Release Notes
DISQOVER 5.0

1. Introduction

This document contains release notes of the DISQOVER version 5.0 and instructions to upgrade. Please make sure you have read this document before updating/installing this new release.

New features:

- The current procedure to load and integrate data sources is extended with a new approach to manage data integration via the user interface, the '**Data Ingestion Engine**'. The existing script-based data conversion and use of virtuoso during the data ingestion process remains fully functional, but the Data Ingestion Engine becomes the default tool for data ingestion. The entire process of accessing a source, converting the data into a semantic web representation, and data inferencing can be done in a **graphical pipeline composed of individual action steps**.
- The end user can configure the layout of the properties of instances in the Visual Analytics Dashboard through the DISQOVER web frontend via an **instance view editor.**
- The default text-based search can be extended with **query plugins** to initiate domain-specific searches. This plugin feature allows to send query requests such as representations of chemical structures, nucleotide or protein sequences to an external service. It is also possible to upload a file containing the query input. DISQOVER processes the response of the external service and shows the search results as linked data.
- The **product customization** is now entirely managed through the web frontend and integrated into the administrator menu. Modifying the default application name and logo, the primary and secondary color schemes, the labels of the main links and buttons, and the general announcement section, can now easily be **configured via the user interface**, removing the necessity for direct access to configuration files via the terminal.

Improvements:

- The data export functionality is more compatible with Microsoft Excel. Field delimiters and line endings are better recognized.
- User account names (email addresses) and search terms are not stored in the permanent logs in DISQOVER.

Notes:

- Not applicable

Known issues:

- Not applicable

2. Key features

2.1 Data Ingestion Engine

The data ingestion procedure prior to version 5.0 is based on a framework of configurable bash scripts, each of them handling the processing of one data source. This process comprises the way to access a data source, executing a data conversion script, loading the data in a virtuoso triplestore, and running inferencing steps. When all scripts are run a final step to produce a Solr index is executed. The new Data Ingestion Engine (DIE) can produce the same results in a **graphical pipeline environment** with many additional capabilities.

The processing of different data sources can be built as a visual pipeline of components linked into a directed acyclic graph. An **extensive library of components** each capable to execute **specific atomic steps** is available. Building and managing such a pipeline can be done via the DISQOVER user interface.

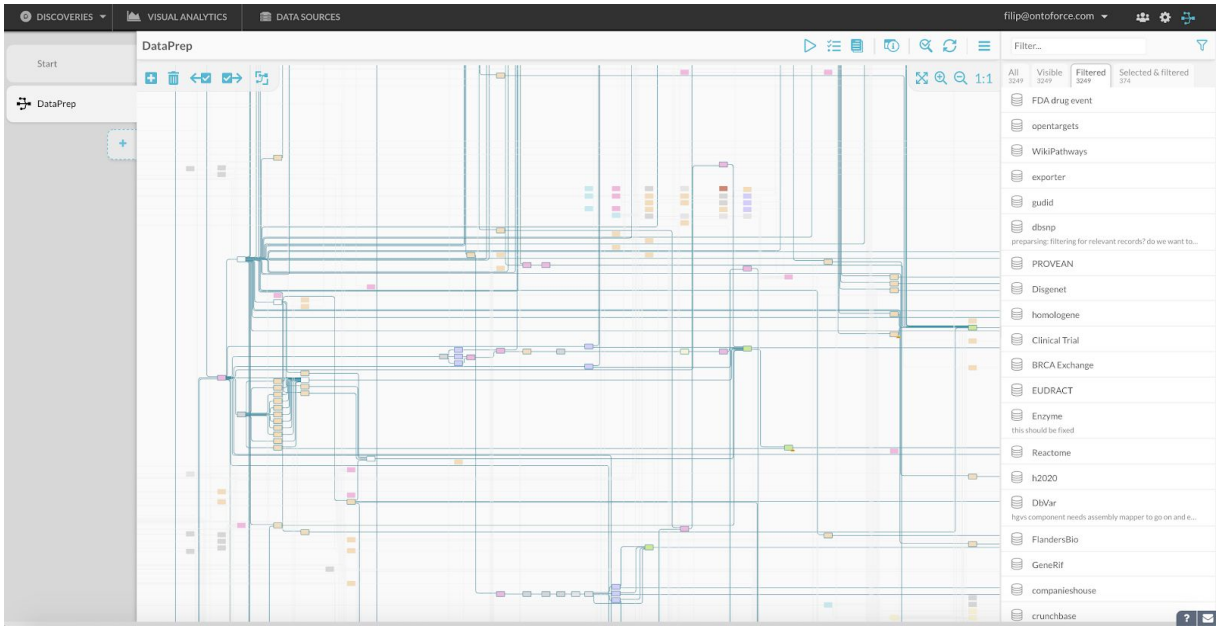

Figure 1: Overview of a pipeline in the Data Ingestion Pipeline user interface.

