus transcends most traditional boundaries. The most significant

[7] 'conventional' outputs = books and articles in peer-reviewed journals

[8] https://docs.google.com/document/d/1sH3JOW5Luki4i37Ve1mOnI2wNZJbaUOx1T42S_7txQ0/edit?hl=en_GB

[9] Notable exceptions exist, such as the PLoS Alternative Metrics project: http://article-level-metrics.plos.org

[10] Fenner, M. Author Identifier Overview. *LIBREAS* 24-29 (2011). Available at http://edoc.hu-berlin.de/docviews/abstract.php?lang=&id=37867

progress on this front was the Dec 2009 launch of the Open Researcher and Contributor ID (ORCID) Initiative. ORCID is the outcome of several years of deliberations between Thomson-Reuters, Elsevier, Nature Publishing Group and many other publishers represented by CrossRef, and numerous research libraries, universities, funders and other stakeholders in the publishing and research domains. The initiative aims to "establish an open, independent registry that is adopted and embraced as the industry's de facto standard" (http://www.orcid.org). Several key aspects of the organization and the upcoming service, scheduled to go live mid-2012, are discussed in a general summary paper recently published[11]. A second article which focuses on the adoption of the service was just published in UKSG Serials[12].

1.4. Common challenges and opportunities

Across the three main identity-related problem areas or communities outlined above - i) unique, persistent contributor identifiers, ii) identity federations and iii) access management - several potentially synergistic areas were identified prior to the workshop as targets for discussion and collaboration.

1.4.1. Institutional validation of ORCID profile data

A major focus for ORCID in early stages of the service will be to collect a "critical mass" of user profiles containing not only self-asserted information (i.e. provided by users themselves) but also organization-asserted information provided by other parties. This includes both biographic information (current and past institutional affiliation, position/role, work address and phone number), and bibliographic information (authorship assertions from publishers). The initial strategy that ORCID has decided on is to "seed" the system: that is, selected institutions will mass-deposit profile data for their researchers, effectively pre-registering them with institution-validated profile data.

Another strategy to consider for later phases of ORCID development might be integration with identity federations. One of the strengths of IDFs is the close relationship between the end user and his/her home organization, which is able to authenticate him/her, ensure his/her identity is up-to-date and provide local support if needed. Enabling users to register and authenticate via their institutional ID would therefore provide another way for ORCID to acquire institution-validated profiles. Such IDF-based integration may well suit some institutions better than mass-deposits. Getting user attribute data from IDFs to ORCID also would be a way to keep the data dynamically up to date.

1.4.2. Bridging identity federations with persistent, global identifiers

The organization-centric identity that underpins IDFs works very well for a range of situations involving a user affiliated with a single organization throughout his/her career. However, the model starts to bend and creak when applied to a number of increasingly common scenarios in the modern e-science era of international, cross-disciplinary research collaborations and digital scholarship.

[11] Fenner, M. ORCID: Unique Identifiers for Authors And Contributors. *Information Standards Quarterly* **23** (3) (2011). http://dx.doi.org/10.3789/isqv23n3.2011.03

[12] Fenner, M., Gómez, C.G. & Thorisson, G.A. Key Issue Collective Action for the Open Researcher & Contributor ID (ORCID). *Serials: The Journal for the Serials Community* 24 (3), 277-279 (2011). http://dx.doi.org/10.1629/24277

The insistence on a single home organization vouching for (by being the sole provider of an online identity for) a researcher is a key limitation. It is not possible to present oneself as belonging to multiple organizations, whether real (e.g. universities) or virtual (multi-national consortia, ad hoc informal research groups etc.) - in other words, having multiple institutional identities which are all equally valid[13]. This is simply a consequence of the breaking down of interdisciplinary walls, and these researchers are sure to become an increasingly important category in terms of impact, and may be building towards a new state of fluid exchange among disciplines.

Another problem is that researchers without a home institution that participates in the federation are effective "identity-less". This is problematic for freelancer researchers and others fitting the "homeless" category, including both affiliated researchers employed by an organization not participating in an IDF, and non-affiliated researchers working independently as freelancers or participating as 'citizen scientists'.

In short, there is an increasing need for a notion of continuous, persistent identifier for researchers as they move between organizations throughout their career. The properties of ORCID-sourced identifiers, if integrated into the IDF ecosystem, will make them appealing as this kind of "bridging" purposes and facilitate cross-IDF traversal. However, in order to avoid ORCID identity thefts, a carefully designed procedure for traversal of ORCID identifiers is needed, when a user changes his/her home organization.

1.4.3. Managing access to online biomedical services

The third area of interest is use of ORCID identifiers for access management purposes. If a permission is coupled to a person as an individual (irrespective where she/he currently is affiliated to), the permission can be coupled to his/her ORCID identifier. This approach has potential, for instance, when the researcher accesses a publisher's service as an author of an article.

However, there are also obvious use scenarios, where the permission is not coupled to a person as an individual, but to a person's affiliation to an organization. For instance, when the person gets a permission to access certain research data as a member of a certain institutional research group, his/her permission to the data ceases when he/she departs from the institution. In this case, coupling the person's permissions to his institutional identifier provides additional security, because (unlike the ORCID identifier) his/her institutional identifier is likely to be revoked when she/he departs.

Separately from the risk of identity theft, data model analysis is needed to better understand the role of ORCID and institutional identifiers in the management of the end user's permissions. The analysis should bring additional understanding on when it is desirable to bind the permission to a person's institutional identifier vs. to a person's ORCID identifier.

1.5. Motivation for the IRISC workshop

There is no shortage of international and regional events dedicated to some of the topics outlined above. However, existing events or event series tend to be broad (e.g. the Internet Identity Workshop,

[13] It is worth noting that recent proposals for identity linking and attribute aggregation (see Chadwick & Inman 2009 and slides here) and an emerging ecosystem around personal data stores (see e.g. http://mydex.org) do address this class of scenarios. These technologies are still in early development stages, however.

RSA Security conference), project-specific (e.g. ORCID participant meetings; eduGAIN workshops), or otherwise too limited in scope or community reach. To our knowledge there was no event dedicated to jointly and comprehensively tackling these important challenges, questions concerning identity and authentication in scientific research. The IRISC workshop was designed to fill this gap.

Building on the success of the IRBW2009 workshop (organized by G.A. Thorisson and A. J. Brookes) which focused on biomedical research, IRISC2011 was convened to bring together identity experts, users, funders and other stakeholders to start a dialogue. Key target communities included: scholarly publishers; service providers (including scientific database managers & developers); ethicists working on data sharing; bioinformaticists; and identity federation service providers and experts. Given the organizers' background and research interests, there was an emphasis on the biomedical domain in the choice of speakers and overall direction of the programme.

