

IRBW2009 Workshop, May 13-14 2009

Meeting minutes

Note takers: Gudmundur A Thorisson
Prof Anthony J Brookes

Workshop attendees:

Myles Axton, Nature	Linda J Miller, Nature
Bharati Bapat, University of Toronto	Barend Mons, NBIC / Concept Web Alliance
Geoffrey Bilder, CrossRef	Francis Ouellette, Ontario Institute for Cancer Research
Anthony J Brookes, University of Leicester	Kent Percival, University of Guelph
Michael Dunn, Wellcome Trust	Steven Petric, Elsevier
John Gallacher, University of Cardiff	Saverio Sgro, Thomson Reuters
Reynold Guida, Thomson Reuters	Lincoln Stein, Ontario Institute for Cancer Research
Jennifer Harris, Norwegian Institute of Public Health	Gudmundur A Thorisson, University of Leicester
Yann Joly, University of Montreal	James Walker, http://walkah.net
Karen Kennedy, Wellcome Trust	Susan Wallace, University of Montreal
Margie Manker, The Centre for Applied Genomics	Kazuto Kato, Kyoto University

Day 1

Workshop Introduction

Tony Brookes gave a brief intro to the GEN2PHEN project and high-level overview of workshop objectives. He identified three main domains where researcher IDs will be important: the complementary areas of i) scientific contributions of various kinds and ii) knowledge discovery, and ii) controlling access to data and other online resources.

Gudmundur Thorisson followed with a use case scenario, demonstrating how a digital identity can streamline the process of applying for access to and retrieving a restricted-access dataset. He followed up with a high-level overview of key projects and technologies in the field, as well as highlighting security concerns and related issues that become increasingly important as users consolidate their online presence to one or a few identities, and in the process put their 'eggs into fewer baskets'.

Session 1: Digital identity

James Walker kicked off the first session by introducing the audience to the concept of 'digital identity', with reference to well-known presentation on YouTube by major Identity 2.0 proponent Dick Hardt. An important part of any scenario involving the use of a digital identity is authentication, or proving who you are. James then introduced OpenID as a mechanism for decentralized authentication which is seeing increased adoption in the social networking world. A major motivation for OpenID and similar frameworks is to enable single sign-on (SSO) across a variety of websites where we use services for one purpose or another (i.e. solves the many usernames/passwords problem).

Discussion:

In response to Saverio Sgro's question on trust and security, James emphasized that OpenID does not have a trust model, and is generally designed to be simple and solve one problem well (distributed authentication). The intent is to make it as simple as possible to implement the protocol and combine with various trust strategies or other components as needed for any given setting, thus hopefully emulating how simple protocols like SMTP and HTTP have become the building-blocks of the Internet. OpenID has certain security weaknesses, but its inherent simplicity and relative ease-of-use (compared with other 'heavyweight' authentication protocols) have been

major factors in helping it gain adoption on the Web.

Geoffrey Bilder, Lincoln Stein and others brought up several security/trust-related issues which were discussed. In particular, Geoffrey Bilder highlighted the future role of Web browsers for managing users' digital identity (whether OpenID, Shibboleth or other) and detecting e.g. compromised OpenID providers who have been taken over by malicious parties.

Kent Percival from the Canadian Access Federation presented digital identity management in research and higher education. He emphasized that requirements in this arena are different from those in social networking. To slightly contrast James Walker's perspective, Kent said that digital identity is a complex thing: it is not 'me': it is what a digital record somewhere (e.g. Ministry of Transport, or University) says 'about me', and that it's important to distinguish between organization-asserted and self-asserted ID attributes. Trust is a major factor in research/hi-ed scenarios, and therefore trust and security are integral parts of solutions in this domain (contrast with 'lightweight' OpenID which has no trust model).

Linked identities will be important, means to link up multiple identities at organizations, professional societies or elsewhere, to avoid 'yet another ID number'. Also membership in transient 'virtual organizations' (VOs). The dominant authentication solutions in this arena are Shibboleth and SAML and host of related technologies. Shibboleth already supports other standards such as SAML and InfoCards, and support for OpenID is said to be on the way. This means Shibboleth users could potentially authenticate on websites supporting OpenID. In general, there seems to be a trend towards websites supporting several identity protocols, so users can sign in with whichever identity they have.