The **pipeline execution process** can be **monitored while running**. The **dependencies** between components are **traceable**. This allows the user to perform **downstream inspections** starting from a data source and **upstream inspections** starting from DISQOVER data objects such as canonical types, facets, and properties. This level of transparency can be used for optimizing pipeline management and auditing.

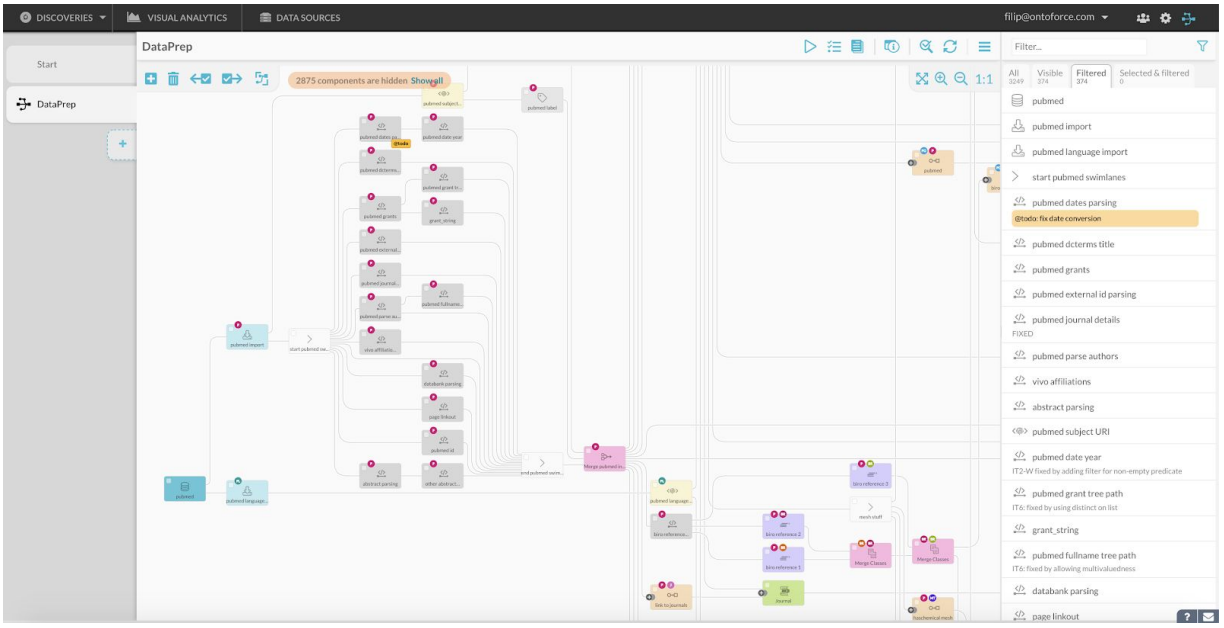

Figure 2: Zooming in on a subset of components in a pipeline.

Data quality control (QC) is an integral part of the Data Ingestion Engine, which has **built-in tools** to perform **unit tests** and **integration tests**. Integration tests can be realised with specific **pipeline QC components** to perform incoming, processing and outgoing data quality assessments. The QC system is **tolerance-based** and uses **warning and error thresholds** as realistic expectations of data quality in the process of data ingestion.

The pipeline layout can be used to map informative **overlays** visualizing different types of information such as component execution time and QC test results.

2.2 Instance view editor

The **layout of the properties** in single and multiple instance views in the **Analytics Dashboard** can be **customized in the user interface**. These layouts can be saved as templates or shared if the user has the right privileges.