The workshop had two main interrelated themes: i) identifying & attributing authors/creators of scholarly works, with a focus on ORCID, and ii) identification for access management purposes, i.e. with a focus on federated identity management. Key workshop aims included:

- Information sharing to raise overall awareness of key technical and non-technical challenges, opportunities and developments.
- Facilitating a dialogue, cross-pollination of ideas, collaboration and coordination between diverse – and largely unconnected – communities.
- Identifying & discussing existing/emerging technologies, best practices and requirements for researcher identification.

The IRISC2011 event was planned as a standalone workshop. But consideration has been given to potentially expanding this into a series of annual events, which could in turn evolve into a longer-term platform for ongoing discussions and collaborations in this space.

2. Workshop participation and proceedings

The IRISC2011 workshop was attended by over sixty participants representing institutional stakeholders, research funders, developers, service providers, publishers and researchers (see list in Appendix I). According to registration data, delegates' background was very diverse, with stated positions or roles ranging from postgraduate students, university professors, bioinformatics consultants and IT administrators to technical managers & scientific directors for research institutions, digital librarians, project managers and security chiefs, to name a few. The great majority were from the academic sector, with very few corporate or other non-academics. The low level of participation from, amongst others, scholarly publishers perhaps suggests that a different event programme and/or marketing strategy is needed for future events to better reach this and other underrepresented sectors.

The majority of delegates were based in Europe, a smaller number on the US East Coast, and some travelled from as far away as the US West Coast, Japan and Australia. Co-location and coordination with the REFEDS IDF workshop was helpful in this regard, as a substantial number of delegates were able to combine the two events into a single trip. There was interest in repeating this co-location arrangement for potential future IRISC events. A number of annual or biannual workshops on the IDF calendar convened by e.g. REFEDS and TERENA, and the biannual ORCID Participant meeting, are candidates now being considered.

2.1. The programme

A key aim of IRISC2011 was information sharing, and thus the main part of the workshop was in the form of traditional plenary sessions, with a pre-arranged speaker programme. Inspired by recent workshops on related topics, such as Beyond Impact, there was also a strong desire to facilitate generation of concrete outputs from the workshop. Therefore, parallel interactive breakout sessions comprised a substantial part of the programme.

Appendix II contains summaries from each plenary presentation, as well as notes from the breakouts. The fully-detailed workshop schedule with links to speaker profiles, slides and video recordings is available at http://irisc-workshop.org/irisc2011-helsinki/schedule/ and also in Appendix III. A brief overall session summary is provided below.

2.1.1. Plenary sessions

The opening session on Monday afternoon focused on identification and attribution in scholarly publishing. Presentations covered some of the core issues around scholarly identity/identifiers and two major projects in this space, ORCID and VIVO, which are tackling the identification challenge from different perspectives. Speakers also talked about attribution and applications of unique identifiers in research evaluation, in new forms of scholarly publishing, and in knowledge discovery.

Presentations in the second session focused on identity federations and access management. Two major European infrastructure projects were presented: one (CLARIN) is well established and with IDF experience; the other (ELIXIR) is still at the planning stage, studying identity/security requirements and evaluating technologies. One speaker gave a general introduction to identity federations and IDF

coordination activities undertaken by REFEDS, followed by a presented on the eduGAIN interfederation project. The final presentation focused on potential application areas, suggested by the GEN2PHEN project, wherein ORCID identifiers could help streamline governance mechanisms for managing access to sensitive biomedical research data.

The final plenary session on Tuesday morning had three presentations. The first was a summary of findings from a recent CERN workshop on federated identity in scientific collaborations, including a broad overview of the current status of IDF deployment in several major projects and communities in Europe and internationally. Delegates also heard about the Publish Trust Project to implement a distributed system for asserting and publishing authorship claims for works published by APA, leveraging the OIX trust framework and established federated identity protocols. The final speaker presented work by the bioinformatics outsourcing company Eagle Genomics to secure their cloud-based bioinformatics analysis platform by applying federated identity technologies.

2.1.2. Parallel interactive breakout sessions

One of the breakouts was titled "*Unique identifiers and the Digital Scholar*" and focused on adoption of ORCID and its utility to the researcher community, following the planned public launch of the identifier service in early 2012. The 20 breakout participants were asked to think about which three features that will drive adoption from researchers and could feasibly be implemented in the next six months. By way of facilitated discussion, first in smaller subgroups and then in the full group, participants came up with a "Top 3" list of features.

The other breakout was titled "*What do researchers need from the IDFs?*". Before the workshop, all IRISC2011 participants were sent a link to a web survey which aimed at studying: (i) the research infrastructures' needs on identity federations and (ii) the IDFs' current service offerings to the researchers. The session started with a summary of the survey findings, followed by discussion. The ~40 session participants were then further divided into three subgroups to continue discussion on specific topics, both those highlighted in the survey results and several others, such as IDF user experience (UX), tools for collaboration and different levels of trust.

2.1.3. Reporting from breakouts, discussion and wrapping up

After the breakouts concluded and all workshop delegates had regrouped after a lunch break, each breakout chair reported findings of his respective group to the whole group. Each round of reporting was followed by an open discussion. The group then took a final break for coffee, before gathering for a final round of discussions and wrapping up.

3. Workshop results

Workshop outcomes fell broadly into three main categories: issues around ORCID, in particular concerning end user adoption and use beyond journals and traditional publications; various types of requirements for identity federations and areas where improvement is needed; and potential interfacing opportunities between ORCID and IDFs.

3.1. Motivating ORCID adoption in the researcher community

3.1.1. "Killer apps" for end users to accelerate adoption

As remarked by Cameron Neylon who chaired one of the breakouts, ORCID is without question the "main game in town", and the discussion around author identifiers will therefore necessarily be focused on that project. There is already broad buy-in from publishers and many other stakeholder organizations. But there has been much less discussion and study of the needs and motivations of end users themselves (i.e. researchers). Some observers have expressed concerns that ORCID will ultimately build an infrastructure that will mostly be useful to publishers, libraries and similar entities. The momentum now being generated around ORCID offers perhaps a once-in-a-lifetime opportunity to establish a *de facto* standard for identification across the entire community of active scholarly researchers.

One of the two breakouts addressed the following question: what kind of kind of features are most likely to appeal to the majority of end users and motivate them to adopt the ORCID service. The Top 3 list of features identified were:

- Integration with manuscript tracking systems (MTSs)
- Integration with data archives and other kinds of scholarly repositories
- Automated CV generation

The planned late 2011 release of the ORCID API sandbox will enable early-adopter developers to create proof-of-principle applications with these features (and others, see below). There is considerable interest in this building up in the community; for example, the Dryad open data repository wants to integrate ORCID into their submission workflow as early as mid-2012.

