Semantic linking and integration of researchers and research organizations in DISQOVER

Filip Pattyn[1], Steven Vanderschaeve[1], Stijn Vermaere[1], Paolo Van Huffel[1], Kenny Knecht[1], and Hans Constandt

ONTOFORCE, Ottergemsesteenweg-Zuid 808, 9000 Gent, Belgium
filip.pattyn@ontoforce.com,
WWW home page: http://www.ontoforce.com

Abstract. **Keywords:** linked data, semantic web, ORCiD, GRID, data aggregation, smart searching

1 Background

ONTOFORCE has developed DISQOVER (http://www.disqover.com), a semantic search engine with faceted search capabilities for life sciences. It currently allows to search automatically across more than 115+ different public data sources that are aggregated, interlinked and contain information about 21 different data types.

This system uses semantic web technologies to embrace the mapping efforts from different projects like Unified Medical Language System (UMLS), SNOMED CT, ICD10, ICD9, MedDRA, Human Disease Ontology (DO), Medical Subject Headings (MeSH) and Human Phenotype Ontology (HPO) amongst others. These projects structure and encode information related to diseases, phenotypes, and clinical signs.

Many of the sources included in DISQOVER aren?t available in a semantic web format. Therefore, we developed a data source update pipeline that constantly checks the update status of the data at its source. It also means that the data needs a conversion step to a semantic web format (e.g. ttl) to be able to be linkable to other data sources.

2 Results

Since we are trying to aggregate the information of identical concepts, we are investigating different mapping strategies. For example data sources describing information about chemicals contain different identifiers depending on the original source of the data. ChEMBL and PubChem could be considered as golden sources and other identifiers like a CAS number or an InChi key could be considered for mapping too.

One of the more challenging items in linking data is to solve the mapping issues of person names in publications, patents, clinical trials and grant applications. Names can be misspelled, initials are sometimes used instead of a full

first name or one person could be annotated with different spellings. We used ORCiD (`http://www.orcid.org`) that provides a persistent digital identifier for researchers, as a golden source for persons. A researcher can make a personal profile on ORCiD and can add his or her scientific output to it. This makes it possible to change names into physical persons based on the claimed scientific output linked to an ORCiD profile. Subsequently, we employ mapping techniques to map the remaining names to these ORCiD Unique Resource Identifiers. As a result user profiles are directly linked to clinical trials, publications, patents, grant applications and indirectly to drugs, chemical, proteins, genes, pathways and more.

Moreover, we try to solve the issue of different layouts and spellings of author affiliations in publications, clinical trials, patents and grant applications. The Global Research Identifier Database (or GRID) (`http://grid.ac`) is a curated catalogue with a worldwide coverage of research organizations. We digitize the author affiliations and map them to other entries of affiliations in public data sources.

Overall, this work has let to a more in-depth linking of persons and organizations with other data types.