Deduplication Update

Introduction

Deduplication refers to a feature of the GovPass digital identity platform, whereby a relying party can be assured of the uniqueness of a digital identity. I.e. that a user cannot, by registering twice, present at a relying party as two distinct individuals.

This contrasts with the broader concerns of identity management, identity matching, and proof of record ownership (PORO), where typically a relying party must implement additional measures in their systems to decide whether there are pre-existing records for the user at the relying party.

(For example, if you are reading this, you probably have a TFN. Were you to use digital identity to log on to the Online TFN Application service, the GovPass program would not know whether you have a TFN – and nor should it know! – but the ATO would certainly have to determine whether you do, before letting you either apply for one, or have access to your existing ATO records.)

The status of deduplication

On 13 September 2018, GPAG agreed the initial design of deduplication across the GovPass ecosystem.

The details of that paper, and the associated design, will not be reiterated here.

The remainder of this paper details the further design considerations that have been identified, the options where there are options, and the current direction or assumptions around how to proceed.

Core deduplication paradigm

The core deduplication paradigm being pursued is that, if a person uses the same documents to register for a second digital identity, these two digital identities are considered to be equivalent, and can both be used to access the same records at relying parties.

The challenge is how to design and implement this deduplication so that users' privacy is not impacted, and so that deduplication can work even when the person has used different identity providers to create their multiple digital identities. The remainder of this paper will address these challenges.

Out of scope

Fraud control is out of scope. For example, when an IP2 digital identity and an IP3 digital identity have been registered using the same driver licence, it may be appropriate for the ecosystem to recognise this, and flag this for investigation, or disable one or both of these identities, or notify the user that this has happened. This is out of scope for the deduplication work.

Identity matching at the relying party is out of scope. For example, should a new digital identity user log in to Centrelink, Centrelink must determine whether or not that person has a Centrelink account, and whether that person is currently claiming benefits – before the new user can be allowed to create a new claim. This will be done by comparing the user's available identity attributes with the data currently held by Centrelink. This process will differ for every relying party, and cannot be obviated by the GovPass platform. This is therefore out of scope for the deduplication work.

Guaranteed global uniqueness of digital identities is not possible, and therefore out of scope. Deduplication should be seen as a risk mitigation approach, and must be understood as such by relying parties.

The ability of the platform to provide an assurance of uniqueness is commensurate with the identity proofing level being used. Lower identity proofing levels make it more likely that users will create multiple indistinguishable digital identities, either intentionally or malevolently. Providing consistent deduplication at all identity proofing levels is therefore out of scope.

Implementation Decision Points

Within the scope of the current deduplication concept, there are some key implementation decision points, with associated trade-offs including privacy, fraud control, assurance, impact to the user experience, and consistency with the program goals. These decision points are detailed in the following sections.

Implementation in the Identity Provider or the Exchange

If the ecosystem was to have only a single identity provider, it would be possible to implement deduplication in the identity provider. In fact, this would be the most effective solution, because the identity provider has the maximum possible information available about the user and the documents they used to verify their identity, and can easily ensure uniqueness across their user cohort.

Extending this to multiple identity providers raises the following decision points.

Decision Point 1: Should identity providers communicate with each other to ensure their users are unique across the ecosystem?

Current Answer: No, because this level of data sharing would violate social licence.

Decision Point 2: Should personal information and/or document details be shared with the Exchange, to allow the Exchange to establish uniqueness across the ecosystem?

Current Answer: No, because it would require the Exchange to store personal information about users, creating a honeypot for hackers.

Where uniqueness fits into the identity proofing levels

It’s not obvious at which identity proofing level uniqueness will be available. Or, to put this another way, how much assurance of uniqueness can the platform provide at each identity proofing level? Deduplication may also come at a cost, and may not be required by all relying parties

These considerations raise the following decision points.

Decision Point 3: Will deduplication be a feature of a specific identity proofing level?

Current Answer: No, because uniqueness cannot be guaranteed at any identity proofing level, yet at all identity proofing levels there is some potential value (with more value and more assurance possible at higher levels, of course).

Decision Point 4: Will deduplication be reflected in the TDIF?

Current Answer: Yes, because it imposes requirements on the Exchange and Identity Provider systems. But this will be done in a way that expresses the limitations of deduplication.

Optional or mandatory

It is possible to make deduplication opt-in for users, and it is possible for relying parties to specify their requirement for uniqueness in addition to their requirements for identity attributes and an identity proofing level. However, the trade-off here is that users may prefer not to be unique, and relying parties may require it for a transaction to complete; thus admitting the possibility of a dead end for the user when these preferences do not align.

