



Funded by the Seventh Framework  
Programme of the European Union



Project full title:

**Hepatic and Cardiac Toxicity Systems modelling**

Project acronym:

**HeCaTos**

Collaborative project

HEALTH.2013.1.3.-1:

Modelling toxic response in case studies for predictive human safety assessment

**FP7-HEALTH-2013-INNOVATION-1-602156-HeCaTos**

### **Deliverable Report D9.4:**

## **Report on data standards and ontologies applied to data curation**

Work package 9

Due date of deliverable: M36

Actual submission date: October 2016

Start date of project: October, 2013

Duration: 60 months

**Maastricht University (UM)**

Project co-funded by the European Commission within the 7th Framework Programme (2013-2018)		
Dissemination Level		
<b>PU</b>	Public	X
<b>PP</b>	Restricted to other programme participants (including the Commission Services)	
<b>RE</b>	Restricted to a group specified by the consortium (including the Commission Services)	
<b>CO</b>	Confidential, only for members of the consortium (including the Commission Services)	

## Contributions to deliverable - Internal review procedure

<b>Deliverable produced by:</b>	<b>Date:</b>
Ines Smit - Partner EBML-EBI	September 2016
Jasmine Minguet - Partner EBML-EBI	October 2016
Ugis Sarkans - Partner EBML-EBI	October 2016
<b>Deliverable internally reviewed by:</b>	<b>Date:</b>
Jos Kleinjans - Partner UM	October 2016

## Contents

Publishable Summary.....	3
Objectives .....	3
Introduction .....	4
Results.....	6
Data standards and ontology-aware search in BioStudies .....	6
Annotation of ChEMBL assay descriptions .....	9
Background .....	9
Methods and annotation system.....	10
General counts of annotated assays.....	10
Making new links using the MeSH indexing .....	13
Concordance between annotations at abstract and assay level .....	13
Finding disease-associated lower-level MeSH terms.....	15
Difficulties .....	16
References .....	17

## PUBLISHABLE SUMMARY

Work package 9 (WP9) is responsible for creating the HeCaToS data management infrastructure. Two of the main components of this infrastructure are:

- 1) Data warehouse, a catalogue of toxicogenomics projects, studies, and datasets providing cross-dataset searching capability, and
- 2) ChEMBL database that contains bioactivity data on nearly 1.5 million drug-like molecules, extracted from the medicinal chemistry literature or from other public domain databases.

This deliverable describes two distinct efforts within WP9:

- First, to address the work package objective to provide indexing and annotation of data within the HeCaToS data warehouse to allow for more flexible querying, we report on the data standards implemented in the BioStudies database that serves as the platform for the HeCaToS data warehouse. We have developed a customised submission template, to ensure that all HeCaToS data submissions use consistent sample naming and descriptions, according to the needs of data analysts. Also, the BioStudies database allows ontology-aware searching using the Experimental Factor Ontology. This makes it possible for the user to query BioStudies using, e.g., disease terms.
- Second, to address the work package goal to index existing assays within ChEMBL to ensure availability of data for HeCaToS modelling efforts, we report on the annotation of ChEMBL assay descriptions with Medical Subject Headings (MeSH), which describe concepts such as diseases and cellular processes. A process was established to annotate the descriptions using a publicly available software system called the Medical Text Indexer, and all 1.2 million ChEMBL assay descriptions have been processed. The preliminary results suggest that the annotation is useful for identifying lower-level processes relevant to disease and toxicity. This is expected to enable queries aimed at discovering relationships between organism-level outcomes and lower-level assays, as well as aid data pooling in support of WP1 deliverables.

## OBJECTIVES

According to the HeCaToS Description of Work:

- Provide deeper indexing and annotation of data within data warehouses to allow more flexible querying;
- Build a map of toxicology assays from available data, and use this clustering to guide model design and generation.

These objectives relate to Task 9.5:

- Curation and indexing of bioassays related to the toxicity associated targets will be performed across both 'omics and phenotypic assay systems to ensure maximal alignment of available data towards toxicity prediction, and also to enable the application of multi-scale modelling. Where available ontologies, such as the BioAssay Ontology (BAO) will be used to describe assays. Tools will be developed to identify data outliers, and inconsistent data, and such data will be flagged in the database, in a similar way to that used in ChEMBL.

## INTRODUCTION

This report describes two distinct efforts within WP9, both addressing the above objectives. First, related to objective 9.4, we describe the data standards used for submissions into the HeCaToS data warehouse (based on the EBI's Biostudies database), as well as the ontology-aware query mechanism implemented. Second, related to objective 9.7, we report on the annotation of ChEMBL assay descriptions with Medical Subject Headings (MeSH) to support modelling efforts.

In the Milestone Report MS.14 "First set of novel experimental data successfully uploaded to data warehouse" we described and analysed the various potential ways of uploading data into the HeCaToS data warehouse. These options were discussed at length during the annual project meeting in 2015, and the consensus was reached to use a HeCaToS-specific template for describing the studies performed and the resulting datasets. Submissions from project partners are further processed by the data manager at EBI, and loaded into the data warehouse. Since the HeCaToS data warehouse runs within the BioStudies database at EBI, HeCaToS benefits from all developments performed for that generic resource. One of the features relevant for this report is ontology-aware search – as an example, users can search for "liver disease" and get not only studies explicitly annotated by that term, but also, e.g., all "hepatocellular carcinoma" studies. This feature will be especially useful after the public release of HeCaToS data in future, when these datasets will be searchable along with the other 600,000+ studies in the BioStudies database.

A substantial part of the ChEMBL database is functional, cellular and phenotypic assays. In contrast to the target binding data in ChEMBL, which is generally well understood, these functional and higher-level assays are still relatively unexplored. The majority of functional assays in ChEMBL is cell- and organism-based, whereas a minority has a tissue-based or subcellular format. Since these assays span levels of biological organisation, they are of primary interest for multi-scale modelling and the discovery of adverse outcome pathways.

One of the tasks of WP9 is to curate and index bioassays relating to toxicity-associated targets, and to ensure maximal alignment of available data towards toxicity prediction and the application of multi-scale modelling (Task 9.5). Some examples of functional assays in ChEMBL are shown in Figure 1 along with their Bioassay Ontology (BAO)<sup>1</sup> classification as currently available in ChEMBL.

