2023-02-7 DMRWG Meeting Notes

Meeting Date

07 Feb 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 - meeting starts at 9:00, discussion prior to that is captured in the notes (below)
  (This link will be replaced with a link to the recording of the meeting as soon as it is available)
- Slides Authentic Provenance Chains for Verifiable Data Registries

Attendees

- Neil Thomson
- Steven Milstein
- Carly Huitema

Main Goal of this Meeting

Authentic Provenance Chains - Functionality and Data/Structure requirements for Verifiable Data Registries.

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 meeting	Chairs	
5 mins	Announcements	TF Leads	News or events of interest to Data Modelling & Representations WG members: • -

20 mins	Authentic Provenance Chains	Chairs	A key requirement for Verifiable Data Registries is an audit trail of all the processing and all the people (producers and governance stewards) who touched the data from the point of data creation/capture through a published dataset. This includes the following requirements:
			• Verifiable proof that the data/data set has not been tampered with
			• Verifiable proof (and identification) of the person responsible for:
			  ○ Processing the data (including collection, clean-up, aggregation or other analytic processing, testing,...)
			  ○ Who audited/ensured governance compliance for data collection and processing
			
			Today's discussion is to look at a model of how that could be implemented in exploring the functionality and data (and data structures) required for Verifiable Trust Registries and their listed trusted entities to be trustable in a robust and secure manner.

The presentation today is concerned with Authentic Provenance Chains - a "Trust chain" (which ACDC supports) using cryptographic signing of data (and all objects are data), counter-signed by data processing authority (the representative of the organization/team that processed the data) and the governance authority (the organization representative responsible for ensuring governance compliance), which is also chained to previous data processing steps and support artifacts (processing scripts applied to the data and testing)

We will look at Verifiable Data Registries - which are registries of objects (which are data), which have been verified and validated by a governed process which performs "due diligence" on the objects and their data and processing lineage - be it with sw/hw, manual processing or some combination.

And Trust Registries are a sub-type - a Trusted SSI Component/Rol Registry, of which the subjects of the first use case are Issuers and Verifiers

The underlying principle for trust chains is the principle of "Authentic Data", where the creator/publisher/owner of the data uses a private key under their control to sign a hash of the data (a crypto-hash), publishing the corresponding public key, which can be used by any consumer of the data, by verifying the crypto-hash of the data (with the public key)

There may be a distinction between a Verifiable Data Registry (Authentic Data Registry?) and a Repository. An V/ADR provides proof of the data's lineage, a Repository may be the database which provides access services for the data (and the two may be combined_.

Data can go through a series of processing from collection/input to publishing:

- Collection (manual data entry or collection via devices) - produces a Raw Data Set
- Consistency dataset - where records are checked for valid attribute/property values
- Cleaned dataset - where outliers, duplicates, and values across multiple record types are "nonsense" values
- Published dataset - where data is checked for "correctness" and is "fit for purpose", possibly validated against related datasets

A data trust chain will link the published data set to the raw dataset and to the personnel or devices that collected the data.

Note: metadata associated with the dataset (e.g. OCA model), plus software licensing and other data administrative artifacts should also be included /linked.

What is different about the Chain Element (in Authentic Data Provenance Chain - ADPC) vs ACDC is

- ACDC - is Authentic Chained Data Container
- Data Provenance Chain - is a container only of links to other links in the chain (which may actually be a mesh or tree) and links/references to the data and supporting information.

The principle here is that Data and supporting documents are not moved to the ADPC but are referenced - where the references are crypto-secured. This allows the same data and artifacts to be included in multiple ADPCs and partitions the chains from the storage management of the referenced data.

Question: - where will ADPCs (the links) be stored, and what structure/model?

Structure & other requirements include:

- Tamper-proof or immutable storage
- Scalable
- "Offlineable" - migrate historic changes to cheaper, lower access speed storage, but make them easily accessible.
- Decentralized friendly - replication, distributed for access
- Authentic: structure, data and relationships are "authentic" - e.g., signed and hashed
- As simple as possible
- Purpose-built
- Repurpose general SSI/Data/Authentic tools, libraries, etc., that will improve over time.

Suggested approach: a variation of the KEL - KERI Event Log, along with its replication/distribution and secure structure approach

Where ADPCs are stored or managed will be up to the Ecosystem.

Persistence of this type of audit trail can take a page from the capture of aircraft and trail system black box logging or any operational system where accident/incident analysis requires records to be kept for potentially years (in some form, not always full fidelity).

The issue of private/sensitive data that may be on the chain was discussed.

The onus on proving that a provenance chain exists is something that will be publically required, but that can be a form of zero-knowlege-proof (not necessarily cryptographic).

Privacy of data in an ADPC: Access to, say, the underlying unpublished data and/or names and identifying information of the data processor and data governance representatives should be tightly controlled and only revealed in the case of a review (e.g., accident, incident). That may suggest that there are public and private views of an AD, possibly in two separate but linked chains.

An example of this would be an Issuer (e.g., GLEIF) who must perform due diligence (Know Your Customer) on a person or organization. In order to issue a verifiable credential, idealy, the issuers should have a record of all the due diligence information captured and referenced in ADPC links, but that information is available only for an authorized external audit. Otherwise, it is sensitive and private data.