Decision Point 5: Will deduplication be mandatory for relying parties?

Current Answer: No. To limit the impact of deduplication, and not to impose it on users where unnecessary, the platform will support relying parties who do not require deduplication, accepting logins from users where deduplication is not possible

Decision Point 6: Will deduplication be mandatory for users?

Current Answer: No, it can be made optional. This is one of the ways the impact can be minimised. Like other attributes, 'uniqueness' can be requested by the relying party, in addition to other attributes and/or the identity proofing level. A request for this scope (in OIDC terminology) can trigger this functionality within the identity providers and at the Exchange. This also has advantages in the user experience, their understanding of deduplication, and social licence. It may also prove useful in excluding users who have verified using obscure documents that are supported by the TDIF, but where the deduplication design cannot provide deduplication, even at high proofing levels.

Consistent with the previous two answers, the current approach is to include the Evanescent Anonymous Identifier (the deduplication identifier), in the response from the identity provider to the Exchange, as an additional attribute, not as the subject identifier. This also allows the Exchange to store the subject identifier for audit purposes, without storing the EAI.

From the Exchange to the relying party, the only additional information needed for deduplication will be a single Boolean: "is unique".

Extension from commencement of identity documents, to other documents

The current approach provides a meaningful assurance of uniqueness only at identity proofing level 3, but that assurance is only as valid as the assumption that no user has more than one of the commencement of identity documents: birth certificate, citizenship certificate, and visa.

Even today, the program is trialling support for passport in place of a birth certificate, creating the possibility of a person using a passport to create a digital identity, and a birth certificate to create a second, duplicated identity (within an IDP, or in a second IDP).

Solving these problems, and acknowledging these limitations, raises the possibility that any solution could be extended to other document types, and other proofing levels. For example, it might be convenient for the user to merge two identities that have both used their driver licence to verify at IP2. Or, at IP3, it might be valuable to identify, flag, and/or investigate multiple identities that happen to have used the same driver licence document in their registration.

Decision Point 7: Will deduplication be extended to identify multiple digital identities that have used the same documents in verification?

Current Answer: No, this is out of scope because the feature is about more than deduplication: identity fraud, investigations, and fraud resolution. It would not be strong enough to allow automatic merging of identities (equivalent to bypassing PORO), and so its effect would either be informational only (notifying the user, or triggering an investigation), or it would be to invalidate the accounts, locking the user out of the platform.

Privacy versus simplicity

There is a very large spectrum between how simple the implementation could be on one side, and how little risk to users' privacy there could be on the other side. Rather than explore every aspect, the following explains the ends of the spectrum. The various decision points and trade-offs should be evident from these extremes.

Extreme Complexity: The EAI is calculated by a separate system (provided as a secure service) and is one-way encrypted with a key stored in a hardware security module. This system is available only to registered GovPass identity providers, and is rate-limited to avoid any possibility of brute-force discovery of any of the original information. The EAI is based on birth certificate, citizenship and visa only. Digital identities are immediately invalidated when these documents are invalidated, for example when a resident becomes a citizen. The EAI is passed to the Exchange as a separate attribute, but is not stored. The Exchange uses a hash of the EAI to look up (or generate) a randomly generated subject identifier for the relying party. EAI is not a mandatory attribute, and is requested only by relying parties that require it for service provision, and the user is given the option to consent to uniqueness, or to terminate the login or transaction.

Extreme Simplicity: The EAI is calculated as a hash of the commencement document identifier, it is passed to the Exchange, and the Exchange hashes it with a relying party specific code, to generate a subject identifier for the relying party.

Current Status: Many of the complex features described above are not required for a minimum viable product, while still allowing the full set to be included at a later date, without breaking the architecture. The core MVP aspects of deduplication are that the EAI be a new (optional) scope and attribute between IDP and Exchange, and that uniqueness be a new (optional) scope and attribute between Exchange and the relying party. Hardware security modules, encryption, remote generation of the EAI, storage in the Exchange, etc., can be added separately.

How to handle the limitations of deduplication

Deduplication will not work for all documents, with all issue dates, at all identity proofing level. As part of the ongoing work of implementation, we will uncover and record the specifics of which documentation details can lead to unavoidable duplication, undesirable deduplication, scenarios where deduplication is not possible, and sufficient information to quantify the failure rate of the deduplication function.