BioMedBridges

Deliverable D7.2

Project Title:	Building data bridges between biological and medical infrastructures in Europe	
Project Acronym:	BioMedBridges	
Grant agreement no.:	284209	
	Research Infrastructures, FP7 Capacities Specific Programme; [INFRA-2011-2.3.2.] “Implementation of common solutions for a cluster of ESFRI infrastructures in the field of "Life sciences"	
Deliverable title:	Development of co-annotated mouse-human datasets	
WP No.	7	
Lead Beneficiary:	11: HMGU	
WP Title	Technical integration	
Contractual delivery date:	31 December 2014	
Actual delivery date:	19 December 2014	
WP leader:	Michael Raess	1: EMBL
Partner(s) contributing to this deliverable:	1: EMBL, 11: HMGU, 12: MUG, 20: CIRMMP	

Authors and contributors: Helen Parkinson, Nathalie Conte, Julie McMurry, Jon Ison, Tony Burdett, Drashtti Vasant, Michael Raess, Philipp Gormanns, Frauke Neff, Natalie Bordag, Heimo Müller, Kurt Zatloukal, Claudio Luchinat, Leonardo Tenori

Contents

Figures

Tables

1 Executive Summary

In order to develop a comprehensive set of terms to describe Type 2 diabetes and obesity phenotypes in mouse and human, Type 2 Diabetes-related phenotypes were mined from the literature for use as new phenotype terms. The mined terms were curated and temporally categorised by expert clinicians/diabetologists. The terms were represented as an ontology in OWL format and the utility of the ontology in the annotation of data resources and partner data sets was evaluated. Using the ontology developed here enabled the annotation of mouse and human datasets with specific terminology representing Type 2 Diabetes progression, which will ultimately support translational research.

2 Project objectives

With this deliverable, the project has reached or the deliverable has contributed to the following objectives:

No.	Objective	Yes	No
1	Identify and develop a set of annotations, necessary terminologies, and mappings between terminologies for human and mouse models of diabetes and obesity	X	
2	Identify and group related interacting parameters in human and mouse which are involved in the development of clinical and molecular phenotypes	X	
3	Formalise rules for phenotypic annotation in human and mouse to work towards automation of phenotypic discovery and develop a related prototype service	X	

3.1 New Diabetes ontology

3.1.1 Aims

Currently, data integration between mouse models and human studies is hindered by fundamental differences in the ontologies used by each respective community to describe the same phenotypes. Our objectives were to develop common standards and ontologies to bridge the phenotype gap between mouse and human; such bridges would afford clinical researchers the use of extensive mouse phenotype data. As previously reported in D7.1, we held an Ontology Mapping Workshop to manually review terms for Type 2 diabetes representation in four disease ontologies: The human ontologies were the human disease ontology (DO), Human Phenotype Ontology OMIM, and the Experimental Factor Ontology (which imports the Orphanet Rare Genetic Disease Classification). It became apparent that these human ontologies poorly describe human diabetes and its sub-phenotypes; moreover they differ in scope to the Mouse Phenotype ontology. In humans, diabetes unfolds in temporal stages: prediabetes, diabetes, and late consequences/complications of diabetes; however, the current disease ontologies do not model this. Based on these observations, we have extended existing ontologies to create a diabetes-specific one which enables users to integrate data resources related to particular diabetes concepts.

3.1.2. Identifying disease phenotype relationship using text mining

In order to leverage knowledge in the literature and to limit the amount of human effort required in generating terms for an ontology we identified Type 2 diabetes phenotype associations by text-mining the PubMed abstracts for terms that could be mapped to Mouse and Human phenotype ontologies (MP[1] and HPO[2]). This text mining exercise generated a list of diabetes related terms including etiology terms, secondary complications, diagnostic terms, and

[1] http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2801442/ (November 2012)

[2] http://www.ncbi.nlm.nih.gov/pubmed/24217912 (build 650)

additional phenotype terms. The nature of this task made precision a higher priority than coverage, so we narrowed the search space in various ways to reduce the noise of false positives.

3.1.2.1. Inputs for text mining tool

Our first step was to search the titles of over seven journals to find those that could be mapped to terms related to diabetes (MeSH term `‘Type 2 Diabetes’` and its MeSH synonyms); this resulted in a list of seven journals.

Table 1 Title and accession of journals used for text mining

Journal title	ISSN accession
Diabetes	ISSN:1939-327x
Diabetes Care	ISSN:1935-5548
Diabetologia	ISSN:1432-0428
Diabetes,Metabolic Syndrome and Obesity	ISSN:1178-7007
Diabetes, Obesity and Metabolism	ISSN:1463-1326
Diabetes Research and Clinical Practice	ISSN:1872-8227
Diabetology & Metabolic Syndrome	ISSN: 1758-5996

We mined the abstracts[3] from these journals due to concerns about limited open access content in medical journals and the potential for introducing noise when mining a few journals as full text. Term negation was not a major concern for the text mining because the results would be manually curated by a domain expert.

[3] EuropePMC API http://europepmc.org/RestfulWebService

3.1.2.2. Outputs from text mining tool

The Whatizit[4,5] system was used to programmatically mine the abstracts for phenotypic terms using a dictionary constructed from the Mouse Phenotype Ontology[6] and from the Human Phenotype Ontology[7]. We further used statistical methodology based on term frequency cut offs to differentiate between noise and putative phenotype. This calculation was based on 'term frequency–inverse document frequency' (tf-idf)[8] which is a standard numerical statistic intended to reflect how important a word is to a document in a collection or corpus. The tf-idf value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general. Code for running the pipeline is available on EBI repository and the process has since successfully been applied to other disease areas in other projects[9]. The resulting set of filtered terms is available in Appendix 2.

