2023-01-10 DMRWG Meeting Notes

Meeting Date

10 Jan 2023 The DMRWG meets bi-weekly on Tuesdays at 12:00-13:00 PT / 16:00-17:00 UTC. Check the ToIP Calendar for meeting dates.

Zoom Meeting Link / Recording

- Recording 10 Jan 2023
  (This link will be replaced with a link to the recording of the meeting as soon as it is available)

Attendees

- Neil Thomson
- Steven Milstein

Main Goal of this Meeting

Data, Repositories and Registries (incl. Verifiable Data Registries, Trust Registries, and other forms) - *Requirements, Similarities, Differences, Use Cases*

Agenda Items and Notes (including all relevant links)

Time	Agenda Item	Lead	Notes

5 min	• Start recording • Welcome & antitrust notice • Introduction of new members • Agenda review	Chairs	• **Antitrust Policy Notice:** *Attendees are reminded to adhere to the meeting agenda and not participate in activities prohibited under antitrust and competition laws. Only members of ToIP who have signed the necessary agreements are permitted to participate in this activity beyond an observer role.* • New Members:
5 mins	Review of action items from previous meetings	Chairs	Work Nov 2022 to end of year • Charter for the group • Burak Serdar - Layered Schema for Source-to-Destination Data Transformation
5 mins	Trust Registries, Repositories & Authentic Data	Chairs	How does the Data Modelling and Representation WG contribute? Trust Registries will be a 2023 focus for ToIP, DIF, and other tech working groups. However, any "registry" is also a repository whose contents are data. • What are the requirements and use cases? • How do they deal with (and leverage) "Authentic Data" and "Authentic Provenance Chains"? • How are they governed, and by what organizations via what relationship between legislation, regulation, and content contributing organizations, etc.? • What are the core, "atomic" level data requirements within these Identity Ecosystem services? Slides: Here
50 min	Discussion on Trust & Data Registries and Repositories	Neil, Carly, Burak, Steven	DMRWG Recording 10 Jan 2023 Slide 3 - Agenda **Neil Thomson:** Trust Registries are bringing forward the first full use of "Authentic Data" in ToIP where an entity, such as an Issuer, is only admitted to being listed in the registry if they can demonstrate that their identity is backed by cryptographic and governance directly traceable (potentially through a chain of data, certificates, etc.) to the cryptographic and governance "root(s) of trust". A presentation was made at Ecosystem Foundry WG on 5 Jan 2023 on the use of verifiable credentials for academic and work experience to support a "smart resume" serving both job applicants and hiring organizations. EFWG Recording 5 Jan 2023

It struck me that their approach to using VCs shows that while it was using crypto-signed Verifiable Credentials, they were "weak" on provenance /traceability as, for example, academic credentials are packaged in VCs, which were crypto-signed, presumably by the SmartResume organization, on the authorization of a University Dean by way of an ink signature. There was also traceability to the marks given out in any course or to the credentials of professors/teachers that taught the courses.

This brings up the question of the credibility and usefulness of a verified credential.

For example, I went to high school and a local lower-level college (CEGEP) in Quebec, Canada, where we wrote province-wide standardized exams.

Arriving in Ontario, I discovered that each high school or school board was setting their own exams.

In talking to the admissions team at the University of Waterloo, they said this non-standardized curriculum/exam as the basis for issuing academic credentials (High School graduate) was a nightmare. While they could effectively compare students from within Quebec for admission and curriculum, the Universities and Colleges in Ontario had to individually or collectively set up a process to evaluate the Ontario schools/boards on a comparative basis, to upgrade or downgrade subject marks of students, based on their school/board to compare them for university admission. A related problem is the lack of consistency in the pre-university curriculum in Ontario high-school. Universities found that many school boards did not meet the expected curriculum minimums, resulting in high failure rates in 1st year - to a large extent because they did not have the education base to understand the 1st university course material.

So the focus (in 2023) for the Data Modelling WG is Authentic Verifiable Data, focusing on support for Trust and Data Registries. This includes pragmatic and requirements-driven use of:

- Authentic Data Chains - where data collection and processing is a series of linked events back to technical (cryptographic) and governance roots of trust
- Data Access Governance for Authentic Data/Chains

The last 24 to 36 months have established the Issuer, Holder, Verifier triangle and mechanisms for creating Verifiable Credentials created by an Issuer for a Holder, which an application or other person in the role of a Verifier can verify.