Discussion (after session + Thu morning)

Myles Axton asked Kent what the smallest unit of virtual organization: does this need to be a managed org. or some ad hoc group? KP says it comes down to trust (within or outside of the org?).

Lincoln Stein pointed out that there's a tendency to conflate several major components/processes, and thus grossly complicate discussions on this topic. Can we not simplify this and try to separate i) identity/authentication, ii) storage of attributes relating to that identity (i.e. assertions), and iii) authorization based on these attributes? Tony Brookes countered by warning against too much simplification, so we do not miss important aspects of the problem.

Barend Mons provided a link to his Concept Web talk the morning after, by pointing out a straightforward mapping to the Semantic Web 'fabric': each attribute assertion is basically an RDF-triple statement about the person represented by a Concept ID.

Finally, Tony Brookes attempted to summarize the session. He said that the simple, easy-to-use/implement OpenID protocol is clearly heavily contrasted by Shibboleth & friends with real-world adoption in very complex use cases in the researcher/hi-ed domain, but comes with very high overhead + complexity in implementation and use. He stressed that an important objective here must be to keep things simple so as to be manageable.

Karen Kennedy asked how ID could be used to facilitate access to controlled data sets?

Francis Ouellette described the situation with dbGaP access control: one has to be in the NIH system already as a principal investigator / grant applicant, OR get certified by an NIH-affiliated institution which is a major hassle.

Kent Percival followed up on this by saying that there is some work done in his organization regarding recognition of e.g. UK-based IDs by the NIH system. But big complications regarding recognition of the ID itself vs other credentials which go with the ID.

Lincoln Stein pointed out that position/job changes more frequently than one's identity, and so better to keep the authentication in one (permanent?) place and distribute attributes/permissions across the relevant organizations.

Day 2

Session 2:

Author name disambiguation

Reynold Guida of Thomson-Reuters (TR) presented the ResearcherID service which lets authors manage their profile

and help disambiguate their publication record. Service is integral part of TR's Web of Science (WoS), created in response to user requests, trouble with ambiguous author names. Problem also highlighted by librarians. Features include secure access, unique author identifier, consistent metadata, persistent location, localized language etc. Also web service API which WoS, researcherid.com and other tools use. Privacy controls: can hide parts of personal profile if desired. General philosophy is to work towards integration with other systems. 'Sponsor' model: organizations such as universities can use the system directly create author profiles for their faculty, which then follows the author to a new job. He also showed cool mashups with geolocation data.

Discussion:

Jennifer Harris asked if there was some tracking of organization contributions, e.g. to track biobanks. Barend Mons said WikiProfessional is complimentary to this and related services, and stressed that this information is mostly already public anyway (in PubMed/CrossRef), and therefore should be fully in the public domain. Companies could add tools to add value, e.g. handle non-public data.
[someone] asked if there was support for Chinese characters, and how many Chinese users they have. Geoffrey Bilder said that clearly each user controls data about him/herself and so 'owns' that bit of data, and then posed the question 'who owns the entire data collection?'. TR own the collection, it seems.
Mummi Thorisson asked if the ID system is connected with TR-associated publisher manuscript tracking systems, and RG replied saying it is not.

Steven Petric of Elsevier presented the Scopus Author Identifier service, released fully last year. Similar to the TR service, again not an authentication/security application but a knowledge discovery app. They use a sophisticated algorithm to attempt to disambiguate. Require 99% certainty in order for a link to be created, otherwise author name is shown separately. Claim 95% recall rate (proportion of authors pubs linked with Author Identifier). Continuous quality assessment, globally and regionally. Different versions: preview vs full, Scopus users vs non-customers.

