Data Provenance

⚠ **Planned feature**

This page describes a feature planned for future midPoint versions.

This feature is roughly designed and it was evaluated as feasible. However, there is currently no specific plan when it will be implemented because there is no funding for this development yet. In case that you are interested in supporting development of this feature, please consider activating midPoint Platform subscription.

Motivation

Identity management systems take data from various sources. This is usually more than one source of data even for simple IDM deployments. Part of the data may be taken from an HR system (authoritative employee data), part of the data is manually entered by identity administrator (contractors, support staff) and there may be even be data provided by users themselves (self-service). The situation may become much complex if several semi-authoritative resources are used. In that case it may not be entirely clear what came from the HR system, what was manually entered, what was synchronized from the company phone book where user manually corrected the data and so on. To use the technical terms: there is no clear information about *data provenance*.

Data provenance is a very useful feature for many related purposes. It is clearly a great tool for diagnostics and troubleshooting. But there are even deeper concerns. Data provenance is one of the mechanisms to implement proper data protection. The data protection mechanisms are best practice of identity management. But it is also mandated by data protection legislation such as European GDPR regulation.

However, implementation of real data provenance is not a simple task. MidPoint already keeps object-level metadata. But the provenance needs to be tracked on an attribute-level - and even on attribute-value-level. This means really fine-grain metadata maintenance. Currently midPoint can provide information about data provenance by using audit trail. But audit records are a time-organized data that may be warehoused, summarized or archived. Audit trail data may expire. Therefore information about data provenience may not be readily available and it may even disappear in time. Clearly a more sophisticated way to manage data provenance is needed.

Data Provenance

MidPoint can implement fine-grained provenance data. MidPoint could maintain the data for every property and every value. The data may contain:

- How the data item originated (manual entry, mapping)
- Who was responsible (user that entered the data, owner of the process, etc.)
- What policy or configuration was used if the data value was computed (object template, role, ...)
- When the data item originated (timestamps)
- The resource where that particular data item came from (if set by inbound mapping)

Data provenance can be quite resource-intensive, especially when it comes to data storage. It may also complicate the data by a significant degree. Therefore there must be a way to control the provenance mechanisms - at least turn it on or off. Compatibility is also a major concern. The data provenance must not break compatibility between midPoint versions.

Implementation

Data provenance are such a fine-grain metadata that it would be a huge impact to implement that using ordinary midPoint schema. In fact, this kind of metadata create a whole new dimension. Therefore a generic and systemic approach is needed here. Fortunately, midPoint is built on top of data representation mechanism that we call Prism. There is almost perfect place in Prism data model to store provenance metadata: *prism values*. However, it is not straightforward to serialize such metadata, store it in the database and generally manage them in midPoint and still keep simplicity and compatibility at the same time. But with improvements to Prism and with a lot of clever midPoint modification this is a feasible way.

See Also

- Management of Lawful Bases for Data Processing (GDPR)