Complexities of Identity Provenance Metadata

Radovan Semančík
Open Identity Summit, June 2021
Introduction and Motivation

- Identity management = management of identity data
- Where the data came from?
  Data protection, transparency, multi-affiliation, data portability, ...
- We need data about the data → metadata
- Whole new dimension of identity management
- MidPoint: open source identity management platform
Identity Data

givenName: Radovan
familyName: Semančík
fullName: Radovan Semančík
Identity Metadata

givenName: Radovan
origin: HR system
created: 28th May 2021 9:16

familyName: Semančík
origin: HR system
created: 28th May 2021 9:16

fullName: Radovan Semančík
origin: HR system
created: 28th May 2021 9:16
Identity Metadata: Complex Origin

givenName: Rado ← Radovan
  origin: user entry

familyName: Semančík
  origin: HR system

fullName: Rado Semančík
  origin: ???
Identity Metadata: Complex Origin

givenName:  Rado ← Radovan
  origin:  user entry

familyName:  Semančík
  origin:  HR system

fullName:  Rado Semančík
  origin:  user entry + HR system
Identity Metadata: Multiple Origins

givenName:  Rado ← Radovan
  origin:  user entry

familyName:  Semančík
  origin:  HR system

fullName:  Rado Semančík
  origin:  user entry + HR system
  origin:  The Institute

Identity federation
Complex Identity Metadata (JSON)

"fullName" : {
    "@value" : "Rado Semančík",
    "@metadata" : [
        {
            "provenance" : {
                "acquisition" : [
                    {
                        "timestamp" : "2020-06-22T10:52:03Z",
                        "originRef" : {
                            ... reference to "User entry" origin ...
                        }
                    },
                    {
                        "timestamp" : "2020-03-06T23:05:42Z",
                        "originRef" : {
                            ... reference to "HR" origin ...
                        }
                    }
                ]
            }
        }
    ]
}
Identity Metadata Complexities

- **Yield (concept)**
  - Single origin of the data: our system, other organization, mapping/transform, ...
  - Acts as a container for rich metadata
  - Multiple yields for each value

- This means complex data structures – in metadata

- Extensibility: no hope to standardize every metadata item

- Big problem: no support in existing data modeling languages
  - JSON Schema, XSD, SCIM, YANG – they deal with data, not metadata
New data modeling language

I know, I know, re-inventing the wheel and all that. But not this time.
We have been living with XSD for 10 years, extending and hacking.
Enough is enough. No, JSON Schema is no better.

Modeling data structures, not data representation (JSON/XML)

Native support for “meta” concepts
Prototype works well
Further development is needed

https://docs.evolveum.com/midpoint/devel/axiom/
model common {
  ...

  object User {
    ...
    property givenName {
      type string;
    }
    property familyName {
      type string;
    }
    property fullName {
      type string;
      minOccurs 1;
    }
    ...
  }

  ...

  ...
}

{
  "givenName" : "Radovan",
  "familyName" : "Semančík",
  "fullName" : "Radovan Semančík",
  ...
}

<user>
  <givenName>Radovan</givenName>
  <familyName>Semančík</familyName>
  <fullName>Radovan Semančík</fullName>
  ...
</user>
Axiom Example: Metadata (a.k.a. new exciting stuff)

model common-metadata {
  ...

  metadata CommonMetadata {
    ...
    container provenance {
      type ProvenanceMetadata;
    }
    ...
  }
}

container ProvenanceMetadata {
  container acquisition {
    type ProvenanceAcquisitionType;
    maxOccurs "unbounded";
  }
  ...
}

Notation is slightly simplified (namespace, prefixes)
MidPoint and MidPrivacy

● MidPoint: open source identity management platform
  Provisioning, identity connectors, RBAC, self service, delegated administration, organizational structure, auditing, access certification, ...

● MidPrivacy: privacy and data protection
  Long-term initiative to implement complete set of data protection features

● MidPrivacy Phase 1: Data provenance prototype
  IDM + Axiom + metadata + mappings + GUI
  NGI_TRUST funding

This project has received funding from the European Unions Horizon 2020 research and innovation program under the NGI_TRUST grant agreement no 825618
Complex Provenance Metadata GUI (Prototype)

- 5th design iteration
- Still not very intuitive
- Suitable for identity administrators
- Obviously, GUI for common users will be a major challenge
Provenance and Data Protection

- Legal basis for data processing
  - Specified by GDPR, requirement for personal data processing
  - Employment contract, business/service requirement, … and consent

- Tracking bases for processing is non-trivial
  - Overlaps: student, employee and volunteer at the same time

- Provenance is closely related to data protection (and portability)
  - However, provenance and basis are not the same

- Further research and development is needed

https://docs.evolveum.com/midpoint/projects(midpoint)/projects(midprivacy)/phases/01-data-provenance-prototype/provenance-origin-basis/
# Conclusion

- **Identity provenance metadata**
  - Complex (meta)data structures

- **Prototype: included in midPoint 4.2**

- **Still limited provenance information**
  - Look from inside of organization (“next hop”), the best we can do now

- **Axiom data modeling language**
  - very promising, but needs further development

- **Avenue to data protection features (basis for data processing)**
Additional Resources

● MidPoint Project
  https://midpoint.evolveum.com/

● MidPrivacy Initiative
  https://docs.evolveum.com/midpoint/projects/midprivacy/

● Axiom Data Modeling Language
  https://docs.evolveum.com/midpoint/devel/axiom/
Thank you for your time

See other talks at https://docs.evolveum.com/talks

Also follow us on our social media for further information!