Real-life Experiences with Federated Search

Kenny Knecht[1] Paul Vauterin[1] Hans Constandt[1]

ONTOFORCE NV, Gent, Belgium,
kenny.knecht@ontoforce.com

ONTOFORCE hosts a public website www.disqover.com containing linked data from 140 different data sources in the bio-medical field, often semi-manually curated en linked. The total volume of this data exceeds 7 billion triples. The software suite behind it, DISQOVER, is also available as a standalone application developed by ONTOFORCE to link and integrate data from different data sources. Customers can host it on-site for data from different private, often sensitive internal sources. A large portion of these customers are major pharmaceutical companies from around the world but ONTOFORCE is also active in other verticals.

In both cases the semantic web platform DISQOVER offers linked data via a user-friendly interface. When a user starts a text query and selects a field of interest from the results (for example a set of genes or diseases, in general a set of entities), a dashboard is shown specific to that field of interest with a number faceted search widgets. This allows the user to see a breakdown of the results in different dimensions, represented by different facets (for example in the field of interest genes: breakdown by chromosome, by gene ontology function or by organism). The user can then drill down and filter using any of these facets. Also available are a limited list of the most relevant individual search results and an overview of all entities linked to the current result set. At each point in the search the user can follow the presented links to new dashboard to continue his or her search. At all times provenance of every piece of information is available. This feature set makes DISQOVER especially fit to query a data knowledge graph in an enterprise setting.

To make the data in the single public data endpoint hosted by ONTOFORCE available at low effort for our customers, ONTOFORCE developed a federated search in DISQOVER with a limited but very practical scope. The goal was to have a seamless integration of public data offered by ONTOFORCE and private data in the user interface and to perform all searches within the scope of an HTTP request. This constraint excluded a solution built on SPARQL, so it was developed entirely in-house. The scope was particularly limited because we only have one federation endpoint (with ONTOFORCE public data) and we focused on the queries which are supported by our API and used by the GUI of DISQOVER. It allows the customer to link private information to information in the public endpoint and allows them to extend public information with private

data. The union of these data sets is fully queryable like a single data set: one query can consist of a combination of filters on public and on private data.

The implementation of the federation by ONTOFORCE with a single public endpoint only requires a limited up-front data synchronization (typically a few million triples or less). This step involves moving data from the public endpoint and adding it to the private endpoint, but not vice versa. It makes sure that the same URI and same display label is used in both DISQOVER endpoints for the same subject. The actual algorithms and implementation of this federation setup are beyond the scope of this document but some of the key ingredients are the efficient breakdown of a query by rewriting it in private and public parts and the extrapolation based on limited randomly sub-sampled batches in order to get approximate results for the facets within the time-frame of an HTTP request. Not only size of the top buckets in a facet is extrapolated but also the number is buckets is estimated.

This solution is available to our customers since 2016 and has shown to be of great value: customers do not have to keep the public data up to date and can concentrate on integrating private data. However in order to use DISQOVER as an enterprise solution for data integration there some pitfalls as well. We will illustrate this with practical examples we gathered in several years of experience with customers in highly regulated industries like pharmaceutical companies. The first is data privacy. Although during federated search no actual data is sent from the private endpoint to the public one, even the search terms in queries, the search path or other meta information could potentially reveal valuable IP of the customer. Another sensitive point is the real-time character of federation. The customer has limited opportunity to pre-approve or to validate a new data releases. Also the real-time dependency on a live external service proves to be a difficult point. The last attention point is the loss of analytic power we experience by performing the searches in a federated setting within the scope of one HTTP request.

Therefore, ONTOFORCE is considering replacing the real-time federation with a more synchronization-based approach. This will allow the customer to keep all queries within the LAN and to choose the moment of integration, e.g. after extra validation. The user will also get exact quantitative results for all possible queries at a speed which can only be offered by in-house searches. There are even additional benefits. For example it will give ONTOFORCE more freedom in offering licensed subsets of the data sources for integration. Although this will require stronger hardware at the customer's site and longer preprocessing times, it will give the customer the amount of control that is required for an enterprise solution plus a better user-experience in terms of query times. Although we are aware that our setup is not the most general federation setup, the experiences we had and limitations we observed in this use case are applicable to a more general federation settings.