MidPrivacy - Identity Provenance as a first step towards personal data protection

Summary

Personal data protection is relatively new and often misunderstood, requiring to be looked at from a completely different perspective. Personal information cannot be handled as an ordinary data item. It is often crucial where the data comes from and what the overall circumstances of data processing are.

The solution is therefore based on the data and metadata - data about the data. When it comes to identity management and personal data protection, perhaps the most important part of the metadata is provenance information. That is why Evolveum's open source identity management and governance platform midPoint seems to be positioned at an ideal place in the personal data flows.

Work on the prototype took place in spring-summer 2020. The functionality enabled development of advanced personal data protection features, removing significant roadblocks. The provenance prototype was implemented as a part of midPrivacy initiative, a long-term effort to extend midPoint with a complete set of personal data protection features. After all, privacy and data protection are an integral part of midPoint design and day-to-day development.

Keywords

midPrivacy, midPoint, Axiom, identity provenance, data provenance, open source, identity management, identity governance, identity data, personal data protection

Actors involved in the project

Evolveum: https://evolveum.com

The NGI_Trust project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825618

The project

MidPoint is the leading open source identity management and governance platform. With its rich feature set, this unconventional system gives organizations full control over identity data, making sure they are copied, synchronized and shared according to the policies. MidPoint is designed to improve information security, automate and improve error-prone activities, enabling organizations to proceed in digital transformation.

Figure 1. Evolveum

MidPoint is professional open source software, created and maintained by Evolveum. With its skilled full-time dedicated development team and valuable community, Evolveum works on maintaining and improving midPoint constantly. Due to the disruptive nature of open source software, a small company based in Slovakia is able to provide its services on a global market. This is no easy feat for a self-funded company with less than 30 employees and a product that is completely open source. That is why an efficient partner network and empowered user community are instrumental in addressing customers' needs all around the world.

The problem

In early days of identity management, the technology was all about cost saving and information security policies. Later on, the focus shifted to identity governance and management of compliance with laws and regulations. While these concerns are undoubtedly important, there is one concern that is unique: personal data protection.

Personal data protection is a special concept in many ways. First of all, it is relatively new and often misunderstood. There are numerous attempts to address personal data protection with existing information security tools, governance and compliance frameworks. However, these attempts are rarely successful, as such tools are lacking the appropriate mechanisms. Personal data protection is not about "can user X access system Y?". We need to look at the problem from a completely different perspective. Personal data protection is mostly about "can we process data A for the purpose B, given circumstances C?". This question is much harder than it seems, and the systems that can efficiently answer such questions are extremely rare.

The core of the problem is in the data, or rather the fact that data are not just data. Traditional systems handle user’s full name as an ordinary data item. They use it and share it without any considerable limitation across the entire organization, and even beyond organizational boundaries. However, personal information cannot be handled as an ordinary data item. It is often crucial where the data came from and what the overall circumstances of data processing are. For example, if user is an employee of our organization, we are entitled to keep this information and use it within our organization as necessary. However, when an employee leaves, the situation is much different. We might still be able (and even required) to keep the name of the user. Yet we are no longer entitled to store it and use it in numerous information systems in our organization. The data item is still the same, it is still stored in the same database, however, the circumstances are all different. Appropriate data minimization and erasure has to take place.

This is still a very simple scenario. We are living in a connected world. The boundaries between employees, partners, customers, students and community are very fuzzy. How can we process user's full name, if the user is a student, they are engaged in two research projects, each of them spanning several academic organizations? How does the situation change when the student graduates and becomes employee of one of the partner organizations, still participating in one of the projects? These questions are not easy to answer.

The solution

The problem has a surprisingly complex and multi-faceted solution. The solution is based on the data - and metadata, which are the data about the data. We need to know a lot of details about user's full name: where it came from, when we have first learned about it, when it was updated for the last time, how it was created from the first name and last name. When it comes to identity management and personal data protection, perhaps the most important part of the metadata is provenance information. Provenance tells us where the data came from, which in turn determines how we can use the data. Once again, provenance is more complicated than it seems, as a data item may come from several overlapping sources, or it may even be a combination of several values.

