



Universidade Estadual de Campinas
Instituto de Computação



Luana Loubet Borges

BioGraph: Linking Biological Bases Across Organisms

BioGraph: Conectando Bases Biológicas de Múltiplos
Organismos

CAMPINAS
2016

Luana Loubet Borges

BioGraph: Linking Biological Bases Across Organisms

BioGraph: Conectando Bases Biológicas de Múltiplos Organismos

Dissertação apresentada ao Instituto de Computação da Universidade Estadual de Campinas como parte dos requisitos para a obtenção do título de Mestra em Ciência da Computação.

Thesis presented to the Institute of Computing of the University of Campinas in partial fulfillment of the requirements for the degree of Master in Computer Science.

Supervisor/Orientador: Prof. Dr. André Santanchè

Este exemplar corresponde à versão final da Dissertação defendida por Luana Loubet Borges e orientada pelo Prof. Dr. André Santanchè.

CAMPINAS
2016

Agência(s) de fomento e nº(s) de processo(s): CAPES, 01P-3501-2014

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Maria Fabiana Bezerra Muller - CRB 8/6162

B644b Borges, Luana Loubet, 1992-
BioGraph : linking biological bases across organisms / Luana Loubet
Borges. – Campinas, SP : [s.n.], 2016.

Orientador: André Santanchè.
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de
Computação.

1. Banco de dados. 2. Ontologias (Recuperação da informação). I.
Santanchè, André, 1968-. II. Universidade Estadual de Campinas. Instituto de
Computação. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: BioGraph : conectando bases biológicas de múltiplos organismos

Palavras-chave em inglês:

Database

Ontologies (Information retrieval)

Área de concentração: Ciência da Computação

Titulação: Mestre em Ciência da Computação

Banca examinadora:

André Santanchè [Orientador]

Julio Cesar dos Reis

Debora Pignatari Drucker

Data de defesa: 05-08-2016

Programa de Pós-Graduação: Ciência da Computação



Universidade Estadual de Campinas
Instituto de Computação



Luana Loubet Borges

BioGraph: Linking Biological Bases Across Organisms

BioGraph: Conectando Bases Biológicas de Múltiplos Organismos

Banca Examinadora:

- Prof. Dr. André Santanchè (*Orientador*)
Instituto de Computação - Unicamp
- Dr. Julio Cesar Dos Reis
Instituto de Computação - Unicamp
- Dra. Debora Pignatari Drucker
EMBRAPA

A ata da defesa com as respectivas assinaturas dos membros da banca encontra-se no processo de vida acadêmica do aluno.

Campinas, 05 de agosto de 2016

*Dedico este trabalho aos meus
pais, Marcos e Nereide, pelo
imenso amor e sacrifício
para que eu chegasse até aqui.*

Agradecimentos

"Mesmo longe dos meus, mesmo na solidão, te louvo, te louvo em verdade". Agradeço primeiramente a Deus, por ter me capacitado, amado e amparado durante toda a minha vida.

Ao professor André Santanchê, o orientador e amigo mais generoso que eu poderia ter. Que trouxe luz para o meu caminho. A quem eu serei eternamente grata.

Ao meu grande amor, a mulher mais doce, carinhosa e amorosa a quem eu tenho a felicidade de chamar de mãe. Ao meu pai, Marcos, o meu maior incentivador, por sempre ter acreditado em mim e por todo o seu amor. Ao meu amado namorado, Junior, pelo seu profundo amor, companheirismo, cuidado e paciência. Aos meus irmãos Marcos Aurélio, Kassio e Kamila. Vocês todos são os pilares do amor que me sustenta.

Ao Matheus, pela sua amizade, por estar sempre disposto a ouvir minhas inquietações e sempre me fazer rir, tornando os meus dias mais leves.

Ao Hércules, meu melhor amigo, por estar sempre presente com sua infinita bondade.

À todos os membros do LIS, por criarem o melhor laboratório, o mais aconchegante e acolhedor do qual eu poderia fazer parte, pelas contribuições ao longo deste projeto e pelas amizades construídas. Especialmente a Fagner, Francisco, Márcio, Ive e Patrícia.

À todos os meus velhos e novos amigos que sempre me acompanham e me transmitem força, em especial Bruna Strack, Victor, Daniela, Mariane, Yumi, Bruna Borges, Priscilla, Jacqueline e Lucas.

Aos professores da graduação, que foram igualmente importantes para mais essa conquista, em especial Liana, Paulo Pagliosa, Débora, Francisco, Luciana e Nalvo.

Finalmente, agradeço à secretaria do IC e às agências de fomento pelo suporte financeiro que, direta ou indiretamente, contribuíram para a realização desta pesquisa: FAPESP (2013/02269-7), FAPESP/Cepid em Ciência e Engenharia da Computação (2013-/08293-7), Instituto Virtual de Pesquisa FAPESP-Microsoft (projeto NavScales - 11/52070-7), CNPq (projeto MuZOO), FAPESP-PRONEX (projeto eScience), INCT em Ciência Web, CNPq e CAPES (01P-3501-2014). As opiniões expressas neste trabalho não necessariamente refletem as opiniões das agência de fomento.

Resumo

A representação de dados como redes tem se mostrado uma poderosa abordagem para análises de dados em biodiversidade, e.g., interações entre organismos; relações entre genes e fenótipos etc. Neste contexto, bancos de dados e repositórios seguindo um modelo de grafo (e.g., RDF) têm sido cada vez mais utilizados para interconectar informações e para dar suporte a análises dirigidas a redes. Usualmente, este tipo de análise requer a coleta e ligação de dados advindos de várias fontes distintas e heterogêneas. Neste trabalho, nós investigamos este desafio no contexto de bases biológicas com foco na caracterização de organismos vivos, especialmente seus fenótipos e doenças. Isto inclui a rica diversidade de Model Organisms Database (MODs) – repositórios especializados em um taxon particular – amplamente usados em estudos médicos e biológicos. Nós exploramos uma abordagem de integração leve, inspirada na iniciativa de Linked Open Data, mapeando várias bases biológicas em um banco de dados de grafos unificado – nosso BioGraph – e interligando elementos-chave para oferecer uma perspectiva interconectada sobre os dados. Apresentamos aqui experimentos práticos para validar a proposta e para demonstrar como o BioGraph pode contribuir para análises de dados biológicos em uma ótica de redes.

Abstract

Representing data as networks have been shown to be a powerful approach for data analysis in biodiversity, e.g., interactions among organisms; relations among genes and phenotypes etc. In this context, databases and repositories following a graph model (e.g., RDF) have been increasingly used to interconnect information and to support network-driven analysis. Usually, this kind of analysis requires gathering together and linking data from several distinct and heterogeneous sources. In this work, we investigate this challenge in the context of biological bases focusing on the characterization of living organisms, especially their phenotypes and diseases. It includes the rich diversity of Model Organism Databases (MODs) – repositories specialized in a particular taxon – widely used in the biological and medical studies. We exploit a lightweight integration approach, inspired in the Linked Open Data initiative, mapping several biological bases in a unified graph database – our BioGraph – and linking key elements to offer an interconnected view over the data. We present here practical experiments to validate the proposal and to show how BioGraph can contribute for biological data analysis in a network perspective.

List of Figures

2.1	Example of Mutation Gene. Source: Kohler [24]	18
2.2	Phenotype Description Approaches.	19
2.3	Structure of a ZP statement.	20
2.4	Example of a molecular function. Source: (http://amigo.geneontology.org/amigo).	22
2.5	Example of a biological process. Source: (http://amigo.geneontology.org/amigo).	23
2.6	Example of a cellular component. Source: (http://amigo.geneontology.org/amigo).	23
2.7	Graph of a Gene Ontology. Source: Marbach et al. [30].	24
2.8	Architecture InterMine. Source: Smith et al. [44]	24
2.9	Identifiers InterMine.	25
2.10	Mines from InterMine.	26
2.11	MouseMine: To obtain data part 1.	26
2.12	MouseMine: To obtain data part 2.	27
2.13	MouseMine: To obtain data part 3.	28
2.14	Intermine Web Services. Source: Kalderimis et al. [22]	28
2.15	Uberon. Source: Mungall et al. [32]	29
2.16	Uberon Classes. Source: (http://pt.slideshare.net/cmungall/uber-on-cl-workshophandel).	29
2.17	Uberon Composition. Source: (http://pt.slideshare.net/cmungall/uber-on-cl-workshophandel).	30
2.18	Uberpheno linking Mouse, Human and Zebrafish. Source: Kohler et al. [25]	30
2.19	Performance test between Neo4j and PostGree by InterMine. Source: (https://intermineorg.wordpress.com/)	31
3.1	(a) ZFIN partial database model of phenotypes. (b) MGI partial database model of phenotypes.	33
3.2	Unified Model.	33
3.3	Mapped Model.	34
3.4	Mapped Model to Graph.	35
4.1	Architecture of project.	37
4.2	Extract Transform Load.	37
4.3	Data downloaded from InterMine.	38
4.4	Uberon and Uberpheno.	39
4.5	Example of terms interlinked by Uberon and Uberpheno.	40
4.6	Example of the interlinks among organisms.	40
4.7	Example of the interlinks among organisms with disease and symptom.	41

4.8	Knowledge generation.	42
4.9	Example of the knowledge generation.	43
4.10	Domain Model.	44
5.1	Graph template for the query that returns organisms with the same phenotype.	46
5.2	Graph template for the query that returns all organisms with the phenotype "iris hypoplastic".	46
5.3	Correlating symptoms in BioGraph.	47
5.4	Example of the correlating symptoms in BioGraph.	47
5.5	PageRank among symptoms that shared disease.	48
5.6	PageRank among symptoms that shared entities.	49
5.7	A screenshot of Xper with zebrafish data.	50
5.8	BioGraph–Xper connection; adapted from Cavoto et. al. [11].	51
5.9	BioGraph to Xper mapping process.	51

Nomenclature

<i>API</i>	Application Programming Interface
<i>BED</i>	Browser Extensible Data
<i>CSV</i>	Comma-Separated Values
<i>DO</i>	Disease Ontology
<i>EQ</i>	Entity - Quality
<i>ETL</i>	Extract Transform Load
<i>GFF</i>	General Feature Format
<i>GO</i>	Gene Ontology
<i>HP</i>	Human Phenotype Ontology
<i>JSON</i>	Javascript Object Notation
<i>MGI</i>	Mouse Genomic Informatics
<i>MOD</i>	Model Organism Database
<i>MP</i>	Mammalian Phenotype Ontology
<i>OBO</i>	Open Biomedical Ontologies
<i>OMIM</i>	Online Mendelian Inheritance in Man
<i>OWL</i>	Web Ontology Language
<i>PATO</i>	Phenotype and Trait Ontology
<i>RDF</i>	Resource Description Framework
<i>SDD</i>	Structured Descriptive Data
<i>SO</i>	Sequence Ontology
<i>SQL</i>	Structured Query Language
<i>SYMP</i>	Symptom Ontology
<i>UML</i>	Unified Modeling Language
<i>Xenbase</i>	Xenopus Anatomical Ontology

<i>XML</i>	Extensible Markup Language
<i>ZFA</i>	Zebrafish Anatomy and Development Ontology
<i>ZFIN</i>	Zebrafish Information Network
<i>ZP</i>	Zebrafish Phenotype Ontology

Contents

1	Introduction	15
2	Foundations and Related Work	17
2.1	Genotype and Phenotype	17
2.2	Describing Phenotypes	18
2.3	Data Sources	19
2.3.1	MOD	19
2.3.2	Disease and Symptom	20
2.4	Interlinking Datasets	20
2.5	Gene Ontology	21
2.6	InterMine	23
2.6.1	Database	24
2.6.2	Web Application	25
2.6.3	Web Services	26
2.7	Uberon and Uberpheno	27
2.8	Graph Databases and Biological Graph Databases	28
3	Unified Model	32
3.1	UML Model	32
3.2	Graph Model	34
4	Building BioGraph	36
4.1	Architecture	36
4.2	ETL Process	36
4.3	Ingest and Linking	38
4.4	Inference	40
4.5	Statistics	42
5	Experimental Results	45
5.1	Searching	45
5.2	Analyzing	46
5.3	Describing	48
6	Conclusion	52
	Bibliography	53
A	XQuery Code	58
B	ZFA Terms in XML	59

C	ZP Terms in XML	61
D	PATO Terms in XML	63
E	Result in SDD	65

Chapter 1

Introduction

The development of computational methods to collect, analyze and store biological data brought unprecedented opportunities to cross data from different organisms. They can support analysis of: phenotypes, connections between diseases and symptoms, and interaction between distinct organisms who are important for research in the biological and medical area. However, there are two main challenges for this kind of analysis. First, data are stored in several distinct datasets, where each repository has its own representation, which is not interconnected with others. Second, it is not trivial to analyse this high amount of data.

This research is concerned with the context in which biologists and researchers work with phenotypic data - i.e., data describing characteristics of living beings - focusing in organisms. We are particularly interested in how we can take advantage of crossing information from several biological bases, which are independently produced, but contain interrelated and complementary information about living beings.

In this context, Washington et al. [50] integrated various biological datasets of different organisms, combining genotypes with their phenotypes. They created a homogeneous model for the source databases, manually discovering and defining association. The authors arrived in a result which confirmed their hypothesis: it is possible to identify ortholog genes¹ crossing phenotypical data in different organisms.

To reach this conclusion, Washington et al. [50], focused in a specific case, defining phenotype descriptions corresponding to the symptoms of diseases related to lack of vision. They used the following organisms: mouse, zebrafish, human, and drosophila. For each organism, it was defined a distinct set of symptoms. Starting from phenotypic symptoms – e.g., characteristics of a blind eye – the authors arrived in the ortholog genes causing blindness of various organisms.

Like Washington et al. [50], scientists often need to cross data from different organisms, resorting to several databases to conduct their research. However, each database contains its particular representation, hampering the data analysis when it involves distinct databases.

This research addresses this problem. It involves creating a database to support the search and analysis of the phenotypic data. Its main goal is to develop techniques to

¹Genes derived from a common ancestor that have the same function in different species or organisms.

transform the phenotypic data from heterogeneous and distinct data sources into a homogeneous format, linking them and crossing phenotype information of different organisms. Its specific goals are: (i) the development of a unified model to support several descriptive approaches for phenotype; (ii) techniques to connect and enrich data from several sources by inference; (iii) the implementation of a unified database based in the main goal.

We have built a unified graph database, that integrates several biological databases, Model Organism Databases (MODs) and ontologies related to phenotypes and diseases. It crosses information among organisms, supporting knowledge discovery and network analysis. MODs are specialized repositories of biological knowledge about model organisms [21]. The main challenges faced in this work is the heterogeneity of distinct data sources and the heterogeneity of descriptive approaches for phenotypes.

We are specially interested in phenotype descriptions and their relations through different organisms, diseases, and symptoms. We imported sets of data of phenotype descriptions from several scattered bases. Then, we connected these data, producing BioGraph, a biological graph database containing 588.237 nodes and 1.790.723 edges, where each node represent a term in an ontology or MOD. Altogether, we collected data from 63 distinct data sources.

BioGraph is the basis to discover new relations and enrich the graph. For example, we can relate characteristics shared by several distinct organisms and their respective properties and descriptions; we can trace which diseases are shared for the same symptoms, discovering interactions among symptoms. Furthermore, the graph provides the possibility of executing network analyses, like: finding recurrent descriptive subgraphs to detect new knowledge. We also created a generalization of phenotype descriptions, which are linked with descriptions of all organisms having the same phenotype. It makes possible the straight association of distinct organisms, enhancing the search.

The remaining of the text is organized as follows: Chapter 2 describes the foundations and related work; Chapter 3 presents the unified model proposed in this work; Chapter 4 details the process to build BioGraph; Chapter 5 presents the experiments over BioGraph and the respective results; Chapter 6 presents the conclusions and future work.

Chapter 2

Foundations and Related Work

This chapter presents foundations and related work about linking biological data. The work started looking the strategies for phenotype and genotype representation in an integration fashion. Section 2.1 describes definitions of genotype and phenotype. We further narrowed our focus to address only phenotypes. Section 2.2 presents several formats to describe phenotypes. Section 2.3 describes the main data sources used in this work. Section 2.4 presents the related work. Section 2.5 portrays the Gene Ontology and its strategy to represent genotypes and phenotypes. Section 2.6 report the InterMine system. Section 2.7 presents the Uberon and Uberpheno ontologies. Section 2.8 presents the graph databases.

2.1 Genotype and Phenotype

Genotype refers to the genetic makeup of the individual, i.e., its set of genes. Gene is the functional unit of heredity, it is a segment in the DNA. It is formed by proteins and nitrogenous bases. The nitrogenous bases are nucleic acids, that have the genetic information. Protein is a macromolecule consisting of small aminoacid molecules [28].

Phenotype is a combination of physical and behavioral characteristics of an individual, resulting from the interaction of their genotype with the environment influences [28]. In the medical context, a phenotype can be physical or biochemical characteristics of an organism, determined by a genotype and the environment. A phenotype can be a mutation from the normal morphology, physiology, behavior, or biochemical characteristics of an organism [38]. Phenotypes have been widely used in the research of interaction among organisms, as well as studies in the medical and biological area.

One important application which involves phenotypes is the study of genetic diseases. In this kind of disease occurs a mutation in a gene, affecting the respective phenotype. Considering the genes alone - without their respective phenotypes - the comparison among these genes is made through alignment algorithms. But in cases in which the gene is mutated, these algorithms can be useless, because they compare genes through similarity between the chains of genes. A mutation can hamper the matching. So, the phenotype can be key for comparison among organisms [38, 24].