The set of filtered terms were provided in two iterations to a review-team consisting of clinical diabetologists and a clinical pathologist with experience of human and mouse data (Andreas Fritsche & Harald Staiger, Universitätsklinikum Tübingen; Frauke Neff, Helmholtz Zentrum München). The review process involved definition of disease stage categories, organisation of phenotypic terms into those categories, deletion of terms (199), and addition of new terms (2) (Appendix 2, 1&2). The fact that curators added only two new terms to the text-mined set implies that the statistical threshold for significance was appropriate to cover nearly all relevant phenotypes.

Clinical input as organised on the original spreadsheet by the domain experts:

[4] http://www.ebi.ac.uk/webservices/whatizit/info.jsf
[5] http://www.ncbi.nlm.nih.gov/pubmed/18006544
[6] http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2801442/ (November 2012)
[7] http://www.ncbi.nlm.nih.gov/pubmed/24217912 (Build 650)
[8] http://en.wikipedia.org/wiki/Tf%E2%80%93idf
[9] The Human Phenotype Ontology: Semantic unification of common and rare disease, Graza et al, Nature Genetics,in review

- Type 2 diabetes has three disease progression stages, as defined by expert diabetologists, (prediabetes, manifest diabetes, consequences/complications).
- A phenotype is *manifest_in* at least one Type 2 diabetes stage.
- A phenotype may be *cause_of* or *symptom_of* Type 2 diabetes.
- A phenotype may *manifest_in* Type 1 diabetes and also be *associated_with* other diseases.

Table 2 Total count of MP/HP terms in all categories

	HP terms	MP terms	New terms	Total terms
Diabetes Cause	45	48	0	93
IFG/IGT (Prediabetes)	98	73	0	171
Manifest Diabetes	237	115	2	354
Diabetes Symptom	102	61	0	163
Consequences/ Complications	200	87	2	289
Type 1 Diabetes	172	75	1	248
Type 2 Diabetes	248	125	2	375
Assoc. w/ other diseases too	230	108	2	340
Total (any temporal stage)	**248**	**125**	**2**	**375**

3.1.3 Ontology Model

Ontologies allow easy computational reasoning over representations of data and its relationships. Thus, the resulting classification of the review process was transformed into an ontology model using the Web Ontology Language

(OWL)[10]. To capture the clinical input as organised on the original spreadsheet each disease stage and each phenotype was modelled as individual classes to describe the complex dependencies between them. The resulting DIAB ontology was published (Appendix 3). Note that in the final ontology there were no HP and MP terms that were synonyms of each other and the MP terms were not previously associated with human disease, the associations of disease to human and mouse phenotype terms and curation by experts therefore represent new knowledge.

3.2 Development of co-annotated mouse and human datasets

3.2.1 Context

The amount and diversity of high scale data has been steadily increasing for the past several years. This increase has enabled integrative translational bioinformatics studies across these datasets. But, in order to develop integrative approaches, there is a strong need to be able to identify all experiments that study a particular disease using a common ontology which model it.

Biological knowledge is distributed among many different general and specialized databases. This sometimes makes it difficult to ensure the consistency of information. Datasets in public repositories are typically annotated with free-text fields describing the state of the studied sample/study. These annotations are rarely mapped to concepts in any ontology, making it difficult to integrate these datasets across repositories. Furthermore granularity of descriptive metadata varies enormously depending on the resource and the data itself.

We have therefore annotated biological datasets coming from diverse public data resources with DIAB ontology to facilitate translational research.

[10] http://www.w3.org/TR/owl-features/

3.2.1 Dataset co-annotation

3.2.1.2 Partner dataset co-annotation

We have annotated datasets from WP7 partners (see Appendix 0) and datasets coming from various public databases. Due to time constraints at the time of annotation, there were three datasets (in two studies) from our WP7 partners available for annotation with the ontology. Both were metabolomics datasets undergoing deposition into the Metabolights database[11]. MetaboLights is a public database for Metabolomics experiments and derived information. The database is cross-species, cross-technique and covers metabolite structures and their reference spectra as well as their biological roles, locations and concentrations, and experimental data from metabolic experiments. As relatively few partner generated datasets were available we also examined other datasets selected from the Metabolights database. The database draws on the ‘Investigation/Study/Assay’ (ISA) framework for capturing experimental metadata and to facilitate curation at source and data was accessed in this format.

The following ISA-tab files were selected for their metadata content:

- The investigation file holds project descriptors and summary information such as title, variables (or factors) and description of the different protocols used;
- The Sample file describes the metadata related to the samples (e.g., organism, sample type, batch number, etc.);
- The assay file captures the contextual data for the Mass Spec analysis (e.g., ion mode, type of mass spectrometry), from data acquisition through to the multiple processing steps (sample names, sample transformation names and derived data files).
- The metabolite file captures information about metabolites identified or “Metabolite Assignment File” or “small molecule assignment” sheet including details such as; the database identifier, formula, InChI, SMILES and metabolite identified.

