

Evolveum

Smart Correlation

Pavol Mederly / March 2024
Senior Software Developer at Evolveum

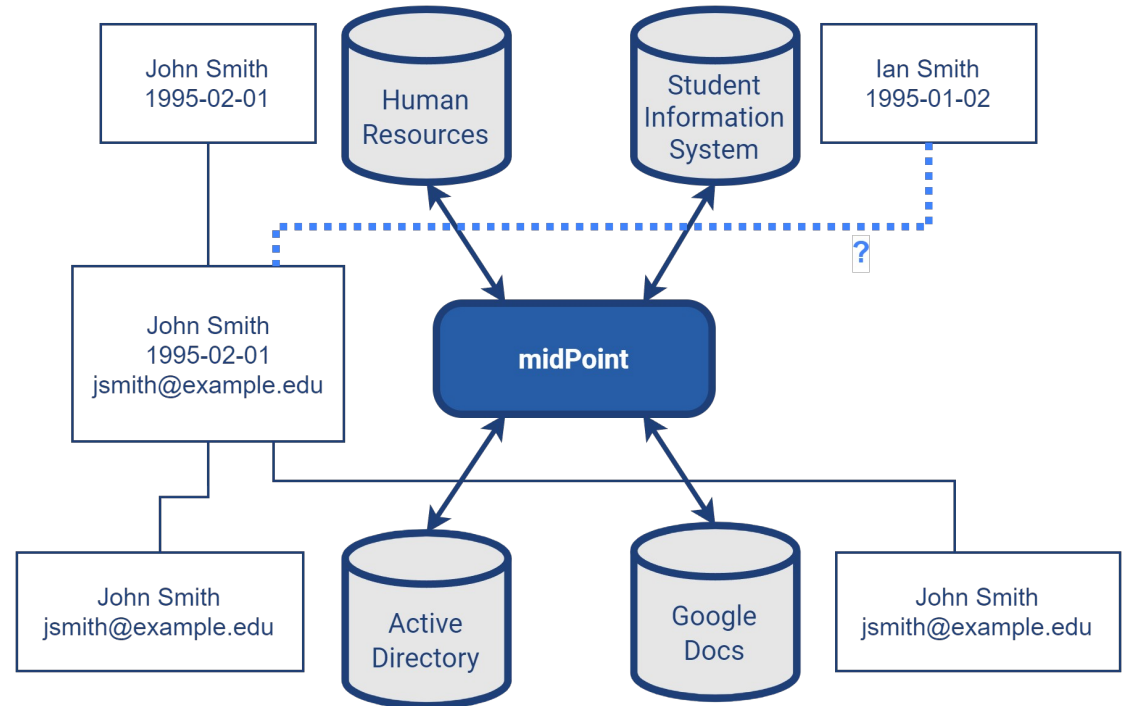
Agenda

- What is (smart) correlation?
- Short demo
- Advanced and experimental features
- Conclusion & future work



What is Correlation?

- Matching an **identity data** coming from the outside to **known identity data** present in the midPoint repository
- Uses:
 - **Synchronization**
 - Identity recovery
 - Registration or self-registration (future)
- Problems:
 - Lack of unique identifier(s)
 - Multitude of data representations
 - Data quality issues



Smart Correlation Features

- Composable **correlation rules**
 - With **confidence levels**
 - Supporting **approximate matching**
 - Levenshtein distance, trigram similarity
- Optional **human involvement** via correlation cases
- **Custom indexing** (experimental)
- Multiple **sources of truth** (experimental)



<https://docs.evolveum.com/midpoint/reference/support-4.8/correlation/>









Composable Correlation Rules

- Given name
- Family name
- Date of birth
- National ID

Rule ID	Description	Confidence
name-date-id	Family name, date of birth, and national ID exactly match.	1.0
names-date	Given and family names match approximately, and the birth date matches exactly.	up to 0.4
id	The national ID exactly matches.	0.4





Configuring the Rules

List of correlation rules

<input type="checkbox"/>	Rule name ?	Description	Weight ?	Tier ?	Ignore if matched by ?	Enabled ?	 
<input type="checkbox"/>	name-date-id	Correlation rule for item(s) familyName, extension/dateOfBirth, extension/nationalId	1			Undefined ?	 
<input type="checkbox"/>	names-date	Correlation rule for item(s) givenName, familyName, extension/dateOfBirth	0.4			Undefined ?	 
<input type="checkbox"/>	id	Correlation rule for item(s) extension/nationalId	0.4			Undefined ?	 