ASSAY_ID	DESCRIPTION	BAO_FORMAT	BAO_ID
469050	Effect on oxygen consumption rate in Wistar Albino rat liver mitochondria ...	mitochondrion format	BAO_0000252
70	Cytotoxicity against human ovarian cancer (1A9) cell lines.	cell-based format	BAO_0000219
226	Cytotoxic activity in a panel of Human ovarian tumor 41M cell line after 9...	cell-based format	BAO_0000219
5590	Inhibition of NFkB (transcription factor) activation in A549 human lung ca...	cell-based format	BAO_0000219
482970	Inhibition of electrically-stimulated contraction in guinea pig ileum at 6...	tissue-based format	BAO_0000221
499118	Antioxidant activity against FeSO4-induced lipid peroxidation in rat liver...	tissue-based format	BAO_0000221
138	Change in heart rate was expressed in percent at a dose 0.002 mg/kg iv ...	organism-based format	BAO_0000218
3155	Compound was tested for the inhibition of 5-HT-induced bradycardia after p...	organism-based format	BAO_0000218
55806	Mean heart rate in conscious dog at 0-1 hr after 5 mg/kg (po) compound adm...	organism-based format	BAO_0000218

**Figure 1:** Examples of functional assays in ChEMBL 21 together with their Bioassay Ontology Assay Format classification, showing assays are indexed across a hierarchy of biological organisation

ChEMBL assays are currently annotated and indexed using a variety of ontologies. The Bioassay Ontology Format classification provides information on the experimental system used, e.g. cell-based format refers to assays that use whole cells, as opposed to biochemical or organism-based format. However, this does not help to define which disease or toxicity an in vitro assay is relevant to. ChEMBL targets are mapped to UniProt<sup>2</sup> identifiers, and cells and tissues are also indexed using various

ontologies. However, currently no indexing of diseases or cellular processes within ChEMBL assays exists. Given the diversity of the functional data in ChEMBL, indexing of related assays with a controlled vocabulary would support identification of data relevant to the modelling efforts in WP1.

The tasks of EBML-EBI within WP1 (Computational Chemistry) include developing approaches to model curated higher order assays, which are linked to lower-complexity target-based models, and to investigate methods for bioassay pooling to extend the scope of models. For example, WP1 has built a graph database using ChEMBL data to facilitate the exploration of relationships between higher-level assays and target-based assays. The first investigations using the graph database have already revealed a need for standardisation and simplification of the network structure, in order to detect patterns among the noise, and to identify predictive lower-level functional assays. More detailed indexing of bioassays in ChEMBL is expected to benefit these tasks.

Medical Subject Headings (MeSH) is a controlled vocabulary provided by the National Library of Medicine (NLM) for indexing and cataloguing biomedical journal articles in PubMed and other health-related information.<sup>3</sup> MeSH terms cover a wide range of subject areas: for example, MeSH concepts include diseases, organ physiological processes and cell physiological process.<sup>4</sup> Given this wide scope, we decided to annotate ChEMBL assays with MeSH terms.

In the results section below we report on the annotation of ChEMBL assay descriptions with MeSH terms. First, we report on the system used to annotate the ChEMBL assay descriptions. We then illustrate how the annotation helps identify relevant lower-level processes for two diseases, and outline the potential benefits for WP1.

## RESULTS

### Data standards and ontology-aware search in BioStudies

The template for HeCaToS data acquisition was developed in Excel (see Figure 2). It follows the “key-value” model – e.g., ‘Submission’ is the key, while the associated value is S-HECA18. Most studies are on the effect of different compounds on cardiac microtissues. Some others are describing functional data, as well as sample descriptions and processes. A single submission template is used across different types of studies.

1	<b>Submission</b>	S-HECA18		
2	<b>RootPath</b>	proteomics-data/Nathalie/Untreated_cardiac		
3	<b>AttachTo</b>	HeCaToS		
4				
5	<b>Study</b>			
6	<b>Title</b>	Protein expression levels in cardiac 3D cells		
7	<b>Description</b>	In Vitro experiment with 3D cardiac cells untreated, 3 replicates, seven time points		
8	<b>Design Type</b>	protein profiling by proteomics		
9	<b>Factor Name</b>	Age		
10	<b>Factor Name</b>	Compound		
11	<b>Factor Name</b>	Dose		
12	<b>Factor Name</b>	Dose Duration		
13	<b>Factor Name</b>	Dose Frequency		
14	<b>Factor Name</b>	Sample TimePoint		
15	<b>Assay Measurement T</b>	Proteomics Profiling		
16	<b>Assay Technology Ty</b>	Mass spectrometry		
17	<b>Assay Technology Pla</b>	Fusion (Thermo)		
18	<b>Organism</b>	Homo sapiens		
19	<b>Organ</b>	heart		
20	<b>Biological Replicate</b>	1		
21	<b>Compound</b>	Untreated		
22	<b>Dose</b>	control		
23	<b>Protocol Type</b>	In solution digestion		
24				
25	<b>Files</b>	<b>Roche ID</b>	<b>Sampling Date</b>	<b>Sampling Time Point</b>
26	20160303_001_400.raw	#0400_F	19/01/2016	T2
27	20160303_031_401.raw	#0401_F	19/01/2016	T2
28	20160303_012_402.raw	#0402_F	19/01/2016	T2
29	20160303_012_403.raw	#0403_F	19/01/2016	T8
30	20160303_015_404.raw	#0404_F	19/01/2016	T8
31	20160303_003_405.raw	#0405_F	19/01/2016	T8
32	20160303_036_406.raw	#0406_F	20/01/2016	T24
33	20160303_032_407.raw	#0407_F	20/01/2016	T24
34	20160303_002_408.raw	#0408_F	20/01/2016	T24
35	20160303_013_409.raw	#0409_F	22/01/2016	T72
36	20160303_026_410.raw	#0410_F	22/01/2016	T72
37	20160303_019_411.raw	#0411_F	22/01/2016	T72
38	20160303_021_412.raw	#0412_F	26/01/2016	T168
39	20160303_022_413.raw	#0413_F	26/01/2016	T168
40	20160303_037_414.raw	#0414_F	26/01/2016	T168
41	20160303_033_415.raw	#0415_F	29/01/2016	T240
42	20160303_008_416.raw	#0416_F	29/01/2016	T240
43	20160303_007_417.raw	#0417_F	29/01/2016	T240
44	20160303_023_418.raw	#0418_F	02/02/2016	T336
45	20160303_027_419.raw	#0419_F	02/02/2016	T336
46	20160303_011_420.raw	#0420_F	02/02/2016	T336
47				
48	<b>Author</b>	a1		
49	<b>Name</b>	Nathalie Selevsek		
50	<b>Contact Mail</b>	selevsek@fgcz.ethz.ch		
51	<b>&lt;affiliation&gt;</b>	o7		
52				
53	<b>Organization</b>	o7		
54	<b>Name</b>	ETH zurich-university Zurich		