***Recommendation**: Developers should use the ORCID API sandbox (available later this autumn) to implement proof-of-concept ORCID integration which adds value to end users of scholarly services.*

***Suggested action**: Identify organizations (ideally including publisher partners) willing and able to participate in early integration work and to promote the ORCID sandbox.*

One course of action discussed in the group to take the MTS integration suggestion forward was to lobby senior journal editors, executives and others who have a major influence on which MTS features are developed by software suppliers (as opposed to lobbying MTS makers directly). Gathering

signature via an online petition was suggested as a potentially useful action to this end. Such a petition may also be of broader use for others within the publishing community (e.g. CrossRef) to make the case for ORCID integration and prioritization.

The remark was also made that, given the larger number of interlocking parts required (compared to journals and MTSs), unique ID based attribution for non-conventional research outputs was less straightforward. But the resulting payoff could be high, especially with respect to tracking research outputs and enabling alternative metrics.

3.1.2. An expanded role for ORCID

In the longer term and beyond the above relatively obvious applications, ORCID is likely to play a vital role in a wide range of scenarios involving identification of contributors to or users of scholarly resources. Specific mentions were made of ORCID as aggregator and mediator of assertions about a contributor, such as validated authorship claims (sourced from publisher) and affiliation (sourced from institution). The APA Publish Trust Project presented in Session 3 showcased some of the technological solutions that might be used for this. Though not in scope for the first phase of ORCID development, such advanced functionality is likely to be important in the longer term. Suggestions for future ORCID-related development included:

- Integration of ORCID with identity federations (see also below)
- Extend VIVO to build features around ORCID
- Pilot ORCID integration in the BioMedBridges project
- ORCID support in nanopublications and microattribution
- Integration with Wikipedia, to attribute contributions to Wikipedia articles
- Support for ORCIDs in descriptions of biological databases as promoted by the BioDBCore project[14] and the Wikipedia Infobox
- Integration with journal commenting systems

3.1.3. Overcoming political blockage

Several delegates asserted strongly that many key obstacles to progress in the above-outlined areas are not technical, but rather policy-related and/or cultural in nature. Overcoming this blockage[15] and achieving a high level collaborative policy is crucial if progress is to be made. Barend Mons proposed the creation of a “Coalition of the Willing”, focused on building a range of value-add end user applications around the ORCID service. No single step by itself will beat the chicken/egg problem, so a pragmatic “just build it” approach is needed to make progress and create momentum.

Suggested action: *Seek the engagement of high-profile researchers and their institutions, to assist in raising awareness and promoting the potential benefits of scholarly identity and identifier infrastructure in the scientific community.*

[14] Gaudet, P. et al. Towards BioDBcore: a community-defined information specification for biological databases. *Nucleic Acids Research* **39**, D7-D10 (2011). http://dx.doi.org/10.1093/nar/gkq1173

[15] ‘Political blockage’ occurs when interest groups in a research community lock up data and access privileges for their own members, and block reforms that would threaten these arrangements, even reforms that are necessary to meet societal needs. Political blockage also arises as patrimonialism, the tendency of academic authorities to favor local institutions, when research in biomedicine now occurs in a global context. Finally, there is the temptation towards privatization and commercialization of research, as a means to secure funding in an uncertain environment, even when the data involved is a form of public property.

3.2. Identity federation requirements from service providers

3.2.1. Retrieval of user information, data protection and scalability

For research infrastructure services such as ELIXIR which target a large, multinational user base, an easy retrieval of the user information form IdPs via identity federations is vital. For instance, CLARIN aims to serve researchers in several hundred institutions across Europe. But if each participating institution must be contacted individually to request attribute release to the CLARIN services, scalability quickly becomes an issue.

The EU data protection directive is largely to blame for IdP's hesitation to release attributes, as institutions are partly responsible for any data protection issues in the services they release use data to. On the other hand, the administrative burden of written agreements between IdPs and services would be extensive. A solution is needed which balances privacy risks with the benefits of easy collaboration. A proposed solution is for IDFs to release a basic subset of available attributes without extensive bureaucracy. This could be implemented relatively easily, at least to countries whose data protection laws fulfill the European standards.

Recommendation: *attribute release should be easy, and service providers should not need to contact individual IdP's to configure attribute release.*

Suggested action: *Lobby IDFs and eduGAIN for developing mandatory policies and practices for attribute release, and for relaxed release of attributes which do not represent significant privacy risks for the user.*

3.2.2. Reliability of Identity and authentication

Currently, IDFs provide authentication services which are mostly based on passwords. Grid and high performance computing have already a need and deployment for strong authentication based on certificates, but otherwise most services seem to be satisfied with password based authentication, provided it is carefully designed. However, research infrastructures showed interest in a reliable user registration process which is based on the end user coming personally to show his/her photo-ID in order to receive his/her credentials. The RIs also asked for a basic level of accountability for the end users; for instance, give up using shared user accounts. Some organizations and IDFs have already specified Identity Assurance Frameworks which cover these issues (e.g. Kantara IAF, inCommon IAP), but they are not deployed widely.

To increase the trust on IDFs and institutional identities, the RIs also wished for some basic requirements for audits of the institutional identity management systems. The audits should focus on verifying the reliability of the user registration process and the quality of the data available in the identity management systems. Peer audits were proposed. However, at that point, the RIs were not asked to share the audit costs.

Suggested action: *Lobby IDFs and eduGAIN for introducing an Identity assurance framework for the home organizations. To support cross-national RIs, a global scheme is preferred.*

3.2.3. Attribute semantics

The number of IDFs is proliferating and nearly 30 countries have already an academic IDF or an IDF project. However, the federations have grown up separately, which has lead them to adopt different practices, for instance, for attributes expressing end user's affiliation. The is no harm of the differences as long as the federations are not interconnected, but when the federations get bridged together and a service starts to receive users coming from different federations, it faces the trouble of needing to adapt to different attribute semantics and practices depending on which federation the end user comes from.

The research services and infrastructures are identified as a potential user of the IDFs. Interfederation initiatives like eduGAIN need requirements and feedback from the research infrastructures to ensure that the results of the project serves the researchers' needs.

Suggested action: *Lobby identity federations to gradually harmonize the attribute availability and semantics, especially the attributes which are useful for authorization in the services.*

3.2.4. Funding model

The funding model of some identity federations is based on the services paying an annual fee (typically, 1.000-3.000 EUR) to the federation for its services. Covering costs is naturally necessary for the federations as well, but invoicing academic and non-commercial service providers may lead to a situation where the funding model kills the collaboration. In the big picture of the academic world, easy collaboration should be encouraged, not discouraged. Federations should not introduce conflicting incentives in their funding models.

Suggested action: *Encourage federations to adopt a "zero-cost" funding model for academic service providers.*

3.2.5. Usability and user awareness

User experience (UX) is a perennial challenge in the IDF domain, in particular communicating to users that A) they can login into a given site with their home institution credentials and B) help them select the right ID provider from a list of hundreds of IdPs (i.e. IdP discovery). Many existing solutions are built as clumsy drop-down menus containing hundreds of items, and often require users to subsequently click through several pages to complete authentication.

