Manual Correlation

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

MidPoint has strong synchronization features that includes ability to automatically correlate identities. Correlation expression is a simple but very powerful mechanism to correlate identities when reliable correlation identifier is present. This is common case in enterprise or government environments. But even in those environments there may be corner cases, omissions and typing mistakes. For example someone mistyped employee number when creating an account, employee numbers missing in older systems and so on. And there are big problems: How do we distinguish new employee from an employee that used to work here, left and came back? This usually cannot be correlated automatically as there is no global correlation identifier. Some countries issue country-wide identifiers of physical persons, but they use is often strictly limited. This problem is even more pronounced for organizations with less tight control over the identities such as universities, libraries and other academic organizations, not-for-profit organizations and so on.

Solution Outline

The situation may seem hopeless as there often is no way to implement completely automated and reliable identity correlation mechanism. But all hopes are not lost. Obviously, there needs be some manual interaction in the correlation mechanism. But midPoint can keep this manual interaction efficient and it can even automate some parts of the correlation process.

Let's consider an example that midPoint gets new identity from an identity source (HR system, academic information system, etc.). Name of this new identity is "John Smith". Given such a name it is likely that there are several user with that name already in midPoint. Is this record a duplicate of any of those? Or is this a completely new record? MidPoint cannot tell that for sure. But it can display the new record alongside existing records that are good candidates for a match. And then human operator can decide. MidPoint will make sure that the information presented to operator make this decision quick and efficient.

And there are cases when midPoint can in fact decide just by itself. E.g. in case that the name of new identity is "Xenophilia Slartibartfast" then it is very unlikely that we have ever seen anyone with such a name. MidPoint can check for approximate matches to rule out typing mistakes and transliteration distortions. But in case that there is no match midPoint can safely assume that this is a new user and a new identity can be created without human interaction.

Sorter: Sophisticated Automatic Correlation

Actually, first step towards improved correlation is already implemented: Synchronization Sorter. Sorter is sophisticated expression that can be used to implement completely custom identity correlation mechanism. The sorter can make several correlation attempts with various criteria over many attributes with optional steps. It can even outsource identity correlation to an external system. There is just one limitation: the correlation mechanism is supposed to be synchronous. Sorter runs in real time and it expects immediate decision. Therefore sorter cannot be used to implement manual correlation. Not yet.

Planned Features

Step 1: Correlation with Multiple Matches

MidPoint can be improved to support manual identity correlation. In fact, this is something that we have expected almost since the beginning of midPoint development. For example there is a `disputed` synchronization situation. This situation is part of midPoint for years. But it was mostly considered to be a situation that indicates an error in a correlation expression. However, it was planned that this situation can be extended to indicate need for manual interaction. This situation can be supplemented with more information, such as (direct or indirect) list of candidate matches. Once again, such an extension was anticipated from the beginning and this was one of the reasons that midPoint has shadow objects. In that case correlation expressions or sorter can be used to determine the candidate matches. In case of multiple matches (or a single low-confidence match) the account will end up in `disputed` situation. Candidate matches will be recorded in the shadow object. MidPoint user interface can be extended to look for disputed correlation cases and present them to operator. The operator can then make manual decisions in an efficient manner.

Step 2: Correlation Cases

Deciding manual correlation cases may be easy. But it may also be difficult. It may require cooperation of several persons. It may take long time. We may need ability to delegate the cases, we may need automatic escalation. Simply speaking, manual correlation needs the same level of management oversight as any other manual activity in midPoint. But there is already a concept that can be used to handle this: concept of a case.

Case is similar to trouble ticket in typical ITSM system. The case describes something that has to be resolved by a human being. This is almost an ideal tool for manual correlation as cases allow cooperation, they could be delegated and managed. Therefore midPoint can automatically maintain a *case*s for manual correlation.

ⓘ **Synergy: Case Management**

Manual correlation is a synergistic feature. It is designed to fit together with another planned midPoint feature: case management. Case management is meant to support cases that describe a unit of cooperative work. Users may delegate cases, may comment on them and may work together to resolve the case. Cases are planned to support many midPoint features where manual work is required from approvals and manual resources to remediation. Manual correlation is just another reuse of the same principle.

Step 3: Machine Learning

Correlation decisions should be made by human operator as they may have far-reaching consequences. But that does not mean that the operator won't have any assistance with the decision. MidPoint can easily provide candidate matches even during the first phases of the development of this feature. But there are several drawbacks with this approach. Firstly, there may be many candidate entries for "John Smiths". Perhaps even too many to reasonably display on screen. It may be nice to have a recommender that can highlight probable matches. And then there may be unexpected ways how names can be distorted with typing errors, transliteration, differences in alphabets, diacritics and so on. It would be nice to have an assistance that goes beyond traditional algorithms.

We hope that machine learning methods can be used to make manual correlation more efficient. For example a recommender could learn by watching operators making manual decisions about identity correlation. And then the recommender may use that knowledge to improve its recommendations in future decision cases.

See Also

- Synchronization
- Synchronization Sorter
- Case Management