MidPoint development team had been aware of the problem for several years. The team has an outline of a solution, as midPoint seems to be positioned at an ideal place in the personal data flows. However, when it comes to practical solution, the problem goes deeper than the technology. Personal data protection is based on fundamental mechanisms, such as auditing for accountability, integration components for data transfer, policies and so on. MidPoint already has most of that mechanisms. However, one crucial piece of the puzzle was missing: data provenance.

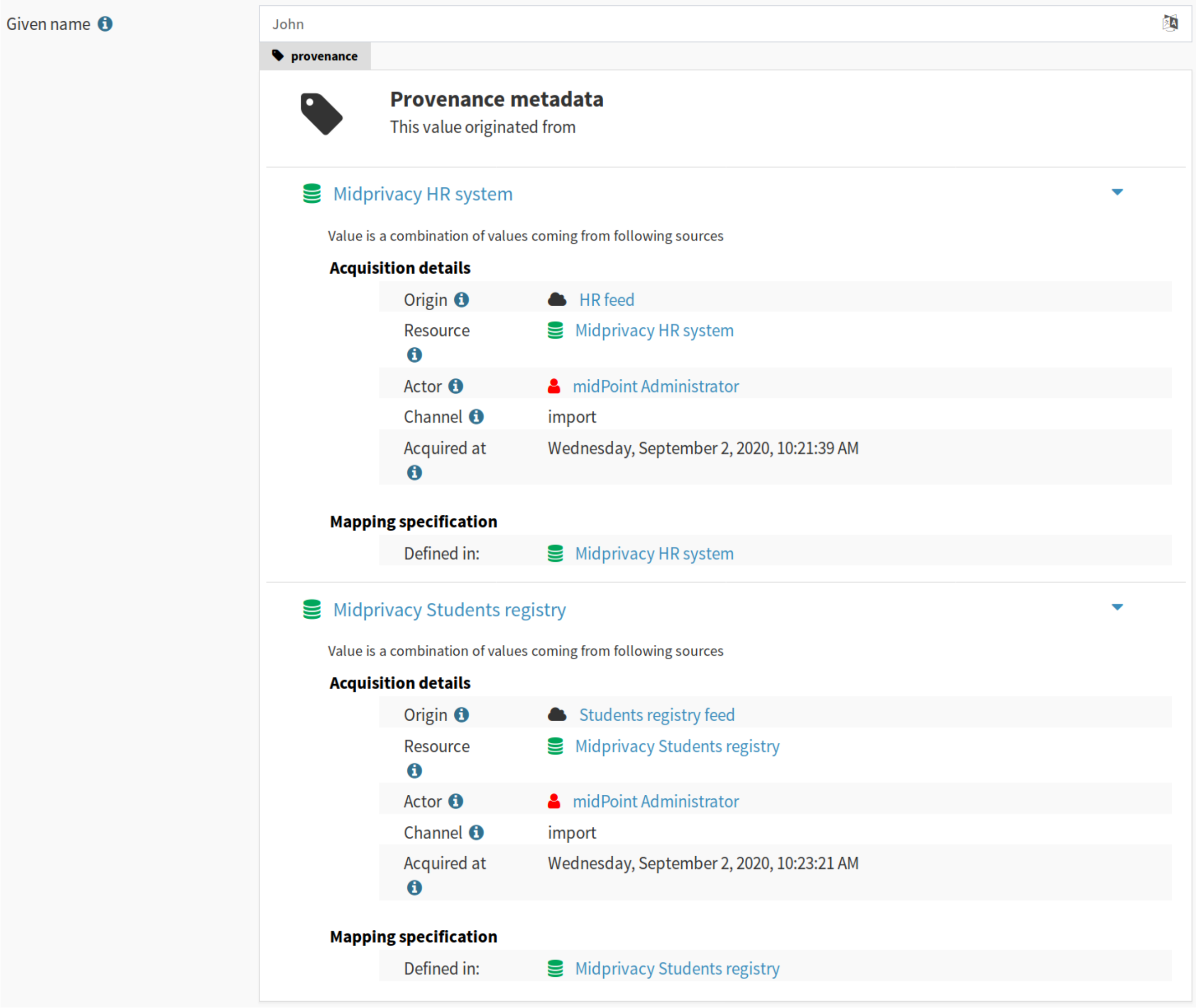
Provenance metadata are not simple, as we have seen. Data provenance is also a problem that is not understood completely. Therefore, there is a need for a complex metadata schema support in the data representation layers. Metadata schemas must be easy to adapt and extend, otherwise the solution would be too limited for practical use. However, none of the popular data modelling frameworks have such support. Simply speaking, there is a lot of groundwork to do, before first practical benefits for users can be done. While the final product features are likely to be commercially viable, the foundation work is a considerable barrier.