The counter-examples shown in the breakout summary session showcased newly developed tools for user-friendly IdP discovery workflow (DiscoJuice, tiqr), which demonstrated what can be done to vastly improve the UX whilst still enabling better security and higher level of authentication (LoA) (e.g. 2-factor authentication on mobile devices). The challenge is to get these and similar solutions adopted widely by service providers in coordination with the IDFs.

Recommendation: *IDFs should pay more attention to usability issues in order to make the federated login easy for the end users.*

Suggested action: *Encourage service providers to introduce usable IdP discovery workflows.*

A related challenge is that many users also have little or no idea what federated identity management is about and how they can use it. Educational materials, training and outreach activities are needed "on the ground", both in a national and international levels.

Recommendation: *IDFs should pay attention to outreach activities among the research service providers and infrastructures.*

3.3. Common challenges/solutions and opportunity for collaboration

3.3.1. Biomedical data services: use cases for IDF integration

As noted by Andrew Lyall from EMBL/EBI (see Appendix II), biomedical data service face implementation challenges related to data security that arise, for example, from data privacy policies based on national or international legislation. Some of the established technical solutions in use by existing IDFs for data access authorization are compliant with these policies and could therefore be adopted as (partial solutions) by service providers. These solutions together with IDF policy frameworks form the backbone of major US-based biomedical research networks such as caBIG (cancer research) and BIRN (neuroscience), but are not very well known and/or widely adopted. Another problem is a lack of good use cases from biomedical sector service providers, which would help to define requirements for federated identities across organizational and country borders.

A suggestion was made to tackle these issues by undertaking a pilot project which concretely demonstrates how a national or regional federation (e.g. Haka) or interfederation (e.g. Kalmar Union, eduGAIN) service can be deployed in the relatively common setting of a biomedical data provider. The specific pilot scenario suggested by CSC involves distribution of (for example) Finnish biobanking data in collaboration the European Genome-Phenome Archive (EGA) that already hosts some of this data. Data users from Finnish/Nordic federations would apply for EGA data access with their home institutions identity, and there would therefore be no need to the EGA service to maintain a record of users who belong to any Finnish/Nordic organization that is part of the federation.

Suggested action: *Establish a pilot on federated access management to a biomedical data provider together with EGA, eduGAIN and related national IDFs.*

The pilot could also involve research of automation of the authorization process by the data access committees as an example of IDF - biomedical service provider collaboration. This could include evaluation of the use of validated information on a researcher's academic record (retrieved from ORCID, see more below) as part of the vetting process, either alongside or in the absence of institutional credentials (the latter would benefit homeless data consumers).

3.3.2. Interoperability between identity federations and ORCID

The need to deal with homeless, nomadic and cross-disciplinary IDF users came up repeatedly in both plenary presentations and discussions. Common to these groups is the need to identify themselves in a way that transcends institutional, geographic and discipline boundaries. For nomads, there is a need for a "proxy" identity linked to multiple institutional IDs for a single person. For homeless users, there is a need for a primary, universally-accepted identity in the absence of institution-provided IDF credentials. These requirements clearly resonate well with the requirements for scholarly author identifiers, and this is probably the area of most overlap between ORCID and identity federations (see more below).

Recommendation: *service providers should investigate possibilities for authenticating homeless users via ORCID or other trusted source of author identifiers.*

Several ideas were discussed concerning how the above and similar scenarios could be implemented, via links between the ORCID service and identity federation or interfederation services. Remarks were made that capturing an IDF user's ORCID identifier as an attribute would in principle be straightforward to implement. This would, for example, be of direct use in the biomedical data services pilot: the author profile and publication list for an IDF user could be retrieved directly from ORCID using the validated identifier. However, this approach would require a mechanism for indicating the provenance of the assertion; i.e some way to verify that the ORCID identifier is in fact under the control of the user (otherwise there's nothing to prevent a user from presenting someone else's ORCID as his own).

Suggested action: *investigate how an ORCID or other author identifier and its provenance can be modelled as an attribute in IDF and interfederation services.*

A further discussion point was whether ORCID could become a IDF participant - as an SP, IdP, attribute provider or both - and what the benefits of this would be. Operating in service provider mode would allow users to register to ORCID by authenticating with their home institution credentials, thereby creating 'pre-validated' contributor profiles (see Introduction). Operating in identity provider mode would allow homeless IDF users to authenticate to e-infrastructure services such as CLARIN with some minimal privileges, using ORCID as a kind of universal or 'common lowest denominator' authentication service as described above. A useful strategy may be for the ORCID organization to operate a virtual organization management service (see Introduction) which releases an extra attribute or attributes for users authenticated by IdPs.

Participation in InCommon, eduGAIN or other IDF infrastructure projects is not on the immediate agenda of ORCID. The organization has ample tasks on its plate as it is, and it could be argued that identity provision via IDFs is simply outside its scope. On the other hand, such work been not been ruled out either and could potentially be included in later phases of development.

Also, judging by the draft of the ORCID Phase I API documentation, it is clear that the OAuth-based delegated authentication[16] to registered client applications via ORCID accounts (in the style of Twitter, LinkedIn and many other mainstream services) will be possible. ORCID as a *de facto* identity provider in this user-centric identity mould (i.e. outside the IDF ecosystem proper) can potentially create

[16] Hammer Lahav, E. *Introducing 'Sign-in with Twitter', OAuth-Style "Connect"*. http://hueniverse.com/2009/04/introducing-sign-in-with-twitter-oauth-style-connect/

numerous integration opportunities across a wide range of use cases, including the “killer apps” already mentioned.

It goes without saying that, whatever integration strategy is used, successful interoperability between IDFs and ORCID will require some measure of metadata standardization. As already noted, creating a new IDF attribute holding a user’s ORCID identifier is a “low hanging fruit” which would be useful on its own. But further standardization would be desirable, in particular for information describing a person’s affiliation (current and past) and role or roles within an organization. Ideally, the data model used by ORCID should be compatible with the relevant IDF attributes. A key challenge in this regard is dealing with the aforementioned semantic interoperability problem in the IDF domain.

Recommendation: *the IDF community and ORCID should work attempt to harmonize core profile fields/attributes which are likely to hold institution-validated information*

4. Final summary and conclusions

The aim of IRISC2011 was to facilitate information sharing and in-depth discussions between communities of stakeholders. With respect to the first goal, the workshop was a clear success, with the quality and diversity of presentations giving the audience a broad view of challenges, opportunities and projects around identity in research and scholarly communication. Interactive discussion sessions also proved fruitful with respect to the set topics covered. However, in hindsight better use could have been made of these focused sessions to jointly tackle some of the key cross-theme issues of interest to both main communities of participants.

