REPOTRIAL

An in silico-based approach to improve the efficacy and precision of drug REPurpOsing TRIALs for a mechanism-based patient cohort with predominant cerebro-cardiovascular phenotypes

D1.1 Diseasome 2.0 network

Project acronym:	REPO-TRIAL
Grant Agreement:	777111
Project Duration:	01 February 2018 – 31 January 2023 (60 months)
Version:	V1
Date:	29/06/2018
WP Leader:	Jan Baumbach (TUM)
Authors:	Elisa Anastasi (UNEW - 3), Jan Baumbach (TUM), Charlotte Brown (UNEW – 3), Simon Cockell (UNEW – 3), Spencer Du (UNEW – 3), Keith Flanagan (UNEW - 3), Tim Kacprowski (TUM), Sepideh Sadegh (TUM), Anil Wipat, (UNEW - 3)
Due date of deliverable	30/06/2018
Actual submission date	30/06/2018

Abbreviations

UM	Universiteit Maastricht
SDU	Syddansk Universiteit
TUM	Technische Universität München
UNEW	University of Newcastle upon Tyne
UKE	Universitaetsklinikum Essen
MHH	Medizinische Hochschule Hannover
UMCU	Universitair Medisch Centrum Utrecht
BIOCRATES	Biocrates Life Sciences AG
SomaLogic	Somalogic Limited
HMPC	Mucke Hermann
concentris	concentris Research Management GmbH

Table of Contents

1. Executive Summary

Diseasome 2.0 contains the information from the OMIM dataset to recreate the Goh and Barabasi 2007 diseasome with updated data from 2017. As part of this deliverable, we have developed an approach to represent data such as OMIM gene and phenotype information in graph format. This approach will make it easier to integrate new datasets in the future. The work in this task has led to the development of a software architecture and system for the standardized parsing and integration of different datasets (RepoDB) (Fig.1.). The preparation of Diseasome 2.0 has acted as a test-case for this architecture.

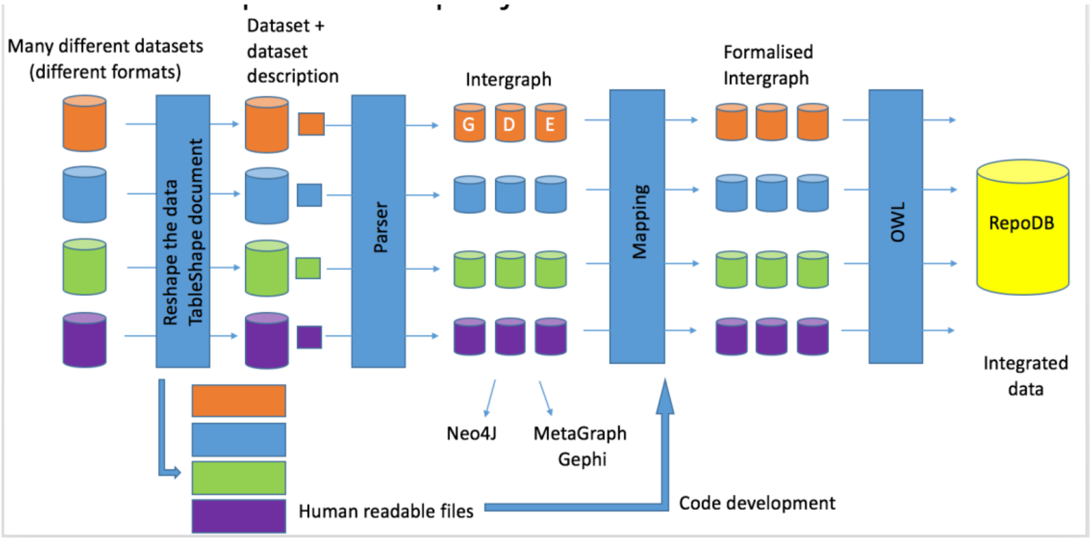

Fig.1. The repotrial integration architecture necessary to produce the final integrated database (RepoDB). Data is parsed in its raw state from many different databases to form datasets and dataset descriptions that ultimately are represented in a graph in the database Neo4J. This graph is then used to build an intermediate graph, called the intergraph, that is ultimately inetgrated through a mapping process to produce views over the data in the form of the final integrated graph.