Figure 2.1 shows a gene mutation of a C base to an A base, reflecting in a different protein modifying the gene function. Even though, the alignment algorithms cannot

correlate the two genes, it is possible to compare phenotypes of diseases with similar effects, e.g., changes that cause blindness.

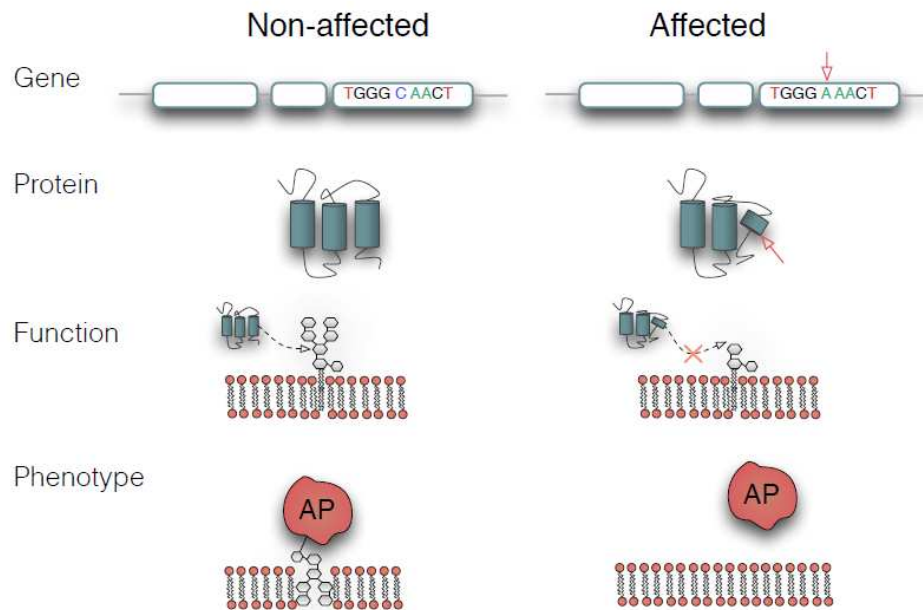


Figure 2.1: Example of Mutation Gene. Source: Kohler [24]

The comparison of organisms departing from their phenotypes has great potential in analyzing and finding correlations among organisms and provides an efficient way, to identify related candidate genes which cause the same disease in several organisms [50].

Another key concept in this context is the profile, which defines a focus of the relevant information to perform searches, analysis and analogy among organisms. In the context of diseases, for example, a profile can be composed of elements for the phenotype description of the disease and its associated genotype. The profile becomes the basic search unit, i.e., the comparison is made between profile searched – e.g., absent eye – and the ones retrieved from the database. Phenotypes may be associated with ontologies.

In the next sections, we present two essential topics for this work: an important base that describes genotypes related to phenotypes, the Gene Ontology, and works that interlink datasets.

2.2 Describing Phenotypes

There are several ways to describe a phenotype. Figure 2.2 shows main phenotypes description approaches. There are datasets that use free textual descriptions to represent a phenotype. This approach hampers the automatic interpretation and use by computational tools. This approach is adopted by OMIM, a disease database [19] and by the FishBase information system (<http://www.fishbase.org>).

Another approach is the C/CS (Character, Character State). It is a semi-structured approach that split the Character of the Character State, where Character represents what is described and Character State is a state of this Character. This approach is still

not the most appropriate, because a Character can contain inside a textual description. This approach is adopted by Xper, a system to edit, store and analyze phenotypes [48].

The most semantically rich approach is the Entity-Quality (EQ) format [27], wherein the entity is a morphological or anatomical structure of organisms, specified through ontologies, and the quality is a property that describes the entity, usually specified in the *Phenotype and Trait Ontology* (PATO) [16]. This approach enables to execute computational analyses and to search more accurately. ZFIN (Zebrafish Information Network) [46] uses the Entity-Quality (EQ) format. It adopts ZFA (Zebrafish Anatomy and Development Ontology) for the entity representation and PATO (Phenotypic Quality Ontology) for the quality representation, as shown in Figure 2.2. Following the Köhler et al. [25] classification, we call this approach a pos-composed EQ.

Some datasets use a variant of the EQ format, where the entity and quality are combined in a single description and it is not possible to distinguish the entity from the quality. We call this variant a pre-composed EQ [25] and it is used by the Mammalian Phenotype Ontology (MP) [43], an ontology for phenotype description adopted by MGI (Mouse Genomic Informatics) [10].

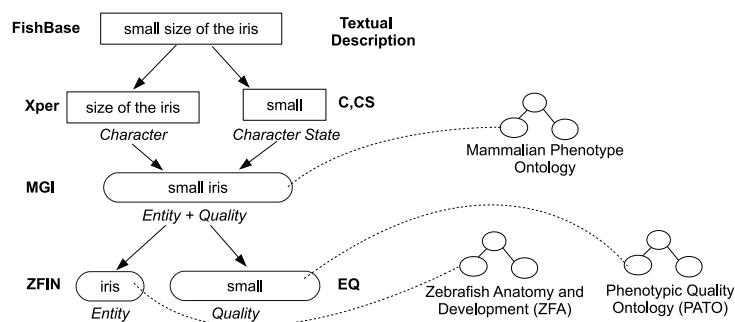


Figure 2.2: Phenotype Description Approaches.

2.3 Data Sources

This section summarizes the main MODs and databases used in this work.

2.3.1 MOD

Model Organism Databases (MOD) are specialized repositories of biological knowledge, whose definition is not strictly established [21]. In the last decades the term model organism referred to a small and select group of species, deeply studied in the laboratory and richly documented [21]. As the mechanisms for genetic mapping have become more affordable, the concept of model organism expanded to a wider range of species [21]. We consider that each MOD stores data of a model organism and may contain data from its genotype and phenotype, providing biological knowledge to conduct research, in domains like genetics, development, and evolution.

The main MODs used in this work are MGI, MP, ZFIN, ZP, and HP. MGI - Mouse Genomic Informatics - is a MOD of mice that contains their genotype and phenotype

data [10]. MGI uses the ontology MP - Mammalian Phenotype Ontology - to describe phenotype, as described in Section 2.2. HP - Human Phenotype Ontology - contains data of the human phenotype [39], describes like MGI in Figure 2.2. Both MP and HP adopt the pre-composed EQ format [25], i.e., they join Entity+Quality in a single atomic concept.

ZFIN - Zebrafish Information Network - is a MOD containing data from Zebrafish [46]. It uses ZFA - Zebrafish Anatomy and Development Ontology - an ontology of the zebrafish anatomical structure and PATO - Phenotypic Quality Ontology.

ZP - Zebrafish Phenotype Ontology - is an ontology for phenotype descriptions. It is post-composed [25], linking a ZFA (entity) term and a PATO (quality) term in a statement containing an entity-quality sentences. Figure 2.3 shows how a ZP statement is built for the statement "abnormal decreased area eye".

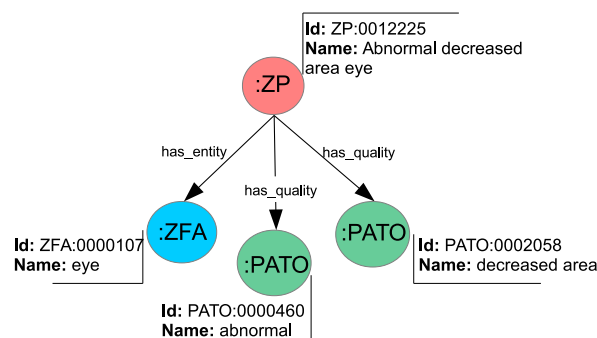


Figure 2.3: Structure of a ZP statement.

2.3.2 Disease and Symptom

The main diseases bases used in this research are OMIM and DO. OMIM - Online Mendelian Inheritance in Man - is a knowledge base of human genes and genetic disorders [19]. OMIM was created in 1966 and is maintained by the Johns Hopkins University. OMIM uses textual description to represent phenotypes. In this context, a phenotype is a symptom. It has 4,746 phenotype descriptions and 82 terms of genes and phenotypes linked. Altogether, OMIM has 23,565 terms and it was updated on June 27th 2016.

DO - Disease Ontology - is a database with data of inherited, developmental and acquired human diseases. The web DO system uses a graph database to store the ontology. DO links diseases with OMIM. It has 1,594 terms from OMIM [40]. DO is maintained by a collaboration of researchers at Northwestern University (Center for Genetic Medicine) and the University of Maryland (School of Medicine, Institute for Genome Sciences). It was updated on June 1st 2016. SYMP - Symptom Ontology - is an ontology with symptoms.

2.4 Interlinking Datasets

Whashington et al. [50] used several MODs to integrate genotypes with their phenotypes and to discover orthologous genes ¹ that mutated in different species. They report an

¹Genes derived from a common ancestral that has the same function in different species.

example of mutated genes that resulted in blindness in different organisms. For this study, they generated a unified model from various heterogeneous MODs containing genes to be considered in the comparison. 11 human genes were chosen which have orthologous genes in mice, zebrafish, and drosophila from the OMIM database. Mouse, zebrafish, and fruitflies genes were obtained from different bases.

Whashington et al. [50] achieved the following conclusions: (i) variant alleles contain more similar phenotypes than other alleles of the same gene; (ii) it is possible to retrieve mutant genes, responsible for phenotypic anomalies through the similarity analysis of the respective phenotypes; (iii) it is possible to identify orthologous genes from phenotypic data that cross different species. These results would not be achieved only looking to the genotypes, due to two major problems: (1) the genetic basis of most diseases are often unknown; (2) although the genetic basis is known, algorithms of sequence alignment are inadequate, as this comparison is based on the similarity of genes along the chains. For this reason, they propose a comparison through the phenotypes, in their case, the disease symptoms.

Whashington et al. [50] faced two major difficulties: (1) they had to manually create a homogeneous model and integrate various MODs focusing in the analyzed profiles; (2) to create a profile encompassing several ontologies, selecting the relevant terms for the search. Similarly, many researchers face the same difficulties, of integrating MODs and setting profiles manually, as there is no computational tool to build a unified model from several different MODs and to support profiles associated with ontologies.

Phenoscape is an ontology-driven database that integrates data from mouse, human, zebrafish, and frog, adopting Uberon for entity terms and PATO for quality terms. Phenoscape has data of genes and phenotypes. The main goal the Phenoscape is to adopt semantic similarity algorithms, e.g., the parsimony algorithm, to discover phenotypic variations among species. It can match similar phenotypes to find related genes in different species [29].

2.5 Gene Ontology

Gene Ontology is an ontology that stores data from genotypes and descriptive related data, it was used in this work like data source. Guarino et al. [18] use three complementary definitions for ontology: (i) Gruber [17] defines ontology "as an explicit specification of a conceptualization"; (ii) Borst [8] includes the concept of sharing, defining an ontology as a "formal specification of a conceptualization shared"; (iii) Studer et al. [47] unified the two definitions: "an ontology is an explicit specification, formal of a conceptualization shared."

The Gene Ontology arose from the need of having consistent descriptions of genes, when they appear in different databases, i.e., the association of different terms that have the same meaning. Ashburner et al. [3] define Gene Ontology (GO) as a structure that contains a controlled vocabulary and defines known genes and proteins. It can support the study of biological information associated with these genes, using structures that record information and assist in data analysis. For example, the Gene Ontology

helps in the study of genetics, where there are works attempting to identify functional relationships between genes and consequent results. The Gene Ontology is composed of three independent ontologies, divided by areas: molecular function, biological process, and cellular component [3]. Every publicly known gene has records representing the biological characteristics within those areas.

- **Molecular function:** describes activities at the molecular level driven by gene. The Figure 2.4 illustrates an example of the molecular function **Cytokine Activity** in the Gene Ontology. This function is part of a taxonomic structure of molecular functions, shown in Figure 2.4(right), indicating the a type of connection as receiver **Receptor Binding**, which in turn is a type of Protein Binding. It is also related with the Receptor-mediated Virion Attachment to Host Cell. This example was taken from the Amigo tool, that is part of Gene Ontology.

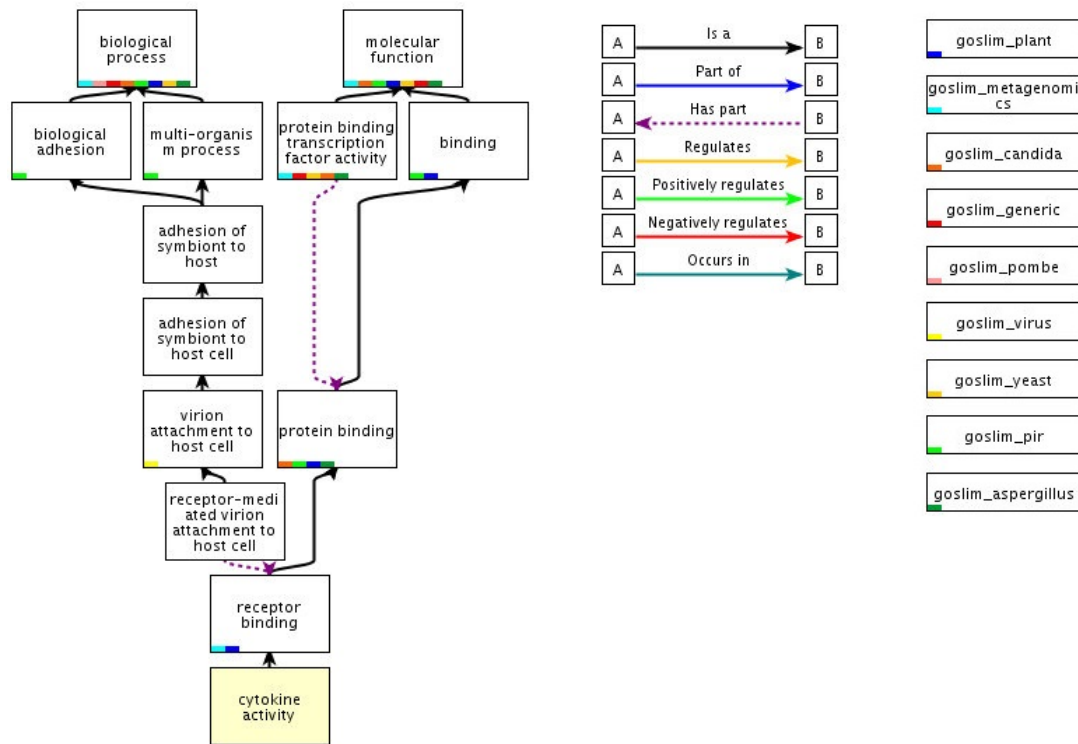


Figure 2.4: Example of a molecular function. Source: (<http://amigo.geneontology.org/amigo>).

- **Biological process:** a series of molecular functions; it defines in which biological processes a gene is involved. Figure 2.5 shows an example of the biological process **Cell Differentiation** in Gene Ontology. It is arranged in a taxonomy of biological processes. **Cell Differentiation** refers to a biological process of cell growth cellular, which is a development process and so on. This example was taken from the Amigo tool.
- **Cellular component:** establishes the location of a gene within the cell. Figure 2.6 shows an example of the cell component membrane, which is a kind of cellular component. This example was taken from the Amigo tool.

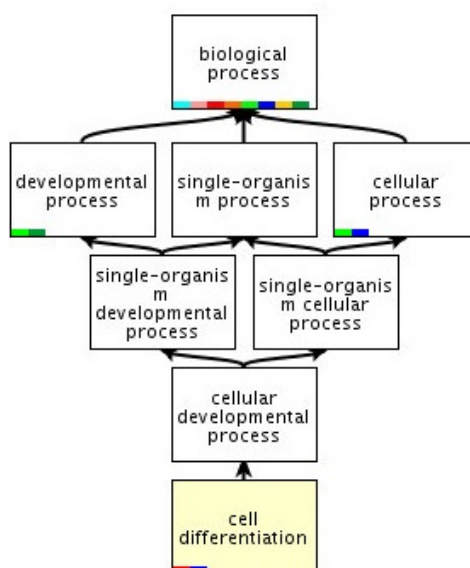


Figure 2.5: Example of a biological process. Source: (<http://amigo.geneontology.org/amigo>).

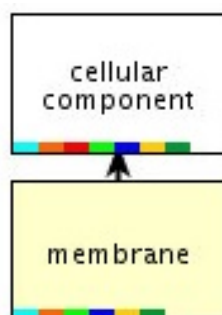


Figure 2.6: Example of a cellular component. Source: (<http://amigo.geneontology.org/amigo>).

According to Bard and Rhee [4], the structure of the Gene Ontology is based on an acyclic graph. As illustrated in Figure 2.7, each term of the Gene Ontology is represented as a node and relationships between terms are the edges between nodes.

One of the main motivations to use the Gene Ontology is due to its a vast network on the web, available to all researchers anywhere in the world. Researchers can put information from their research in the network and acquire data from studies of other people. Gene Ontology can be linked to related research data, even though they are in different databases. Rat Genome Database (<http://rgd.mcw.edu/>) and VCmap (<http://www.animalgenome.org/VCmap/>) interlink their descriptive structures to Gene Ontology, allowing to associate additional data of different sources.