Figure 2: Layout of the HeCaToS data acquisition template

A submission template includes a header that describes some technical information, followed by a section describing the metadata of the study: title, factor names, the technology used, the study description, references of the samples used, compounds and doses, and a few others. All these annotations are constant across the study.

This is followed by a table with file names, and the information that helps interpret the individual files, i.e., information that is not constant across the study. For a compounds effect study, this information would include sampling date, time points, and dosage regime. At the end of the template the information about the submitter can be entered.

Figure 3 illustrates how the data is represented in the web interface.

The screenshot displays the BioStudies web interface. At the top, there is a navigation bar with links for Home, Browse, Submit a study, Help, and About BioStudies. A search bar is located on the right side of the header. Below the header, the BioStudies logo is prominently displayed. The main content area shows a study titled "Protein expression levels in cardiac 3D cells" by Nathalie Selevsek, created on June 3, 2016. The study is marked as private. The description states: "In Vitro experiment with 3D cardiac cells untreated, 3 replicates, seven time points". The design type is "protein profiling by proteomics". The factor name is "Age, Compound, Dose, Dose Duration, Dose Frequency, Sample TimePoint". The assay measurement type is "Proteomics Profiling". On the right side, there is a section titled "Download data files" which shows a table of data files. The table has columns for Name, Size, Roche ID, and Sampling Date. The files listed are:

Name	Size	Roche ID	Sampling Date
20160303_001_400_re.raw	1 GB	#0400_P	19/01/2016
20160303_002_408.raw	1 GB	#0408_P	20/01/2016
20160303_003_405.raw	1 GB	#0405_P	19/01/2016
20160303_007_417.raw	1 GB	#0417_P	29/01/2016
20160303_008_416.raw	1 GB	#0416_P	29/01/2016

Below the table, it indicates "Showing 1 to 5 of 21 entries" and "No file selected". There is also a section for "Similar Studies" which lists related studies.

**Figure 3:** Example of a HeCaToS study, as available via the data warehouse user interface.

In addition to a list of studies (as in Figure 5), which is the default representation of a set of studies in the BioStudies user interface, we have developed a HeCaToS-specific study overview page, which provides an easy to use, 2-dimensional overview of what data is available. The two dimensions are for the compound, and the technology used.


EMBL-EBI

Services

Research

Training

About us



Search

Examples: [hyperplasia](#), [PMCS16016](#)

☒ Search in hecatos only

Home

Browse


Submit a study

Help

About BioStudies

Feedback


Logout [Hecatos]




The HeCaToS project (Hepatic and Cardiac Toxicity Systems modelling) aims at developing integrative in silico tools for predicting human liver and heart toxicity

Assay Technology Type → NMR	MeDIP-seq	total RNA-seq (with ribo-depletion)	Mass spectrometry
Compound ↓			
Idarubicin	<a href="#">S-HECA15</a>	<a href="#">S-HECA7</a>	<a href="#">S-HECA12</a> <a href="#">S-HECA19</a> <a href="#">S-HECA20</a>
DMSO	<a href="#">S-HECA16</a>	<a href="#">S-HECA5</a>	<a href="#">S-HECA1</a> <a href="#">S-HECA24</a>
Doxorubicin	<a href="#">S-HECA17</a>	<a href="#">S-HECA6</a>	<a href="#">S-HECA10</a> <a href="#">S-HECA2</a> <a href="#">S-HECA3</a>
Epirubicin		<a href="#">S-HECA11</a>	<a href="#">S-HECA22</a> <a href="#">S-HECA21</a>
Untreated			<a href="#">S-HECA18</a>

Figure 4: HeCaToS compound study overview page

EMBL-EBI				Services	Research	Training	About us
				<input type="text"/> <input type="button" value="Search"/>			Examples: <a href="#">hyperplasia</a> , <a href="#">PMCS16016</a>
Home	Browse	Submit a study	Help	About BioStudies	Feedback	Logout [hecatos]	

Page [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) .. [24547](#)
 Showing [1 - 25](#) of [613668](#) results
 Page size [25](#) [50](#) [100](#) [250](#) [500](#)
 Sort by: [Released](#)

private
 
[HeCaToS \[HeCaToS\]](#)

private • 53 data files  
[NMR Extracellular Metabolomic Profiles of Cardiac Micro Tissues Exposed to Idarubicin \[S-HECA15\]](#)

private • 34 data files  
[NMR Extracellular Metabolomic Profiles of Cardiac Micro Tissues Exposed to 0.1% DMSO as Vehicle Control \[S-HECA16\]](#)

private • 59 data files  
[NMR Extracellular Metabolomic Profiles of Cardiac Micro Tissues Exposed to Doxorubicin \[S-HECA17\]](#)

private • 14 data files  
[Effect of DMSO on the methylome of cardiac microtissues \[S-HECA5\]](#)

private • 28 data files  
[Effect of doxorubicin on the methylome of cardiac microtissues \[S-HECA6\]](#)

private • 28 data files  
[Effect of idarubicin on the methylome of cardiac microtissues \[S-HECA7\]](#)

private • 42 data files  
[Effect of DMSO on the transcriptome of cardiac microtissues \[S-HECA9\]](#)

private • 84 data files  
[Effect of doxorubicin on the transcriptome of cardiac microtissues \[S-HECA10\]](#)

Figure 5: HeCaToS studies – list view

Finally, Figure 6 illustrates the ontology-aware search in BioStudies, which is available in the HeCaToS context. Users can use a term hierarchy when typing in their search terms, and the search results include not only direct text matches (2<sup>nd</sup> study in the example), but also hits found via traversing the ontology (1<sup>st</sup> study).