On the whole, organizers and delegates were pleased with the execution and outcome of the workshop and plans are already being laid for a repeat IRISC2012 event next year. There is interest in developing IRISC into an annual event series which would provide a much-needed forum for cross-community discussions on these important topics.

5. Acknowledgements

IRISC receives funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 200754 - the GEN2PHEN project - and is further supported by University of Leicester, CSC - IT Center for Science, Institute for Molecular Medicine Finland and Wellcome Trust.

Appendix I: Workshop participants

Juha Herrala, Tampere Univ. of Technology
Chris Phillips, CANARIE Inc.
Ma'n H. Zawati, Centre of Genomics and Policy - McGill University
Kimmo Koskenniemi, University of Helsinki
Chris Brown, JISC
Enrico M. V. Fasanelli, INFN
Vera Hansper, CSC/NDGF
Steven Newhouse, EGI.eu
Michal Prochazka, CESNET
Brook Schofield, TERENA
Valter Nordh, NORDUNET / GU
Ludek Matyska, CESNET
Pekka Järveläinen
Jani Heikkinen, CSC
Pascal Panneels, BELNET
Hideaki Goto, NII
Dieter Van Uytvanck, Max Planck Institute for Psycholinguistics
Mikael Berglund, ITS/SWAMI
Danny Sternkopf, CSC - IT center for science
Krister Lindén, University of Helsinki
Pieter van der Meulen, SURFnet
Antti Pursula, CSC - IT Center for Science Ltd
Jarno Laitinen, CSC
Pablo de Castro, Carlos III University Madrid
Sabita Behari, SURFnet
Hideaki Takeda, National Institute of Informatics
Mauno Vihinen, Institute of Biomedical Technology
Kari Laalo, CSC
Kalle Happonen, CSC - IT Center for Science
Thomas Neidenmark, Stockholm University Library
Juha Muilu, FIMM
Myles Byrne, FIMM
Daan Broeder, Max-Planck Institute for Psycholinguistics
Gudmundur Thorisson, University of Leicester
Barend Mons, NBIC
Joost van Dijk, SURFnet
Geoffrey Bilder, CrossRef
Alex Reid, University of Western Australia
Hildegunn Vada, UNINETT
Acácia Reiche, Fundació IMIM
Tommi Nyrönen, CSC
Maria Laura Mantovani, GARR
Anne Leinonen, FIMM
Heath Marks, Australian Access Federation
Andrew Lyall, EBI
Lucy Lynch, Internet Society
Licia Florio, TERENA
Hal Warren, American Psychological Association
Alasdair Gray, University of Manchester
Nicole Harris, JISC Advance
Urban Andersson, Chalmers University of Technology
Aamir Shahzad, University of Vienna
Anthony Brookes, University of Leicester
Todd Harris, WormBase
Henri Mikkonen, Helsinki Institute of Physics

Cameron Neylon, STFC
Jason Priem, University of North Carolina at Chapel Hill School of Librarly and Information S
Stefan Lueders, CERN
William Spooner, Eagle Genomics
Mikael Linden, CSC
Brian Lowe, Albert R. Mann Library, Cornell University
Kevin Dolby, Wellcome Trust
Martin Fenner, ORCID
Laurence Mabile, INSERM
Arthur Smith, American Physical Society
Michael Taylor, Elsevier

Appendix II: Workshop session proceedings

Plenary sessions

The opening plenary session was dedicated to identification and attribution in scholarly publishing. Given its prominence in this area, the ORCID initiative was the topic of the first two presentations. The first was given by **Geoff Bilder** from CrossRef who is also ORCID's interim Technical Director. Although the presentation was not specifically focused on ORCID, Geoff touched on several important issues around identity and identifiers that are
central to the ongoing design & implementation of ORCID, which is squarely focused on solving knowledge discovery problems rather than security and access control problems. He talked about why provenance metadata (including identity of authors) is important to engender trust in published works, the fragility of tightly coupling an identifier to a location, dangers of embedding (intentionally or unintentionally) semantic information in a identifier string, social issues in achieving persistence, and the difficulty in finding a balance between distributed and centralized identifier service provision.

Next up, **Martin Fenner** of Hannover Medical School in Germany and currently on the ORCID Board of Directors, talked more specifically about the ORCID project and the prospects of adoption once the identifier service is operational early in 2012. Potential metrics for judging its success could be, for example, total no. users, no. supporting organizations, no. supporting services or no. attributions catalogued. A useful strategy may be to focus on specific disciplines (following the model of arXiv which started in physics), rather than aiming for adoption across the entire scientific community. Martin highlighted several characteristics of both the ORCID service and the not-for-profit organization behind it, including its global reach across all discipline, geographic, national and institutional boundaries, the planned interaction with other author identifier systems, and a commitment to open participation, open data and open-source software. He also pitched the ScienceCard pilot project as an example of an author-level metrics service which utilizes author identifiers rather than names.

Brian Lowe, from Cornell University and semantic development lead on the VIVO project presented an introduction to VIVO. VIVO shares with ORCID many of the same over-arching scholar identification and attribution aims, but from an institution-based rather than global perspective. Brian briefly described the technical basis of the system which is built as a distributed architecture, with each participating institution operating its own VIVO repository into which they collect rich information about their researchers and works associated with them. The system is heavily based on Semantic Web technologies and relies on Linked Data published as machine-readable RDF, shared URIs and ontologies for data integration, with emphasis on reusing and extending existing ontologies (e.g. FOAF and BIBO) wherever possible. VIVO's approach allows for straightforward handling of multiple URIs in use for the same person (by asserting equivalence) and will be able to incorporate and utilize ORCID identifiers as soon they are available. Brian also talked about the extensible semantic model employed in VIVO which provides great flexibility in capturing a person's contributions beyond publications and grants, such as clinical roles and (via collaboration with the Eagle-i project) connect people with the scientific resources they create (e.g. protocols, reagents, biological specimens).

The main message from STFC's **Cameron Neylon** (organizer of the Beyond Impact workshop) was that major changes are needed in the way research is communicated if funding is to continue. Traditional publications remain the primary scientific currency that is counted, but the various stakeholders (the public, government, funding bodies, industry, others) actually are looking for *outcomes* that have impact, that change the world. The Web provides a previously non-existent platform for the researcher community to engage with stakeholders about the output of their work and the impact of those outputs. If this opportunity isn't used and we don't progress towards more

transparency and openness and start having conversation with those stakeholders about the research that public money is being invested in, then researching funding will start to dry up. Cameron focused on data publication as a specific example of where this has been gaining a lot of traction in the last 5-10 years. There is increasing expectation that data generated with public money should be made publicly available for reuse. Identity is a crucial piece of provenance metadata to index those research outputs on, enabling assessment of data quality, data discovery, and crucially to encourage data publication through authorship and contributor recognition.