2.6 InterMine

InterMine is a data warehouse that integrates diverse biological databases to support data analysis [36]. We used InterMine in this work to obtain data from mouse, zebrafish

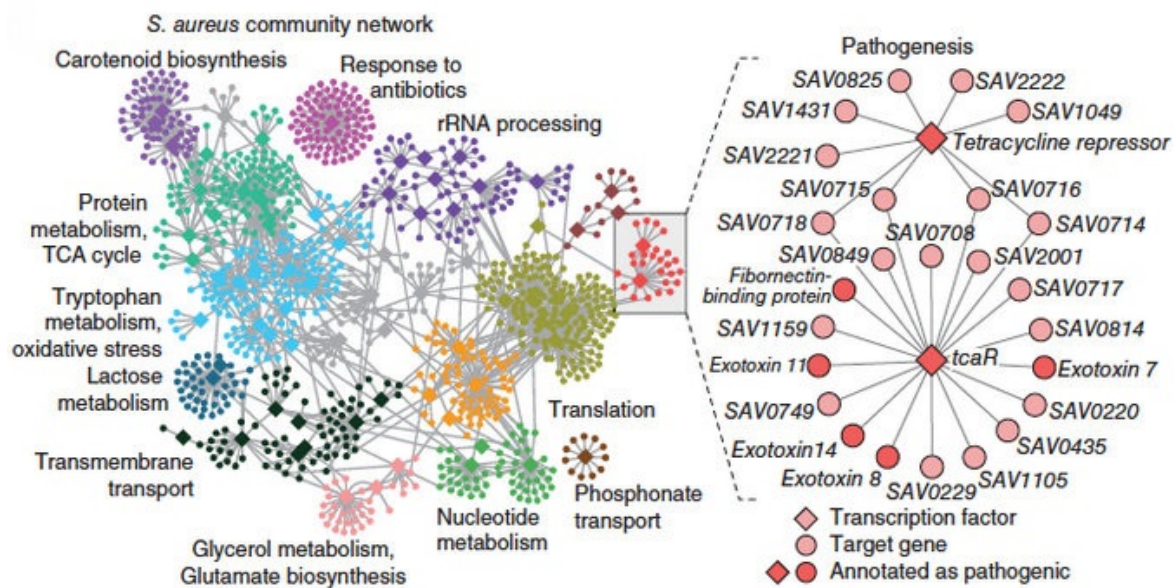


Figure 2.7: Graph of a Gene Ontology. Source: Marbach et al. [30].

and human. It was developed by the Micklem lab at the University of Cambridge, since 2002. Initially InterMine was called FlyMine. Figure 2.8 represents the architecture of InterMine, that has three parts further described.

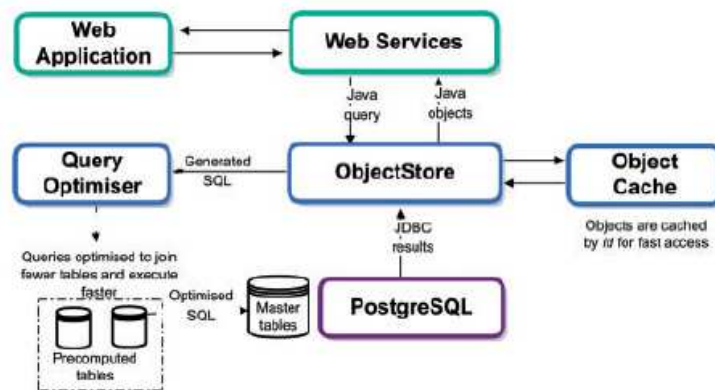


Figure 2.8: Architecture InterMine. Source: Smith et al. [44]

2.6.1 Database

In InterMine, data from each database to be integrated are loaded and stored in a local relational database, represented by the PostgreSQL box in Figure 2.8. It has a core data model based on the Sequence Ontology [15, 44].

To address the outdated data problem, InterMine uses the concept of identifiers, wherein each ontology has an open unique identifier, that is used as a reference in others ontologies. By convention, this identifier is formed by a namespace and a unique number inside the ontology. For example, one identifier of the ZFA ontology is ZFA:0000001, where the namespace is ZFA, followed by a unique number inside the ZFA ontology. This

unique identifier is represented by the fields ID in the respective ontologies in Figure 2.9. InterMine has a mechanism to replace outdated data by current ones, consistently updating even identifiers, when the data are loaded again from the source databases. Furthermore, the same datatype may conflict in different bases. In this case, InterMine decides which one is the most relevant, based on a priority score defined for each data source, giving precedence to the more reliable [44].

This kind of identifier also enables to link resources distributed in several bases, as they have been adopted by several ontologies and knowledge bases. The identifiers inside InterMine are equivalent to those adopted by the referred bases.

To address performance problems, InterMine has a module called **Query Optimiser**, which reuses results of previous requests whenever it is possible. Moreover, the query results are stored in a smart cache system represented by the **Object Cache** box in Figure 2.8. It can get results of a previously executed query or part of it to improve the response time [44].

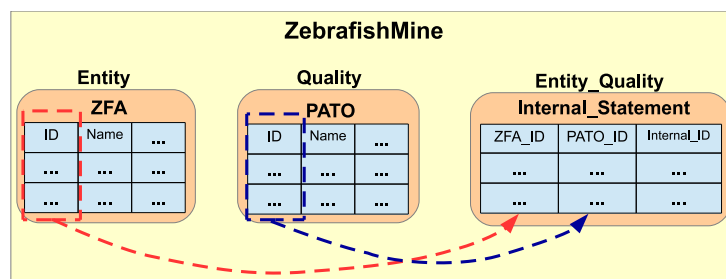


Figure 2.9: Identifiers InterMine.

2.6.2 Web Application

InterMine offers a standard interface to wrap existing MODs, to homogenize the way in which users access data. Each MOD has its own InterMine data warehouse, they are MouseMine, ZebrafishMine, RatMine, YeastMine, and Wormbase. Figure 2.10 shows the three MODs from InterMine used in this work. The column (a) shows that each database has a distinct model; Column (b) shows the original source MODs: MGI, ZFIN and HP. Column (c) represents the ETL (Extract Transform Load) process conducted by InterMine to transform the data in a standard format. Column (d) is the InterMine data warehouse of each MOD. This process corresponds the Web Application box in Figure 2.8 [36].

We further present an example to access data from the MouseMine web application (www.mousemine.org/mousemine). Through the interface will be possible to grasp the InterMine potential of homogenizing interfaces. Figure 2.11 shows the **QueryBuilder** tab, where it is possible to select the data type of the wanted data. In Figure 2.11 we are selecting the MP Term.

The following tab is illustrated in Figure 2.12. On the left side in the **Model browser** the user chooses the fields. On the right side in **Query Overview**, the system shows the fields chosen. In the example, they are: Description, Identifier and Name. The result is shown in Figure 2.13. On the Export button it is possible to download the returned data

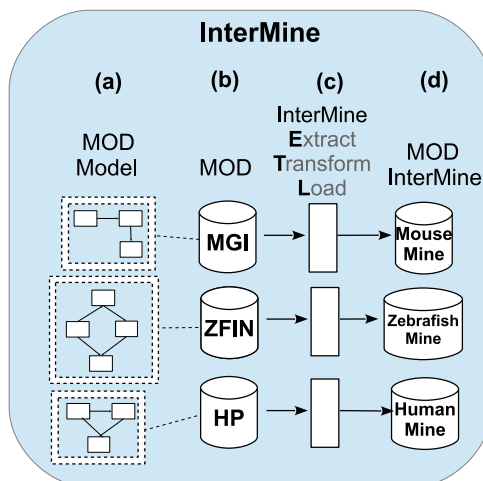


Figure 2.10: Mines from InterMine.

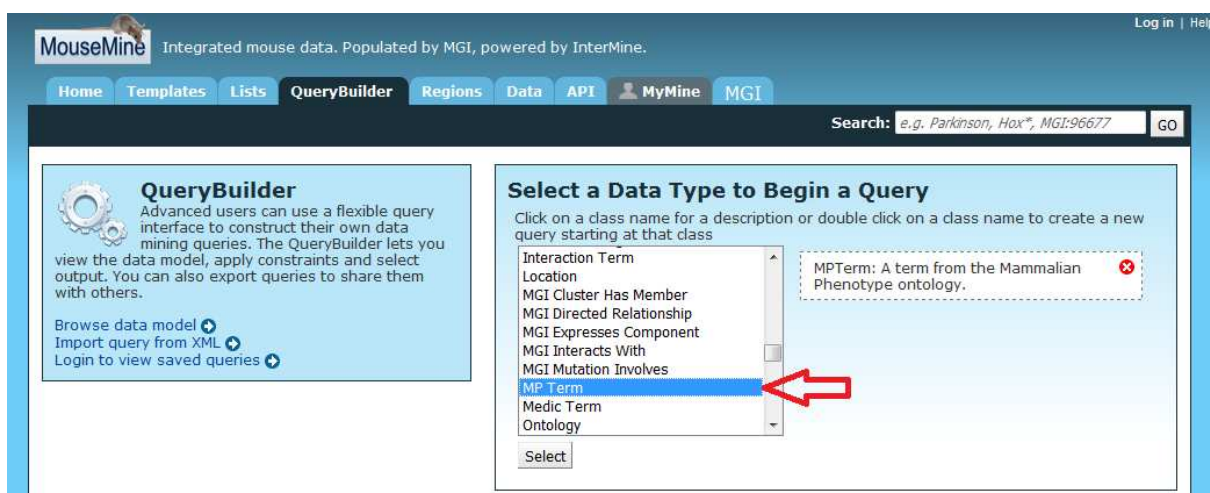


Figure 2.11: MouseMine: To obtain data part 1.

in CSV, TSV and other formats. Users can choose as input their own list of identifiers and queries can be saved in their 'MyMine' space [44].

2.6.3 Web Services

Complementary to the web application, the search described in Subsection 2.6.2 is available via web services API to be used by computational tools, see Figure 2.14. Querying the data via web services is similar to search via web application, where the data are returned in following formats: JSON, XML, CSV, TSV, GFF, BED and FASTA.

InterMine offers support for API, in the following languages: Python, Perl, Java, Ruby, JavaScript and XML [22]. Next, we show a code in Python for get the same data described.

```
1 from intermine.webservice import Service
2 service = Service("http://www.mousemine.org/mousemine/service")
3 query = service.new_query("MPTerm")
4 query.add_view("identifier", "name", "description")
```

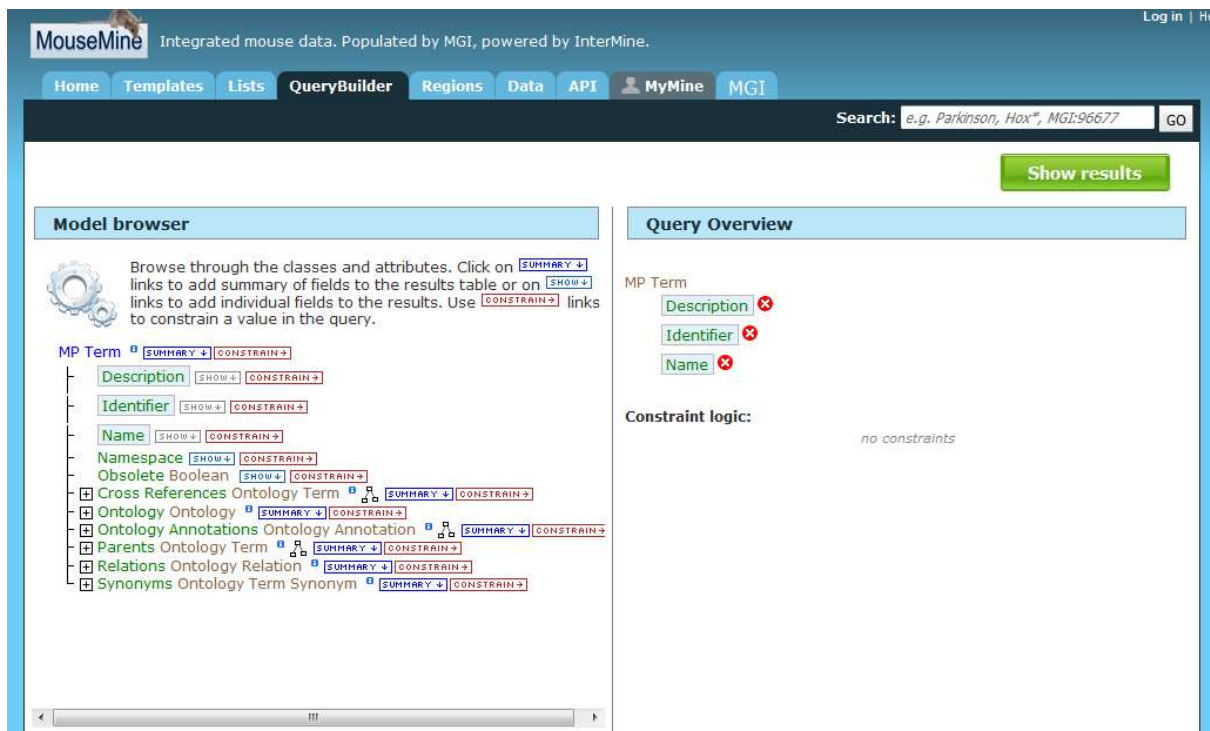


Figure 2.12: MouseMine: To obtain data part 2.

```

5
6 for row in query.rows():
7     print row["identifier"], row["name"], row["description"]

```

2.7 Uberon and Uberpheno

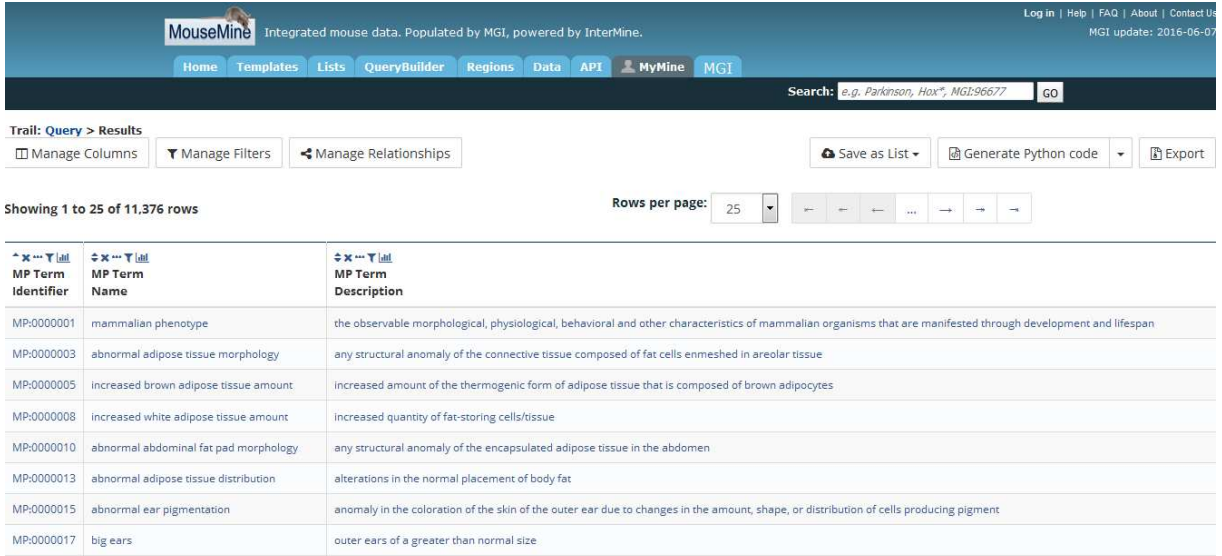
Uberon

Uberon (Uber-anatomy ontology) is an ontology with integrates entity terms of several anatomic ontologies defining anatomical structures of different organisms. It enables crossing data among organisms, see Figure 2.15 [32].

Uberon links the ontologies creating classes to generalize terms of several ontologies. The concepts in the integrated ontologies are specializations of these generic classes or their equivalents.

Uberon works in an integration chain progressively integrating modules, i.e., it starts integrating a more specialized groups of ontologies in modules – e.g., mammals – and progressively merge them in modules that reach broder groups – e.g., vertebrates. When overlapping classes exist in Uberon modules, a merge is made avoiding a new class, i.e., Uberon only creates a new class when it is necessary to generalize the merged ontology classes. Uberon uses axiom based relationships among the classes to integrate the ontologies. The Figure 2.16 synthesizes the process of creating generic classes from various ontologies.

This approach present some disadvantages: the resulting ontology is very large, and highly latticed, making difficult the navigation [32]. Actually, Uberon has 8.000 classes



MouseMine Integrated mouse data. Populated by MGI, powered by InterMine. Log in | Help | FAQ | About | Contact Us
MGI update: 2016-06-07

Home Templates Lists QueryBuilder Regions Data API MyMine MGI

Search: GO

Trail: Query > Results

☐ Manage Columns

Showing 1 to 25 of 11,376 rows Rows per page: 25

MP Term Identifier	MP Term Name	MP Term Description
MP:0000001	mammalian phenotype	the observable morphological, physiological, behavioral and other characteristics of mammalian organisms that are manifested through development and lifespan
MP:0000003	abnormal adipose tissue morphology	any structural anomaly of the connective tissue composed of fat cells enmeshed in areolar tissue
MP:0000005	increased brown adipose tissue amount	Increased amount of the thermogenic form of adipose tissue that is composed of brown adipocytes
MP:0000008	increased white adipose tissue amount	Increased quantity of fat-storing cells/tissue
MP:0000010	abnormal abdominal fat pad morphology	any structural anomaly of the encapsulated adipose tissue in the abdomen
MP:0000013	abnormal adipose tissue distribution	alterations in the normal placement of body fat
MP:0000015	abnormal ear pigmentation	anomaly in the coloration of the skin of the outer ear due to changes in the amount, shape, or distribution of cells producing pigment
MP:0000017	big ears	outer ears of a greater than normal size

Figure 2.13: MouseMine: To obtain data part 3.