[11] http://nar.oxfordjournals.org/content/41/D1/D781

We first examined the metadata associated with each dataset (Investigation, Sample, Assay and Metabolite) to decide what metadata type was best fitted for mapping with DIAB ontology. The different metadata sets[12], were examined in term of granularity and specificity relative to DIAB ontology. Metabolite and Assay files were respectively linked to metabolites/molecule identified (i.e. “a-Hydroxyisovalerate”) and experimental factors/parameter values (i.e. “2:1 urine/deuterated 0.2 M phosphate buffer”), these metadata were too specialised and out of context for the DIAB ontology. The Sample file contains information linked to sample characteristics (i.e. “Urine”) and described low granularity phenotypic data (i.e. ”Type 2 Diabetes”). Interestingly, the investigation file contain details about the study and a study description field where authors describe the details/aims of the study, often coming from the abstract of the publication thus containing detailed/granular data.

Table 3 Study description example from MTBLS1 dataset in Metabolights database

Study description example for MTBLS1 dataset	"Type 2 diabetes mellitus is the result of a combination of impaired insulin secretion with reduced insulin sensitivity of target tissues. There are an estimated 150 million affected individuals worldwide, of whom a large proportion remains undiagnosed because of a lack of specific symptoms early in this disorder and inadequate diagnostics. In this study, NMR-based metabolomic analysis in conjunction with uni- and multivariate statistics was applied to examine the urinary metabolic changes in Human Type 2 diabetes mellitus patients compared to the control group. The human population were unmedicated diabetic patients who have good daily dietary control over their blood glucose concentrations by following the guidelines on diet issued by the American Diabetes Association. Note: This is part of a larger study, please refer to the original paper below."

We decided to annotate free text present in the study description field with DIAB ontology using NCBO annotator. The annotator software matches words in the text to terms in the DIAB ontology by doing an exact string comparison (a “direct” match) between the text and ontology term names, synonyms, and IDs.

[12] e.g. http://www.ebi.ac.uk/metabolights/MTBLS1

We mapped the study description field to DIAB ontology using NCBO annotator tools, and reviewed the annotations manually, we compiled the results in Table 4.

Table 4 Partner dataset mapping to DIAB ontology results, columns 3-10 correspond to stages of the disease as classified by the experts

Ontology ID	Term	Pre-Diabetic	Manifest Diabetes	Consequences Complications	Associate with other disease (and diabetes)	Diabetes Cause	Diabetes Symptom	Type 1 Diabetes	Type 2 Diabetes	WP7 Data
HP_0001394	cirrhosis	x	x	x	x				x	Graz Mouse
MP_0002038	carcinoma		x	x	x	x		x	x	Graz Mouse
HP_0001396	cholestasis		x		x		x		x	Graz Mouse
HP_0001397	hepatic steatosis	x	x		x	x			x	Graz Mouse
HP_0001394	cirrhosis	x	x	x	x				x	Graz Human
MP_0002038	carcinoma		x	x	x	x		x	x	Graz Human
HP_0001396	cholestasis		x		x		x		x	Graz Human
HP_0001397	hepatic steatosis	x	x		x	x			x	Graz Human
HP_0012115	hepatitis	x	x		x	x			x	Graz Human
MP_0001261	obese	x	x		x	x	x		x	Florence Human
HP_0001513	obesity	x	x		x	x	x		x	Florence Human
HP_0000822	hypertension	x	x	x	x		x	x	x	Florence Human
MP_0002055	diabetes		x	x	x	x	x	x	x	Florence Human

We were able to map the text contained into the description field to the diabetes ontology for all three WP7 partner Mouse and Human datasets, each dataset was mapped to a minimum of 4 terms. This type of annotation allows to flag this data as relevant for Type 2 diabetes phenotypes and help translational research.

3.2.1.2 Metabolight dataset co-annotation

We then use a similar approach on Human and Rodent (Mouse and Rats) Metabolights datasets. We used custom Perl scripts to automatically pull from Metabolights database, study description fields from each datasets and manually mapped the text to DIAB ontology using NCBO annotator. We selected 25 Mouse/Rats and Human datasets we considered to be in scope for annotation and mapped 13 of them to the ontology using both Annotator and manual curation. The average number of mapped terms was 1.57 with a minimum of one to 4 mapped terms per datasets. We then carefully looked at the relevance of these datasets to diabetes and obesity. 5/13 of these datasets were relevant to Diabetes and/or Obesity, furthermore the average number of mapped terms was 2 which is higher than the average of 1.57 for all 13 mapped datasets (see table in Appendix 4-1, gray colored). The unmapped data didn’t correspond to data related with metabolic syndrome and/or obesity and were not expected to map. The complete Metabolight mapped dataset is available in Appendix 4 (1).