EMBL-EBI Services Research Training About us

# BioStudies.

Home Browse Submit a study Help About BioStudies

The HeCaToS project (Hepatic and Cardiac Toxicity Systems model) human liver and heart toxicity

## BioStudies results for cancer

2 results

Sort by: Relevance

private • 53 data files

[NMR Extracellular Metabolomic Profiles of Cardiac Micro Tissues Exposed to Idarubicin \[S-HECA15\]](#)

...). IDA is an anthracycline antibiotics, which is used as a potent chemotherapeutic agent in **acute myeloid leukemia (AML)** and **myelodysplasia (MDS)**. Cells were cultured in 96 well plates and were treated either with therapeutic or toxic dosage of IDA for over 2 to 336 hrs. Metabolomic profiling by...

private • 59 data files

[NMR Extracellular Metabolomic Profiles of Cardiac Micro Tissues Exposed to Doxorubicin \[S-HECA17\]](#)

... (DOX). DOX is topoisomerase inhibitor and widely used anticancer drug. Usage of DOX has been shown to induce cardiac toxicity in **cancer** patients. Cells were cultured in 96 well plate and were treated either with therapeutic or toxic dosage of DOX for over 2 to 336 hrs. Metabolomic profiling by NMR...

**Figure 6:** Ontology-enabled auto-complete and search in BioStudies

## Annotation of ChEMBL assay descriptions

### Background

Two forthcoming HeCaToS tasks for WP1 are...

- T1.10: Develop approaches to extend scope of models through data pooling (cell lines, species, target families etc.);
- T1.11: Develop predictive models for higher-level curated functional assays, linked to lower complexity, target based models.

These tasks are interrelated, as pooling of functional assay data from ChEMBL will most likely be required in order to enable it's modelling using target-based models. Pooling of data for a given protein target is routine in QSAR modelling exercises. However, it is much more difficult to do this with cell-, tissue- or organism-based functional assays, as it is rarely obvious which assays are appropriate to merge. This is due to the heterogeneity of assay systems, readouts, cell lines, animal models *etc.* used in these assays.

One approach to solving this problem might be to annotate ChEMBL assays with ontology terms. Assays tagged with common (at some level of the ontological hierarchy) terms might then be candidates for pooling. One ontology that is of interest is MeSH, as it is widely used in the biomedical community and has good coverage of the relevant knowledge domain. For these reasons, and because a tool for automatically annotating texts is available (see below), we decided to explore this ontology first.

### Methods and annotation system

MeSH terms are hierarchically organised within the MeSH hierarchy, and each descriptor has a list of corresponding 'entry terms', which are synonyms and closely related terms that can be used for indexing purposes.<sup>5</sup> An example of the MeSH descriptor for cholestasis and its entry terms is shown in Figure 7. There are more than 27,000 unique MeSH descriptors in MeSH version 2016, each having its own list of entry terms. In addition, there are more than 230,000 MeSH supplementary concepts, which include chemical names and rare diseases.

	DESCRIPTOR_UI	DESCRIPTOR_TEXT	ENTRY_TERM
1	D002779	Cholestasis	Bile Duct Obstruction
2	D002779	Cholestasis	Bile Duct Obstructions
3	D002779	Cholestasis	Biliary Stases
4	D002779	Cholestasis	Biliary Stasis
5	D002779	Cholestasis	Cholestases
6	D002779	Cholestasis	Cholestasis
7	D002779	Cholestasis	Duct Obstruction, Bile
8	D002779	Cholestasis	Duct Obstructions, Bile
9	D002779	Cholestasis	Obstruction, Bile Duct
10	D002779	Cholestasis	Obstructions, Bile Duct
11	D002779	Cholestasis	Stases, Biliary
12	D002779	Cholestasis	Stasis, Biliary

**Figure 7:** An example of a MeSH descriptor showing its unique ID, preferred term, and entry terms (synonyms).

ChEMBL 21 has more than 1.2 million assays with assay descriptions available for annotation. Designing systems for large-scale automated assignment of MeSH headings to biomedical text is a field of research on its own, and machine learning methods continue to be investigated for addressing the task.<sup>6-8</sup> One of the first MeSH indexing systems to have been published is the Medical Text Indexer (MTI) by the National Library of Medicine (NLM).<sup>9</sup> The MTI software is used in-house by the NLM to suggest MeSH terms to curators who index and catalogue new citations.<sup>10</sup> The NLM MTI is used as a baseline for evaluating the performance of other indexing systems.<sup>6,8</sup> Given it is publicly available and provides a convenient Application Programming Interface (API), we used the MTI for annotating ChEMBL assays.

Two major components of the MTI system are MetaMap Indexing, and PubMed Related Citations. MetaMap Indexing identifies and ranks biomedical concepts in the text supplied, which are subsequently mapped to MeSH concepts. Additionally, MeSH terms are extracted from PubMed Related Citations, which are the documents most similar to the supplied text or abstract. When using the MTI API, MeSH headings recommended from both pathways are returned.

### General counts of annotated assays

We developed a workflow to submit assay descriptions from ChEMBL to the Medical Text Indexer API for annotation. We submitted all assay descriptions from ChEMBL version 21 and retrieved at least one MeSH descriptor or supplementary concept for more than 1.2 million assays. The system did not suggest any terms for 4555 ChEMBL assays. An example of a ChEMBL assay description and the MeSH terms annotated to it by the MTI is shown in Figure 8.