We have met the month five deliverable as described and will use this as basis of future work to increase the size and complexity of the functionally integrated database.

2. Deliverable report

The parser

We have standardized and distributed our parser documentation so that different team members are focusing on their own datasets and parsing them into the intergraph. Team members have also compiled human readable files describing their assigned dataset, which we will use to inform the process of creating a formalized Intergraph, where the disease terms and gene IDs from different datasets are integrated together. There are currently seven datasets parsed into the Intergraph and a further five datasets are in progress. The parser, as detailed in Milestone 1.1, has been updated to accommodate differences in each dataset (e.g. commented out lines, different table structures). At the mapping stage, the dataset genes are mapped to Entrez Gene IDs and the dataset disorders will be mapped to MeSH disease terms. We have also completed a parser that will convert XML documents to the Intergraph structure, therefore enabling us to import MeSH disease terms as well as gene-disease datasets structured in the XML format.

Diseasome 2.0

Diseasome 2.0 is composed of an updated version of the OMIM dataset to replicate the Goh *et al* 2007 Diseasome. We have extracted the gene information (Gene Symbols) and Phenotype strings (Phenotype) from the file MorbidMap.txt. As per Barabasi, the phenotypes tagged with a (3) were retained as they indicate strong evidence that at least one mutation in the particular gene is the cause of the disorder. Some curation (both automated and manual) of the phenotype strings was necessary to merge the same diseases together (e.g. Immunodeficiency 13, Immunodeficiency 16, Immunodeficiency 22 etc.).

Diseasome 2.0 consists of 3736 gene nodes, 2523 disorder nodes, and 5987 edges. In future versions, as the integration software matures, we will integrate more datasets into diseasome 2.0, with diseasome 2.1 consisting of OMIM and CTD data.

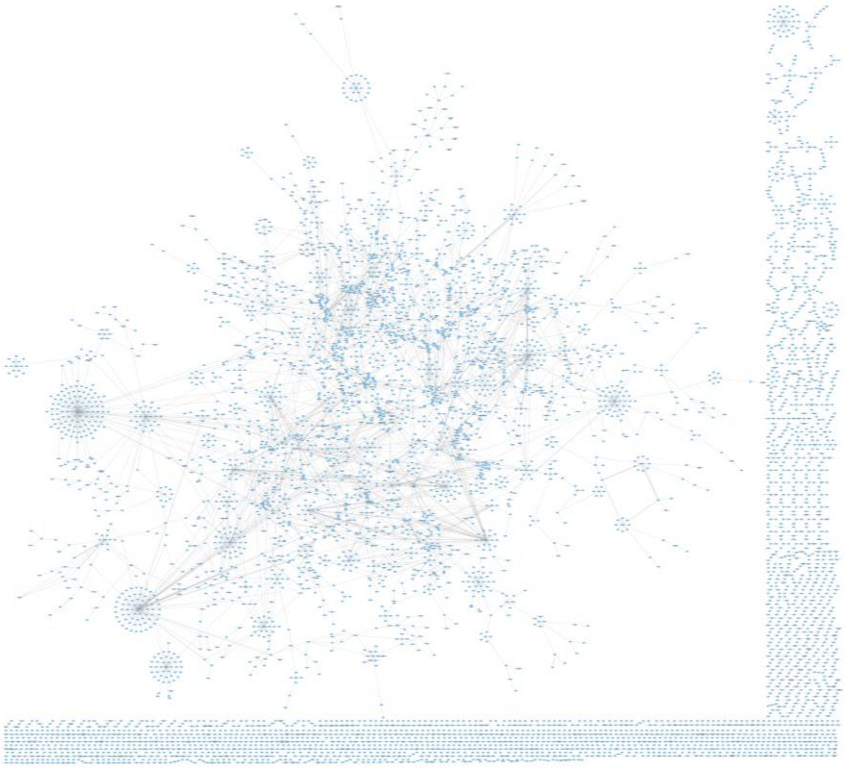

Fig.2. Screenshot of the network Diseasome 2.0

3. Other supporting documents

Document 1: The parser documentation that has been rolled out to team members, who will each analyse and parse their individual datasets.

The Diseaseome 2 network is available on GitHub at https://github.com/repotrial/networks

4. Acknowledgement and Disclaimer

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777111.

This report reflects only the author's views and the European Union is not liable for any use that may be made of the information contained therein.