3.2.1.2 Biosamples dataset co-annotation

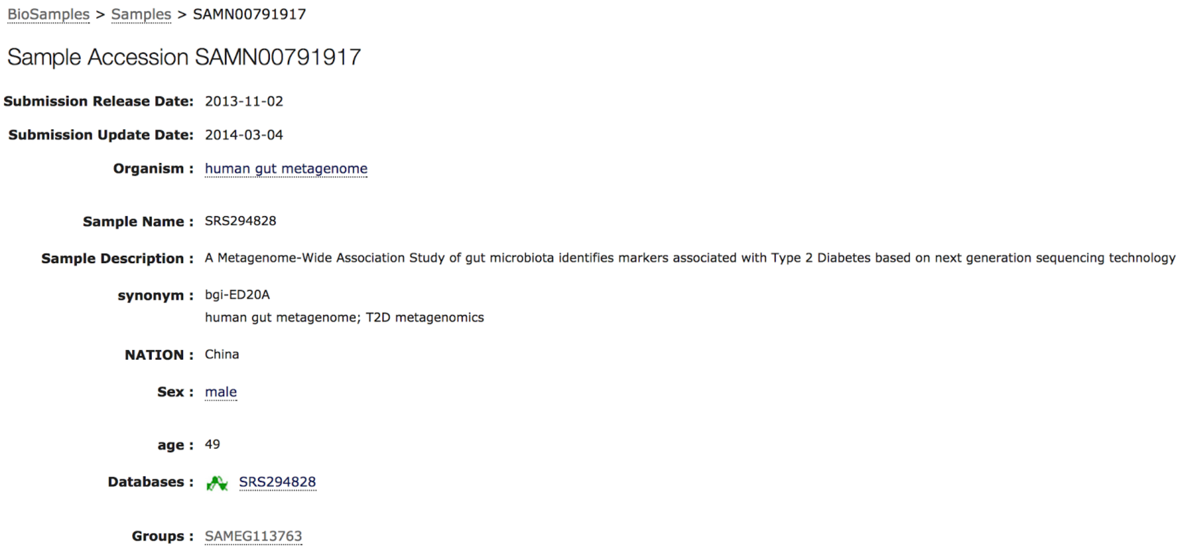
We then annotated datasets from Biosample database. The BioSample Database[13] (http://www.ebi.ac.uk/biosamples) is a database that stores information about biological samples used in molecular experiments, such as sequencing, gene expression or proteomics coming from multiple assay databases such as ArrayExpress (http://www.ebi.ac.uk/arrayexpress), the European Nucleotide Archive[14] or PRoteomics Identificates DatabasE[15] and many more. It contains nearly 4 million samples. Each samples will display attributes types (i.e. ‘organism’) and associated metadata (i.e. ‘Homo Sapiens’) describing the sample. These attributes will vary with the sample, annotator/data owner and source databases. We manually accessed attribute

[13] http://nar.oxfordjournals.org/content/42/D1/D50

[14] http://www.ebi.ac.uk/ena

[15] http://www.ebi.ac.uk/pride

types coming from different samples and sources for their relevance to the Diabetes ontology. Example of attributes types and values for a particular sample from the Biosample database is show below:



BioSamples > Samples > SAMN00791917

Sample Accession SAMN00791917

Submission Release Date: 2013-11-02
Submission Update Date: 2014-03-04
Organism : human gut metagenome

Sample Name : SRS294828
Sample Description : A Metagenome-Wide Association Study of gut microbiota identifies markers associated with Type 2 Diabetes based on next generation sequencing technology
synonym : bgi-ED20A
human gut metagenome; T2D metagenomics
NATION : China
Sex : male

age : 49
Databases : SRS294828

Groups : SAMEG113763

Figure 1 Example entry in the BioSamples database used as input for the co-annotation task

To match the mapping we performed using Metabolight datasets, we looked for text description of a particular sample and mapped the free text found in the Sample Description field. We used custom Perl scripts and BioSD Perl API[16] to automatically pull Sample Description field from Biosamples database via the Rest interface. Code is available on request.

We first select samples linked to ‘Diabetes Mellitus’ and obtained 29004 samples in total aggregated from different databases like ENA, array express and more. From these samples we extracted the Sample Description text and mapped to DIAB ontology using NCBO Annotator and manual curation. Depending on the datasets, there were multiple samples with the same description depending on how many samples were part of the same experiment or group. Unique description were pooled and counted and mapping was processed on those unique descriptions.

[16] https://github.com/EMBL-EBI-GCA/BioSD

Mapping was possible on 15,455 samples (53% of the total number of samples) corresponding to 73 unique Sample Description with a number of mapping ranging from one to 4, the average mapping value was 2.57. Sample groups were also retrieved for mapped datasets and indicates which samples belong to the same study. Full data mapping is available in Appendix 4. We looked carefully at the mapped sample datasets for its relevance to diabetes, the "veracity" of the mapping was compiled under Mapping interpretation column:

- Type 2 represented samples truly linked to Type 2 Diabetes Mellitus.
- Type 1 represented samples truly linked to type 1 Diabetes Mellitus.
- Diab represented samples truly linked to Diabetes Mellitus without clear distinction between Type 1 or Type 2.
- False Pos represented samples not linked to Diabetes Mellitus (false positive).

The number of 'true hit' was very high (93%), true hits were defined by belonging Type 2, Type 1 and Diab mapping category defined above.

