SCRY: Enabling quantitative reasoning in SPARQL queries

Bas Stringer[1,4], Albert Meroño-Peñuela[2,3], Antonis Loizou[2], Sanne Abeln[1], and Jaap Heringa[1]

[1] Centre for Integrative Bioinformatics, VU University Amsterdam, NL
[2] Knowledge Representation and Reasoning Group, VU University Amsterdam, NL
[3] Data Archiving and Networked Services, KNAW, NL
[4] To whom correspondence should be adressed (b.stringer@vu.nl)

The inability to include quantitative reasoning in SPARQL queries slows down the application of Semantic Web technology in the life sciences. SCRY, our ***S**PARQL **c**ompatible se**r**vice la**y**er*, improves this by executing services at query time and making their outputs query-accessible, generating RDF data on demand. The power of this approach is demonstrated with two use cases, where we use SCRY to calculate standard deviations and to find homologous proteins.

More and more biological knowledge is being made available through the Semantic Web as Linked Open Data. However, not all knowledge is easily captured or stored in static triple repositories. This is especially true for knowledge derived from quantitative reasoning.

For example, BLAST is commonly used to predict homology, but its input parameters and interpretation of its output vary greatly depending on the research question. Precomputing the outcomes of all parameter combinations and every conceivable input is practically impossible. Thus, using BLAST results in a SPARQL query, for example to look up information about a protein's homologs, requires these results to be generated, analysed and incorporated at query time. No currently available tools facilitate this effectively.

Likewise, the standard deviation of a set of measurements can be used to detect outliers. Thus, if an observation is classified as an outlier depends on the other observations in the set. Many triplestores contain data which a simple query can group into sets, but SPARQL does not support a straightforward way to calculate their standard deviation. This makes it difficult to use outlier selection in SPARQL queries.

More generally, many research questions require some form of quantitative reasoning. Such reasoning is currently poorly supported within SPARQL queries and is often infeasible to capture through precalculation; this hinders the application of Semantic Web technology to research questions in the life sciences.

We present SCRY, the **S**PARQL **c**ompatible se**r**vice la**y**er, which is a lightweight SPARQL endpoint that interprets parts of basic graph patterns as calls to user defined services. These services are evaluated at query time, generating triples needed to resolve the query on the fly.

SCRY allows users to incorporate algorithms of arbitrary complexity within standards-compliant SPARQL queries, and to use the generated outputs directly within these same queries. Unlike traditional SPARQL endpoints, the RDF graph against which SCRY resolves its queries is generated at query time, by executing services encoded in the query's graph patterns. Figure 1 illustrates the dataflow of a typical query using SCRY, where its services are accessed through federated queries from a primary endpoint.

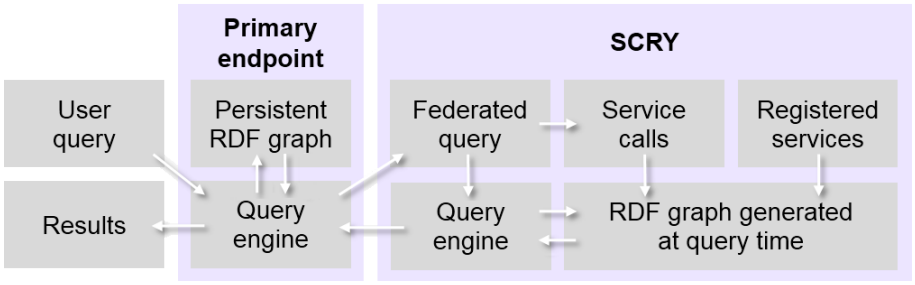

Fig. 1. Dataflow diagram of a typical SPARQL query using SCRY, through federated queries from a primary endpoint. The primary endpoint will send a federated query to SCRY, optionally interrogating a persistent RDF graph to retrieve inputs for services. SCRY will parse service calls from the federated query's basic graph patterns, and generate an RDF graph by executing them. It then resolves the federated query like a traditional endpoint, and returns its results to the primary endpoint.

The federation-oriented design allows for easy integration with existing SPARQL endpoints and tools. Moreover, SCRY facilitates the execution of arbitrary services within SPARQL queries, which enables quantitative reasoning as well as other forms of data processing and analysis. Ultimately, this allows Semantic Web technologies to be more effectively applied to research questions in the life sciences.

We demonstrate the power of SCRY's on-the-fly triple generation with two use cases. First, we select outliers among GO-annotated proteins from the UniProt SPARQL endpoint, using SCRY to calculate the standard deviation of their sequence lengths. Second, we combine SCRY, BLAST and the Human Protein Atlas to find homologous proteins which are expressed in the same tissues as a query protein.