The issue of Trust Registries is to provide a set of authentic, verifiable Entities (- people, organizations or devices) that can be trusted. The intent is to provide the digital identity equivalent of aircraft and airlines being certified by an aviation regulator (e.g., FAA (US) or EASA (Europe))

One of the first pilots by the British Columbia (Canada) government's SSI team was to set up digital identities to be registered and verified within an early Trust Registry for small and medium businesses.

As such, it becomes a "lookup" service for small/medium businesses, regulated and registered with the province.

At the most basic level, a registry is just a data structure - a variation on an LDAP server. To become a trust registry, it needs to be verifiable as to the data, the organization that produced the data and the governance that stands behind that data (and metadata) being true.

While crypto-signing of data, including by the producer or the data and the governance authority for the data, is only part of the requirements - the governance - how it is done, the qualifications of those who do it, which ideally are verified by an accreditation, certification and auditing process, is equally important.

While some aspects of Governance can (and will) be automated, in general, throwing additional code at governance does not increase its trustability unless trustworthy people verify it.

Side 5 - Authentic Data: Simple in Principle

An example of data traceability - the BC government mandates reporting of greenhouse gas emissions data from in-province organizations based on the amount of such gases released untreated each year. The data and governance of gas emissions go through data collection, processing, aggregation and auditing through several organizations before it arrives as a gas emissions report at a BC government ministry. Currently, there is no verifiable way to trace the report results back through all the touch points in the data collection and processing (e.g., IoT devices/sensors or human-operated/monitored emissions testing).

What is the difference between Verifiable Credentials (VC) vs. Authentic Data? If a VC is signed using a private/public key pair such that the VC can be verified as produced by the private key owner, is that not sufficient?

A VC is an end-product of potentially a series of data collection and verification steps to validate different claims within a VC. A VC has an **Authentic Provenance Chain (APC)** when each of those steps is traceable via cryptographic signatures of each collection or processing of data back to the source. For example, a VC for a driver's license needs to be traceable to the organization that administered the driving test, which must be able to produce the certification of the examiner that administered the test, plus verification that the driver is, in fact, qualified to have a drivers license (citizenship, age requirement, driving license records (e.g., suspended)), etc.

And it is not sufficient that there is only a cryptographic root, there needs to be a governance root that verifies the process that created those cryptographic signatures (e.g., are the private keys used to sign a data structure secured)

Slide 6 - Trust Registry vs. Data Registry vs Data Repository

Trust Registries are a more formal approach to establishing a catalogue of trustable entities, as has been routine in the banking industry for years through a process known as Know Your Customer. The banking standard is that a bank will collect a great deal of personal data, including financial history, citizenship, where you live, etc., before issuing a bank account. Having a large sack of cash to deposit doesn't qualify you for an account. In 2023, that is a red flag for potentially suspicious financial operations (e.g., money laundering), which is a persistent problem.

Slide 7

A question on data publishing will need to resolve the Data Subject (who the data is about or associated with), who owns the Data, who controls the Data (e.g., guardian/delegation case), and who will govern access to the Data once is published - this could be the same individual or multiple individuals, organizations and roles.

There is also how data changes over time. A VC is essentially a token (albeit a token that may need a periodic refresh) which is very different from a data source (e.g., greenhouse gas emissions) that is continually being updated with new information. Management and governance of that data will be different depending on whether it is raw, real-time data, or periodic, aggregated updates. These pose different challenges from the perspective of Authentic traceability and governance.

Another question is, who gets access to the data supporting traceability and verifiability? A case in point is data about legal firms or legal information about an individual or an organization. That information may or may not be contained in the Trust Registry. It may be stored in a separate location/data store depending on the type of entities registered. For example - in the legal case - see sensitive and private data - such as criminal records and arrest history.

Slide 9 - Use Case - University Data Reuse/Publishing

One issue is that self-attested qualifications, skills, and history can be very different from verifiable claims by authoritative entities in credibility and form.

Carly: Today, I may find a data source/block of data attributed to a researcher that I download. I may not know the researcher, but I would trust (or expect to trust) that the data repository governance did some level of verification on the data and researcher before agreeing to include the data in their repository/register.