Evolveum had been trying to secure funding for the metadata groundwork for years, until an opportunity was provided by NGI_TRUST. NGI_TRUST provided funding to build a personal data provenance prototype to lie the foundations and demonstrate feasibility of the technology.

Work on the prototype was done in spring-summer 2020. The objective was a demonstration the feasibility of complex data provenance metadata in a practical and established identity management system. The code is an evolutionary prototype, which is a native part of midPoint source code, continually improved during regular midPoint development cycles. Modelling of complex metadata structures proved to be a major challenge, which was quite expected. The challenge was addressed by designing and developing Axiom, a new data modelling language with native support for metadata modelling. Axiom was a success; its capabilities were proven several times during the project. The flexibility of Axiom was a real benefit, as metadata schemas changed dramatically several times during the project. The team has encountered unforeseen difficulties, caused by limited experience of identity management community with data provenance metadata concepts.

Results

Despite the difficulties, there is a complete working prototype at the end of the project. The prototype has demonstrated the ability to attach complex provenance metadata to any individual value in the identity management system. MidPoint knows where every value came from, even in case it came from several independent sources. The metadata are processed together with the value. For example, when user's full name is computed from first name and last name, the resulting full name value will reflect provenance metadata from the first name and last name values.



Given name

John

provenance

Provenance metadata

This value originated from

Midprivacy HR system

Value is a combination of values coming from following sources

Acquisition details

Origin	HR feed
Resource	Midprivacy HR system
Actor	midPoint Administrator
Channel	import
Acquired at	Wednesday, September 2, 2020, 10:21:39 AM

Mapping specification

Defined in:	Midprivacy HR system

Midprivacy Students registry

Value is a combination of values coming from following sources

Acquisition details

Origin	Students registry feed
Resource	Midprivacy Students registry
Actor	midPoint Administrator
Channel	import
Acquired at	Wednesday, September 2, 2020, 10:23:21 AM

Mapping specification

Defined in:	Midprivacy Students registry

Figure 2. MidPrivacy prototype GUI

MidPoint user interface is metadata-aware, it is able to display provenance information for all the values. Axiom, the metadata-native modeling language, shows a great potential for further development and practical use by a broad community. Overall, the prototype has reached and exceeded the expectations.

Testimonial

The help from NGI_TRUST was an essential part of the success. NGI provided the funding, yet the guidance and coaching were also very helpful. This is one-of-kind opportunity, an impulse without which the functionality might never get developed.

Future plans

Being a prototype, the functionality is still in a nascent stage, yet it shows a commercial potential. This functionality enabled development of advanced personal data protection features, removing significant roadblocks. With the prototype in place, further development of production-ready functionality becomes commercially-viable.

Unfortunately, further development of the functionality was slowed down by the events related to the pandemic. The customers have shifted their priorities towards more pressing concerns, displacing personal data protection. However, we believe that the interest will be revived as the economy recovers. We already have preliminary plans to continue development of both Axiom and the provenance functionality.

The provenance prototype was implemented as a part of midPrivacy initiative, a long-term effort to extend midPoint with a complete set of personal data protection features. Privacy and data protection is not just an after-though in the midPoint world, it is an integral part of midPoint design and day-to-day development. After all, personal data protection is not just a legal requirement, it is the right thing to do.