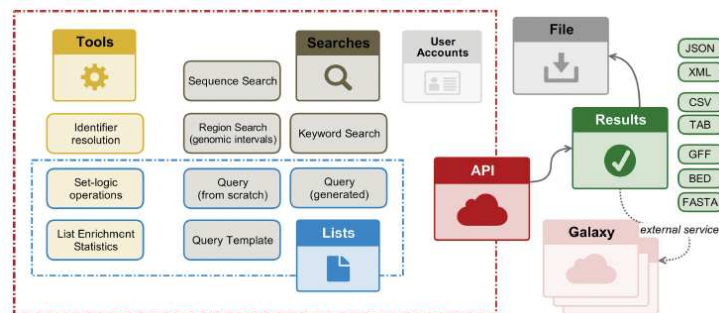


Figure 2.14: InterMine Web Services. Source: Kalderimis et al. [22]

and 13.629 relationships of type Xref (Uberon with external ontologies) to 4.087 integrated classes. Figure 2.17 presents the participation of the ontologies in the Uberon composition. Uberon has a web interface where users can download the data in OBO and OWL format.

Uberpheno

Uberpheno is a cross-species ontology, i.e., it focuses on bridging elements of existing ontologies to produce a single integrated result. It links phenotype descriptions from MP (mouse) to HP (human) and ZP (zebrafish) in HP (human).

Uberpheno does not have a web interface, but it's possible to download data in OWL and OBO formats [25]. The Figure 2.18 shows a MP (mouse) term linked in an HP (human) term and a ZP (zebrafish) term linked in an HP term.

2.8 Graph Databases and Biological Graph Databases

Graph databases natively store data as graphs. Graphs are mathematical models, consisting containing a triple of a non-empty set of vertices or nodes, a set of edges (relations)

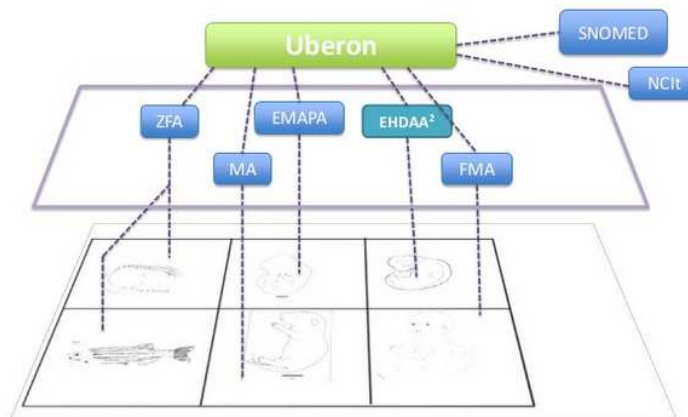
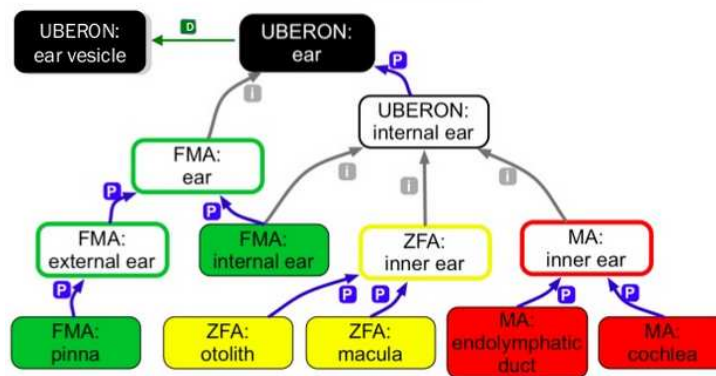


Figure 2.15: Uberon. Source: Mungall et al. [32]

Figure 2.16: Uberon Classes. Source: (<http://pt.slideshare.net/cmungall/uber-on-cl-workshophaendel>).

and a function that associates each edge with a pair of vertices [7]. Figure 2.3 shows a directed graph [2], i.e., the edges have a direction as indicated by the arrow. Over this basic definition, graph models can vary according to the intended application and the database management system.

There are several approaches to model the data in a graph database. We have adopted in this work the property graph, where it is possible to create properties in nodes and/or relationships [1]. In Figure 2.3 the nodes have the properties *Id* and *Name* to describe their terms. The model adopted here also defines a special property named *Label*, which is used to classify nodes and edges. In the Figure, labels are prefixed by colors. The nodes are labeled according to their original ontologies (:ZP, :ZFA, :PATO) and the edges according to their roles (*has_entity*, *has_quality*).

Studies between relational and graph databases have shown that graph databases can be better in some cases, like bioinformatics for example. It has many fields that can exploit the graph representation, including, metabolic networks, protein-protein interaction networks, chemical structure graphs, gene clusters, genetic maps and genotype-phenotype interaction [20, 49].

According to Vicknair et al. [49], graphs are the most useful structure for modeling interactions, like: protein, genes, organisms, among others.

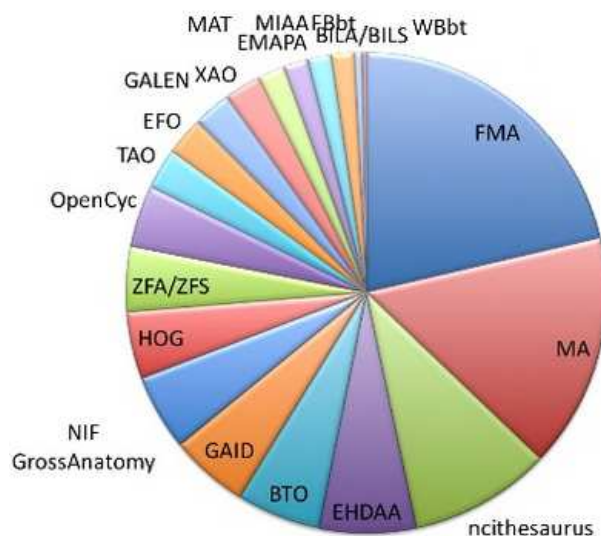


Figure 2.17: Uberon Composition. Source: (<http://pt.slideshare.net/cmungall/uber-on-cl-workshopaendel>).

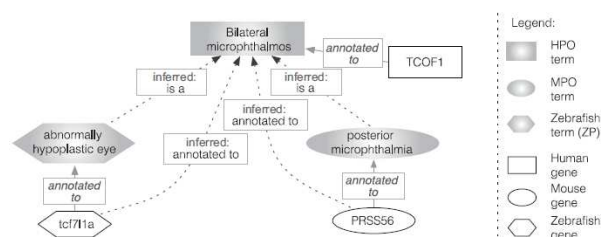


Figure 2.18: Uberpheno linking Mouse, Human and Zebrafish. Source: Kohler et. al. [25]

With graphs is simpler traversing long paths, while in a relational database, this query can be time inefficient due to the number of joins.

In this work we adopted graph databases due to its natural vocation to represent relations. Linking things is the main task of this project.

Neo4j

Neo4j is a graph database that implements a property graph data structure. It was chosen based on a comparison of different models of graph databases. It has its own query language – Cypher [2, 14]. Neo4j also provides graph algorithms, e.g., to determine the shortest path between two nodes.

The InterMine system has initiated studies with Neo4j, as an alternative to the current relational database, to handle complex biological data and relations. They conducted performance tests with queries in Neo4j and Postgree and in almost all the tested queries the Neo4j showed better response time – see Figure 2.19.

Bio4j

Bio4j is a bioinformatics graph data platform, that uses Neo4j. It contains protein data integrated from UniprotKB (a database of protein), Gene Ontology (GO), UniRef (data of

psql (SQL)	Neo4j endpoint (Cypher)	Notes
1200 ms	5 ms	Return all properties
1400 ms	1400 ms	Return all properties order by primary identifier
360 ms	12 ms	Return primary identifier and symbol
85 ms	5 ms	Return genes count

Figure 2.19: Performance test between Neo4j and PostGree by InterMine. Source: (<https://intermineorg.wordpress.com/>)

clusters of UniProtKB sequence), NCBI Taxonomy², and Expasy Enzyme DB (enzyme³ database) [35].

²NCBI taxonomy database contains names and classifications for the organisms present at the GenBank database

³Enzymes are the kind of proteins responsible for catalyzing chemical reaction

Chapter 3

Unified Model

In this chapter, we present our unified model which is a main contribution of this research. It is represented in Unified Modeling Language – UML. Among several challenges, we faced the heterogeneity of approaches to describe phenotypes.

3.1 UML Model

In this research, we approached the pre-composed and post-composed EQ models for phenotype description. To solve the heterogeneity problem here, we developed a unified model to support pre-composed and post-composed EQ models and their alignment. To generate a unified model we analyzed two MODs widely used and cited in related work: ZFIN and MGI. The analysis was based in the study of the database schemes, which are published by ZFIN and MGI. Its results are further detailed.

As described in the previous section, ZFIN is a MOD that contains both data genotype as phenotypes of the zebrafish, wherein the phenotypes are described by post-composed EQ model [46, 50]. The partial model of phenotype description from ZFIN is presented in Figure 3.1. A phenotype description is formed by a statements (**Phenotype_statement**) involving an entity (**ZFA_term**) and a quality (**PATO_term**) from external ontologies: ZFA (Zebrafish Anatomy Ontology) or GO (**Gene Ontology**) and PATO. Entities and qualities are generalized as terms (**term**) that have a self-relationship type (e.g., is-part-of), as one can build a taxonomy of terms.

MGI is a MOD with genotype and phenotype data of the mouse [6]. Figure 3.1(b) shows a partial model of phenotype descriptions from MGI. As ZFIN, the phenotype descriptions are treated as a set of statements set. Each statement corresponds to a term in MGI (**voc_term**). Each term is associated with the MP ontology (*Mammalian Phenotype Ontology*). As mentioned in the previous section, they are using a pre-composed EQ model, as each ontology concept is already an indivisible composition of entity plus quality [43]. The class **voc_vocab** corresponds to the class **ontology** of ZFIN and enables the use of terms of several ontologies.

Figure 3.2 presents our unified model, wherein a phenotype (**Phenotype**) is composed by a set of statements (**Statement**). Each statement generically defines an EQ, without discern an entity and quality. Therefore, it corresponds to a pre-composed EQ, like the

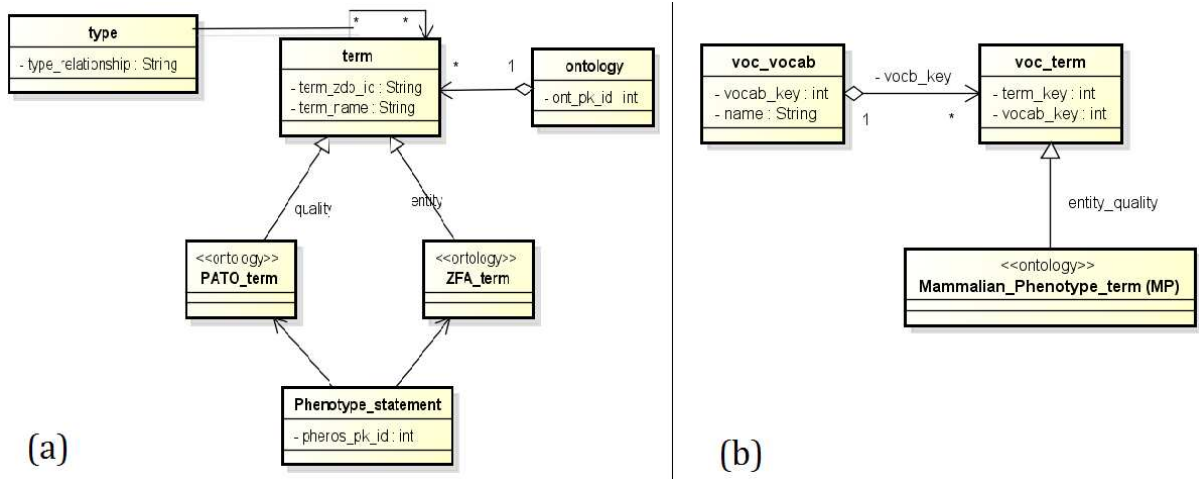


Figure 3.1: (a) ZFIN partial database model of phenotypes. (b) MGI partial database model of phenotypes.

voc_term of MGI. The class **Statement_EQ** specializes the **Statement** to represent the post-composed EQ, in which entity and quality are discriminated, as the ZFIN **term** class. The **voc_vocab** class from MGI and the **ontology** class from ZFIN match our **Ontology** class. Furthermore, there is a self-relationship in the **Statement** class whose type is defined by the **Association Type** class – an Association class. The classes **Entity** and **Quality** have also a self-relationship to record synonyms.

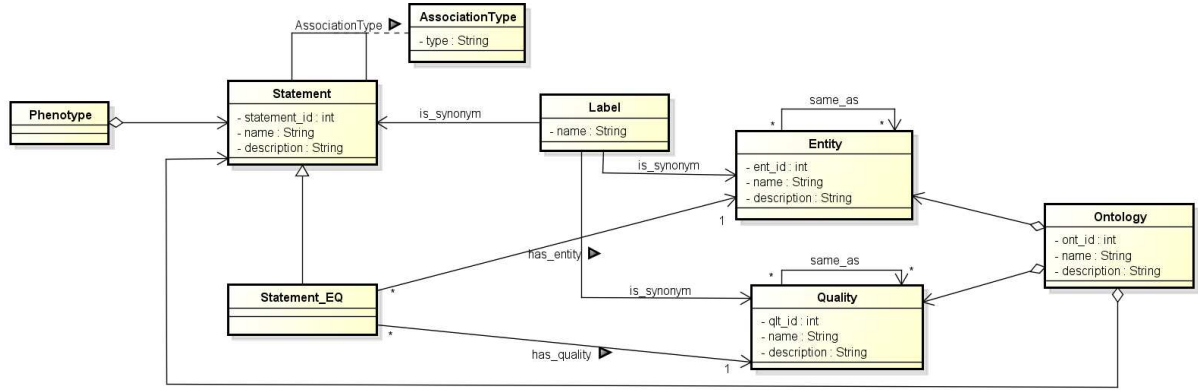


Figure 3.2: Unified Model.

The main MODs used in this work are MP from MGI (mouse), HP (human) and ZP (zebrafish). As shown in the previous section, HP has pre-composed EQ terms like MP.

Figure 3.3 shows the main classes of Figure 3.2 plus the classes of external ontologies – which we integrate – as specializations of our core model. The type attribute defines the relationship type. The classes with orange color represent classes of external ontologies, i.e., HP and MP. The type attribute defines the relationship type. The classes with orange color represent classes of external ontologies. MP and HP are terms specializations of **Statement**. ZP is a specialization of **Statement_EQ**. PATO is a specialization of **Quality**. Uberon and Uberpheno are auxiliary ontologies used to connect several bases. Uberon contains entity terms and generalizes entities from other ontologies. Therefore, Uberon is an **Entity**, ZFA and XAO are ontologies with entity terms from zebrafish and frogs

respectively. ZFA and XAO are related to a Uberon terms. Uberpheno is an association class, since it links statements without creating a new term.

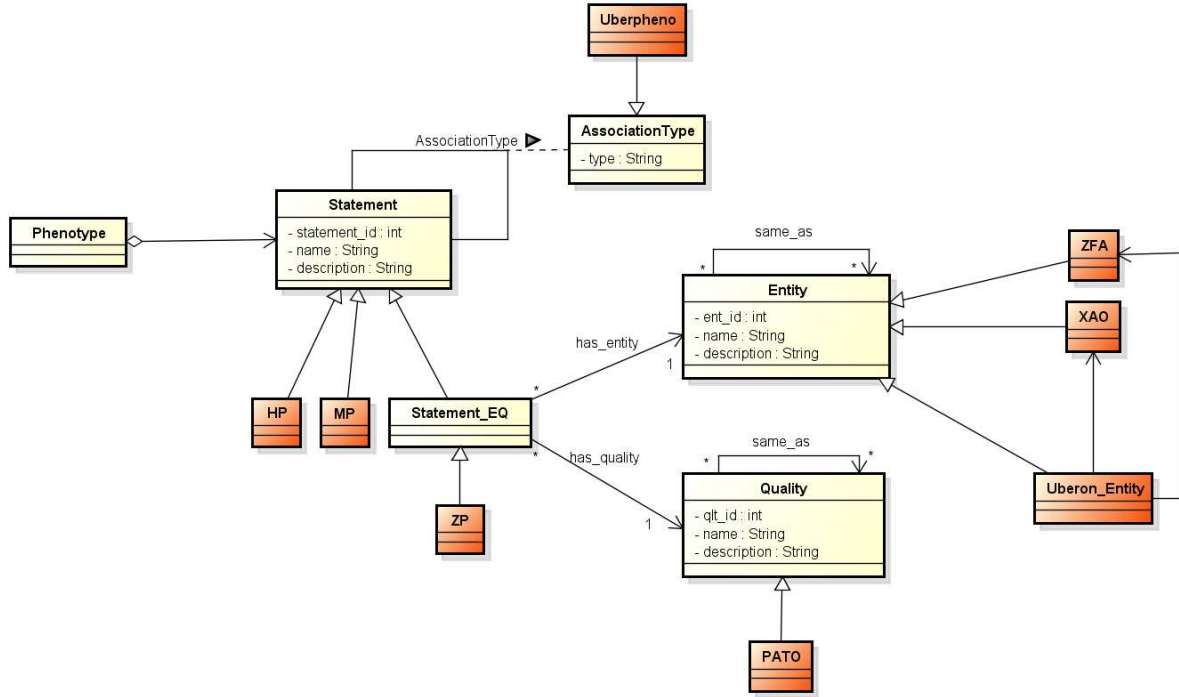


Figure 3.3: Mapped Model.