+ Add rule

List of correlation items

<input type="checkbox"/>	Item ?	Search method ?	Match threshold ?	Inclusive ?	
<input type="checkbox"/>	givenName	Levenshtein distance	1	True	
<input type="checkbox"/>	familyName	Levenshtein distance	1	True	
<input type="checkbox"/>	extension/dateOfBirth	Exact match			

+ Add correlator

Rows per page 20

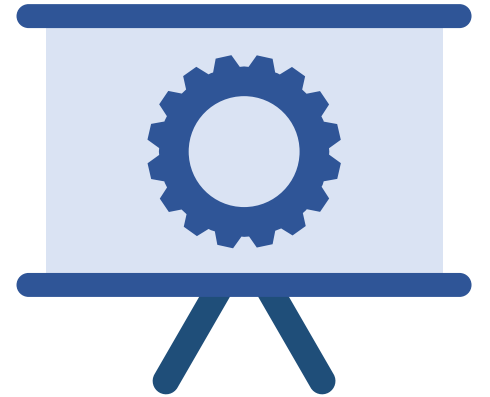
1 to 3 of 3

<< < 1 > >>

Short Demo

- We run midPoint at a fictitious university with three campuses
- We have imported 10,000 students from Campus 1
- Now we are connecting the systems from Campus 2 and Campus 3

- All names are randomly generated by <https://www.fakenamegenerator.com/>
- Used under **CC BY-SA 3.0 US** license



Advanced Features of Correlation Rules

- **Ignore if matched by**

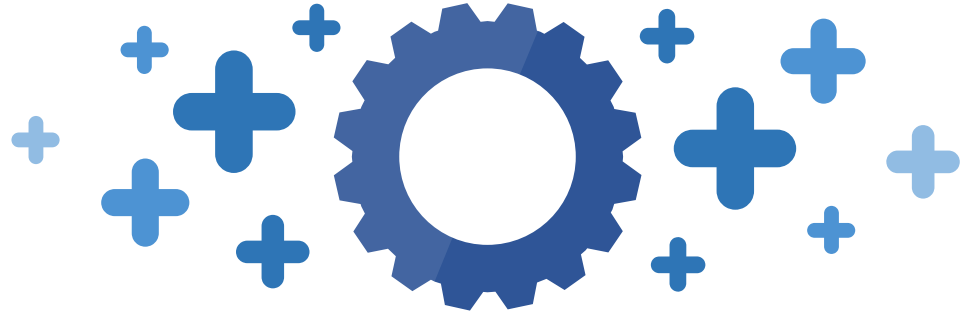
- Rule 8: givenName + familyName: 20%
- Rule 9: familyName: 10%
- John Smith vs. John Smith = 20%? No.
 - ignore rule 9 if rule 8 matches

- **Tiers and rule ordering**

- After #1 finds the match, we can stop

- **Custom thresholds**

- Default: show all $0\% < c < 100\%$, take automatically if $c = 100\%$
- Custom: show all $20\% \leq c < 90\%$, take automatically if $c \geq 90\%$