Assumptions:

- SAIDS + GUIDs for ADPC link elements and contents
  - SAIDs - The published data, the support docs, signed components - all the artifacts, including the query-able data - will be identified by SAIDs - Self Addressable IDentifiers - meaning a crypto-hash of their content is the identifier. What has not been solved (that the note taker is aware of) is if the use of SAIDs has some resolving mechanism (as for DIDs) that can locate an object anywhere online just with the content-based SAID as an identifier. Not clear how that scales to trillions of online 'things" with SAIDs, DIDs, etc.
  - GUID DIDs - as link elements or the contents they refer to are immutable (create a new, updated object vs. modify an existing one) this implies versioning so that for a given object, each object has a SAID, based on its content, which will be unique for each version, plus a stable GUDID (Globally Unique DID) which is consistent across each version, and most likely a version number (to provide sequential order which a SAID will not provide), requirement TBD.

Observation - the most important part of the ADPC are the signatures and their provenance - back to the people authorized to sign.

APIs - some requirements

- For walking the chain, including to the roots of trust (crypto and governance).
  - Proof-only chain walking, which only confirm that the links are all verified/validated, and the roots of trust are registered in some Trust list
  - Detail walk, which can access the underlying identifiers and owners who signed the data and as representatives of processing and governance

More on walking to the root of trust. What are the trust checkpoints in interacting with any entity in an SSI? Some observations on trust checkpoints for the other party in a two-party exchange:

- Prove you have a validated and compliant :
  - Identifier
  - Key management system
  - Key rotation system
  - VCs that I require (context - the type of party and the type of interaction)
  - Verify your Wallet or other executables with the crypto-hash of the actual .exe
- If interacting with a Verifier or some other Role/Service component
  - Are a member of a trusted trust list

Unknown:

- How does Reputation fit into all this? Can it be tamed to have reliable rules that give fair a 'non-bot-able' ways of artificially increasing or decreasing someone's reputation
  - Papers by Sam Smith on Reputation High Level, Low Level

Academic reputation (of data & data sets is based on)

- Cite the Author, City the Paper, Cite the Dataset
- How to track use of data
  - Can get a count of downloads of the data
  - Also attributions to data in other academic papers
- Rules vary from person to person on their opinion of other academics and trust tends to be a weighted score

Observation on a verifiable credential application for resumes, which consist of "signed" affidavits of an academic degree and of time, skills and job description for employment.

- In this case the provenance trail for the VC was literally signed (including scanned) document from a University Dean's office attesting to the earned academic degree of a student and presumably similar to attestation for an employer verifying position and time with the organization.

This is a starting point, which could be extended from the paper evidence to crypto-signed replacements, which get deeper into the backing information over time (e.g., from a degree to marks on each course, plus links to the teacher/professor of the course, etc.

Looking at these different trace paths - they are a labelled graph, a top verified object - the academic degree which follows to multiple roots.

Observation - any SSI object (entity, dataset, VC, etc.) should have an API to allow any object to provide proof of its own history and provenance on data and trust checkpoint criteria

Assumption: there is a partition between data about an SSI object in a Trust List/Registry or in a ADPC chain vs. data access through the object.

> There may be information about the history and provenance of medical data records, but the actual medical data is not accessible by the Trust List/Registry API, only it's proof chain and data hash. The hash of the actual data would be delegated to the SSI object, preserving privacy control with the data and data controlling component.

Summary (by one of the meeting members)

So the components for this kind of system

- someplace (many places) where the data is stored.

- a provenance chain that is either stored on a distributed ledger OR stored locally but occasionally hashed onto a distributed ledger

- some kind of VDR where the keys that were used to sign chain elements can be verified.

- some kind of governance of the ecosystem (components of this ecosystem) which establishes the necessary standards to ensure trust in the provenance.

Observations:

- There is a case to be made for having PKI sets only for crypto-signing of data vs. PKI sets used for every day messages and interactions. There may be a difference in the level of management and the level of security. The characteristics would be:
  - Signing keys are used rarely and are long-lived (e.g. the public key for a signed data block may well be effectively permanent)
  - Interaction keys are used often and are short-lived - the public keys may deliberately have a short shelf-life for security reasons
- <u>There is no reason that Authentic Data, trusted data lists and Authentic Data Provenance Chains can't be used with legacy and other non-SSI systems.</u>

15 mins	SSI Risks		A core principle of Governance is to prioritize based on risk for an application, service or data source, the organization, its personnel and its financial /legal state/reputation. The current state of cybersecurity points to a number of key risks to which SSI is (also) vulnerable: • Message interception and decoding • Message and service hijacking (imposter, identity theft - personal and organization) • General impersonation, including falsified credentials/claims, etc. • Capture and exploitation of data (undetected) Some examples pertinent to SSI: • Use of a compromised identifier by yourself or other parties • Hijacking PKI management or private key theft • Accepting VCs from a compromised Issuer • Use of compromised (including fictional) data • Interacting with compromised trust components (Trust Registry, other ledgers) • Replacement of cryptographic or other security-related components and services with a compromised version For example, if you are a root administrator on a server, and a service relies on the server's cryptographic libraries, the admin can swap out the libraries for a compromised set. "Authentic" methods, e.g., crypto-signing an executable and related support files (the libraries), controlled by the actual administrator (or other authority), can provide tamper resistance as the service can verify the library by using the public key of the signing authority by rehashing the executable's signature and verifying w the public key. Another example: "If you are not the person who is building the compiler, anyone can add a back door and you would never know".
5 mins	Any other business		
5 mins	• Review decisions/action items • Planning for next meeting	Chairs	Plan for the next meeting (Feb 21, 2023 - Sam Smith will present on the experience for authentic provenance chains with the GLEIF project using ACDC.

Screenshots/Diagrams (numbered for reference in notes above)

#1

Decisions

- Sample Decision Item

Action Items

- [ ] Sample Action Item