Discussion:

Tony Brookes points out law of diminishing returns: one can only automate up to a point, presumably need user feedback to sort out the rest. Barend Mons supported this and said that no single entity can do this job: need to engage other organizations and the community.
Tony Brookes also asked why the author ID part of the system (i.e. not user interface) is not completely open? Steven Petric replied that there's a balance to be struck, with respect to resources spent to maintain system and get good query performance for paying customers vs non-paying ones.
Mummi Thorisson again asked if system is linked to publishers' MS tracking/submission systems, and as with ResearcherID the answer is no.

Geoffrey Bilder's presentation started by emphasizing that we're talking about a 'general Internet identity problem', rather than something restricted to researchers. But this is very difficult to solve: many authentication technologies, often difficult to deploy, hard to get organizations to agree on what to use and how to use it, different requirements. He also talked about trust: how trustworthy is a given piece of personal information? We base this on provenance: evidence to support 'trustworthiness' of information.

A key requirement of many ID systems is persistence. Internet domain names, and anything based on them (e.g. website URLs, OpenID), is at risk because we do not own the domain, we lease it (case in point: <http://nathan.torkington.com>). The scholarly record needs far more persistence than this, hence the creation of DOI system for publications. Same is true for authors: system needs to accommodate inactive authors, including deceased ones. But ultimately persistence isn't a technical issue, it's a social issue. Geoffrey highlighted two major requirements for persistence: A) data portability policy, and B) a living will, i.e. what will happen when the entity running the service (organization, person) is no longer around.

Geoffrey stressed that it's important to realize that the entire spectrum of applications/problem domains for an Internet identity is very broad, and next to impossible to solve it all in one go. Different problems/applications have different requirements; e.g. data access control is largely an authentication/authorization problem and not knowledge discovery. We need to 'segment' the problem domain, try and get something useful running using available information (e.g. mostly public author/pub info) to solve a particular problem.

He then went on to talk about CrossRef's use case analysis for an author ID system. Publishers wanted solutions for common back end scenarios where a unified author ID profile would streamline things, including SSO to manuscript

tracking systems. Their Contributor ID project is centered on a centralized author profile and unique identifier with CrossRef, which could serve as an author's professional profile, photostream etc., and even OpenID provider if user so chooses. They are interested in letting users link their profile with OpenID, to e.g. let people do SSO with publishers and share their affiliation and other information publisher needs for MS tracking. They are exploring options to 'prime the pump, i.e. use Scopus or ResearcherID author-to-publication data to jump-start the system.

Session 3:

Data access control

Tony Brookes talked about access to aggregate, or summary-level data from genetic association studies. He briefly introduced the HGVbaseG2P web resource which is a catalogue of such studies, as well as future plans for an 'in-a-box' version for remote installation, and a federated network of such 'franchise' databases all searchable through a central portal. He then described how fallout from the the Homer et al publication¹ has severely crippled HGVbaseG2P operations. As a result he and collaborators have looked into which data elements can be considered 'safe' to release with no restrictions (vs 'unsafe', e.g. individual-level genotypes), and ways to streamline sharing of the 'mildly unsafe', sensitive data, possibly via a researcher ID based registration system.

Discussion:

Adressing the 'data sensitivity' issue, Lincoln Stein pointed out that if one has acquired somebody's biosample, there are much easier ways to get phenotypic information (e.g. independent diagnostic labs) and identify the person than it is to genotype & mine databases. Protection of data should be proportional to risk and economics, and points to credit card business analogy: rather than massive prevention efforts, fraud is detected 'after the fact' with transaction-monitoring algorithms, and perpetrators pursued.

Michael Dunn made the point that funders and data providers are being super-careful not the least because of concern over the future of the field: e.g. supply of volunteer study participants might dry up if there was a major incident.

Lincoln Stein in turn said that in reaction to Homer et al funders/data producers went after the parties they DO have influence over (secondary data producers, e.g. HGVbaseG2P), rather than journals who do publish aggregate genotype/phenotype information but which they DO NOT have control over.

Tony Brookes again stressed that the current cumbersome data sharing mechanisms create barriers to reuse of data, and therefore hurts scientific progress.