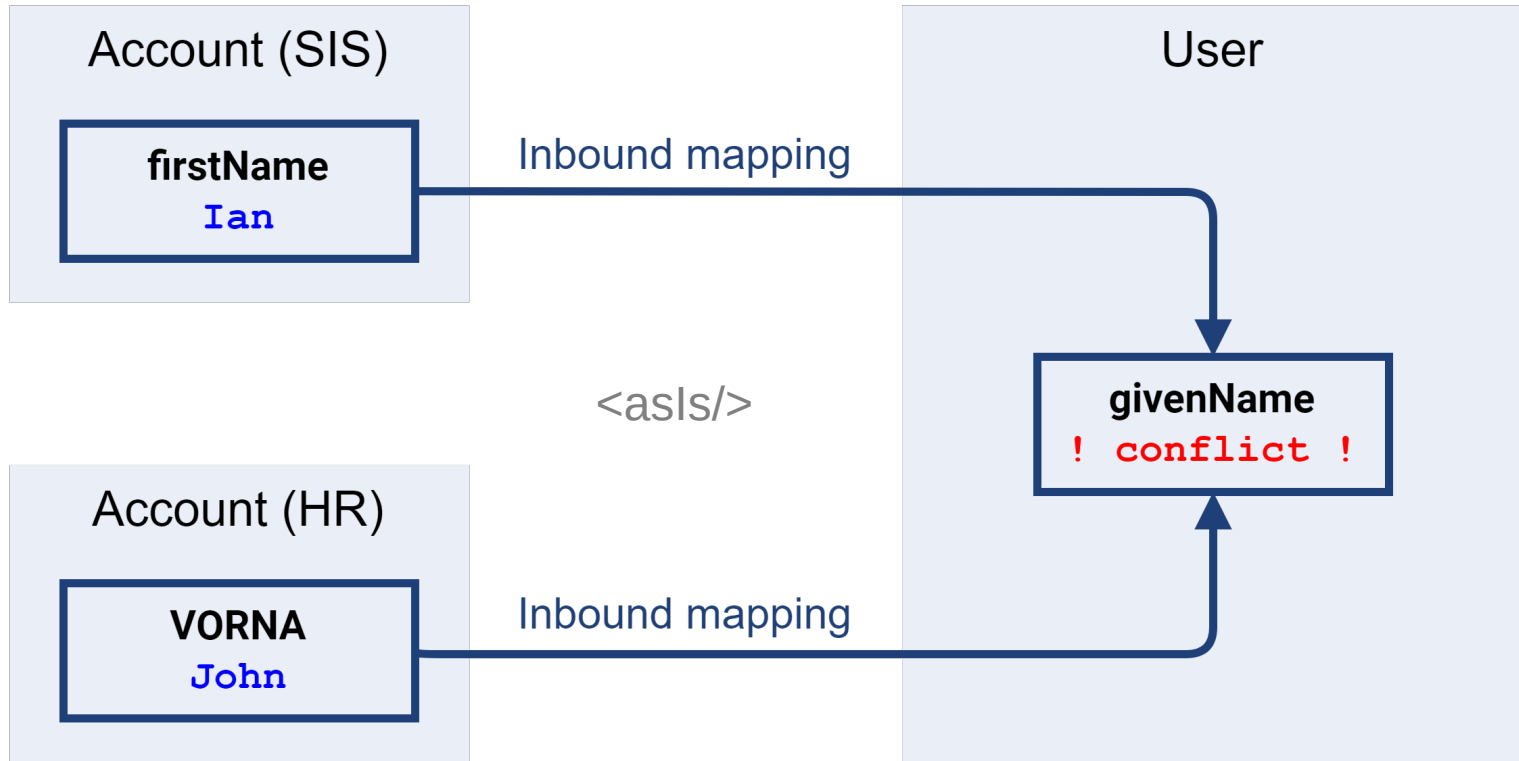
Custom Indexing

- Extra normalization of the data + using properties not indexed by default
- Options available:
 - Prefix (first N characters)
 - Custom PolyString normalization
 - Custom script
- Maintained automatically
 - Like invisible mappings

```
<item>
  <ref>familyName</ref>
  <indexing>
    <normalization>
      <!-- default = PolyString normalization -->
      <default>true</default>
    </normalization>
    <normalization>
      <steps>
        <polyString>
          <order>1</order>
        </polyString>
      </steps>
    </normalization>
  </indexing>
</item>
```

