Sticky Fingerprints on Personal Data

Is Federated Digital ID Enough?

WHITE PAPER

Version 0.3
Released: Sep 2020

Cox Financial Consulting, Australia http://www.coxfinancialconsulting.com.au/

Who is Who in the Identity Zoo(m)?

The surging move to digital ID has caused academics, government and practitioners to revisit the very essence of identity. Not only is the application of identification overwhelmingly not in-person, but now it must be virtual, portable and regularly revalidated and refreshed.

In addition, identification is now clearly separated from verification. Both, ultimately, are circumstantial, but while identity just piles up artefacts (even with biometrics) about the person or other legal entity, verification looks for means to indicate that they apply to the very person right now proffering their data, or requesting a status or service.

With TrustID, connectID, MyGovID and so on moving to interoperability, new universal, flexible, risk-defined validation services are becoming available, soon no doubt to be mandatory.

Terrific.

That potentially takes care of the top-down assurance of the source and ownership of the data and its custodian-ship by the recipient and the basis of access for the 'relyer' or downstream consumer of the data. What about the data itself? What means is there to follow the data through its lifecycle so that the Consumer Data Right is really fulfilled?

Behind Closed Doors

We all know that our data is fairly thoroughly exploited, especially in the private sector where the business model often relies hard on it. This may be due to a fat EULA or broad privacy statement going in, where the recipient/service provider reserves pretty-much all rights and the user has no power to negotiate. Take it or leave it.

There often seems to be a second level where every organisation is tempted to collect every possible scrap of data about an individual to feed into the dripping maw of an AI engine. This flies in the face of privacy principles that (paraphrased) mandate only the data required to satisfy to individual's need or request, plus legal or regulatory requirements should be collected. Nothing extra. Line of sight.

The Consumer Data Right seeks to change this at least in part by putting

legal weight behind the non-expiring ownership of one's data and a means to ensure that receivers set their structures and products up to endorse that and provide means to show they are compliant.

With that in place there are still at least two problems remaining. One is the fact that what an organisation signs up to and what actually it does diverge serially. Companies get paid to use and sell data so the motives are misaligned, mistakes are made, and data is misused, corrupted or exposed. Individual personnel may also exploit or collude to compromise their custodianship regularly. Anyone who has undertaken, or undergone, a systems and control audit, knows that practice, even in the best of companies, always falls short of promise. No one gets it perfect.

The second is the value and classification of data itself. Everyone would agree that name, address, and date of birth are key examples of personal identity data especially sensitive when coupled with anything else personal and individual like your pregnancy test result, your political party membership or your salary. But who considers, say, the humble postcode and even gender, when de-identified from the individual's name, to be sensitive? Well for one thing, that data, when aggregated, has substantial value in a range of contexts, about representation, use, predilection and so on.

Also, possibly more importantly, there is always the risk of data uniques, when a de-identified dataset is overlayed by other datasets, for example electoral, or rate payer and the de-identification rolls back. That medical procedure you had recently. Very private to you. But is there only one person of say your age, gender, earning band, rental status and postcode in the last three months that has had it? Who also has an Android phone and drives an hybrid? Add more databases and it just gets narrower. Visa claims that it checks 260 data points about your card transaction to prevent fraud. In real time, as you swipe your card. They can see an awful lot about you.

We want to corral our cattle carefully at the entry end but all the smaller animals can duck under the single strand at the other end and head for the open range.

Bottoms Up

Just maybe we need a bottom-up approach alongside the top down of identity controls. This is where the idea of fingerprinting the data at first input bears some thought. We want to encrypt data at rest anyway, as best practice. Why not make that a permanent personal trace-point back to the owner. Some practitioners of the TrustID framework already encrypt each incoming datum, using the individual's TrustID. We would endorse that, using a base key derived from the TrustID on the originator, but not the ID itself. Each datum is encrypted individually, but using the same derivative key for every datum in this group of data.

Why? The idea here is not to make the data inaccessible, it is to ensure that an electronic request to the custodian, by way of drawing the key, is required to access the data. This then forms a notification event that the data is being used as authorised. The scheme also needs a law that requires to ensure the data is always stored in this form – 'fingerprint on'.

The process then does five things:

1. The requestor is tested for permission to use the data from the data owner or their custodian.

2. It discloses/traces who actually accesses it (down potentially to role and even the individual, depending on the arrangements made).

3. It notifies the owner or custodian that an expected process step is underway.

4. It ensures each personal record stored encrypted in a recipient's database has a unique key. That means that a hacker doesn't have to crack the code once, but separately for each personal data record present. A big incentive not to bother.

5. Perhaps most important of all, it means that after the data has been sent, the owner can still 'turn it off' or rescind access to it well after the initial submission, by issuing no further keys.

What's next?

As well as simple data exchanges, this model is intended to operate across complex transactions involving potentially many participants or providers. If the recipient of the data changes the data, or adds further data, such as in a processing step of a multi-step transaction, they request a new derivative key from the custodian to re-encrypt the data. This is done for two reasons.

1. It keeps the modified dataset discrete and traceable by the original custodian representing the owner.

2. It enables the recipient to pass on the data to another recipient with an inherent authorisation of assent for them to use this personal data. This is revalidated as the new recipient draws a key to access the data.

How does this work as a system?

The intention of this model is to use the TrustID framework and its compatibles in a complex transactional setting.

a) It requires that all recipients of the data are previously registered with the TrustID provider representing the individual. Nationally. Overtime, globally. Since there will be a limited range of trust providers, this is just getting a ticket to the game, like a Certificate Authority or a DNS registration.

b) At the commencement of an engagement, the recipient will receive data and knows how to retrieve a key from the respective trust provider. Even orphaned data will remain traceable as the key metadata will indicate which trust provider it emanates from.

c) The recipient will be set up to store each datum at rest encrypted by its TrustID-derived key as standard procedure.

d) If the data is modified, the recipient draws a new derivative key from the custodian and sends the data on directly to next recipient in the transaction chain. Their permission to have the data, to have modified or added to it and the grant to the next recipient, are all embodied in the system itself by way of key issue.

e) The Trust provider manages when to inform the individual about the use

of their data (or possible monetisation) and progress in a multi-stage transaction.

Why all the fuss?

In the end, we are trying to return control and visibility to the individual as to the disposition of their personal data, that they may have submitted at any level. That extends to their potential control of benefits both monetary and other outcomes that are produced by the use of that data.

Establishing the unique identity and mandate of the individual is a crucial part of that, but needed also is the ability to follow the data wherever it goes. All current business models fail mightily in this regard. We know from experience it is not enough for a recipient organisation to say they will protect the information and use it only for agreed purposes.

1. Those agreed purpose terms are heavily influenced and controlled by the recipients.

2. They are generally unable to guarantee who in their organisation will access the data and for what purpose.

3. The individual owner never knows the who, when and why that their data is touched.

4. Derivative value from the data is not visible to the owner.

CARRIED ASSENT

We have notionally called this model ‘Carried Assent’. That is because it is intended to reorient the way the digital paradigm can embed values of status that include ownership, put also permission to use.

While the assent is revalidated (and enacted in blocking if not confirmed), the delivery of the data already has the context of permission in the delivery as part of an unbroken chain of transactional assent, prescribed by the custodian at the outset.

It is also intended to assist maximum dispatch and to capture as much value in the process reengineering to digital as possible.