ASSAY_ID	DESCRIPTION
315071	Alanine transferase (GPT) enzyme level in serum upon single dose administration was determined as measure of hepatotoxicity; Normal range = 18–30 IU

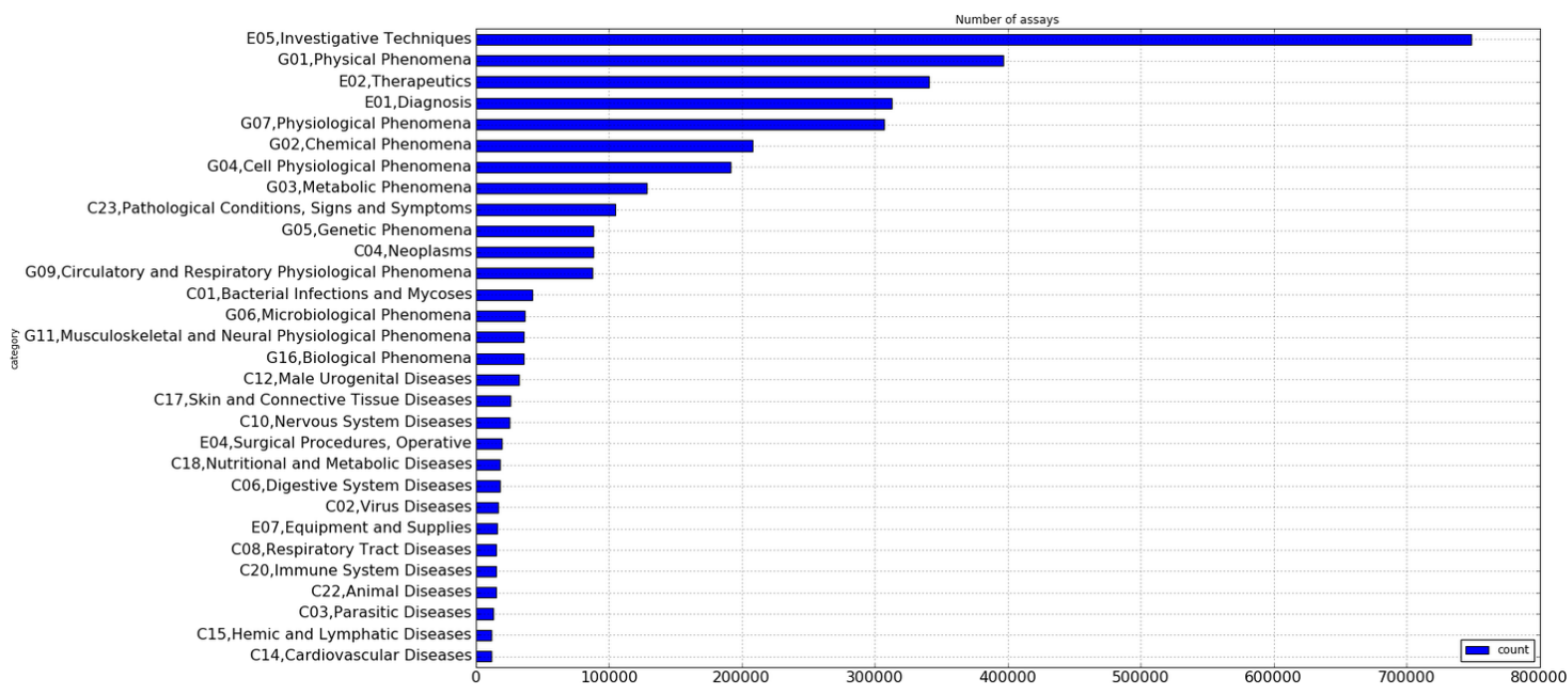
ASSAY_ID	DESCRIPTOR_UI	DESCRIPTOR_TEXT
315071	D000409	Alanine
315071	D000410	Alanine Transaminase
315071	D001219	Aspartate Aminotransferases
315071	D004796	Clinical Enzyme Tests
315071	D005767	Gastrointestinal Diseases
315071	D008099	Liver
315071	D012016	Reference Values
315071	D044967	Serum
315071	D056486	Drug-Induced Liver Injury

**Figure 8:** An assay description from ChEMBL and its corresponding MeSH terms annotated by the Medical Text Indexer system

The highest-level categories in the MeSH hierarchy are shown in Figure 9. The most relevant categories are highlighted, including diseases, investigative techniques, and processes. The underlying idea is to discover, e.g., biological processes and experimental techniques used in ChEMBL assays which are associated with diseases or toxicity. Category F is highlighted because some mental diseases fall under this category (category F03). The number of ChEMBL assays annotated with terms from these categories and all its child terms is shown in Figure 10.