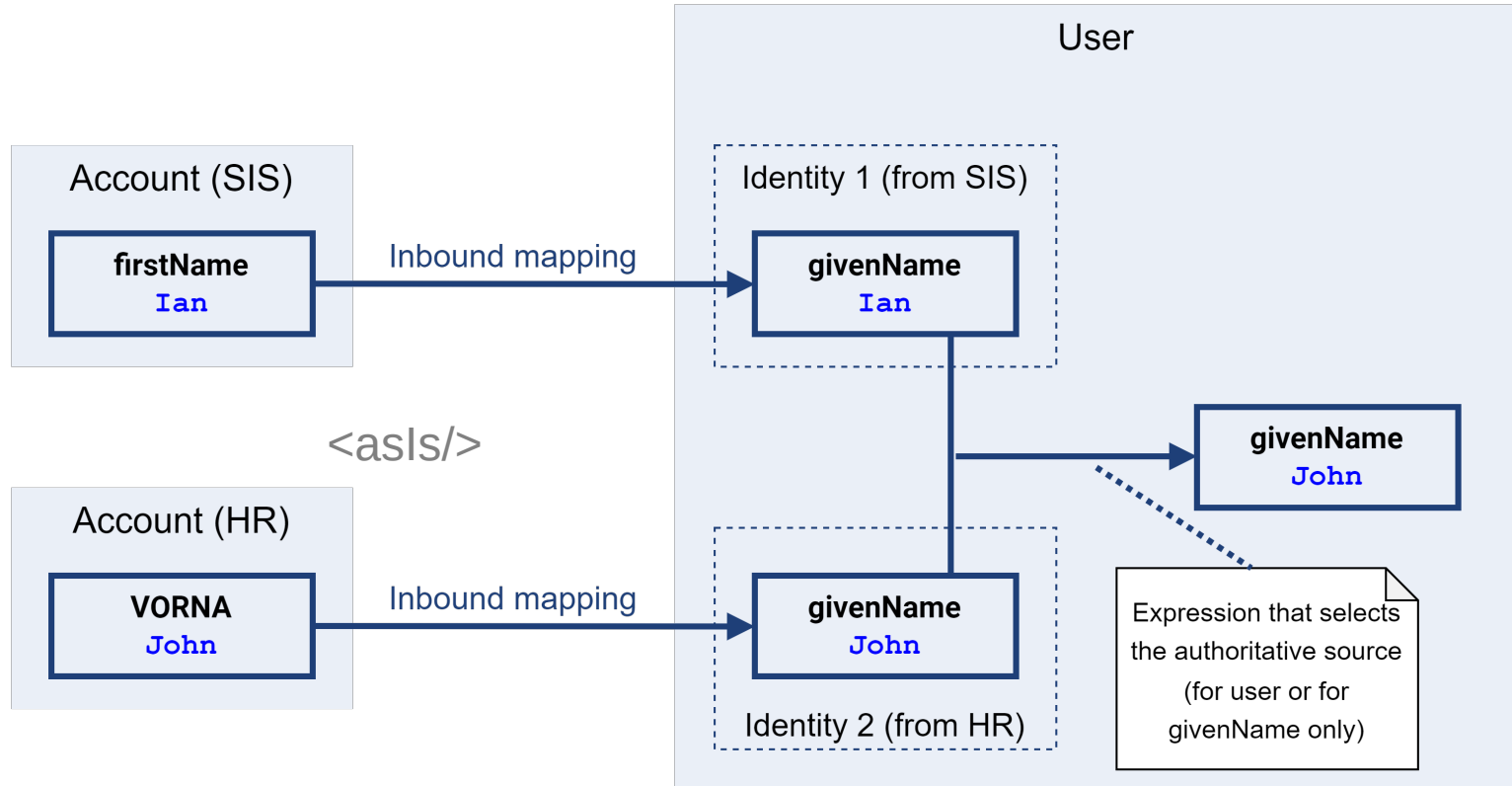
```
<item>
  <ref>extension/nationalId</ref>
  <indexing>
    <normalization>
      <name>digits</name>
      <steps>
        <custom>
          <expression>
            <script>
              <code>
                // Removes everything except for digits
                basic.stringify(input).replaceAll("[^\d]", "")
              </code>
            </script>
          </expression>
        </custom>
      </steps>
    </normalization>
  </indexing>
</item>
```

Multiple Sources of Truth: Current State



The correlation considers only the “selected” given name.

Multiple Sources of Truth: Proposed Solution



When correlating, the *identities* data are matched as well.

Conclusion & Future Work

- Powerful and flexible definition of correlation rules
- Fuzzy matching
- Experimental: custom searching, multiple sources of truth
- **Give it a try and let us know**
- Improving based on your feedback



Next Webinars

- **Securing midPoint deployments**, April 4, 2024
- **Preparing for NIS2 directive**, April 18, 2024
- **Introduction to flexible authentication**, May 16, 2024
- with more coming!



Thank you for your attention

Do you have any **questions**? Feel free to contact us at info@evolveum.com

Follow us on social media or **join us** at GitHub or Gitter!



/Evolveum



@Evolveum



/Evolveum



/Evolveum



/Evolveum

Evolveum

© 2024 Evolveum s.r.o. All rights reserved.