If I wanted more data, I might Google the researcher's name and see if their name comes up in citations or on university websites or other data repositories. I might also use an (academic) publication search engine (such as PubMed) and view some of the researcher's publications and who their co-authors were, including looking for names I recognize from my personal network.

And then there is the case of the author, who may have worked at a university 10 years ago but can't be found in recent searches.

Then I may connect with them personally - via email or phone - to ask some qualifying questions as the current level of documentation on published online data is not enough to trust.

So today, the reliability of an academic data source is largely based on the academic reputation of people you know, which also tends to restrict the use of published data to academic researchers who personally know each other.

Information and reputation are subjective as well because one person's crazy-eyed theories are some other person's idea of a genius new hypothesis. Reputation is not a single, quantifiable value (across a spectrum of peers and others who provide input on an individual's reputation.

And the techniques/methodologies that were used to collect and/or process the data may be unique and need to be verified and validated, which may be labour-intensive, up to and including attempting to independently reproduce the data (which has always been the ultimate test of academic research).

There is also the issue of a researcher who is publishing data outside of their professional expertise.

The net result, you end up with cliques of researchers

Slide 10 Use Case - University Data Reuse/Publishing - second user need

Carly: A problem with academic research data repositories is they may restrict access based on the academic field of the requester vs research publisher and within a university or country. Consequently, it is a manual process to give people access to your data. Certainly, a global repository for academic researchers would greatly simplify sharing academic data.

Academic research identities are very much tied to personal names, which can be ambiguous, particularly in Asia. There is a self-attested registry for academic researchers called Orcid, attempting to address this via a unique identifier.

Yes, there is an issue with individuals producing academic data, but then there are also organizations such as the Hubble Telescope or Atlas (mega physics projects). This is very project specific. Then there is the case where the register collects data from many different sensors. Do you trust me to have vetted the raw data correctly?

There is also that many of the data collectors/processors are trainees - grad students - who are doing this for the first time. And compliance by grad students will vary from class to class, year to year.

And as all things published on the web, the feedback from other researchers looking to use the data is not necessarily going to be constructive. Criticism (whether justified or not) is not unknown. That's a risk for which grad students can be high (for their reputation), which makes many very reluctant to publish for public access. Which can be career-destroying. Yes, that should be covered by a good supervisor, but all supervisors are not necessarily careful, or they may lack the math/processing skills to spot processing errors.

The consequence is that someone will only be credible after publishing a series of sets of data and analyses that prove useful.

Then there is a potential situation where the conditions of the data collection or experiment are not fully controlled. Example: a lab testing for nitrogen impact, not realizing that they were being impacted by a fertilizer plant upwind, which invalidated work, which was only validated when they attempted to reproduce the experiments and testing elsewhere.

Validation of academic data may come much later - after cross-verification by other researchers (or other means).

Case in point - the recent reports on gas stoves is a case of data that has been recognized as valid after repeated tests by different researchers.

Burak Serdar - there are repositories for health care data in the US, where perhaps the data quality and reliability problem is not so much a factor due to data collection, but where information about the organization publishing the data is leading to the main uncertainty on the data. There are steps to start addressing those concerns, but there are also many such repositories being managed separately (without an agreement on governance/quality).

An issue with medical data is that information collected and processed under HIPAA guidelines (and the metadata for the data) may be different than through other guidelines, such as SAMHSA (42 CFR Part 2) known in the industry as Part 2, which is concerned with substance abuse. So data about a patient contracting Flu (which is not a substance abuse-specific issue) may be reported quite differently using HIPAA vs. Part 2.

Another example is, say, birthdate - get that from HIPAA can't disclose due to privacy regulations, but if you get it from a public library, it's not governed as strictly on privacy.

Carly - Then there are standards as applied to data collection and processing, of which standards are not generally applied for academic data - except for data repositories that have specifically stated standards for applicants to meet.

The problem is that historical data is used extensively (e.g. 30-year-old field trial of alfalfa fields), which will not have been done with current standards. So there an apples/not quite apples/oranges on data sets that will be a fact of life even after standardization (if, in fact, that happens).

5 mins	• Review decisions/action items • Planning for next meeting	Chairs	

Screenshots/Diagrams (numbered for reference in notes above)

#1

Decisions

- Sample Decision Item

Action Items

- [ ] Sample Action Item