DISQOVER: A Paradigm Shift in Semantic Data Processing using a Fast, Scalable, Transparent and Easy to Use New Semantic Data Ingestion Engine

Filip Pattyn[1][0000−0003−0858−6651], Hans Constandt[1][0000−0002−9685−5016], Bérénice Wulbrecht[1][0000−0002−9444−1709], Kenny Knecht[1][0000−0002−1049−3684], and Paul Vauterin[1][0000−0003−3665−4519]

ONTOFORCE nv, Technologiepark 19, 9052 Gent, Belgium
filip.pattyn@ontoforce.com
http://www.ontoforce.com

Abstract. Semantic web technologies are gaining renewed interest since data indexing, layered on top of a traditional semantic triplestore, has been adopted. This greatly improved the speed of semantic applications and opened new opportunities. DISQOVER (http://www.disqover.com) is a web-based semantic search, exploration and analysis platform for linked data sources. The platform allows to ingest and harmonize a wide spectrum of public, private and third-party data which are glued together via an overarching DISQOVER configuration ontology. The system supports data federation between different DISQOVER installations and is capable to prepare and create visual analytics dashboards directly based on the data (see Fig. 1).

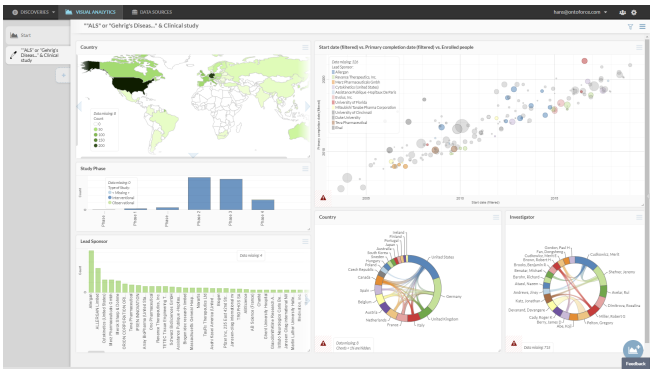

Fig. 1. A dashboard showing an overview of Amyotrophic Lateral Sclerosis (ALS) clinical studies

We will zoom in on binning and quantifying semantified data resulting in lightning fast visual analytics. Slicing and dicing a multitude of data sets

in an easy way became simple. The latest 5.0 release extends traditional text search with alternatives like chemical or protein structure search. In this demonstration we introduce another development focusing on lowering the threshold for semantically integrating, enriching and linking new data sources.
ONTOFORCE filed a patent on a new data ingestion engine as this is a novel concept in the field and allows data conversion processes to not only be managed in a visually attractive web-based application but eliminates the need to write and maintain data conversion scripts (see Fig. 2). This new semantic data ingestion framework comes with a huge gain in speed, stability and scalability. It consists of data processing components that perform atomic data manipulations traditionally done via scripting or as an inferencing step via a SPARQL insert statement in a triplestore. Combining these components allows to easily create and maintain data conversion pipelines. This engine is especially designed to process complex many-to-many relationships more efficiently compared to traditional ETL (Extraction, Transformation and Loading) pipeline tools that process line per line. Every process step can be fully inspected and allows monitoring of upstream or downstream processing impact. This level of transparency contributes to efficiently handle data provenance and full data processing quality assessment.

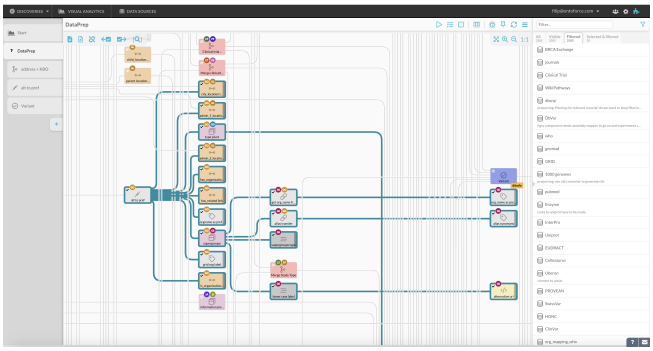

Fig. 2. View on a data ingestion pipeline highlighting components downstream of one component.

The plugin and pipeline architecture strategically fit the need to enhance the ability of life science companies to harness their data assets better and more easily and to create a structured data foundation ready for artificial intelligence (AI) applications. A typical life science use case will be presented integrating a set of data and metadata to create a data catalog of research data with the capacity to be combined with a downstream AI application.

Keywords: Semantic web · Linked Data · visual analytics · data integration.