3.2 Graph Model

Our unified model is based on a graph structure, so we will map the model to a property graph model [37]. We mapped our model of Figures 3.2 and 3.3 to a graph model in Figure 3.4. We defined the following mapping rules, to produce our UML model:

- A class in the UML model (Figure 3.3) becomes label of the node in the graph model (Figure 3.4). This labels will be replicated in each node which is an instance of the respective classes. The classes are Statement, Statement_EQ, Entity, and Quality.
- A class instance in the relational model becomes a node in the graph model.
- The inheritance generates multiple labels, e.g., MP is subclass of Statement, which results in two labels for the same node Statement and MP (see Figure 3.4). The same occurs with HP, generating a node with the Statement and HP labels. In our model, the Statement label will always have another label associated with it.
- Each attribute of these classes becomes a node property in the graph.
- Each relation between two labels becomes an edge in the graph, connecting the node containing the label of the origin class to the node containing the destination class.
- Each relation type among classes becomes a label of the edge. For example, the `has_entity` and `has_quality` relation types became node labels.

- Each association class instance define a label in the respective edge. For example, the `equivalent_to` label in the edge is an instance of `AssociationType`. We indicate through a dashed contour that this edge applies for the all `:Statement` nodes.

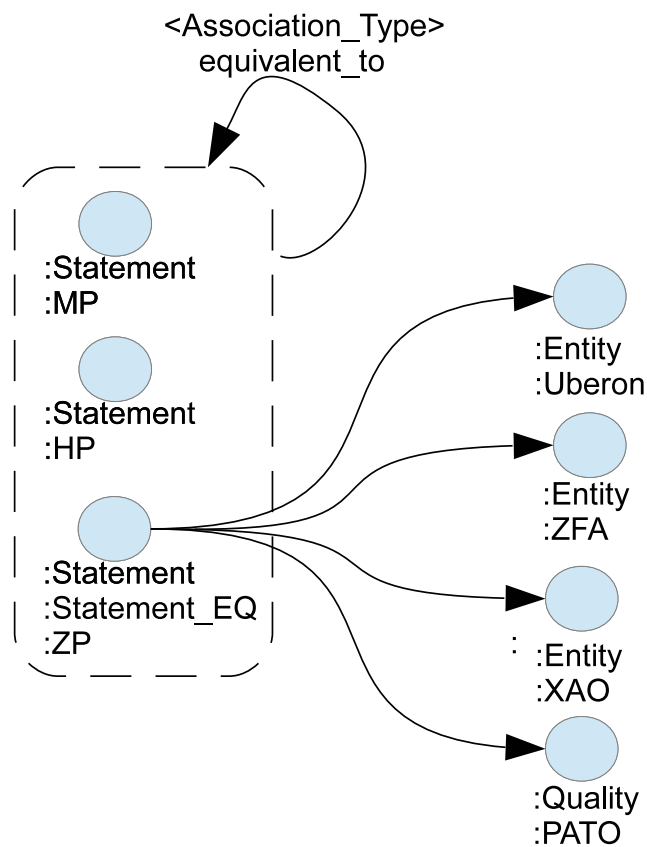


Figure 3.4: Mapped Model to Graph.

Chapter 4

Building BioGraph

In this research, we are interested in phenotype data. These kind of data is an access key for integration of biological data. Through phenotypes, it is possible to discover orthologous genes in different organisms. It helps the research in the medical human area, like, to study new drugs and diseases. Phenotype data also allow linking anomalous behaviors, e.g., the same disease in distinct organisms. In addition, we can integrate different datasets of the same organism, for example, it is possible to link diseases with symptoms which in turn is linked with body parts reaching the genotype from the phenotype – in this case, reading the genotype departing from the diseases.

This section details the process that we designed and implemented to integrate several biological knowledge bases, mainly focusing in phenotypes and model organism databases (MODs). As we have introduced in this section, we are also interested in integrating organisms and diseases. The section starts presenting our architecture, including the applications of our project, which are detailed in the next section.

4.1 Architecture

Figure 4.1 represents the architecture of our project. In a nutshell, our architecture is organized in four parts. First, we have the data sources used in this research, like: intermine, MODs, Uberon, Uberpheno and other ontologies. Second, we show the steps to ingest, link, and build BioGraph in a graph database. Third, we present the resulting unified database containing data from all these data sources. It is produced according to the unified model presented in the previous section. Fourth, we have the data access, where the user can search and analyze data in BioGraph.

In the following subsections, we detail the three steps inside the Ingest and Enrich process (see Figure 4.1), which involves the parts one to three of the architecture. In the next section, we detail the fourth part.

4.2 ETL Process

As shown in Figure 4.1 there are several formats of data sources and their heterogeneity is a challenge. Each MOD and dataset used to build BioGraph have its specific for-

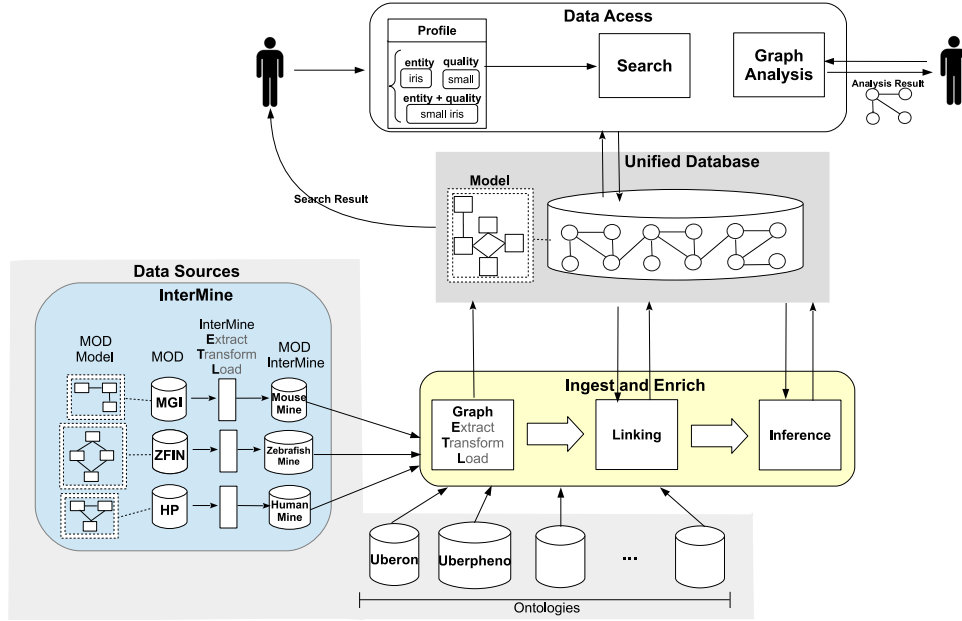


Figure 4.1: Architecture of project.

mat. Thus, we obtained data in various and distinct formats. It is illustrated in the box Graph Extract Transform Load of Figure 4.1. This process is portrayed in Figure 4.2. We downloaded data from data sources in several formats, like: JSON, OBO, OWL, RDF and XML. Extensible Markup Language (XML) [9] and Javascript Object Notation (JSON) [12] are formats to describe documents as hierarchies of elements. Resource Description Framework (RDF) [23] and Web Ontology Language (OWL) [31] and Open Biomedical Ontologies (OBO) [42] are designed to describe knowledge bases and ontologies. RDF and OWL have a graph model and OBO can be handled as a graph. Comma Separated Values (CSV) is a simple tabular format, which can represent relational tables.

We have converted all the files to an internal homogeneous format. This process is represented as the ETL (Extract Transform Load) box in Figure 4.2. We further load the data in the Neo4j graph database, where we are building BioGraph.

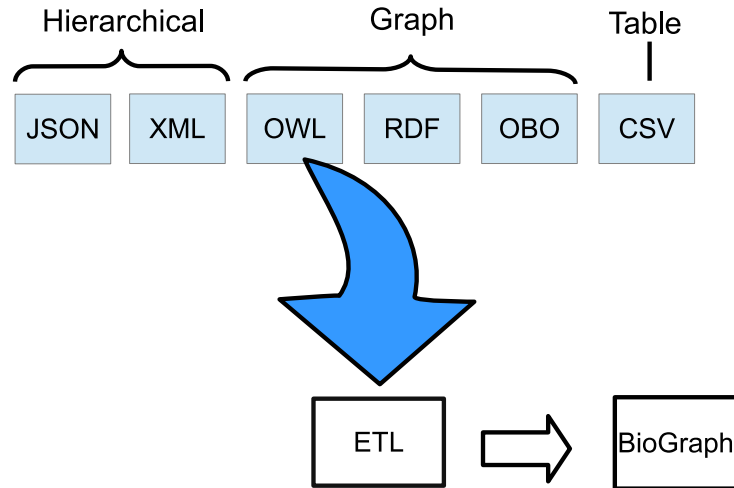


Figure 4.2: Extract Transform Load.

We have obtained data from several sources, among them, from the InterMine system in CSV format. As described in the previous section, InterMine generates a standard web interface for each MOD, represented as *InterMine Extract Transform Load* in Figure 4.1. It creates an InterMine data warehouse for each MOD. In this work, they are: MouseMine, ZebrafishMine, HumanMine. We used InterMine as a way to solve part of the problem of files with various formats, see Figure 4.1. Using InterMine, we obtained data from MGI, ZFIN and HP in a standard CSV format.

4.3 Ingest and Linking

This section presents how we have ingested data in our graph database and how we further linked them.

Through the standard way of identifying using namespaces, as described in the previous section, it was possible to cross data among organisms. This unique identifier is represented by the fields ID in Figure 4.3, which shows the files obtained from InterMine according to the MOD. In ZebrafishMine, we got data from ZFA that contains anatomic entities of zebrafish and PATO that contains qualities. Post-composed phenotype descriptions, connecting entities and qualities (statements), adopt an internal identifier designed for local affairs and not to be linked with external resources. We addressed this limitation aligning them with the ZP ontology, which also defines ZFA and PATO post-composed EQs, but defining open public identifiers, based on namespaces.

From MouseMine, we got data from MGI, which are related to MP pre-composed phenotype EQs. The same occurs with HumanMine, which is related to the HP pre-composed phenotype EQs. This process is represented in Figure 4.1 in the InterMine frame.

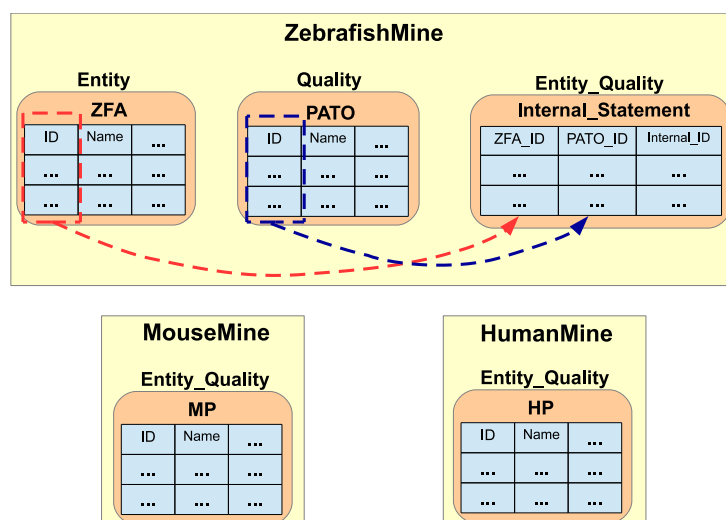


Figure 4.3: Data downloaded from InterMine.

After loading the data in a graph, we used Uberon and Uberpheno to link the description, represented by the linking box in Figure 4.3. Uberon is an ontology integrated with entities terms from several ontologies.

Uberpheno to link phenotype descriptions of mouse (MP) to the human (HP) and phenotype descriptions of zebrafish (ZP) to the human (HP). We have linked the unique identifiers. Uberpheno publishes the data in OBO format [25]. ZP post-composed EQs have been linked with the MP and HP pre-composed EQs. Figure 4.4 shows the result in the Entity-Quality cloud. It is important to emphasize that ZP is the only one which is able to distinguish the entity (ZFA) from the quality (PATO).

Uberpheno does not define its own entities, it just links pairs of terms, whereas Uberon creates generic entities to generalize existing entities from distinct ontologies, as show in Figure 4.4. For each generalized entity, Uberon defines a unique open public identifier, which is linked with the external ontologies. For example, the identifier "UBERON:0001769", of the Uberon entity "iris", is linked with the term "ZFA:0001238", of the zebrafish entity "iris". Uberon links only anatomic entities. The general schema of connections is represented in Figure 4.4, in the Entity cloud, and its respective instance example in Figure 4.5. Uberon is the red entity.

Besides ZFA, Uberon also links other anatomical ontologies like Xenopus Anatomy and Development Ontology (XAO) and Mouse Adult Gross Anatomy Ontology (MA), which we also ingested and linked in our graph. When we interconnect all these elements in a unique graph, as shown in Figure 4.4, it becomes possible to infer new relations described in the next subsection. The Figure 4.5 shows how the entity iris from Uberon is linked with the same entity of other ontologies in Entity cloud. The Entity_Quality cloud shows how Uberpheno link statements. We have combined fragmented interlinking strategies like Uberon and Uberpheno in our unified BioGraph.

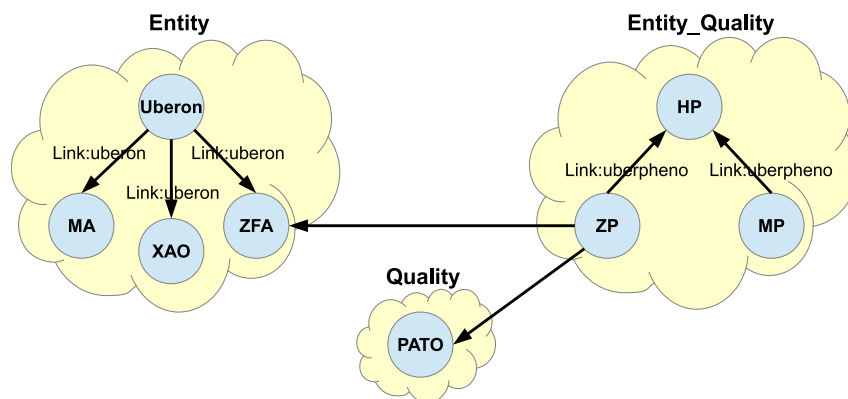


Figure 4.4: Uberon and Uberpheno.

The Figure 4.6 shows an example of how these three organisms became interlinked. The label of the node (prefixed by a colon) represents their ontology, e.g., :ZP identifies that this term refers a zebrafish term. Each node has the **Id** and **Name** of the term as properties. A search started in the statement **abnormal cornea** in zebrafish (ZP), can arrive in cornea of Uberon that is linked with astigmatism in human (HP) and mouse (MP).

Several research projects are using phenotype data to study the interaction between phenotype, organisms and diseases [50, 33, 5, 13, 45, 26, 41].

The National Research Council [33] raises important aspects of having a biological knowledge network with the interaction among several areas like: diseases, phenotypes,

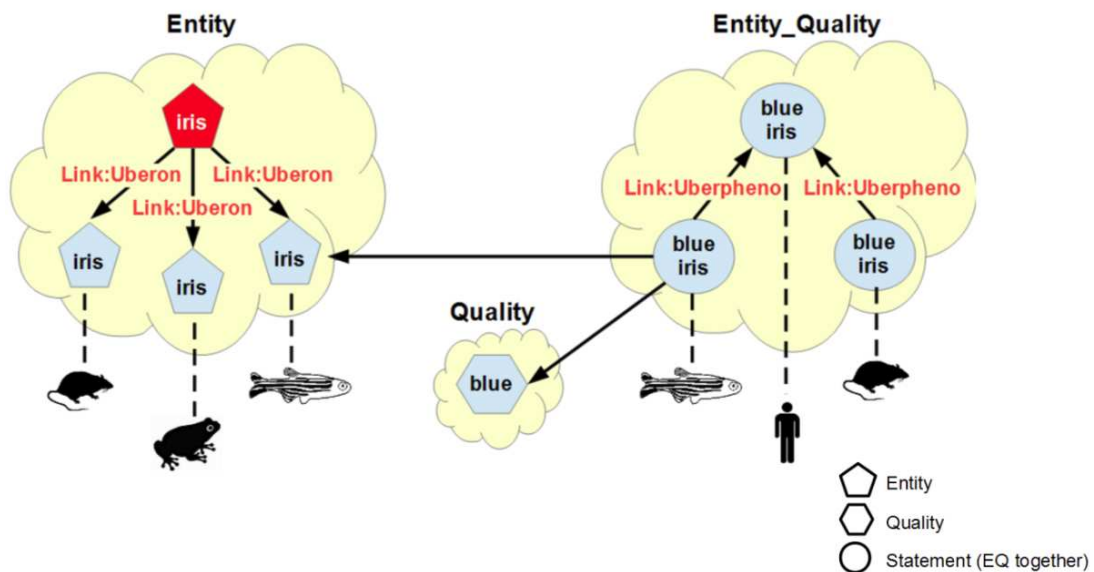


Figure 4.5: Example of terms interlinked by Uberon and Uberpheno.

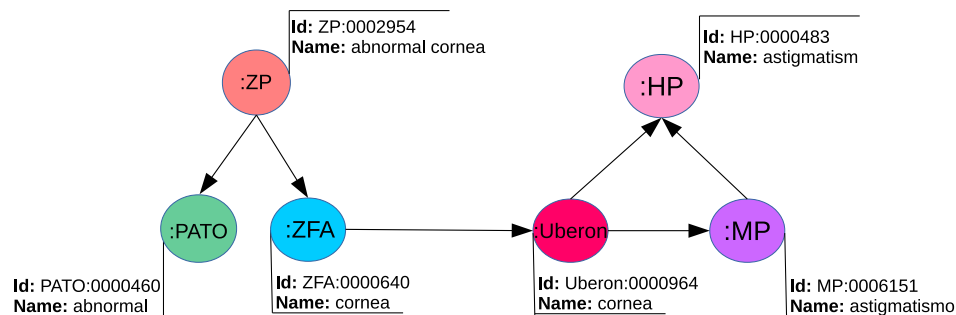


Figure 4.6: Example of the interlinks among organisms.

environment, behavior, drugs, among others to make precise diagnostics and to classify the variations of each disease. It provides the answers like: which type of treatment and prevention is most suitable for the type of detected disease. PhenoDigm is a database that provides the connection between model organisms and human diseases based on the phenotype description. It finds gene candidates for human genetic diseases through the phenotype [41].