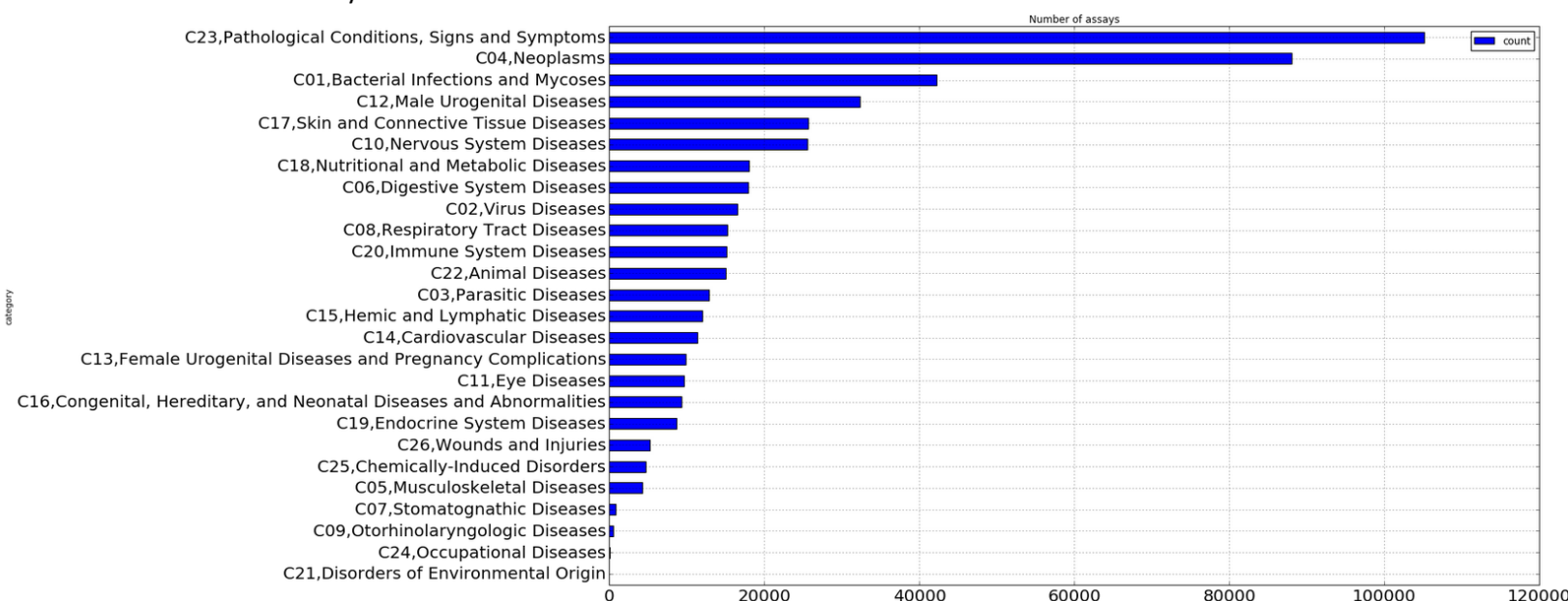
1. + Anatomy [A]
2. + Organisms [B]
3. + Diseases [C]
4. + Chemicals and Drugs [D]
5. + Analytical, Diagnostic and Therapeutic Techniques and Equipment [E]
6. + Psychiatry and Psychology [F]
7. + Phenomena and Processes [G]
8. + Disciplines and Occupations [H]
9. + Anthropology, Education, Sociology and Social Phenomena [I]
10. + Technology, Industry, Agriculture [J]
11. + Humanities [K]
12. + Information Science [L]
13. + Named Groups [M]
14. + Health Care [N]
15. + Publication Characteristics [V]
16. + Geographicals [Z]

**Figure 9:** Categories in the MeSH hierarchy. Categories of interest for the ChEMBL assay annotation are highlighted.



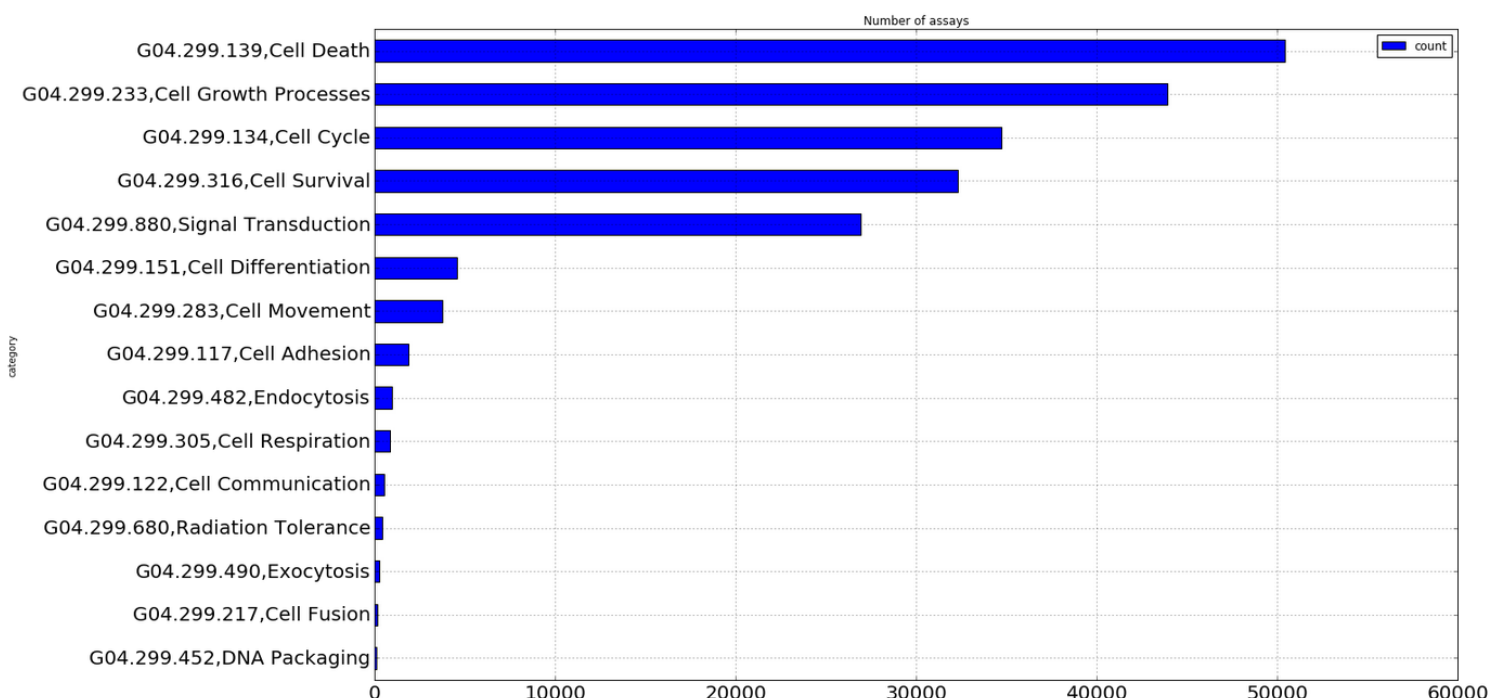
**Figure 10:** The number of ChEMBL assays annotated with the MeSH headings listed (or any of its child terms). Counts have been grouped by categories at one level below the highest level in the MeSH hierarchy. Only the top 30 categories are shown.

The number of assays annotated with terms within each disease category is shown in Figure 11. In total, more than 300,000 assays were annotated with at least one term from the disease categories in the MeSH hierarchy.



**Figure 11:** The number of ChEMBL assays annotated with MeSH headings for diseases (or any of its child terms).

The top 15 counts for the number of assays annotated with 'Cell Physiological Processes' are shown in Figure 12.



