# DISQOVER: Through the large data landscape of Life Sciences

Berenice Wulbrecht[1] [0000-0002-9444-1709], Filip Pattyn[1] [0000-0003-0858-6651], Kenny Knecht[1] [0000-0002-1049-3684], Hans Constandt[1] [0000-0002-9685-5016]

[1] ONTOFORCE NV, Ottergemsesteenweg-Zuid 808, 9000 Gent, Belgium
berenice@ontoforce.com

DISQOVER (http://www.disqover.com) is a semantic search platform that integrates disparate data sources in life sciences. The application makes data easily navigable. Complex queries can be run intuitively and are delivered at speed, enabling you to find smart data faster. More than 120 open data sources (e.g. Medline, NCBI Gene, CHeMBL, clinicaltrials.gov) are integrated in DISQOVER and together with an integration ontology this creates a portal to more than 25 data types (e.g. Chemical, Active Substance, Clinical Trial, Publication, Gene, Protein) and 85 million single concepts. The DISQOVER platform not only contains open data sources but also licensed data only accessible when the user is a member of a user group of licensees.

The different data sources are harmonized during the data ingestion process by converting them into a semantic web format representation if that's not the case already. Where possible existing ontologies are used and all data is loaded into a triplestore. Together with the data, an integration ontology can be loaded to control how the data can be visualised in DISQOVER. A data scientist has control how to group concepts into data types. Per data type, the visualisation of the properties and the necessary filters can be defined. This allows to truly integrate different data sources and to combine the properties of these sources per concept. For example, DISQOVER integrates 38 databases focusing or related to diseases (e.g. UMLS, SNOMED CT ICD-10, MedDRA, DrugCentral). It results to 809862 disease references, that are consolidated to 305905 disease objects, during the integration process. These diseases are contextualized and linked to various object types : 5012070 clinical trials, 2583 pathways, 84784 cell lines, 38 publications.

Moreover, the visualisation of the relations between concepts can be configured in a similar way. A full configuration opens a wealth of information which can be tailored into different user views to avoid data overload. A user view can be added to the configuration to visualise a subset of the available data needed for a specific use case.

The whole software suite can be installed locally to integrate private data and to enrich it with all data available in the public version thanks to a proprietary DISQOVER federation approach. This is possible thanks to the linked data principles where each concept is identified by a URI (Unique Resource Identifier). Private data can be linked to public data via the use of globally used URIs in the public DISQOVER platform.

DISQOVER is providing an intuitive and easy access to a large amount of Life Science data. Semantic data types and linked data result in a powerful model that allows to easily search relevant hits and navigate across data sources. We present here the visualisation of the DISQOVER data landscape : the data sources and their imbrication, the concepts and their interactions. The DISQOVER user interface

clearly shows the advantages of linked data and how these principles can help to get new insights from integrated data.

Keywords: Semantic link, Data integration, Ontology, Linked Data