BioGraph also comprises diseases and the respective symptoms. We imported and linked data from the Human Disease Ontology (DO) and the Symptom Ontology (SYMP). As shown in Figure 4.7, diseases (labeled as DOID) are linked to the human phenotype (HP) and symptoms are linked to the diseases.

4.4 Inference

With the interlinked graph we can infer new edges and nodes, generating knowledge. While Uberon generalizes only anatomical entities, this work goes beyond exploiting the existing links to infer new links and to produce generic post-composed EQs, which we call `Generic_EQs`.

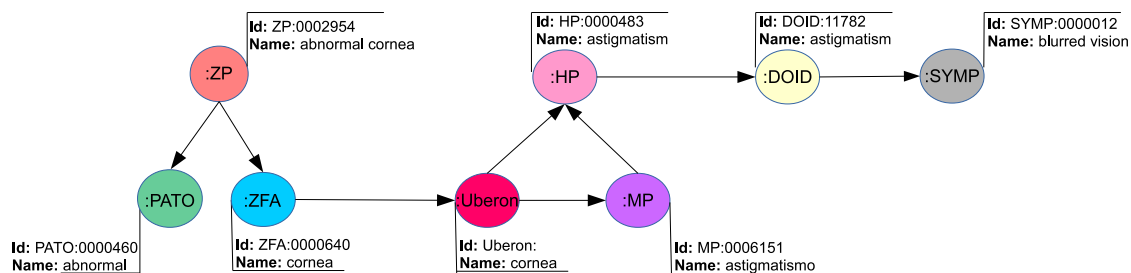


Figure 4.7: Example of the interlinks among organisms with disease and symptom.

Generic_EQs are generic statements able to cross phenotypes of organisms. The Generic_EQ have been produced through inferences, as presented in Figure 4.8. In the first column, we present the statements with relations in their original format as extracted from the sources and presented in the previous subsection. In the second column, we present the inferred Generic_EQ derived from the respective original form. In the first case, we have a ZP term statement linking with PATO and ZFA, which in turn is associated with Uberon by a relation of equivalence. We have created a Generic_EQ that makes a join of the quality (PATO) and the Entity (Uberon) originally linked with ZP, producing a statement, that generalizes it. Furthermore, it points for the ZP term as one of its specializations. This Generic_EQ is showed in the second column in Figure 4.8. The same procedure is applied to other MODs like MP and HP as illustrated in Figure 4.8.

In the rightest column of Figure 4.8, is presented a General Vision frame of the Generic_EQ connected with all the organisms, who shares the same statement, and their entities and qualities. This graph put together those of the Inference box in Figure 4.1. As can be seen in the figure, whenever a generalized statement points to the same entity and quality it will refer to the same **Generic_EQ**, i.e., our algorithm always checks if there is an existing **Generic_EQ** pointing to a given entity and quality (to be reused) before creating a new one. Figure 4.9 presents an example of the process, in which there is a Generic_EQ term, whose label is "blue iris", for example, and this term is linked with the same statement "blue iris" in the ontologies of human, mouse and zebrafish.

These inferences can enhance analyses and searches in the graph. Since both PATO and Uberon, linked by a **Generic_EQ** statement, are independent of a given organism, it is possible to formulate statements (specifying a link among entities and qualities) independently of a specific organism. Furthermore, with the Generic_EQ, a search can return all organisms which are linked by the same statement in a profile. For example, in a search for "blue iris", BioGraph returns all organisms that have this phenotype. Compared to our work, Uberon generalizes only entities, Uberpheno links existing statements without a generalization. BioGraph produces a generic statement that takes advantages of the Uberpheno links, but also joins, in the same network, generic qualities (PATO) and generic entities (Uberon), to produce a descriptive system, which is independent from specific organisms. In addition, it is possible to access and download (in CSV or JSON) data from BioGraph, with all the integrated data and the inferences to use in other works.

Figure 4.10 shows an overview of BioGraph and how it is organized; it contains: descriptions of phenotypes; Uberon entities, terms of gene ontology; diseases; and symptoms.

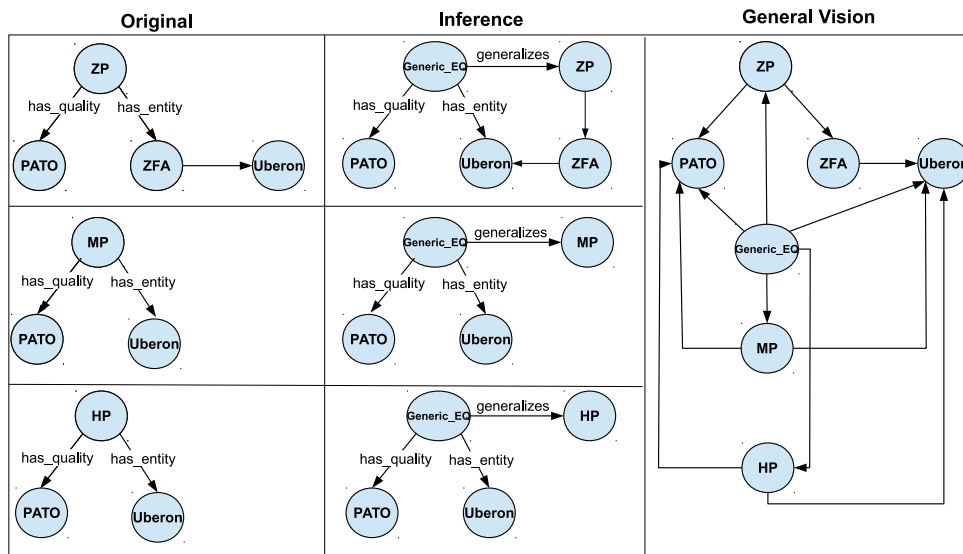


Figure 4.8: Knowledge generation.

Edges with the "link:uberon" and "link:uberpheno" labels indicate that these edges are derived from Uberon and Uberpheno respectively. In red, we highlight nodes and edges which we created by our inference process. Besides the links and inferences that are preciously detailed, we also ingest links of Uberon with diseases of DO (Human Disease Ontology) and SYMP (Symptom Ontology). From those links, we related Uberon straight with symptoms.

BioGraph is available at <http://pandora.lis.ic.unicamp.br:7474/>. Currently, it contains 588,237 nodes from 63 distinct databases, with 1,790,723 relations of different types, being a complex network that interlinks a big volume of data. In the next section we show how the graph produced here can be exploited for searching and to support data analysis based on the produced network.

4.5 Statistics

This section presents the statistics of the results obtained in the interconnection provided by BioGraph. It has the purpose of showing the interconnection potential achieved by our work, i.e., how many phenotype statements we are able to link departing from a given kind of organism.

For the sake of simplicity, we related only the more expressive bases. The three first rows contain the number of phenotype statements of each kind of organism, extracted from the original bases. BioGraph has 21,192 statements of human, 17,091 statements of mouse and 19,872 statements of fish. The following rows show the number and the percentage of the interlinked statements departing from a specific kind of organism. We were able to link 4,161 human statements with mouse ones (19% of the total), 5,982 mouse statements with human ones (35% of the total) and so on. Part of these links have been made by the inference process, detailed in Section 4.4.

As can be seen, the interlinked organisms are directionally computed and the number

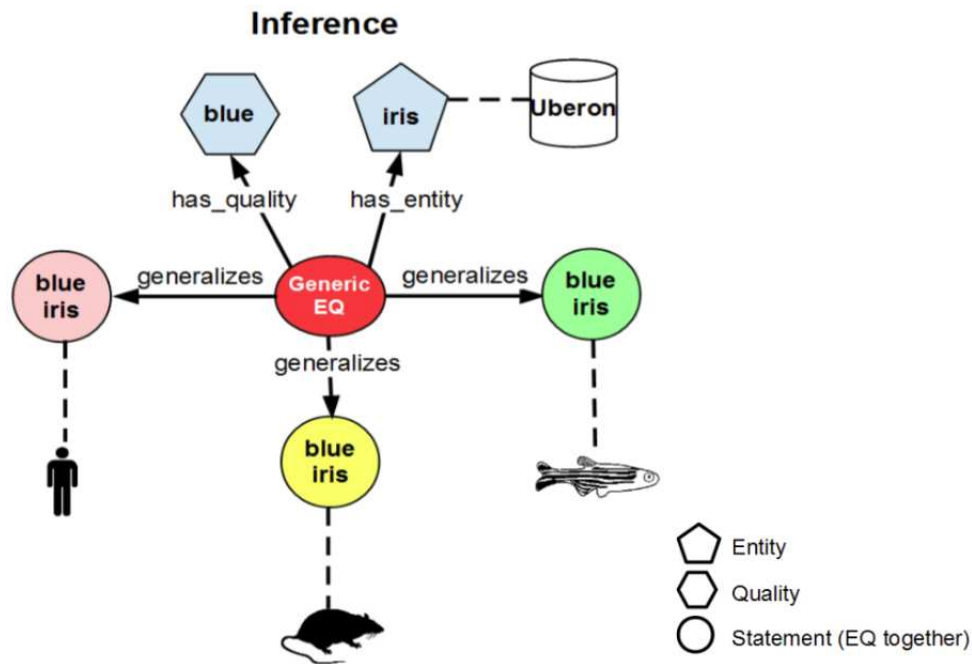


Figure 4.9: Example of the knowledge generation.

of linked statements are different in each direction due to the fact that one statement can be linked with several statements of the other organism and vice-versa.

Table 4.1: Statistics of interconnection among organisms.

Source	Amount Statements	Percentage (%)
Human only	21,192	
Mouse only	17,091	
Fish only	19,872	
Human -> Mouse	4,161	19%
Mouse -> Human	5,982	35%
Human -> Fish	885	4%
Fish -> Human	5,044	25%
Mouse -> Fish	1,041	6%
Fish -> Mouse	2,500	12%

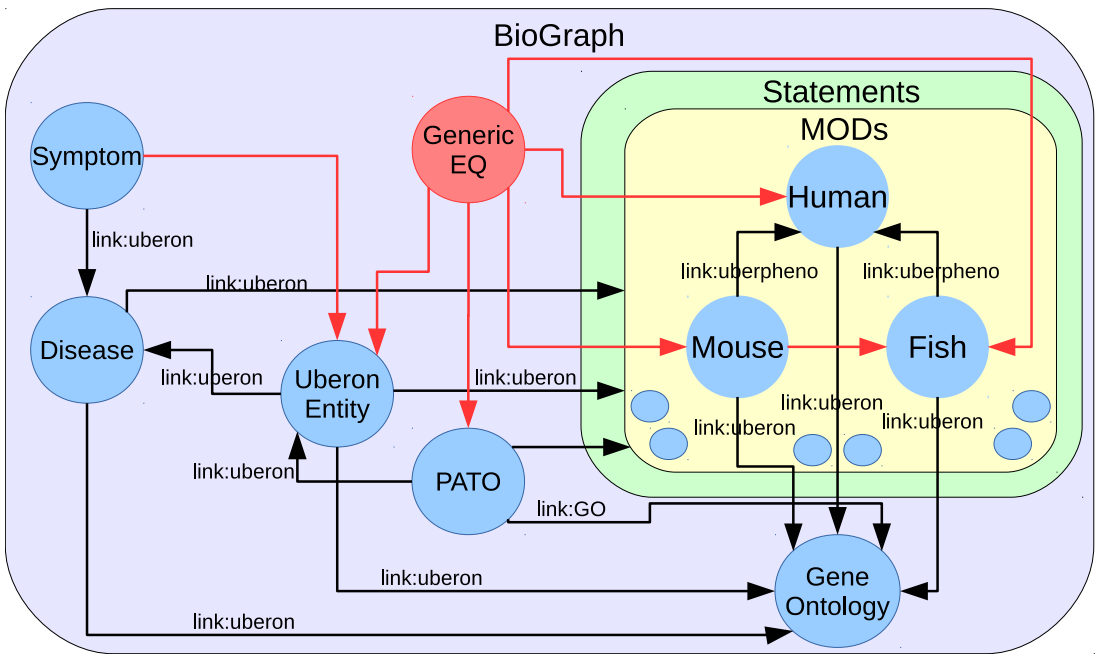


Figure 4.10: Domain Model.

Chapter 5

Experimental Results

This chapter presents some practical applications of our BioGraph in contexts like searching (Section 5.1), analyzing (Section 5.2) and describing (Section 5.3). Each section shows experimental results of querying examples or preliminary prototypes aimed at illustrating and evaluating the potential of our network.

5.1 Searching

This section illustrates the querying potential provided by interconnection of our network. BioGraph enables searches across organisms in single queries – an operation that would require integrating and interconnecting several sources without BioGraph. We further show two questions that can be asked in BioGraph and the respective queries answering them:

1. Which phenotype statements (a composition of Entity and Quality) are shared by the human, mouse, and fish?

Figure 5.1 shows the graph template to answer this query. The requested statements are specializations of the `Generic_EQ`. The figure also shows an instance returned by the respective query in Cypher:

```
1 match (z:ZP) -- (g:Generic_EQ) -- (h:HP)
2 match (m:MP) -- (g)
3 return z,g,h,m
```

The Cypher sentence presented here has two parts: `MATCH` (a sentence with a template that would match with some subgraph) and `RETURN` (the elements of the sentence to be returned). The `MATCH` reflects the graph of Figure 5.1, with: nodes represented between parenthesis; edges represented by two dashes `--`; labels preceded by colons; and variables before the colons. In this case, `z`, `g`, `h`, `m` are variables that receive instances of nodes which match the template.

2. Which organisms have the phenotype "iris hypoplastic" or equivalents?

Figure 5.2 shows the graph template to answer this query and an instance returned by the respective query.

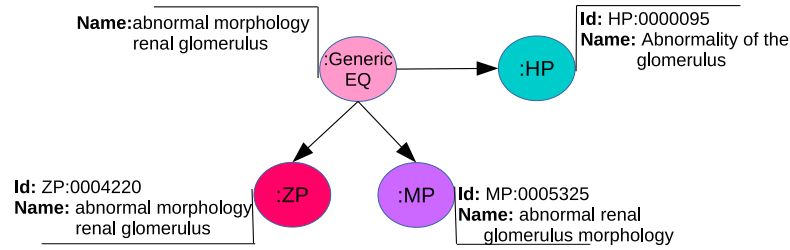


Figure 5.1: Graph template for the query that returns organisms with the same phenotype.

```

1 match (g:Generic_EQ) - [] - (o) where g.Name="iris hypoplastic"
2 return g,o

```

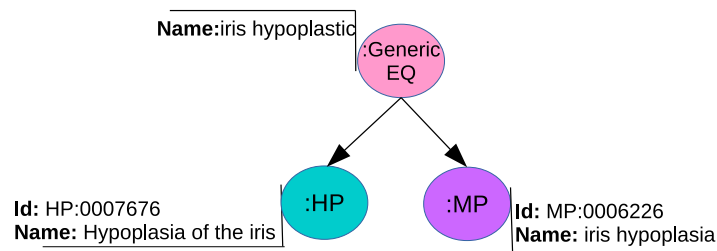


Figure 5.2: Graph template for the query that returns all organisms with the phenotype "iris hypoplastic".

Besides the presented questions/queries there are several other possibilities of questions, as the three following examples:

3. Which disease manifests itself in a determined part of the body (entity)? For example, which diseases attack the heart?
4. Which diseases have a given symptom?
5. Which symptoms belongs one specific disease?

5.2 Analyzing

The interconnections produced in BioGraph can be the basis to discover latent knowledge, based on the analysis of correlations in the network. We show this potential here through an experiment to discover relations among symptoms of diseases. The network enables us to correlate symptoms occurring in the same disease or in the same entity as follows:

1. **Correlating symptoms through diseases:** Figure 5.3 shows in the left side a fragment of our original BioGraph. Departing from one disease (DOID – Human Disease Ontology) shared by two symptoms (SYMP – Symptom Ontology), we have created a new edge between the symptoms. Since the same two symptoms can share several diseases, the edge has a property that indicates the number of diseases shared by them.

2. **Correlating symptoms through entities:** The second part of Figure 5.3 (down) shows an entity (Uberon) shared by two diseases having different symptoms. We have created an edge linking the symptoms transitively related by the entity, with a property that indicates the number of entities shared by these symptoms.

Figure 5.4 shows an example with terms of ontologies of Figure 5.3 and the respective inferences.

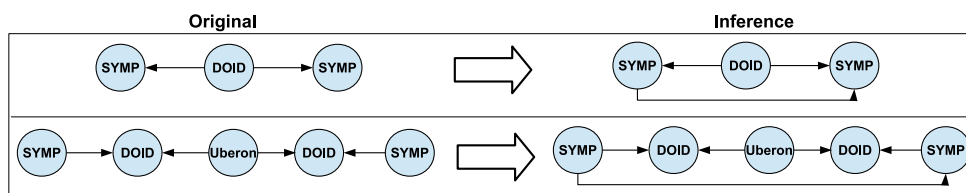


Figure 5.3: Correlating symptoms in BioGraph.