[someone] made the counter-argument to the aggressive data sharing principle, saying that this would destroy the compel/compete/collaborate economy: investigators already sharing data and combining for meta-analysis for certain disorders, this is better than everyone trudging through the data freely.

John Gallacher presented the UK Biobank project which is designed to study common, complex diseases, targeting 500K participants in the UK. They are using sophisticated paper-less tracking systems; e.g. participants get a USB key to carry with them, all electronic. A major focus with the project is on 'keeping noses clean' and avoid incidents, because any blunders become big news, participants withdraw and entire project is a waste of time. Confidence in this and similar project would be destroyed for the current generation.

Nearly all of the data collected is sensitive, as it can be used to identify the sample donor. He contrasted UK Biobank with concepts in Tony Brooke's and others' slides: it is an epidemiological resource, not merely a 'data repository'.

John then described different access models for blood/tissue and other depletable resources: i) resource model => general access, data enclave (common ground) vs ii) hypothesis model => call for proposals for e.g. cancer studies when cohort is ready for this sort of studies (e.g. dementia 30y from now). Either way, all genotyping will be done high-throughput locally, no giving out samples to others, approved researchers will be provided with genotypes and other data from UK Biobank itself.

¹ Homer et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet (2008) vol. 4 (8)
<http://dx.doi.org/10.1371/journal.pgen.1000167>

Discussion:

Lincoln Stein points out that access to study results is different from physical resources: data cannot be used up, and therefore different access constraints should apply.

John G says their interest in Internet IDs mainly relates to researcher access to the resource (samples, data), as opposed to IDs for participants so they can e.g. track bioresource usage.

Regarding data sharing, Susan Wallace said she'd looked through the UK Biobank consent forms and they say project will share data with other projects.

Tony Brookes suggested(asked?) whether using NHS IDs to track participants is critical to make project work, to which John G replies that it would be possible but far more hassle and more expensive than what they ended up implementing.

Lincoln Stein presented data sharing in the International Cancer Genome Consortium (ICGC). ICGC is sequencing samples from both healthy tissue and tumor tissue in 500 patients, compare normal vs cancer genomes to find anomalies which may help to explain disease. Major informatics challenge: datasets too large for HapMap-style centralization, so they went with 'franchise database' model: at each participating research centre, a subset of data is copied into local ICGC database for validation & QC. All local 'franchises' are then queried in a federated way from a central portal at ICGC, so user only sees a virtual front end.

ICGC provides open, unrestricted access to genotype frequencies and minimal patient info (e.g. cancer histology, gender), while there's controlled access for individual genotypes, gene expression data etc. Data access control system will be OpenID based: user applies for access to data access committee (DACO) and provides his/her OpenID as credentials. If access is granted, flag is set on OpenID at the DCC, and then user can log in via OpenID and access the data. They have audit requirements from regulators, and so need to have paper trail for data access requests (which DACO takes care of).

Discussion:

Tony Brookes asked about re-distribution of ICGC data. Lincoln replied that this would involve a legal contract, but couldn't see why not.

Barend Mons asked about possible user here about a 2nd verification step for data access, for extra safety. Lincoln replied that the risk is not to the researcher (e.g. career-wise, for posing inappropriate comment on wiki), but to the patient and his/her privacy. All access requests for each OpenID will anyway be tracked and shown to user on a report page.

Tony Brookes notes that this tracking would in effect enable ranking datasets by no. times downloaded by users, and therefore sort of assessment of which datasets are deemed interesting by the community

In response to Saverio Sgro's question on OpenID security, Lincoln said they will have OpenID provider whitelisting to try and avoid fraudulent access requests.

Session 4:**Contribution recognition**

Myles Axton of Nature NY talked about contributor recognition. He started by outlining the problem of large consortia and generally authors with many affiliations. Currently some journals have a free-text 'author contributions' section at the end of a paper, but Nature experimenting with structured, 'atomic' statements for each author and his/her contribution: author X did this, author Y did that, and so on.