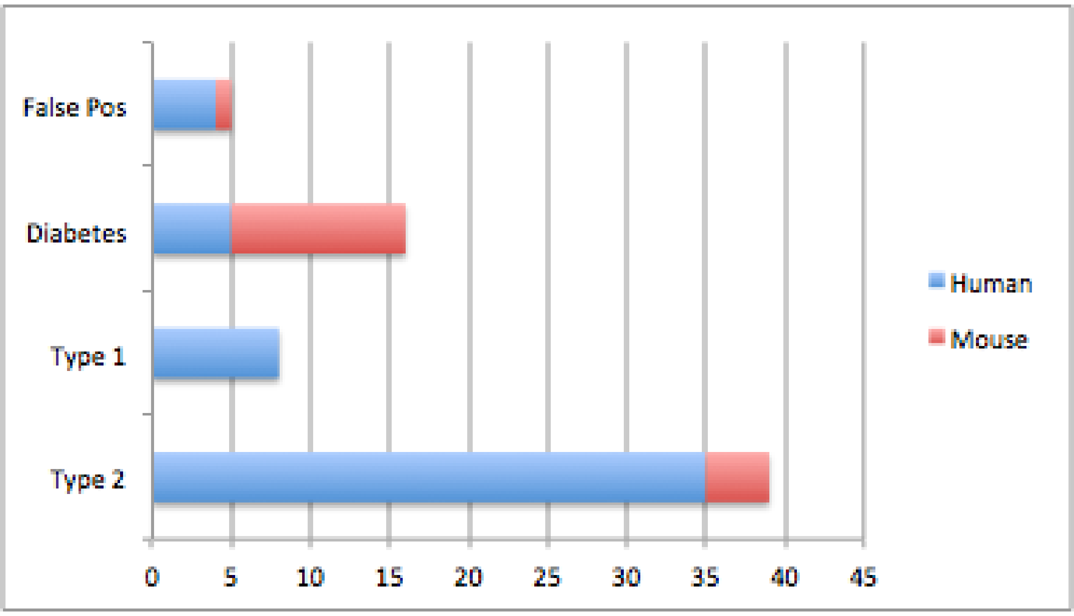


Figure 2 Mapping type breakdown by species. X axis: number of unique dataset descriptions, Y axis: type of mapping ('True mapping': Type 2, Type 1, Diabetes and 'False mapping': False Pos). Blue: human origin dataset, red: mouse origin dataset

The breakdown of mapping was calculated and is represented in the bar chart below. We mainly obtained Human and Mouse Samples, there were 3 Rat samples unique description, one dog and one bacteria. Only Mouse and Human unique datasets description counts are represented. .

55% of the mapped samples were corresponding to Type 2 Diabetes samples all species represented, Human and Mouse being the most frequent ones, respectively 87.5% and 5%. There was one bacteria sample description corresponding to a Metagenome Association Study of gut microbiota to identify markers associated with Type 2 Diabetes. 11% of the mapped samples were corresponding to Type 1 Diabetes samples which was expected as Type 1 and Type 2 Diabetes share common phenotypes and are represented in DIAB ontology. We retrieved only Human samples for Type 1 Diabetes. Finally 27% of the mapped samples were corresponding to diabetes in general with no clear indication of type 1 or 2 in the description. In this case, 25% of the samples were Human and 55% were Mouse.

As expected the average number of mapped samples decreased with the specificity of the mapping from a value of 2.65 for the ‘true’ Type 2 diabetes samples to a value of 2 for ‘true type 1’, to a value of 1.65 for ‘true Diabetes’ samples and finally 1.2 for false positive samples.

Table 5 1Statistical summary of results for Human and Mouse only

Type of Mapping (true or false)	Human sample % per type	Mouse sample % per type	% of each type	Average mapping number per type
Type 2–True	87.5	10	54.8	2.65
Type 1–True	100	0	10.9	2
Diabetes–True	25	55	27.4	1.6
False pos–False	80	20	6.8	1.2

Unmapped Biosample data is presented in Appendix 4. It corresponded to 47% of the Samples obtained from Biosamples database. These samples were initially selected using the Perl API as they contained ‘Diabetes Mellitus’ terminology within their metadata structure. No DIAB ontology terminology was retrieved at the level of Sample description. On manual inspection he level of sample description was often very poor (i.e. :’NAU140gut’) or of very low granularity (i.e. American Gut Project Left hand sample) hence not semantically useful and the fact that we did not obtain mappings was a result of the quality of the input data and unlikely to be due to the ontology.

3.3 Conclusion and Future work

We developed the DIAB ontology representation of expert knowledge about Type 2 diabetes etiology/development and used the ontological representation of this to integrate Type 2 diabetes datasets obtained via multiple techniques from multiple sources and at multiple levels of granularity.

From our text mining it is apparent that individual phenotypic information is usually not captured in databases using an “isolated” terminology, annotators/data owners will describe a particular sample or experiment with a disease terminology rather than an individual or set of phenotypes (i.e. Type 2 Diabetes Mellitus vs abnormal glucose homeostasis, obese). More information/details about the phenotypes may be captured in the user-supplied, text-based description of the study or samples, or linked to the abstract of the related publication.

We searched within open access databases like Metabolights and Biosamples biological datasets corresponding to Diabetes Mellitus with which to test our ontology. Using this ontology we were able to annotate mouse and human datasets with specific terminology representing Type 2 Diabetes progression, which will ultimately support translational research.

During this exercise, data was automatically mapped using the NCBO Annotator and reviewed and revised by manual curation. We noticed that, when the term was not perfectly matched to the ontology label, no match was found using automatic annotators (i.e. “diabetic vs diabetes”, “DM2” vs “Type 2 Diabetes Mellitus”, “impaired insulin secretion” vs “abnormal insulin secretion”). This type of mismatch will not be recognised by the NCBO annotator or Whatizit. Nevertheless, it would be important to capture these imperfect matchings. One solution for this would be to reference in the DIAB ontology synonyms an expanded set of ‘non exact synonyms’ e.g. Type 2 diabetes mellitus synonymous with Type 2 diabetic for such text mining cases.