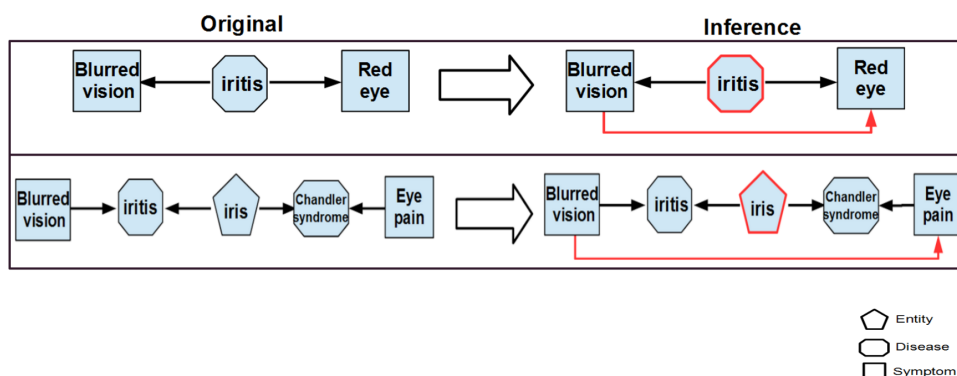


Figure 5.4: Example of the correlating symptoms in BioGraph.

The correlation of symptoms enabled us to build a homogenous network of symptoms, as shown in Figure 5.5, where the nodes represent symptoms and edges represent correlations of symptoms sharing a disease. The edges have weights according to the number of correlations. Figures 5.6 represents the correlations of symptoms sharing an entity.

This network was the basis for the network analysis that we conducted here. We analyzed the topology of the network, i.e., the characteristic way in which the links are organized in the network. This kind of analysis is usual in the Complex Network domain, which defines metrics to discover knowledge observing the characteristic way in which the topology is organized in the network.

Figures 5.5 and 5.6 show the result of a classic centrality algorithm called PageRank [34], applied to the network correlating symptoms by disease. In simple terms, centrality is a measure of “importance” in the network. PageRank measures this importance of a given X node by computing the number of incoming edges of X , weighted by the importance (PageRank) of the nodes which point to X . Therefore, it is a recursive measure, since one must also evaluate the PageRank of the nodes pointing to X .

The nodes in Figures 5.5 and 5.6 have their size according to the value of the PageRank. Therefore, Figure 5.5 shows that high fever (first) and chills (second) are the two most

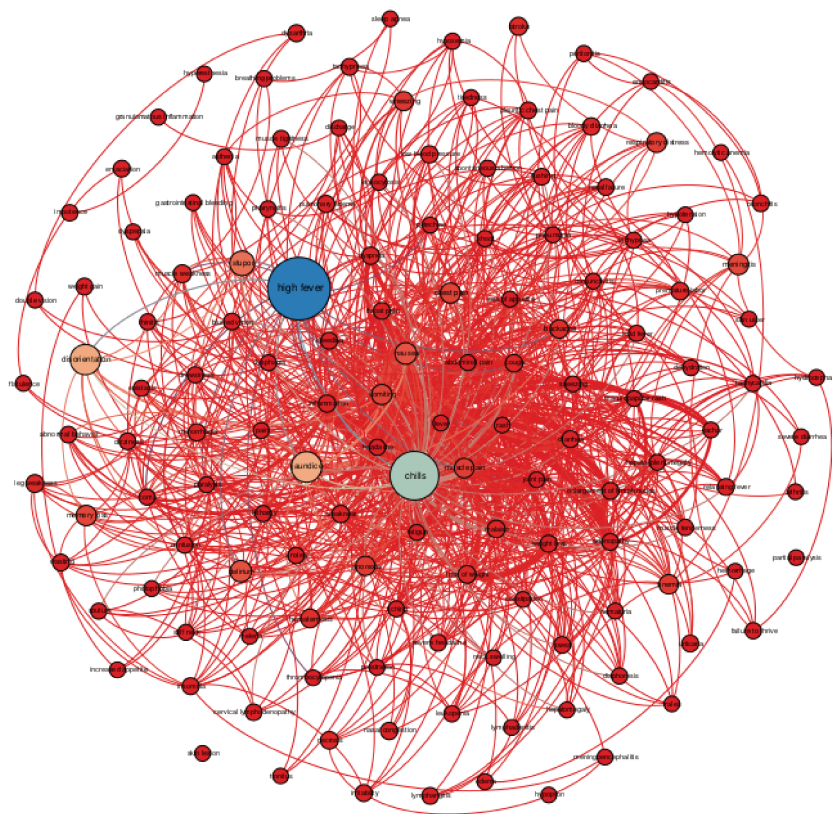


Figure 5.5: PageRank among symptoms that shared disease.

central when considering the correlations by disease. Figure 5.6 shows that high fever is also the most central when considering the correlations by entity.

5.3 Describing

In this subsection, we illustrate how BioGraph can be exploited to help in the homogenization and interlinking of phenotype descriptions carried by systems. The BioGraph base comprises phenotype descriptive elements encompassing several organisms. They are interlinked and generalized in such a way that it is possible to describe organisms using generic statements, e.g., to describe a fish it is possible to use generic terms which can be mapped to several other organisms. Our experimental test was conducted in the Xper system.

Xper is a system to create, store, edit, manage, analyze and distribute (even online) descriptive data using the Character/Character State (C/CS) format to describe phenotypes (see details in Section 2.2). It creates keys for identifying specimens through the phenotype description [48]. *Xper*³ is the online version and is divided into five modules [48] (see Figure 5.7):

1. **item:** enumerates the members of the taxon to be described (e.g., species that are part of a taxon);

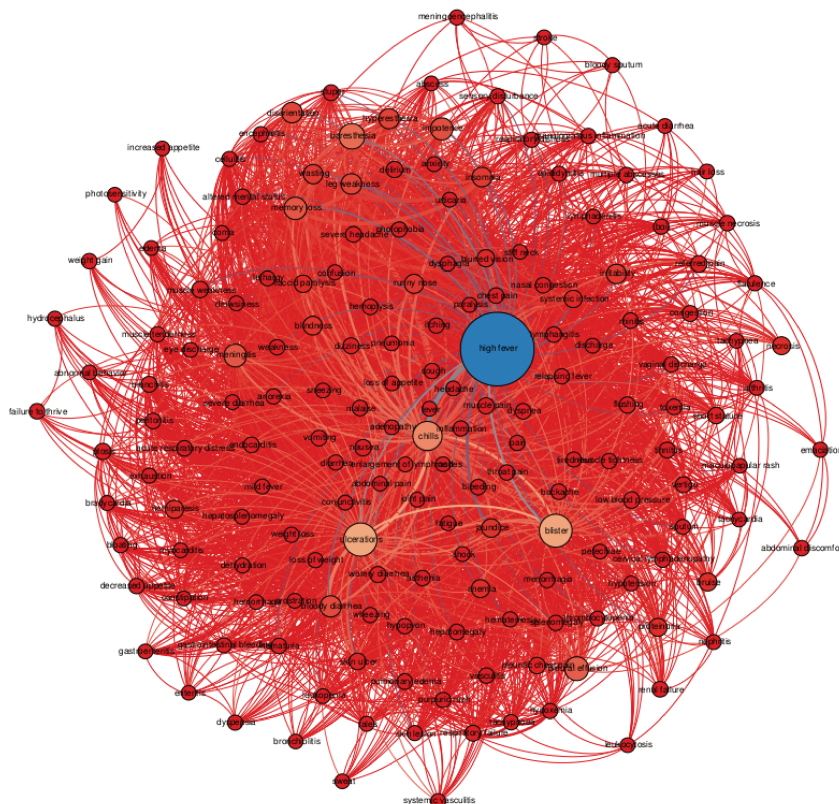


Figure 5.6: PageRank among symptoms that shared entities.

2. **descriptive model:** creates and edits elements of a standard phenotype description for the taxon to be described, structured in characters, states, groups and logical dependencies;
3. **description:** relates the members of the taxon (e.g., species) with the elements of the descriptive model, defining values for the states, i.e., for each member, it is defined the state value(s) of each element of the descriptive model;
4. **identification:** produces an identification key based on the previous data;
5. **management tools:** tools to control and prevent inconsistencies in the descriptions.

In Xper, each user creates her own independent database. Thus, if several researchers study the same subject – e.g., fish – each one will create one different database, which is not connected with other, even for the same taxon. We envisage several benefits in the integration of these data. The experiment shown here is based on a previous work of Cavoto et al. [11].

BioGraph can be exploited to interlink databases in Xper. As shows Figure 5.8, the basic principle is to import to Xper descriptive elements of BioGraph. Therefore, whenever a user adopts these elements, she will be implicitly connecting her description to the others available in BioGraph or connected to it. Figure 5.8 shows the elements involved in the interconnection. In our experiment, we have imported the statements present in BioGraph

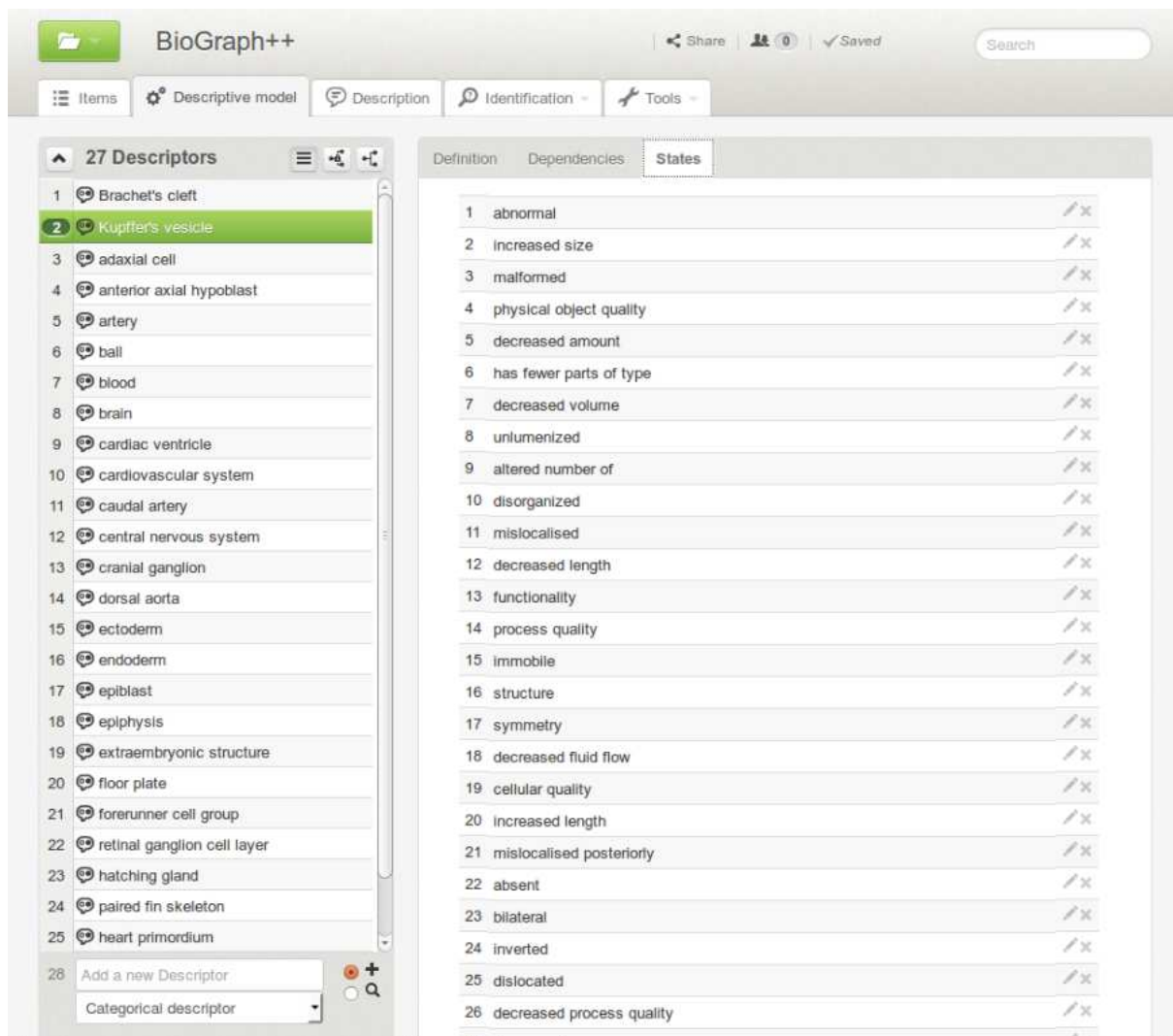


Figure 5.7: A screenshot of Xper with zebrafish data.

and mapped them transforming Entity-Quality (EQ) statements in Character/Character State (C/CS) sentences of Xper (see details in Section 2.2). Therefore, each Entity has been mapped to a Character and each Quality has been mapped to Character State.

The clusters of subgraphs in Figure 5.8 represent different ontologies that are inter-linked by Uberon and Generic_EQs in BioGraph. Whenever a user adopts these terms in a description (a Xper relational database in the figure), the independent bases become implicitly interlinked.

To validate our proposal, we exported the zebrafish data from BioGraph and imported in Xper, as previously described. Figure 5.9 illustrates this process. Data have been exported from Neo4j in XML format, containing elements from ZP (see Appendix C), ZFA (see Appendix B), and PATO (see Appendix D) – see Figure 2.3. These files were the input of our XQuery code (see Appendix A) to produce only one Structured Descriptive Data (SDD) file (see Appendix E). SDD is an open format to share data of phenotypes and identification keys. We have imported this file in Xper. The result is shown in Figure 5.7. In the tab description model, the character (entity) **Kupffer's vesicle** is selected on the left side and, on the right side, it is showed all the states (qualities) that this

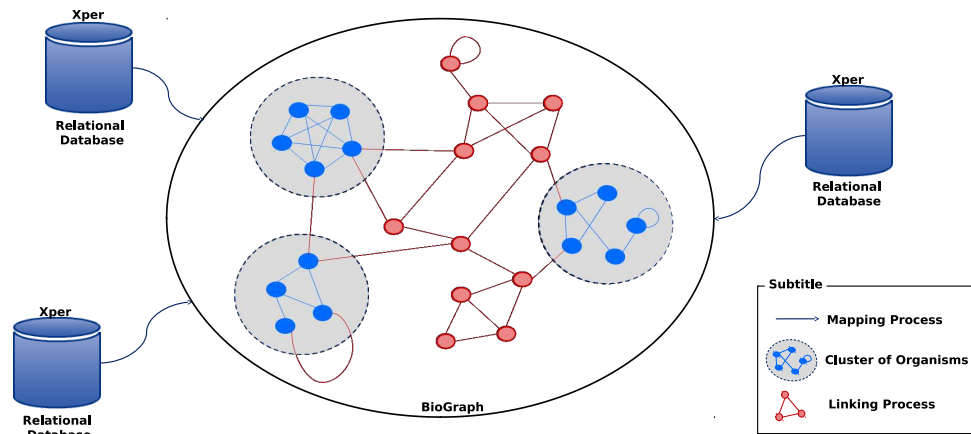


Figure 5.8: BioGraph-Xper connection; adapted from Cavoto et al. [11].

character can present.

As this experiment illustrates, BioGraph can be used as a knowledge base to enrich and to interconnect systems adopted to describe phenotypes.

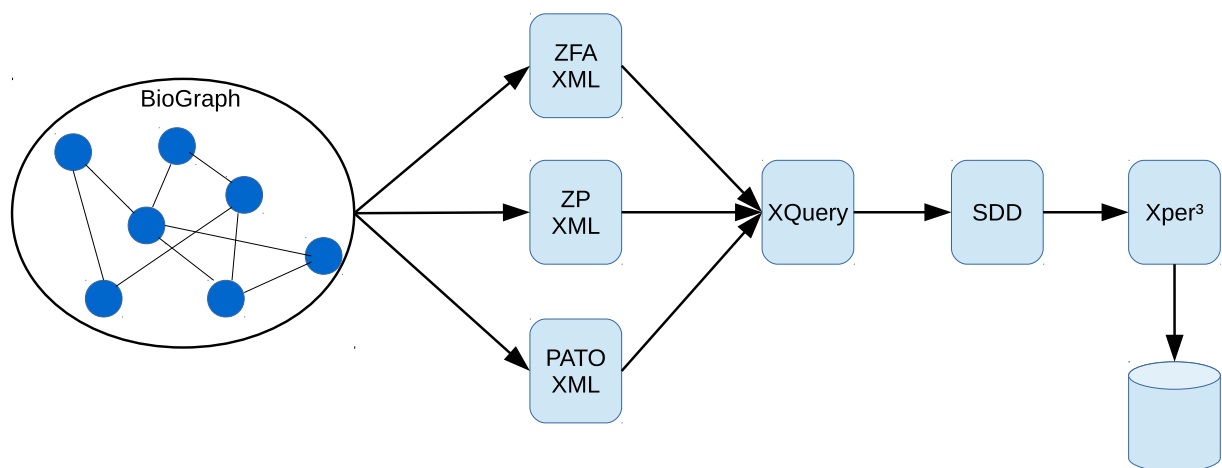


Figure 5.9: BioGraph to Xper mapping process.

Chapter 6

Conclusion

Cross data of different organisms is crucial in medical and biological research. However, each organism is described in a distinct base. The bases are represented in heterogeneous formats, hampering their interconnection.

Sharing phenotypic descriptions is a key to cross data among organisms, due to the possibility of comparing phenotypes. Furthermore, through phenotypes becomes possible finding candidate genes for diseases. It can also help in the diagnosis of diseases. However, even though there are several fragmented initiatives to describe phenotypes and to interconnect them, there is a lack of a solution to combine them in a single unified base. This is the main challenges faced in this work, which addressed the heterogeneity of distinct data sources, having different formats and descriptive approaches for phenotypes.

The main contributions of this work are: the unified model to support several descriptive approaches for phenotypes and the unified graph database, containing descriptions of phenotypes from 63 distinct data sources. The main limitations of this work are: its domain dependency, i.e., the proposed model needs to be extend in order to be applied to other scenarios than phenotypes and there is no mechanism to import data automatically in Biograph.