Myles said that the Fort Lauderdale community agreement on data release was excellent in principle, but incomplete in its execution. Its failure is that it provides no mechanism to give data generators credit for uses of their data, so they still try to control use of data they have already made public. Very hard to enforce, and no provision for crediting data generators (i.e. genomics 'factory' laboratories), hence little incentive for people to share data.

Myles introduced the concept of 'microattribution', or tracking people's contributions down to the very smallest meaningful unit (database record, gene), as opposed to whole papers as is tradition. He also discussed the idea of peer review/curation for genetic variants by locus: journal editor could commission annotation of a given locus, curators to get publication credit for this (author X curated gene FOX2). To do the this, we would probably require a central (or several) independent attribution tracking services, rather than individual journals running this each in their own corner. Others could develop additional services/views/metrics on top of the base system.

Anne Cambon-Thomsen spoke about collective authorship vs microattribution, in the context of use of bioresources (biobanks). Different kinds of scientific contributions, different way of representing this (order in manuscript authorlist). How does 'collective' authorship link into this, such as a consortium? There is a difference between saying that 10x individuals contributed to the work vs consortium of 10x individuals did the work.

Anne talked about the primary objective of assessing the use of biobanks (the BRIF concept)^{2,3}

Benefits from having a unique, traceable identifier assigned to the biobank 'bioresource' as a whole (NOT individual participants) range from accreditation of contributions to the resource (PIs, nurses, clinicians, data managers) over to empowering study participants who would be able to track how their study samples and data generated from samples are used. She suggested that the World Health Organization (WHO) may possibly sponsor this and/or assign IDs to resources.

Bharati Babat presented the InSiGHT Consortium and the Human Variome Project (HVP). The HVP is a global initiative to collect and curate all genetic variation affecting human health, from a variety of sources including gene-specific and disease-specific collections (locus-specific databases, or LSDBs), and also national collections. InSiGHT is a pilot for the HVP strategy, focusing on gastro-intestinal hereditary tumors. Progress so far includes all parties agreeing to housing mutation data in three instances of the open-source LSDB software toolkit (LOVD) and acquiring some new members (NIH colon family registry).

Discussion:

Myles Axton noted that it is not just about federating all these databases: an important factor is the add-on pieces of information like contributions, annotations and so on, which create additional value for science. We also need to show this in context of the reference genome sequences, in genome browsers and other portals.

Barend Mons presented the Concept Web, starting with explaining how simple 'triples', or short logical statements about concepts (e.g. 'Barend Mons' <published> 'paper X'), form the basis of a knowledge graph which can be consumed by and reasoned over by automated agents. An essential building block of this is the 'concept', a unique unit of thought, which can be anything from a physical object (protein, person) to more abstract entities (protein function). The WikiProtein knowledgebase contains millions of triples mined from publicly available information in PubMed and other sources, including LOVD instances at Leiden.

Barend proceeded to demonstrate how the WikiProteins system can, with the help of a little embedded Javascript, automatically parse the content of any web page to find known concepts, highlight these concepts inline and show popup-windows on mouseover. Users can create additional annotations (i.e. triple statements) about anything in the text on-the-fly. All such annotations are credited to the user via his/her profile in WikiProfessional.

He also outlined different classes of triples: observational ('hard' evidence) vs curational (somebody interprets some evidence) vs hypothetical (e.g. co-occurrence, found automatically?), with respect to the level of belief we have in a given statement.

Discussion:

Tony Brookes expressed his concern with possible over-interpretation of statements: e.g. variant <something> disease interpreted as 'this variant CAUSES some disease', drawing comparison with controversy over PhenCode project, where LSDB information was overlaid on the UCSC genome browser.

General discussion / wrapping up:

In the last session, Tony Brookes asked everyone to say 2-3 sentences on what they thought was most important as the take-home message from the workshop.

² Cambon-Thomsen. Assessing the impact of biobanks. Nature Genetics (2003) vol. 34 (1) <http://dx.doi.org/10.1038/ng0503-25b>

³ Kauffmann and Cambon-Thomsen. Tracing biological collections: between books and clinical trials. JAMA (2008) vol. 299 (19) <http://dx.doi.org/10.1001/jama.299.19.2316>

Anne CT highlighted the connection between IDs for resources and IDs for individuals, in the context of tracking bioresource usage.