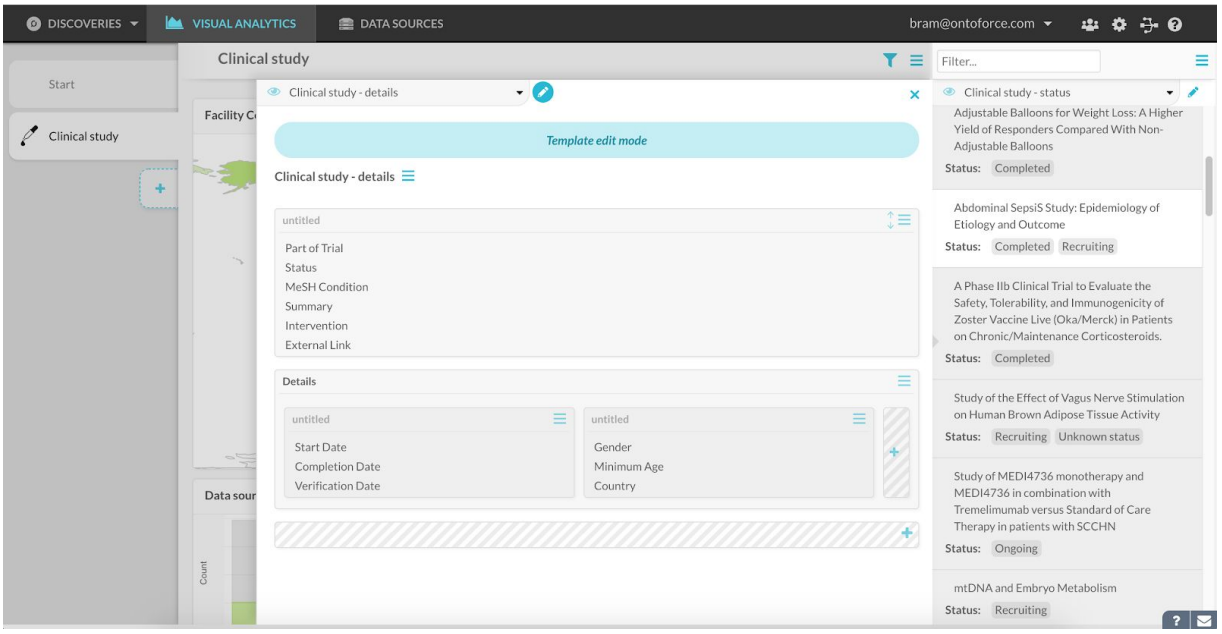

Figure 3: Instance view in the Analytics Dashboard in edit mode.

Single instance views can be visualized in a **print-friendly** way in the browser and printed as a **report**.

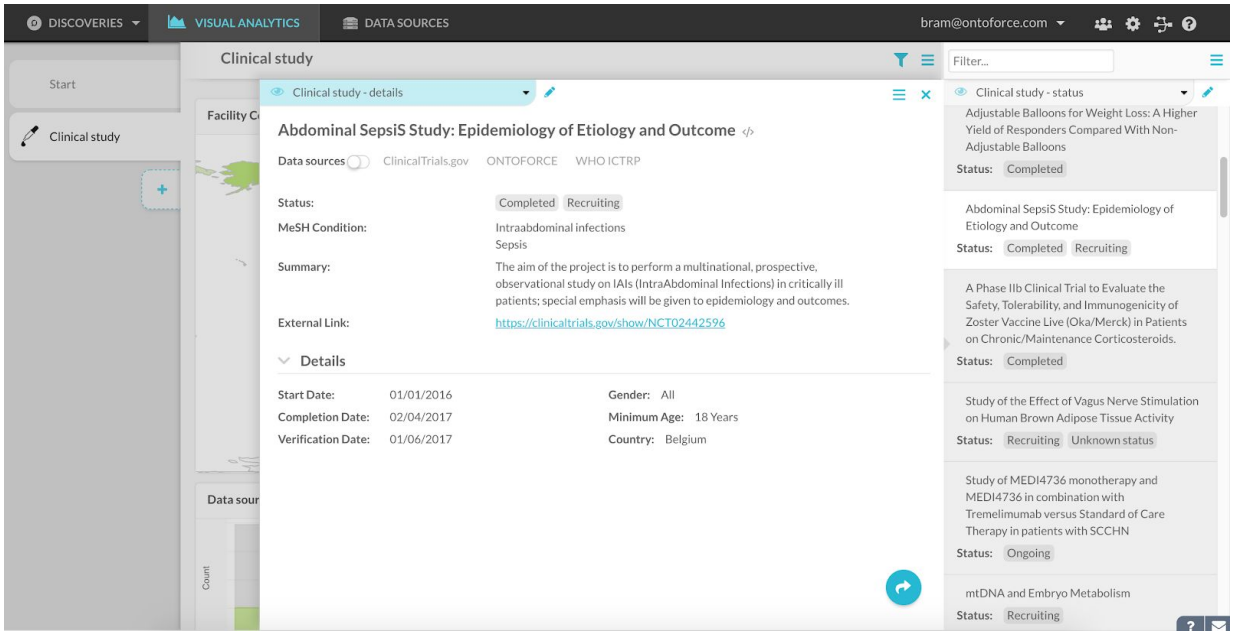

Figure 4: Instance view in the Analytics Dashboard in view mode.