Future plans:

- Develop an automatic pipeline to fetch from annotation resources, description fields and or publication abstract when available and mine those text for DIAB annotation using text mining software like NCBO Annotator or Whatizit. The results of the co-annotation could be fed back to the annotation resources which would tag those mapped datasets for their potential relevance in Type 2 diabetes disease.
- Capture gene/metabolites/variation information in DIAB ontology annotated datasets in a second steps to help better understanding molecular causes of Type 2 diabetes.
- Model loosely related phenotypes as well as association to non-diabetic disease with OBAN (https://github.com/jamesmalone/OBAN)
- Test the ontology with PhenoDigm (This work will be included in Del 7.3)
- Import DIAB ontology or cross reference to existing relevant Disease/ phenotypes ontologies:
    - DO (Disease Ontology)
    - EFO (Experimental Factor Ontology)
    - MP (Mammalian Phenotype Ontology)
    - HPO (Human Phenotype Ontology)
- Publication of a paper (the draft outline is available in Appendix 5)

5 Delivery and schedule

The delivery is delayed: □ Yes ☑ No

6 Adjustments made

No adjustments were made to the deliverable.

7 Background information

This deliverable relates to WP 7; background information on this WP as originally indicated in the description of work (DoW) is included below.

WP 7 Title: PhenoBridge-crossing the species bridge between mouse and human
Lead: Michael Raess (HMGU)
Participants: EMBL, HMGU, MUG, CIRMMP

This demonstrator project tackles a major challenge related to the available mouse phenotype and human clinical data: different ontological phenotype descriptions hinder researchers from both sides to cross the species bridge between mouse models and human. We will make use of comparable diabetes and obesity-related large-scale datasets in mouse and human provided by Infrafrontier, BBMRI and ELIXIR.

To achieve integration at the level of phenotypes in these species interaction with the wider community is required, several resources are currently in use by different resources and we require expert input to describe the phenotypes, but also to formalize the phenotypic descriptions. In order to make maximum use of existing terminologies (mouse phenotype, human phenotype ontology etc.) we will work with these communities to map between existing terms, provide new terms and also to annotate our datasets.

Work package number	WP7	**Start date or starting event:**	month 13
Work package title	PhenoBridge-crossing the species bridge between mouse and human		

Activity Type	RTD			
Participant number	1:EMBL	11:HMGU	12: MUG	20: CRIMMP
Person-months per participant	21	27	14	15

Objectives

Identify and develop a set of annotations, necessary terminologies, and mappings between terminologies for human and mouse models of diabetes and obesity

Identify and group related interacting parameters in human and mouse which are involved in the development of clinical and molecular phenotypes

Formalise rules for phenotypic annotation in human and mouse to work towards automation of phenotypic discovery and develop a related prototype service.

Description of work and role of participants

Task 1:

Analysis of existing mouse phenotype and human disease ontology terms, leading to submission of proposals for new terms to gather a comprehensive set of terms to describe diabetes and obesity phenotypes in mouse and human. As a first step, potentially relevant phenotypic parameters available from mouse high-throughput screening and in-depth phenotyping studies as well as from human studies across technologies such as gene expression and GWAS studies from the BioSD and metabolome profiles will be listed. For analysed parameters associated with diabetic/obese conditions that are not yet described appropriately in ontology terms, new terms have to be defined. A many to many mapping of phenotypic or diagnostic parameters onto ontology terms will be developed using the clinical expertise of scientists coming from the mouse or human diabetes and obesity fields. (EMBL-EBI, HMGU, MUG, CIRMMP).

Task 2:

Mapping of mouse and human phenotypes of diabetes and obesity. Based on the ontology terms developed in Task 1, observation patterns will be defined that describe clinical and molecular characteristics of diabetes and obesity phenotypes in mouse and human (EMBL-EBI, HMGU, MUG, CIRMMP). The identification of rules and criteria for identifying diabetes and obesity phenotypes (and how they map in mice and humans) will lead to a prototype for an automated procedure to identify phenotype matches across mouse and human (EMBL-EBI, HMGU, MUG, CIRMMP).

Appendix 1: Descriptions of WP7 partner datasets

Study 1	
Name	**The metabolomic fingerprint of severe obesity is dynamically affected by bariatric surgery in a procedure-dependent manner**
Authors	Leonardo Tenori and Collaborators
Location	Università degli Studi di Firenze, Florence, Italy
Taxon	*Homo sapiens*
Study description:	
We applied a 1H-NMR-based metabolomics strategy on human serum samples that were collected before and repeatedly up to one year after distinct bariatric procedures (sleeve gastrectomy, proximal and distal Roux-en Y gastric bypass; RYGB). Serum samples from 106 severely obese patients of the Interdisciplinary Obesity Center, St. Gallen, were collected before and 3, 6, 9, and 12 months after bariatric surgery. Our data indicate that bariatric surgery - irrespectively of the specific kind of procedure - reverses most of the metabolic alterations associated with obesity and suggest profound changes in gut microbiome - host interactions after the surgery. The average age of the recruited patients was 43.6 ± 1.0 years. 31.2% of them are males. Other phenotypic data available are the following: height (cm) 166.2 ± 8.7, weight (kg) 129.4 ± 25.8, BMI (kg/m2) 46.2 ± 7.8, fat mass (kg) 56.6 ± 15.7 (43.9 ± 8.0%), lean body mass (kg) 72.4 ± 18.3, body water (l) 53.8 ± 13.7, extracellular water (l) 28.9 ± 19.1, phase angle (¡) 7.1 ± 1.2. The average diastolic blood pressure was 89 ± 12.8 and the systolic was 135.76 ± 16.1 Sixteen of them experienced psychiatric disorders, 33 had hypertension, 13 had dyslipidemia, 17 were affected by Type 2 Diabetes Mellitus, 39 were smokers, 5 were in menopause. The patients cohort is characterized by higher (with respect to normal weight individuals) levels of alanine, phenylalanine, tyrosine, leucine, isoleucine, valine, acetoacetate, citrate, formate, lactate, pyruvate, glucose, VLDL, methanol, isopropanol, and by lower levels of glutamine and histidine.	
Study 2, Dataset 1	
Name	**Determination of metabolic signatures of murine and human serum and liver tissue**
Authors	Natalie Bordag and Collaborators
Location	Medical University of Graz, Graz, Austria
Taxon	*Mus musculus*