Francis Ouellette said it was important to distinguish between the 'author name problem' and the 'data access problem', i.e. the needs of 'data accessors' vs needs of 'contributors'. Both involve IDs for people, but conflating these two problem domains makes things unnecessarily complicated (see also discussions after 1st session). Francis also mentioned the 'slippery slope' of different levels of access: unless strictly required, best to keep things simple and only have open vs controlled access. Tony Brookes disagreed with this and stressed that we'll need at least 3x levels, which Francis ended up agreeing with. Lastly, Francis said we still have an open question: what is a 'bona fide researcher'?

Geoff Bilder yet again stressed the separation of requirements for i) knowledge discovery and ii) authentication, to keep things manageable. Authentication requirements can range from not-so-critical (relatively easy to do, e.g. OpenID model may suffice) over to super-mission-critical (very hard to do, need more 'heavyweight' tools). Geoff also said that CrossRef's Contributor ID plans to support several auth protocols (OpenID, Shibboleth) rather than focusing on just one. He also wants to build in a 'delegate' feature, so when say a member of faculty passes away, the University dean or other representative can 'disconnect' or disable the profile. He also pointed out the need for a generic mechanism for digital signing access permissions: e.g. somebody's identity could be signed by 4x organizations.

Mummi Thorisson said it's good that people generally agree on trying to keep a separation between authentication and authorization. Also points out that it would probably be fine to use OpenID for data access (as ICGC will do), even if they will not 'last forever' (as per Geoff Bilder's criticism), because access permissions will likely be only transient anyway (e.g. 1y) or require periodic renewal by user. Geoff Bilder disagrees with this: often a permanent audit trail is needed, so important that the ID you give access to now refers to the same person 10y from now.

Francis Ouellette said we need some way to sanction 'evil' researchers who subvert the system, e.g. give sensitive data to people who are not authorized. Saverio Sgro asked which party/parties should do said sanctioning. Francis also called for more granularity/flexibility in granting access permissions: for instance, a PI or head of department should be able to grant/revoke access to students or postdocs, and take responsibility for them. Sgro pointed out that such vouching could perhaps be decentralized.

Margie Manker said that given she is redesigning the DBGV structural variant database infrastructure now, it was very valuable to get insight into data access issues at this stage.

Myles Axton revisited the point Geoff Bilder made about pre-populating Contributor ID profiles (in future CrossReg system) with available public data. If this were propagated across the publishing domain, publishers would love it, would enable big savings in editorial time+cost. He also again raised the question of whether we need a central reputation/accreditation tracker service.

Geoff Bilder mentioned that some of this could be done with microformats and emerging mechanisms for claiming any web resource (e.g. regular web page) via one's digital signature. Myles said this could possibly be done on the publisher's website.

Karen Kennedy said that technology shouldn't drive policy, we shouldn't set up a 'cool ID-based systems' just because we can. There should always be ongoing assessing and re-assessing/revisiting which data should be shared freely and which ones not, in a risk/benefit analysis process.

Michael Dunn said he came into this knowing almost nothing about the identification situation, and will return from the workshop to Wellcome Trust as an expert-of-sorts on these matters. Lots of potential for making things work smoother.

Barend Mons suggested a lexicon for technical terms / jargon (e.g. authentication vs authorization). He also reiterated a previous recommendation for a model where public, non-commercial information is fully open, and companies build added-value services around this, and the whole thing linked together with data standards,

globally-unique IDs of concepts etc.

Tony Brookes said in closing comments that it was striking how inter-disciplinary the workshop had been, so in that respect the meeting had succeeded in bringing people together to learn more and discuss these important issues. He said further education is needed, we need to communicate this to the community. He proposed that the natural next step would be to write a short paper for the general reader, and asked who would be interested in contributing to this.

Tony also suggested that people come up with some small-scale proof-of-principle demo/pilot projects that could help to move things forward, including possibly engaging EBI, NCBI and other major players and encourage them to participate on some level.