Barend Mons from the Netherlands National Bioinformatics Centre closed the session by talking about "big data" challenges and how we can hope to make use of the massive amounts of scientific knowledge and data already published, and will be generated in the future. He focused on the important role of unique identifiers in the context of a newly proposed form of scientific publishing called "nanopublications". The idea is to publish core scientific facts extracted from journal article (or, in the future, published 'natively') and/or data in a structured, machine-readable format, along with critical provenance information (including attribution) in a self-contained bundle. A major aim of the scheme to provide a scientific reward scheme which credits researchers for annotation, curation and data publication, not just for mainstream journal articles. To this end, unique identifiers for persons and an infrastructure for tracking their contributions to nanopublications are core building block. Barend concluded with a list of 'Killer apps' for author identifiers (several of which were subsequently discussed in a breakout session the next day). He strongly suggested a 'just build it' approach in this area: instead of succumbing to "political blockage" and seeking perfect technical solutions, focus on testing feasibility of many different software tools that can help us progress. He suggested the organization of one or more 'crackathon' events dedicated to implementation of demonstrator applications.

The second plenary session focused on early experiences from the use of institutional identities for access control and on potential use in other disciplines. EBI's **Andrew Lyall** kicked things off by presenting ELIXIR, a large-scale ESFR-funded infrastructure project to accelerate life science research in Europe, and identity challenges relating to sharing of biomedical data on human subjects. At the heart of the problem is the person-identifiable nature of many datasets being generated and various Ethical, Legal or Societal Implications (ELSI) issues surrounding their access and use. Key obstacles include requirements for obtaining consent from subjects of clinical studies, EU data protection directives, addressing needs of the diverse stakeholders involved (clinicians, researchers, analysts, technicians, patients themselves etc.). Secure access to sensitive data and privacy protection is therefore a crucial requirement for the ELIXIR e-infrastructure now being planned. Andrew highlighted a key use case: the European Genome-Phenome Archive (EGA) operated by the EBI. ELIXIR is in the process of evaluating identity technologies and strategies (see also Stefan Lueder's presentation in Session 3 below).

A very useful overview was next provided by REFED's **Licia Florio,** who talked about the basics of identity federations and some of the coordination activities undertaken by REFEDS. A federation requires two main components: i) a common technology (most IDFs have standardised on SAML) and ii) a trust framework based on legal agreements between participating organizations. Over 20 federations are now either already in production or being piloted, serving a total of ~16 million end users on all continents (except Africa), mostly in HER. REFEDs is a TERENA-run initiative to coordinate IDFs globally (aka "networking the networkers"). Aims include cross-IDF harmonization and influencing directions taken by a diverse community which includes IDFs themselves as well as identity experts, user groups and service providers.

Licia's overview was nicely followed by **Valter Nordh** from University of Gothenburg who presented the eduGAIN project. As a core part of the Europe-wide GEANT data network initiative, eduGAIN is an

"interfederation" service which will join together several national and regional IDFs into a single "confederation" across which trustworthy exchange of identity and authentication/authorization information can take place. Some of the challenges encountered in eduGAIN are service provider discovery, attribute release in an interfederated environment and providing a satisfactory user experience when things break. The scheme is opt-in for individual providers, so even if a federation has joined eduGAIN at a high level its constituent institutions and service providers "on the ground" are not necessarily participating. Outreach is therefore required to encourage providers to join.

The role of federated identity in the CLARIN project was presented by **Daan Broeder** from the Max-Planck Institute for Psycholinguistics. CLARIN aims to enable "eHumanities" by creating unified e-infrastructure for various resources and technology used by language scholars, such as audio and text corpora, text mining technology. 25 major CLARIN centres are planned. Daan outlined the "holy grail" user scenario for CLARIN involving institutional authentication: user accesses various assets of interest (some of which are access-controlled) from digital repositories and collects these into a virtual collection for use in his research. CLARIN created their own federation of 9 service providers in 2009 linked up with three regional federations. They are working to utilize the eduGAIN interfederation, but the opt-in scheme is not likely to scale well. Attribute release is a major challenge, as some IdP's have very restrictive policies. Another problem concerns "homeless" users (from countries with no proper IdP, or freelance "nomadic" researchers), currently dealt with via a special "homeless IdP" service with limited identity verification.

To close the session, **Anthony Brookes** from University of Leicester talked about data sharing challenges in the context of GEN2PHEN's mission, which is to create a seamless online "knowledge environment" for biomedical information. Sharing of primary research data allows published findings to be reproduced and verified, and also enables data reuse/repurposing. But researchers face many obstacle to sharing data (lack of tools/processes/repositories) and have little or no incentives to do so. They also have a wealth of reasons not to share in many cases, notably ELSI complications for fully detailed, individual-level, person-identifiable data as noted by a previous speaker. Data are commonly categorizes as either i) "sensitive" data that requires stringent access controls (e.g. via EGA), or ii) non-sensitive data that can be published open-access (including metadata). Tony pointed out that there is a third category of aggregate representations of primary data carrying theoretical subject-reidentification risk, which can and should be released with much less stringent controls. He suggested a "speed pass" scheme for rapid, automated approval for data access, based on verified affiliation with a known research organization. He emphasized the potential for using researcher IDs in this context and for ORCID as a global mediator of the required information. He finished by briefly mentioning several projects now in the works, including Cafe Variome, GWAS Central and DataSHIELD, and the new EU-funded BioSHaRE initiative to create distributed computing infrastructure for biobanks.

On the second day, CERN computer security officer **Stefan Lueders** started the final plenary session by reporting on a recent workshop on federated identity in scientific collaborations held at CERN. He first described CERN's challenges in dealing with ~15K users annually, most of whom are guest researchers using the shared research facilities for high-energy physics. Creating local user accounts for everyone doesn't scale, so they are exploring federated identity solutions, not just for managing access to HPC resources and storage power and storage but also to support cross-institution research collaborations more generally.
Stefan proceeded to review the diverse projects/communities represented at the CERN workshop (see Table 2): the European γ/n Facilities, the Worldwide Large Hadron Collider (LHC) computing grid, earth sciences, life sciences (e.g. ELIXIR) and social sciences & humanities (e.g. CLARIN). Some communities have a well-established identity infrastructure, e.g. LHC HPC security is based on standard x509 grid certificates (though certificates are is not a solution suited for non-technical end

users). Others are in the very early stages (e.g. ELIXIR). Key common requirements across communities/projects were SSO, usability for non-tech users, access management across communities, support for homeless users and emphasis on smooth transition from existing systems. Stefan also highlighted the importance of trust. The bulk of technical solutions required already exist and are used in various national/international approaches, such as SURFnet, InCommon, the International Grid Trust Foundation, TERENA Certificate Service and others. More cross-IdF discussion/coordination and a common roadmap are urgently needed.