2.3 External search plugin architecture

One implementation of an external query plugin is a **chemistry search** with the inclusion of a chemical drawing canvas This plugin is available for local deployment upon request. A search starts with drawing a molecule or molecular structure or pasting a .mol file in the canvas. After choosing a search type such as identity, similarity or substructure search, the query is sent to a third-party cheminformatics service that is installed back-to-back with a local deployment of DISQOVER. The service responds with chemical identifiers matching the search criteria. The tiles of the canonical types ‘chemical’ or ‘active substance’ show the number of hits.

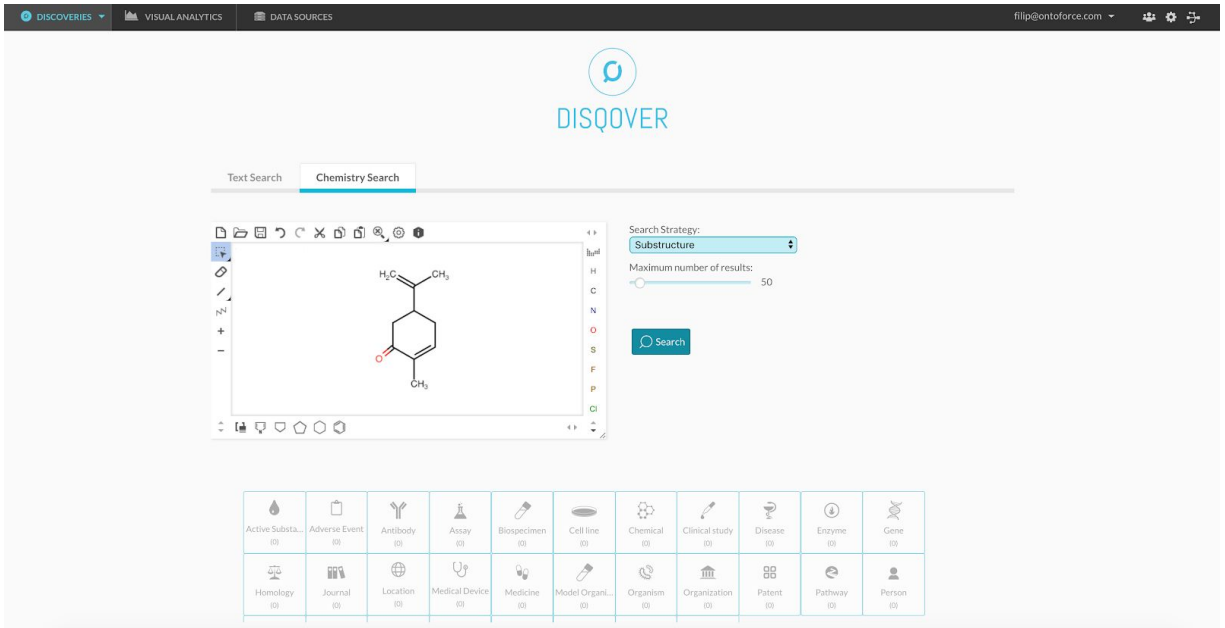

Figure 5: Layout of a chemistry search implementation. A chemical drawing canvas is included in the frontend and is accessible via selecting the ‘Chemistry Search’ tab.

Another implementation could for example be a **nucleic acid, a protein alignment or a similarity search** starting with uploading or pasting a nucleic acid or amino acid sequence or identifier. The sequence or identifier is sent to a third party service which executes a BLAST-like or a protein 3D-model similarity search on a library of sequences. The identifiers of the matching results are sent back to DISQOVER and can be visualized as linked data in DISQOVER.

3. Upgrade instructions

Upgrading from version 4.1 to version 5.0 by executing the default upgrade command is possible but not recommended since that causes the new Data Ingestion Engine not to be functional.

Upgrading to version 5.0 requires to migrate to a new installation using the most recent DISQOVER 5.0 machine image for Amazon Web Services (AWS) or Microsoft Azure, or using the DISQOVER install script on a server with a freshly installed CentOS or RHEL version 7.5 operating system.

Product customization management is now stored in a database. In addition to previous migrate instructions, existing customizations need to be saved before upgrading and re-applied via the frontend after installation.

More details about migrating to DISQOVER 5.0 are available in the DISQOVER 5.0 Data Science & Administrator Manual and via contacting our support team at support@ontoforce.com.