

Identifying Researchers on the Biomedical Web

IRBW2009 Workshop, Toronto, May 13-14 2009

Executive Summary

The Internet has revolutionized the way science is conducted and scientists increasingly rely on the Web in their daily activities. Given the pace and direction of evolution of scientific research and the Internet, there is an urgent need for unambiguous and secure ways for researchers to be identified as they use and contribute to electronic publications and online data resources. This need is especially pronounced in the following key areas and activities:

- Creating practical solutions for managing access to protected online resources, particularly sensitive biomedical data.
- Disambiguation of author names in the literature and establishing/validating relationships between authors and publications, both of which are critical to assessing scientific impact.
- Accrediting non-traditional scientific contributions, such as database submissions/curation, scientific blogging and community curation efforts which currently go largely unrecognized and unrewarded.
- Biobanking applications, including services enabling individuals to track how data from studies they have participated in are used.
- Knowledge discovery applications involving some or all of the above.

This executive summary reports on the IRBW2009 workshop which was organized to promote the necessary collaboration, coordination, and awareness to address the challenges outlined above. The workshop was funded jointly by the GEN2PHEN project and Genome Canada, with participation from HUGO, P3G and the Wellcome Trust. Full workshop details, presenter slides, meeting minutes and other materials are available at <http://www.gen2phen.org/event/irbw2009-workshop-may-13-14-toronto>.

Digital identity on the Internet and authentication technologies

- Digital identity, or digital IDs for short, will enable scientists to streamline interactions with various online services which require access to or hold their personal information.
- Digital IDs facilitate *federated authentication*, the process of signing into an Internet site using *ID credentials* from another site, the ID provider. This typically involves the following steps:
 1. On the target website, the user indicates that he wishes to use his digital ID to sign in.
 2. The user's web browser is redirected to his ID provider website, where he is asked to prove his identity with a username and password (or some other, stronger form of authentication such as a smart card).
 3. If the authentication step is successful, the user is returned to the target site, already signed on.
- *Single sign-on* (SSO) across many websites and the provision of a portable Internet identity profile are among the chief benefits of federated authentication, helping to minimize the 'many usernames/passwords' problem and increase security and privacy in Internet transactions.
- Most existing federated identity systems deployed in the higher-education (HE) and research domain have focused on managing and using organization-issued IDs for cross-institutional SSO and access to protected resources. The emphasis on handling a wide range of security-focused '*enterprise-centric*' use cases has tended to make these systems complex and costly to implement and maintain, and difficult to use.
- By contrast, the concept of '*user-centric*' federated identity is currently attracting a great deal of interest in the online community, as a means for Internet users to link together their profiles on various Web 2.0 social networking websites ('identity silos') and control when and how they share their personal information.
- Key to these Web 2.0 developments is widespread adoption of Web technologies and standards such as OpenID, a decentralized, open authentication protocol designed to be simple to implement, which greatly eases the task of building user-friendly identity-enabled applications with modest security requirements.

- A single, universal ID-system for researchers is unlikely to successfully address the broad range of use cases. A more feasible approach is to 'segment' the problem domain and focus on creating practical, standards-based solutions for particular sets of use cases, thus minimizing complexity of each individual system.

Author names, scientific contributions and accreditation/rewards

- Ensuring systematic accreditation is a major motivating factor in encouraging new forms of scientific contributions and integrating them into the scientific discourse proper. To achieve this, robust identity-enabled systems for tracking and unambiguously attributing these contributions to individuals are required.
- The challenge of disambiguating author names is primarily a knowledge discovery problem. Unique author identifiers, generally agreed to be central to solving this problem, need to fulfill certain key requirements critical to the long-term scholarly record, in particular that they be i) *persistent* and ii) *never recycled*.
- There are strong arguments for a centralized system specialized for assigning and managing identifiers for authors and contributors in general, analogous to DOIs for scholarly publications. A central profile linked to a contributor ID, pre-populated with existing author-publication data, and basic tools to manage this profile would be a major asset to scientists and publishers alike. Given that contributors would need to be able to claim and manage their profile, authentication will be a key requirement of the system.
- To ensure community support and trust, a global contributor ID system should preferably be operated by a neutral, non-profit international organization rather than commercial for-profit entities.
- CrossRef, the operator of the DOI resolver system, is currently developing a Contributor ID system and is in discussions with commercial providers of existing services ResearcherID and Scopus Author ID about connecting their systems. This system seems poised to serve as a major focal point for authors and other contributors in the scholarly publishing domain and scientists will have numerous incentives to join the system. Moreover, if CrossRef also serves as an ID provider as an opt-in service for users of the system, this could drive awareness and adoption of digital IDs in the scientific community.

Digital IDs, security and access control

- Leveraging digital IDs for controlling access to protected online resources, such as sensitive datasets, involves at its core three distinct processes which usually take place in physically separate locations on the Internet:
 - i. *Authentication* with an ID provider as described above.
 - ii. Storage and retrieval of access privilege attributes, or assertions, relating to an identity.
 - iii. *Authorization*, or determining based on ID attributes whether or not to grant access to the resource.
- Digital ID transactions usually involve *trust*: the resource provider requires a method for assessing the certainty that the ID credentials presented are trustworthy. An important distinction is made between credentials issued by the user (self-asserted) and those issued by a trusted party (e.g. government, HE organization).
- For use cases where trust and identity assurance are important, enterprise-level features provided by 'heavyweight' identity frameworks such as SAML will be needed. But in many other cases where security requirements are much less onerous, the simpler, 'lightweight' OpenID model may be sufficient.
- OpenID-based data access control systems are currently being explored by several projects, including the International Cancer Genome Consortium and GEN2PHEN. Other projects (e.g. UK Biobank) are also showing interest in utilizing digital IDs for the same purpose.

Recommendations for next steps

- Funding opportunities are needed to explore further options in this important field.
- General awareness of the issues needs to be raised in the scientific community.
- Small-scale pilot projects which could demonstrate the potential of identity-enabled systems to address key problems, ideally in collaboration with EBI/EMBL, NCBI or other trusted vendors in the relevant domains.