Representing the American Psychological Association (APA) and the Open Identity Exchange (OIX),

user community	other projects	# users	chosen technology	status	IGTF
photon/neutron	EUROFEL, PanData, CRISP	10,000	Shibboleth/SAML	Umbrella prototype	no
Social Sciences and Humanities	DARIAH, CLARIN, CESSDAH, (DASISH)	hundreds now, potential for 10000+ across SSH	Shibboleth/SAML	CLARIN SP federation - will see if they can use eduGAIN	yes
WLCG	WLCG	5900 globally	X509	production	yes
earth sciences	Earth System Grid Federation, GENESI-DEC, CMIP5, Metafor, IS-ENES	5000+ for CIMP5	OpenID, X.509 and SAML	production - earth system grid	not yet but foresee for EGI integration
life sciences	ELIXIR & potentially 10 BMI ESFRI projects	millions access data via EBI website	no chosen yet	security included in BioMedBridges project workplan	no

Table 2: Status of identity infrastructure in several scientific communities (from the CERN workshop report).

Hal Warren opened his talk with the very quotable assertion “trust is the new gold”, referring to the importance of trust in online identity transactions. He presented a timeline of developments in the identity space since the creation of OpenID and InCommon in 2005, leading up to the creation of OIX in 2009 and later NSTIC (both are US-focused initiatives with government participation). He next presented recent work by the APA on the Publish Trust Project to build a framework for asserting and publishing authorship claims for works published by APA. Strong authentication is used: APA validates authors’ identity by snail-mailing a confirmation code to the institutional address which the author then uses to complete the online registration process. Authors can then start to manage their profile and choose to export or “extend’ some or all of their claims from the closed APA community out to other trusted services (currently based on the InCommon federation) as trust attributes. Claim assertions are exported as RDF/XML Linked Data which are then imported into a modified VIVO instance (used by APA as the external system in this pilot). In the VIVO author publication list, APA-sourced authorship assertions are decorated with a “trustmark” icon which visually indicates their APA-validated provenance.

The final speaker was **Will Spooner** from Eagle Genomics who presented recent work by the bioinformatics outsourcing contractor to secure their cloud-based bioinformatics analysis platform by

applying federated identity technologies. Eagle is doing this work as part of the Pistoia Alliance, a pre-competitive pharma/academic collaboration to build a common data & analysis environment for support research & development. A key question for them was whether federated identity was ideal for such a cross-organizational collaboration. In this pilot project Eagle used the open-source OpenAM toolkit to add federated identity-based authentication and authorization to a multi-tiered architecture platform hosted on the Amazon cloud. Will outlined the main benefits of the approach: convenience to users (no need to create yet another account); enhanced security; less maintenance on the service side by "outsourcing" security and placing responsibility with the client organization, and ability of the platform to scale in no. organizations and no. users without adding cost or complexity.

Parallel interactive breakout sessions

One of the breakouts was titled "*Unique identifiers and the Digital Scholar*" and focused on adoption and utility of ORCID to the researcher community, following the planned public launch of the identifier service in early 2012. The aim was to A) explore the potential of ORCID-integrated services or tools that can offer a compelling value case for researchers, and B) identify courses of action that could deliver that value case in a way that will increase the awareness and engagement of researchers in the development and discussion of identifier infrastructure.

In the week before the workshop, session chairs **Cameron Neylon** and **Jason Priem** sent preparatory notes to participants and asked them to consider the following question:

What are the three features that will drive adoption from researchers, and can be implemented in the next six months?

In the session, participants were split into three subgroups. Each group was asked to discuss the above main question amongst themselves and come up with ideas for three features, giving consideration to two key factors:

- **Cost of change:** Is the feature a big enough improvement over current workflows to cover the cost of changing? Jason noted that only incremental improvement is unlikely to provide enough incentives for people to make the effort to adopt.
- **Bootstrap problem:** Will this feature benefit a given researcher at adoption, or does he/she need to wait until a critical mass also adopts?

To seed the discussion, Cameron suggested several broad areas of functionality likely to be of importance to the researcher end user:

- Not having to do stuff (i.e. saving people time, such as filling out online forms)
- Systems that notify me of things
- Systems that bring opportunity to me
- Discovery tools

Functionality of features discussed in the subgroups spanned a wide range, including impact metrics, citation, authentication and access management, discovery, publicity/promotion, journal and dataset submission, scholarly profile management and more.

Each subgroup came back to the main group with their agreed-on "top three" features. The whole group then voted on the combined set of proposed features, resulting in the following final Top 3:

- ***Integration with manuscript tracking systems***
- ***Integration with data archives and other kinds of scholarly repositories***
- ***Automated CV generation***

This set of Top 3 was then presented to all workshop participants in the breakout summary session.

The other breakout chaired by **Brook Schofield** was titled "*What do researchers need from the IDFs?*". Before the workshop, all IRISC2011 participants were sent a link to a web survey which aimed at studying: (i) the research infrastructures' needs on IDFs and (ii) the IDFs' current service offerings to the researchers. The survey (i) to research infrastructures was filled in by 11 participants from linguistics, bioinformatics, grid/high performance computing and scientific publishing. The survey (ii) for IDFs was filled in by 7 participants representing identity federations in Europe and Australia.

To start the session, **Mikael Linden** of CSC summarized the key survey findings, profiling the responses based on the respondents representing linguistics, bioinformatics, grid/HPC or scientific publishing. Currently, the research infrastructures make use of IDFs mostly for identifying the authenticated end user (to avoid having to issue separate usernames and passwords) but typically not for authorization purposes. The main reason given by survey respondents was the lack of harmonized attribute semantics, such as semantics of the eduPersonAffiliation attribute which can have the values 'staff', 'faculty', 'employee', 'student', 'member', 'affiliate', 'alumn' and 'library-walk-in'.

Research infrastructures that aim at a large user community spanning dozens or hundreds of institutions (such as, CLARIN) felt that easy attribute retrieval from IdPs via IDFs is vital for them. It is a showstopper if the research infrastructure needs to contact hundreds of IdPs individually and negotiate on retrieval of attributes they need. IdPs are often hesitant to release attributes because of European data protection laws which make providers accountable for any personal data they decide to release.

With the exception of grids/HPCs and some bioinformatics services, strong passwords were usually seen as good enough means for authenticating the end users. However, the requirement for new users to present a photo ID when they receive their institutional credentials got wide support. Audits, if any, should focus on ensuring the user registration process, attribute semantics and data quality. A basic level of accountability was also expected; for instance, that a user account on an IdP service belong to an individual person.