Study description:	
Steatohepatitis is a severe liver disease with a general incidence of 3-5% an eventually leads to cirrhosis or hepatocellular carcinoma, depending on individual susceptibilities. Both in human and mouse settings, genetic modifiers are regarded as the most likely cause of the individual susceptibilities for developing steatohepatitis instead of simple steatosis. Diagnosis of steatohepatitis relies on biopsies, which is often performed only after the disease has become clinically manifest, which is clearly too late. Therefore reliable, non-invasive biomarkers are of high interest, which requires also a solid understanding of the mechanisms of disease development. To facilitate identifying the susceptibility-conferring genes and the associated disease-relevant pathways, an experimental model for steatohepatitis (feeding DDC to mice) was combined with the use of mouse strains where individual chromosomes from a strain sensitive to DDC-feeding (i.e. developing the steatohepatitis phenotype) were placed into a background of a DDC-insensitive strain. For the 21 individual mouse strains (19 consomics and 2 founder strains) an exceptionally large set of transcriptomic, proteomic and metabolomic data was collected and fed into a systems biology model of the pathways relevant for steatohepatitis. A similar approach was undertaken for a mouse model of cholestasis, an impairment of bile secretion and flow, (feeding cholic acid, CA to mice). Here, nine individual mouse chromosome substitution strains and the parental mouse strains were analysed histologically to identify chromosomes harbouring modifier genes for cholestatic liver disease. In parallel, transcriptomic, proteomic and metabolomic data from clinically well-characterized human samples were obtained. These data serve to identify similarities in pathways of the mouse model and the human disease, and to identify biomarkers which reliably, and noninvasively allow distinguishing between simple steatosis and steatohepatitis. The project aimed at identifying genes and associated pathways which determine susceptibility to steatohepatitis distinguishing between simple steatosis and steatohepatitis, both regarding mechanism and biomarkers exploiting systems biology for the analysis of the complex interrelations of the disease-relevant pathways.	
Study 2, Dataset 2	
Name	**Determination of metabolic signatures of murine and human serum and liver tissue**
Authors	Natalie Bordag and Collaborators
Location	Medical University of Graz, Graz, Austria
Taxon	*Homo sapiens*
Study description:	
Steatohepatitis is a severe liver disease with a general incidence of 3-5% an eventually	

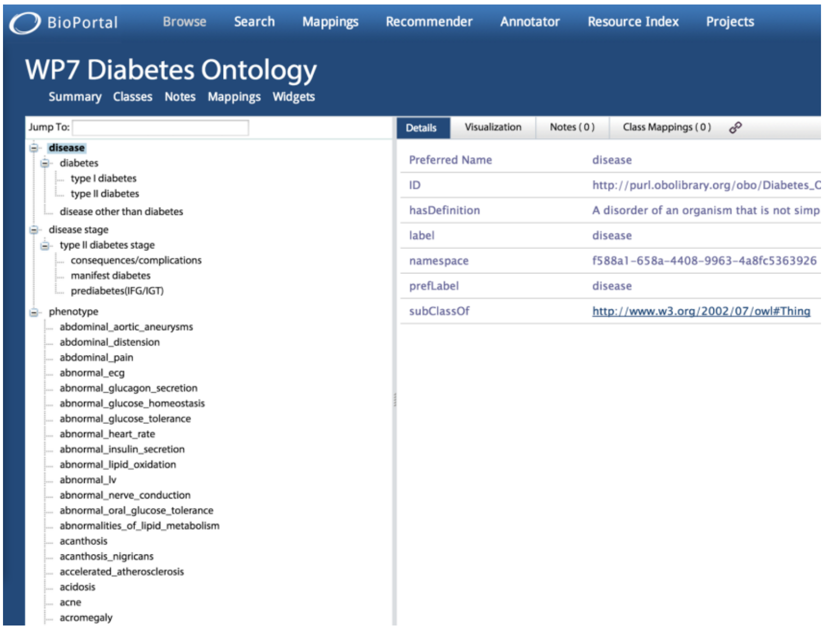
leads to cirrhosis or hepatocellular carcinoma, depending on individual susceptibilities. Both in human and mouse settings, genetic modifiers are regarded as the most likely cause of the individual susceptibilities for developing steatohepatitis instead of simple steatosis. Diagnosis of steatohepatitis relies on biopsies, which is often performed only after the disease has become clinically manifest, which is clearly too late. Therefore reliable, non-invasive biomarkers are of high interest, which requires also a solid understanding of the mechanisms of disease development. Global microarray gene expression analysis was applied to unravel differentially expressed genes between steatohepatitis compared to steatosis and control samples. For functional annotation as well as the identification of disease-relevant biological processes of the differentially expressed genes the gene ontology (GO) database was used. Selected candidate genes (n = 46) were validated in 87 human liver samples from two sample cohorts by quantitative real-time PCR (qRT-PCR). The GO analysis revealed that genes down-regulated in steatohepatitis were mainly involved in metabolic processes. Genes up-regulated in steatohepatitis samples were associated with cancer progression and proliferation. In surgical liver resection samples, 39 genes and in percutaneous liver biopsies, 30 genes were significantly up-regulated in steatohepatitis. Furthermore, immunohistochemical investigation of human liver tissue revealed a significant increase of AKR1B10 protein expression in steatohepatitis (see DOI: 10.1371/journal.pone.0046584). From the set of 80 patients with either chronic Hepatitis C (CHC), non-alcoholic fatty liver (NAFL), or cirrhosis metabolic profiles from serum samples were acquired. In total 17 bile acids, 13 energy metabolism intermediates, 41 acylcarnitines, 100 lipids, 21 amino acids, 20 biogenic amines, 18 oxysterols and 17 prostaglandins were quantified. Statistical analysis of the human data is ongoing and was found to be more complicated than the mouse data due to higher biologically variability of the metabolites and unbalanced group sizes. In parallel, transcriptomic, proteomic and metabolomic data from mouse steatohepatitis and cholestasis models were obtained. These data serve to identify similarities in pathways of the mouse model and the human disease, and to identify biomarkers which reliably, and noninvasively allow distinguishing between simple steatosis and steatohepatitis.The project aimed at identifying genes and associated pathways which determine susceptibility to steatohepatitis distinguishing between simple steatosis and steatohepatitis, both regarding mechanism and biomarkers exploiting systems biology for the analysis of the complex interrelations of the disease-relevant pathways.