**Figure 12:** The number of ChEMBL assays annotated with Cell Physiological Processes from the MeSH hierarchy. The top 15 processes are shown.

### *Making new links using the MeSH indexing*

An important motivation behind the MeSH classification is to aid discovery of links between compounds, diseases and/or toxicities, and lower-level (in vitro) assays and cellular processes. We are only beginning to draw such links, and this will need more time to be explored in collaboration with WP1. However, a few preliminary views on the data can give a useful impression.

### *Concordance between annotations at abstract and assay level*

MeSH terms associated with abstracts, which are indexed manually by NLM curators, are available from PubMed. We have downloaded these terms as provided by PubMed for all articles in ChEMBL.

The tables below (Figures 13 and 14) show the MeSH terms associated with an abstract next to the pooled MeSH terms for all assays extracted for this paper. A random set of PubMed articles is shown, for either articles annotated with 'Asthma', or a MeSH term for any form of diabetes, at the abstract level.

	pmid	abstract_terms	overlap	assay_terms
0	13214	Asthma   Passive Cutaneous Anaphylaxis	Passive Cutaneous Anaphylaxis	Anaphylaxis   Histamine Release   Hydrogen-Ion Concentration   Passive Cutaneous Anaphylaxis
1	423210	Airway Resistance   Asthma   Chemical Phenomena   Stereoisomerism   Structure-Activity Relationship	Airway Resistance   Asthma	Airway Resistance   Anoxia   Asthma   Cough   Foodborne Diseases
2	1969485	Asthma   Chemical Phenomena   Passive Cutaneous Anaphylaxis   Structure-Activity Relationship	Asthma   Passive Cutaneous Anaphylaxis	Asthma   Histamine Release   Passive Cutaneous Anaphylaxis
3	2329571	Asthma   Chemical Phenomena   Lipid Peroxidation   Structure-Activity Relationship	Lipid Peroxidation	Inhibitory Concentration 50   Lipid Peroxidation   Muscle Contraction
4	2415706	Anaphylaxis   Asthma   Chemical Phenomena   Histamine Release   Hypersensitivity   Passive Cutaneous Anaphylaxis   Structure-Activity Relationship	Anaphylaxis   Histamine Release   Passive Cutaneous Anaphylaxis	Anaphylaxis   Histamine Release   Leukocyte Count   Passive Cutaneous Anaphylaxis
5	2769691	Asthma   Chemical Phenomena   Platelet Aggregation   Stereoisomerism   Structure-Activity Relationship	Platelet Aggregation	Biophysical Processes   Bronchoconstriction   Muscle Contraction   Platelet Aggregation
6	2903244	Asthma   Structure-Activity Relationship		Bronchial Spasm   Cyanosis   Histamine Release   Muscle Contraction
7	2999403	Asthma   Blood Pressure   Chemical Phenomena   Molecular Conformation   Passive Cutaneous Anaphylaxis   Structure-Activity Relationship	Asthma   Blood Pressure   Passive Cutaneous Anaphylaxis	Asthma   Blood Pressure   Cerebrovascular Circulation   Dose-Response Relationship, Drug   Histamine Release   Hypotension   Molecular Weight   Passive Cutaneous Anaphylaxis   Phosphorylation

**Figure 13:** A random set of articles the abstract of which is annotated with the MeSH term ‘Asthma’. The MeSH terms at the abstract level and the pooled MeSH terms of all assays extracted from that publication are shown. Terms in **red** are from the Disease and Pathological Signs categories of the MeSH hierarchy, whereas the terms in **blue** are from the category Phenomena and Processes of the hierarchy. Only terms from these branches were included in the table.

In Figure 13, the term ‘Asthma’ itself is not always in the overlap column, but several processes relevant to the disease do appear in the overlap. Relevant processes also occur in the assay terms, such as Histamine Release, Airway Resistance, Leukocyte Count, and Bronchoconstriction. In Figure 14, for the case of diabetes, the disease seems to appear more often in the overlap column. Again, the assay terms include several relevant processes, such as Carbohydrate Metabolism, Obesity, Diet, and Hyperglycemia.

	pmid	abstract_terms	overlap	assay_terms
0	2089120	Diabetes Mellitus, Experimental	Diabetes Mellitus, Experimental	Biochemical Processes   Diabetes Mellitus   Diabetes Mellitus, Experimental
1	2113948	Diabetes Mellitus, Experimental   Molecular Structure   Pregnancy   Stereoisomerism   Structure-Activity Relationship	Diabetes Mellitus, Experimental	Carbohydrate Metabolism   Diabetes Mellitus, Experimental   Inactivation, Metabolic   Neural Conduction
2	2329563	Chemical Phenomena   Diabetes Mellitus, Type 2   Structure-Activity Relationship		Body Weight   Diabetes Mellitus, Experimental   Dose-Response Relationship, Drug   Fasting   Kinetics   Obesity   Protein Binding
3	2614424	Diabetes Mellitus, Experimental		Body Weight   Diabetes Mellitus   Dietary Proteins   Urinary Retention   Urologic Diseases
4	2642552	Diabetes Mellitus, Experimental   Structure-Activity Relationship	Diabetes Mellitus, Experimental	Body Weight   Diabetes Mellitus, Experimental   Obesity
5	2788743	Chemical Phenomena   Diabetes Mellitus   Obesity   Structure-Activity Relationship		Body Weight   Diet   Dietary Fats   Fatty Liver   Food Habits   Hypercholesterolemia   Lipid Peroxidation   Lipogenesis   Oxidation-Reduction
6	3039134	Binding, Competitive   Chemical Phenomena   Diabetes Mellitus, Experimental	Diabetes Mellitus, Experimental	Amino Acid Sequence   Biochemical Processes   Biophysical Processes   Diabetes Mellitus, Experimental   Enzyme Activation   Half-Life   Hyperglycemia   Kinetics   Structure-Activity Relationship
7	3121857	Diabetes Mellitus, Experimental   Structure-Activity Relationship	Diabetes Mellitus, Experimental	Diabetes Mellitus, Experimental   Inhibitory Concentration 50   Kinetics   Neural Conduction   Substrate Specificity

**Figure 14:** A random set of articles the abstract of which is annotated with a MeSH term mentioning ‘diabetes’. The MeSH terms at the abstract level and the pooled MeSH terms of all assays from the same paper are shown. Terms in **red** are from the Disease and Pathological Signs categories of the MeSH hierarchy, whereas the terms in **blue** are from the category Phenomena and Processes of the hierarchy. The **orange** terms occur in both these branches of the hierarchy. Only terms from these categories were included in the table.