The ~40 session participants were further divided into three subgroups to continue discussion on specific topics, including those highlighted in the survey results. Other topics tackled were the IDF end user experience (UX), collaborative tools and different levels of trust. IdP discovery tools like DiscoJuice improve UX by helping to redirect the end user to the correct IdP, the open-source authentication tool tiqr for smart phones and web apps, and OpenConext is a solution for federated group management. A subgroup also reminded that an institution's identity management system may carry user identities with several levels of trust: a researcher may need to show up personally on campus and present his/her photo-ID to receive his/her user credentials, whereas remote students may be able to register on line. An attribute can be introduced to demonstrate the user account's level of trust.

Appendix III: Workshop schedule

Overview

See also overview on http://lanyrd.com/2011/irisc/schedule/

Monday, September 12, 2011

10.30 – Registration opens
10.30 – 13.00 Lunch
13:00 – 13:30 Welcome and introductions by organizers
13:30 – 15:00 **Session 1:** Challenges in identifying and attributing knowledge contributors
15:00 – 15:45 Coffee/tea break
15:45 – 17:15 **Session 2:** E-infrastructure possibilities for authenticating and authorizing researchers whose identity is known/confirmed
17:15 – 18:00 Discussion
18.00 Bus transportation from CSC to the dinner
22.00 The first bus to the city center
23.30 The second bus to the city center

Tuesday, September 13, 2011

9:00 – 10:00 **Session 3:** Internet identity e-Infrastructure – use cases, applications and vision for future
10:00 – 13:00 **Breakout session #1:** Unique identifiers and the Digital Scholar
10:00 – 13:00 **Breakout session #2:** What do researchers need from the authentication and authorization infrastructure (AAI)?
12:00 – 13:00 Working lunch during breakout sessions
13:00 – 14:00 Breakout subgroups report back to main group
14:00 – 16:00 Discussion, conclusions and wrapping up

Session details

Session 1: Challenges in identifying and attributing knowledge contributors

Summary: *Identification of authors and other contributors to scholarly works is a prerequisite for reliable, accurate attribution and research evaluation. This opening session focuses on long-standing challenges concerning scholarly identity, emerging solutions to these challenges, and opportunities presenting themselves, both for conventional publications and for new forms of digital research outputs which are increasingly important in today's scientific research.*

When: 13:30 – 15:00, Monday September 12, 2011
Chair: Gudmundur A. Thorisson
Topics / speakers:

- The scholarly identity ecosystem – Geoff Bilder, CrossRef / ORCID - slides
- ORCID – tackling an identity crisis in scholarly communication - Martin Fenner, Hannover Medical School / ORCID - slides
- VIVO – Semantic Data for Scholar Identification and Attribution – Brian Lowe, Cornell University / VIVO Collaboration – slides
- Evaluation of research, supported by researcher identifiers – Cameron Neylon, STFC / Beyond Impact
- Nanopublication and microattribution – Barend Mons, NBIC / Concept Web Alliance / http://nanopub.org – slides

Video recording from session: http://csc-fi.adobeconnect.com/p9ladf60aor/
See also http://lanyrd.com/2011/irisc/sdwph/

Session 2: E-infrastructure possibilities for authenticating and authorizing researchers whose identity is known/confirmed

Summary: *Researchers need to be authenticated and authorized to access only those resources they are permitted to use. This session studies the existing work relying on the use of institutional identities for access control purposes and presents some early experiences from different disciplines.*

When: 15:45 – 17:15, Monday September 12, 2011
Chair: Myles Byrne
Topics / speakers:

- Linking identity to Research Infrastructure services – Andrew Lyall, ELIXIR / EBI – slides
- Academic identity federations – Licia Florio, TERENA / REFEDS – slides
- Institutional identity: the eduGAIN service – Valter Nordh, eduGAIN project – slides
- CLARIN e-infrastructure – lessons learned – Daan Broeder, CLARIN – slides
- Challenges in sharing sensitive biomedical data – Tony Brookes, University of Leicester / GEN2PHEN– slides

Video recording from session: http://csc-fi.adobeconnect.com/p9fm74fes46/
See also http://lanyrd.com/2011/irisc/sdwpk/

Session 3: Internet identity e-Infrastructure – use cases, applications and vision for future

Summary: *This session studies the approaches taken elsewhere on leveraging other trust frameworks and interacting with other sectors in authentication and authorization.*

When: 9:00 – 10:00, Tuesday September 13, 2011
Chair: Mikael Linden
Topics / speakers:

- CERN June 2011 workshop in identity federations – conclusions and next steps – Stefan Lueders, CERN – slides
- User-centric identity and trust frameworks – Hal Warren, APA / OpenID Society / Open Identity Exchange – slides
- Security and identity in the cloud: a real-world experience of securing academic software for industry – Will Spooner, Eagle Genomics – slides

Video recording from session: http://csc-fi.adobeconnect.com/p85w1cjxshh/
See also http://lanyrd.com/2011/irisc/sdwpm/

Parallel breakout sessions

When: 10:00 – 13:00, Tuesday September 13, 2011

Breakout #1: Unique identifiers and the Digital Scholar

Chaired by Cameron Neylon and Jason Priem.

Summary: *The case for effective and unique researcher identifiers has been made by many stakeholders including institutions, funders, publishers, policy advocates, and technical developers. However, despite this there is at best limited interest in the potential uses and implications of researcher identifiers from a key group, the researchers themselves. In this session we will explore the potential of services or tools that can offer a compelling value case for researchers, and seek to identify courses of action that could deliver that value case in a way that will increase the awareness and engagement of researchers in the development and discussion of identifier infrastructure. The workshop will be discussion driven throughout.*

Programme:

- Introduction

    - Main question: what are the three features that will drive adoption from researchers, and can be implemented in the next six months?
    - present some sample ideas
- Breakout groups: develop and list potential tools and services for researchers
- Report back and discussion: selection of the top three ideas
- Breakout groups: one group to work on implementation of each idea
- Final report back and summary for report back to main group

Notes from session: http://piratepad.net/irisc11-breakout2
See also http://lanyrd.com/2011/irisc/sdwpp/

Breakout #2: What do researchers need from the authentication and authorization infrastructure (AAI)?

Chaired by Brook Schofield.

Summary: *What functionality do scientific services expect from the AAI? This breakout focuses on the requirements such as strength of authentication, harmonization of attributes, compliance and audits of the AAI, ease of adoption to the scientific service and and ease of use for the end user.*

Programme:

- Introduction (Brook Schofield), 15 min
- Results of the pre-workshop surveys + discussion (Mikael Linden) – slides
    - *A survey to the Research Infrastructure representatives*
    - *A survey to the AAI service representatives*
- breakout groups
- Wrap-up and conclusions (Brook Schofield) – slides

See also http://lanyrd.com/2011/irisc/sdwpq/

Breakout results

Video recording: http://csc-fi.adobeconnect.com/p4a8890rydx/

Final conclusions and wrapping up

Video recording: http://csc-fi.adobeconnect.com/p5nhzyw341p/