Appendix 2: Text mining outputs

1. **Text Mining Output Summary**: http://tinyurl.com/diabont1
2. **First expert curation**. The result of the text mining output summary above was reviewed by a clinical expert who associated each term to a temporal stage of clinical diabetes. http://tinyurl.com/diabont2
3. **Second Expert curation**. The result of the first expert curation was then further reviewed by Diabetologist experts who refined the temporal stages and re-assessed each term. http://tinyurl.com/diabont3

Appendix 3: Published diabetes ontology

The Diabetes ontology was published in Bioportal: http://purl.bioontology.org/ontology/DIAB



- disease
  - diabetes
    - type I diabetes
    - type II diabetes
  - disease other than diabetes
- disease stage
  - type II diabetes stage
    - consequences/complications
    - manifest diabetes
    - prediabetes(IFG/IGT)
- phenotype
  - abdominal_aortic_aneurysms
  - abdominal_distension
  - abdominal_pain
  - abnormal_ecg
  - abnormal_glucagon_secretion
  - abnormal_glucose_homeostasis
  - abnormal_glucose_tolerance
  - abnormal_heart_rate
  - abnormal_insulin_secretion
  - abnormal_lipid_oxidation
  - abnormal_lv
  - abnormal_nerve_conduction
  - abnormal_oral_glucose_tolerance
  - abnormalities_of_lipid_metabolism
  - acanthosis
  - acanthosis_nigricans
  - accelerated_atherosclerosis
  - acidosis
  - acne
  - acromegaly

Preferred Name	disease
ID	http://purl.obolibrary.org/obo/Diabetes_C
hasDefinition	A disorder of an organism that is not simp
label	disease
namespace	f588a1-658a-4408-9963-4a8fc5363926
prefLabel	disease
subClassOf	http://www.w3.org/2002/07/owl#Thing

Appendix 4: Annotated datasets

1. Metabolight annotated datasets: http://tinyurl.com/MetData
2. Biosamples annotated datasets: http://tinyurl.com/bioSDM
3. Biosamples non mapped datasets: http://tinyurl.com/bioSDNM

Appendix 5: Ontology paper outline

Title

Building a hierarchical cross-species diabetes ontology

Introduction/Context

1. Current problem
   - No comprehensive presentation of Type 2 Diabetes in ontologies available.
2. Area of focus
   - Ontology development
   - Clinical research
3. Key terms
   - Ontology
   - Diabetes
   - Disease progression
   - Text mining
4. Thesis statement

   Comprehensive hierarchical ontology development through text-mining driven expert knowledge curation.

Background

1. Current ontological Diabetes representation
   - Complexity of diabetes.

- Ontologies covering diabetes on human side.
- Ontologies covering diabetes on mouse side.

2. Ontology development

- Current approaches

3. Gaps

- Phenotype variety of complex disease like Diabetes is poorly described in current ontologies.
- No disease stages present in current approaches.
- Cause and symptom of diseases a not covered at all.
- Ontologies usually biased and incomplete.

Major and Minor Points

Major Point 1: Text Mining supports ontology development

Minor Point 1: Phenotype text-mining search enables broad and detailed disease representation.

Major Point 2: First comprehensive description of phenotypes present in diabetes

Minor Point 1: Novel integration of mouse and human on ontology level.

Minor Point 2: Our ontology improves existing ontologies like HPO, DO and EFO.

Major Point 3: Temporal and causative classification

Minor Point 1: New kind of classification improves matching of mouse-models to disease as well as existing tools in this field (e.g. PhenoDigm).

Minor Point 2: Disease stage categories allow novel dataset comparison.

Minor Point 3: Phenotype subset identification enables definition of new potential mouse models.

Conclusion

1. Restatement of thesis

2. Next steps

- Incorporation with HPO, DO and EFO.
- Integrate hierarchy into PhenoDigm.