### *Finding disease-associated lower-level MeSH terms*

The tables in Figures 13 and 14 use randomly selected assays for illustration, so the next step is to investigate whether some diseases are particularly associated with certain lower-level processes across all assays annotated with the disease. This can be done at the abstract level, using abstracts annotated with a certain MeSH disease and taking the corresponding assays, or at the assay level, using assays that have been annotated with both a disease and lower-level processes.

Figure 15 lists the top 20 MeSH terms most associated with assays from papers annotated with ‘Asthma’. The list was sorted on the Fisher’s Exact p-value (not corrected for multiple comparisons yet). Several relevant lower-level processes appear.

	descriptor_ui	term
0	D001249	Asthma
1	D016535	Bronchial Hyperreactivity
2	D016084	Bronchoconstriction
3	D004195	Disease Models, Animal
4	D002448	Cell Adhesion
5	D012130	Respiratory Hypersensitivity
6	D010323	Passive Cutaneous Anaphylaxis
7	D000403	Airway Resistance
8	D005541	Forced Expiratory Volume
9	D008213	Lymphocyte Activation
10	D007249	Inflammation
11	D006636	Histamine Release
12	D011657	Pulmonary Eosinophilia
13	D007964	Leukocytosis
14	D012143	Respiratory Physiological Phenomena
15	D009119	Muscle Contraction
16	D002450	Cell Communication
17	D006967	Hypersensitivity
18	D001986	Bronchial Spasm
19	D000375	Aging

**Figure 15:** The top 20 MeSH terms most significantly associated with assays from publications annotated with the MeSH term for Asthma.

It is reassuring that the most associated term in assays from papers annotated with ‘Asthma’ is Asthma itself. However, we are of course interested in lower-level processes. It is possible to restrict the results to specific branches of the MeSH hierarchy, and this strategy is being investigated further.

In conclusion, we have established a process for annotating ChEMBL assay descriptions with MeSH terms and submitted all ChEMBL assay descriptions. The preliminary results suggest that the annotation is useful for identifying lower-level processes relevant to disease and toxicity. It is hoped the annotation will enable more complex queries aimed at discovering relationships between organism-level outcomes and lower-level assays, and aid data pooling in support of WP1 deliverables.

Overall, the MeSH indexing provides a standardised annotation of diseases in the assays, which can now be used by WP1 for HeCaToS modelling objectives.

## DIFFICULTIES

The data warehouse work of HeCaToS would benefit from a more direct way for project partners to submit data into the database, via a user-friendly submission tool. Such a tool has been released for generic BioStudies submissions, but it is not yet possible to impose a submission template for a particular set of submissions (such as HeCaToS ‘omics studies). This improvement has been planned (not funded by HeCaToS), and HeCaToS will be able to benefit once this functionality is ready.

The MTI tool was designed by the NLM to annotate article abstracts. However, ChEMBL assay descriptions are much shorter than the typical abstract, typically consisting of a single sentence. The method might thus not be expected to perform as reliably on the assay descriptions as on the abstracts.

Some further actions are needed to clean up the annotations, including flagging false positive assignments. For example, sometimes wrong annotations happen, such as the assignment of the MeSH term for ‘Negotiating’, in its conflict resolution sense, in cases where the assay description mentions e.g. “MRP3-mediated drug transport”. Rules need to be developed to exclude such cases or flag these types of outliers, and restrict to the appropriate parts of the hierarchy.

The sheer number of assay descriptions in ChEMBL (over 1.2 million) precludes any large-scale manual annotation with ontology terms. While the MTI tool allows the automated annotation of text with MeSH terms, other ontologies of interest might not have similar tools available. Thus, other techniques, such as keyword matching and text mining algorithms might need to be investigated if this exercise is to be extended to other ontologies.

## REFERENCES

1. BioAssay Ontology. [accessed 2016 Sep 20]. <http://bioassayontology.org/>
2. UniProt. [accessed 2016 Sep 20]. <http://www.uniprot.org/>
3. National Library of Medicine, Medical Subject Headings (MeSH) Fact Sheet. [accessed 2016 Aug 15]. <https://www.nlm.nih.gov/pubs/factsheets/mesh.html>
4. MeSH Browser 2016. [accessed 2016 Aug 15]. <https://www.nlm.nih.gov/mesh/MBrowser.html>
5. National Library of Medicine, Medical Subject Headings Entry Terms and Other Cross-References. [accessed 2016 Aug 15]. [https://www.nlm.nih.gov/mesh/2016/mesh\\_browser/MBrowser.html](https://www.nlm.nih.gov/mesh/2016/mesh_browser/MBrowser.html)
6. Tsatsaronis G, Balikas G, Malakasiotis P, Partalas I, Zschunke M, Alvers MR, Weissenborn D, Krithara A, Petridis S, Polychronopoulos D, et al. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*. 2015;16(1):138.
7. Liu K, Peng S, Wu J, Zhai C, Mamitsuka H, Zhu S. MeSHLabeler: Improving the accuracy of large-scale MeSH indexing by integrating diverse evidence. *Bioinformatics*. 2015;31(12):i339–i347.
8. Peng S, You R, Wang H, Zhai C, Mamitsuka H, Zhu S. DeepMeSH: deep semantic representation for improving large-scale MeSH indexing. *Bioinformatics*. 2016;32(12):i70–i79.
9. The NLM Indexing Initiative’s Medical Text Indexer. *MEDINFO*. 2004:268–272.
10. Mork JG, Yepes AJ, Aronson AR. The NLM Medical Text Indexer System for Indexing Biomedical Literature. *BioASQ*. 2013.