Future work includes:

1. to import genes, linking them with their phenotypes and diseases;
2. to implement an interface for our system;
3. to improve the integration with Xper, even including features in the system that takes advantage of BioGraph – e.g., the system can recommend phenotype elements from BioGraph;
4. to further explore the possibility of searching and analysis provided by the network.

Bibliography

- [1] Renzo Angles. A comparison of current graph database models. In *Data Engineering Workshops (ICDEW), 2012 IEEE 28th International Conference on*, pages 171–177. IEEE, 2012.
- [2] Renzo Angles and Claudio Gutierrez. Survey of graph database models. *ACM Computing Surveys (CSUR)*, 40(1):1, 2008.
- [3] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.
- [4] Jonathan BL Bard and Seung Y Rhee. Ontologies in biology: design, applications and future challenges. *Nature Reviews Genetics*, 5(3):213–222, 2004.
- [5] Àlex Bayés, Louie N van de Lagemaat, Mark O Collins, Mike DR Croning, Ian R Whittle, Jyoti S Choudhary, and Seth GN Grant. Characterization of the proteome, diseases and evolution of the human postsynaptic density. *Nature neuroscience*, 14(1):19–21, 2011.
- [6] Judith A Blake, Joel E Richardson, Carol J Bult, Jim A Kadin, Janan T Eppig, Mouse Genome Database Group, et al. Mgd: the mouse genome database. *Nucleic acids research*, 31(1):193–195, 2003.
- [7] John Adrian Bondy and Uppaluri Siva Ramachandra Murty. *Graph theory with applications*, volume 290. Citeseer, 1976.
- [8] Willem Nico Borst. *Construction of engineering ontologies for knowledge sharing and reuse*. Universiteit Twente, 1997.
- [9] Tim Bray, Jean Paoli, C Michael Sperberg-McQueen, Eve Maler, and François Yergeau. Extensible markup language (xml). *World Wide Web Consortium Recommendation REC-xml-19980210*. <http://www.w3.org/TR/1998/REC-xml-19980210>, 16:16, 1998.
- [10] Carol J Bult, Janan T Eppig, James A Kadin, Joel E Richardson, Judith A Blake, Mouse Genome Database Group, et al. The mouse genome database (mgd): mouse biology and model systems. *Nucleic acids research*, 36(suppl 1):D724–D728, 2008.

- [11] Patrícia Cavoto, Victor Cardoso, Régine Vignes Lebbe, and André Santanchè. Fish-graph: A network-driven data analysis. In *e-Science (e-Science), 2015 IEEE 11th International Conference on*, pages 177–186. IEEE, 2015.
- [12] Douglas Crockford. The application/json media type for javascript object notation (json). 2006.
- [13] Sandra C Doelken, Sebastian Köhler, Christopher J Mungall, Georgios V Gkoutos, Barbara J Ruef, Cynthia Smith, Damian Smedley, Sebastian Bauer, Eva Klopocki, Paul N Schofield, et al. Phenotypic overlap in the contribution of individual genes to cnv pathogenicity revealed by cross-species computational analysis of single-gene mutations in humans, mice and zebrafish. *Disease Models and Mechanisms*, 6(2):358–372, 2013.
- [14] David Dominguez-Sal, P Urbón-Bayes, Aleix Giménez-Vanó, Sergio Gómez-Villamor, Norbert Martínez-Bazan, and Josep-Lluis Larriba-Pey. Survey of graph database performance on the hpc scalable graph analysis benchmark. In *International Conference on Web-Age Information Management*, pages 37–48. Springer, 2010.
- [15] Karen Eilbeck, Suzanna E Lewis, Christopher J Mungall, Mark Yandell, Lincoln Stein, Richard Durbin, and Michael Ashburner. The sequence ontology: a tool for the unification of genome annotations. *Genome biology*, 6(5):1, 2005.
- [16] Anais Grand, Regine Vignes Lebbe, and Andre Santanche. From phenotypes to trees of life: A metamodel-driven approach for the integration of taxonomy models. In *e-Science (e-Science), 2014 IEEE 10th International Conference on*, volume 1, pages 65–72. IEEE, 2014.
- [17] Thomas R Gruber. A translation approach to portable ontology specifications. *Knowledge acquisition*, 5(2):199–220, 1993.
- [18] Nicola Guarino, Daniel Oberle, and Steffen Staab. What is an ontology? In *Handbook on ontologies*, pages 1–17. Springer, 2009.
- [19] Ada Hamosh, Alan F Scott, Joanna S Amberger, Carol A Bocchini, and Victor A McKusick. Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. *Nucleic acids research*, 33(suppl 1):D514–D517, 2005.
- [20] Christian Theil Have and Lars Juhl Jensen. Are graph databases ready for bioinformatics? *Bioinformatics*, 29(24):3107–3108, 2013.
- [21] S Blair Hedges. The origin and evolution of model organisms. *Nature Reviews Genetics*, 3(11):838–849, 2002.
- [22] Alex Kalderimis, Rachel Lyne, Daniela Butano, Sergio Contrino, Mike Lyne, Joshua Heimbach, Fengyuan Hu, Richard Smith, Radek Štěpán, Julie Sullivan, et al. Intermine: extensive web services for modern biology. *Nucleic acids research*, page gku301, 2014.

- [23] Graham Klyne and Jeremy J Carroll. Resource description framework (rdf): Concepts and abstract syntax. 2006.
- [24] Sebastian Köhler. *Phenotype informatics: network approaches towards understanding the diseasome*. PhD thesis, Berlin Freie Universität, 2013.
- [25] Sebastian Köhler, Sandra C Doelken, Barbara J Ruef, Sebastian Bauer, Nicole Washington, Monte Westerfield, George Gkoutos, Paul Schofield, Damian Smedley, Suzanna E Lewis, et al. Construction and accessibility of a cross-species phenotype ontology along with gene annotations for biomedical research. *F1000Research*, 2, 2013.
- [26] Sebastian Köhler, Uwe Schoeneberg, Johanna Christina Czeschik, Sandra C Doelken, Jayne Y Hehir-Kwa, Jonas Ibn-Salem, Christopher J Mungall, Damian Smedley, Melissa A Haendel, and Peter N Robinson. Clinical interpretation of cnvs with cross-species phenotype data. *Journal of medical genetics*, 51(11):766–772, 2014.
- [27] Jacques Lebbe. *Représentation des concepts en biologie et en médecine. Introduction à l’analyse des connaissances et à l’identification assistée par ordinateur*. PhD thesis, 1991.
- [28] Martin Mahner and Michael Kary. What exactly are genomes, genotypes and phenotypes? and what about phenomes? *Journal of Theoretical Biology*, 186(1):55–63, 1997.
- [29] Prashanti Manda, James P Balhoff, Hilmar Lapp, Paula Mabee, and Todd J Vision. Using the phenoscape knowledgebase to relate genetic perturbations to phenotypic evolution. *genesis*, 53(8):561–571, 2015.
- [30] Daniel Marbach, James C Costello, Robert Küffner, Nicole M Vega, Robert J Prill, Diogo M Camacho, Kyle R Allison, Manolis Kellis, James J Collins, Gustavo Stolovitzky, et al. Wisdom of crowds for robust gene network inference. *Nature methods*, 9(8):796–804, 2012.
- [31] Deborah L McGuinness, Frank Van Harmelen, et al. Owl web ontology language overview. *W3C recommendation*, 10(10):2004, 2004.
- [32] Christopher J Mungall, Carlo Torniai, Georgios V Gkoutos, Suzanna E Lewis, Melissa A Haendel, et al. Uberon, an integrative multi-species anatomy ontology. *Genome Biol*, 13(1):R5, 2012.
- [33] National Research Council (US) Committee on A Framework for Developing a New Taxonomy of Disease et al. *Toward precision medicine: building a knowledge network for biomedical research and a new taxonomy of disease*. National Academies Press (US), 2011.
- [34] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: bringing order to the web. 1999.

- [35] Pablo Pareja-Tobes, Raquel Tobes, Marina Manrique, Eduardo Pareja, and Eduardo Pareja-Tobes. Bio4j: a high-performance cloud-enabled graph-based data platform. *bioRxiv*, page 016758, 2015.
- [36] Lyne Rachel, Sullivan Julie, Butano Daniela, Contrino Sergio, Heimbach Joshua, Hu Fengyuan, Kalderimis Alex, Lyne Mike, Smith N Richard, Štěpán Radek, et al. Cross-organism analysis using intermine. *genesis*, 2015.
- [37] Ian Robinson, Jim Webber, and Emil Eifrem. *Graph databases*. O’Reilly, 2013.
- [38] Peter N Robinson. Deep phenotyping for precision medicine. *Human mutation*, 33(5):777–780, 2012.
- [39] Peter N Robinson, Sebastian Köhler, Sebastian Bauer, Dominik Seelow, Denise Horn, and Stefan Mundlos. The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. *The American Journal of Human Genetics*, 83(5):610–615, 2008.
- [40] Lynn Marie Schriml, Cesar Arze, Suvarna Nadendla, Yu-Wei Wayne Chang, Mark Mazaitis, Victor Felix, Gang Feng, and Warren Alden Kibbe. Disease ontology: a backbone for disease semantic integration. *Nucleic acids research*, 40(D1):D940–D946, 2012.
- [41] Damian Smedley, Anika Oellrich, Sebastian Köhler, Barbara Ruef, Monte Westerfield, Peter Robinson, Suzanna Lewis, Christopher Mungall, et al. Phenodigm: analyzing curated annotations to associate animal models with human diseases. *Database*, 2013:bat025, 2013.
- [42] Barry Smith, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J Goldberg, Karen Eilbeck, Amelia Ireland, Christopher J Mungall, et al. The obo foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, 25(11):1251–1255, 2007.
- [43] Cynthia L Smith, Carroll-Ann W Goldsmith, and Janan T Eppig. The mammalian phenotype ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome biology*, 6(1):R7, 2004.
- [44] Richard N Smith, Jelena Aleksic, Daniela Butano, Adrian Carr, Sergio Contrino, Fengyuan Hu, Mike Lyne, Rachel Lyne, Alex Kalderimis, Kim Rutherford, et al. Intermine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data. *Bioinformatics*, 28(23):3163–3165, 2012.
- [45] Jamie Soul, Timothy E Hardingham, Raymond P Boot-Handford, and Jean-Marc Schwartz. Phenomeexpress: A refined network analysis of expression datasets by inclusion of known disease phenotypes. *Scientific reports*, 5, 2015.
- [46] Judy Sprague, Leyla Bayraktaroglu, Dave Clements, Tom Conlin, David Fashena, Ken Frazer, Melissa Haendel, Douglas G Howe, Prita Mani, Sridhar Ramachandran,

- et al. The zebrafish information network: the zebrafish model organism database. *Nucleic acids research*, 34(suppl 1):D581–D585, 2006.
- [47] Rudi Studer, V Richard Benjamins, and Dieter Fensel. Knowledge engineering: principles and methods. *Data & knowledge engineering*, 25(1):161–197, 1998.
- [48] Visotheary Ung, Guillaume Dubus, René Zaragüeta-Bagils, and Régine Vignes-Lebbe. Xper2: introducing e-taxonomy. *Bioinformatics*, 26(5):703–704, 2010.
- [49] Chad Vicknair, Michael Macias, Zhendong Zhao, Xiaofei Nan, Yixin Chen, and Dawn Wilkins. A comparison of a graph database and a relational database: a data provenance perspective. In *Proceedings of the 48th annual Southeast regional conference*, page 42. ACM, 2010.
- [50] Nicole L Washington, Melissa A Haendel, Christopher J Mungall, Michael Ashburner, Monte Westerfield, and Suzanna E Lewis. Linking human diseases to animal models using ontology-based phenotype annotation. *PLoS biology*, 7(11):e1000247, 2009.

Appendix A

XQuery Code

```
let $zfadoc:=doc('http://www.ic.unicamp.br/~santanch/temp/zp/zfa.xml')
let $zpdoc:=doc('http://www.ic.unicamp.br/~santanch/temp/zp/zp.xml')
let $patodoc:=doc('http://www.ic.unicamp.br/~santanch/temp/zp/pato.xml')
return
<Datasets><Dataset xml:lang="en">
<Characters>
{
  for $zfa in ($zfadoc//ITEM)
  return
    <CategoricalCharacter id="{data($zfa/ID)}">
      <Representation>
        <Label>{data($zfa/NAME)}</Label>
        <Detail>{data($zfa/DESCRIPTION)}</Detail>
        <States>
          {
            for $zppato in distinct-values($zpdoc//ITEM[ZFA=$zfa/ID]
              /PATO),$pato in ($patodoc//ITEM[ID=$zppato])
            return
              <StateDefinition id="{concat(data($zfa/ID),';',
                data($zppato))}">
                <Representation>
                  <Label>{data($pato/NAME)}</Label>
                  <Detail>{data($pato/DESCRIPTION)}</Detail>
                </Representation>
              </StateDefinition>
          }
        </States>
      </Representation>
    </CategoricalCharacter>
}
</Characters>
</Dataset></Datasets>
```

Appendix B

ZFA Terms in XML

```
1
2 <ZFA>
3 <ITEM><NAME>Brachet's cleft</NAME><DESCRIPTION>The visible division
   between epiblast and hypoblast in the gastrula.</DESCRIPTION><
   ID>ZFA:0000000</ID></ITEM>
4 <ITEM><NAME>Kupffer's vesicle</NAME><DESCRIPTION>Small but
   distinctive epithelial sac containing fluid, located
   midventrally posterior to the yolk cell or its extension, and
   transiently present during most of the segmentation period.
   Kupffer's vesicle has been compared to the mouse embryonic node
   .</DESCRIPTION><ID>ZFA:0000001</ID></ITEM>
5 <ITEM><NAME>adaxial cell</NAME><DESCRIPTION>Muscle precursor cell
   that is adjacent to the notochord and part of the presomitic
   mesoderm.</DESCRIPTION><ID>ZFA:0000003</ID></ITEM>
6 <ITEM><NAME>anterior axial hypoblast</NAME><DESCRIPTION>Anterior
   portion of the axial hypoblast.</DESCRIPTION><ID>ZFA:0000004</ID
   ></ITEM>
7 <ITEM><NAME>artery</NAME><DESCRIPTION>Blood vessels that carry
   blood away from the heart.</DESCRIPTION><ID>ZFA:0000005</ID></
   ITEM>
8 <ITEM><NAME>ball</NAME><DESCRIPTION>The anterior round region of
   the yolk cell present after the yolk extension forms during the
   segmentation period.</DESCRIPTION><ID>ZFA:0000006</ID></ITEM>
9 <ITEM><NAME>blood</NAME><DESCRIPTION>A complex mixture of cells
   suspended in a liquid matrix that delivers nutrients to cells
   and removes wastes.</DESCRIPTION><ID>ZFA:0000007</ID></ITEM>
10 <ITEM><NAME>brain</NAME><DESCRIPTION>Cavitated compound organ which
   is comprised of gray and white matter and surrounds the
   cerebral ventricular system.</DESCRIPTION><ID>ZFA:0000008</ID></
   ITEM>
11 <ITEM><NAME>cardiac ventricle</NAME><DESCRIPTION>Cavitated compound
   organ that receives blood flow from the atrium and delivers
   blood to the body via the aorta. Valves are present to direct
   flow. There are only two chambers present in the fish heart.</
```

```

DESCRIPTION><ID>ZFA:0000009</ID></ITEM>
12 <ITEM><NAME>cardiovascular system</NAME><DESCRIPTION>Anatomical
    system that functions in circulation and has as its parts the
    heart and vasculature. The lymphatic system is considered part
    of the cardiovascular system.</DESCRIPTION><ID>ZFA:0000010</ID>
    </ITEM>
13 <ITEM><NAME>caudal artery</NAME><DESCRIPTION>Extension of the
    dorsal aorta in the post-vent region.</DESCRIPTION><ID>
    ZFA:0000011</ID></ITEM>
14 <ITEM><NAME>central nervous system</NAME><DESCRIPTION>The brain and
    spinal cord.</DESCRIPTION><ID>ZFA:0000012</ID></ITEM>
15 <ITEM><NAME>cranial ganglion</NAME><DESCRIPTION>Ganglion which is
    located in the head.</DESCRIPTION><ID>ZFA:0000013</ID></ITEM>
16 <ITEM><NAME>dorsal aorta</NAME><DESCRIPTION>Principal unpaired,
    median artery of the trunk, leading from the paired roots (
    radices) of the dorsal aorta to the caudal artery.</DESCRIPTION>
    <ID>ZFA:0000014</ID></ITEM>
17 <ITEM><NAME>ectoderm</NAME><DESCRIPTION>The outer layer of the
    embryo derived from the epiblast. The definitive ectoderm will
    give rise to such tissues as epidermis, the central nervous
    system, neural crest, and sensory placode.</DESCRIPTION><ID>
    ZFA:0000016</ID></ITEM>
18 <ITEM><NAME>endoderm</NAME><DESCRIPTION></DESCRIPTION><ID>
    ZFA:0000017</ID></ITEM>
19 <ITEM><NAME>epiblast</NAME><DESCRIPTION>The outer of the two layers
    of the blastoderm that form during gastrulation, corresponding
    to primitive ectoderm during gastrulation and to the definitive
    ectoderm after gastrulation.</DESCRIPTION><ID>ZFA:0000018</ID></
    ITEM>
20 <ITEM><NAME>epiphysis</NAME><DESCRIPTION>A circumscribed swelling,
    includes the pineal primordium that appears late in the
    segmentation period in the dorsal midline of the diencephalon.</
    DESCRIPTION><ID>ZFA:0000019</ID></ITEM>
21 <ITEM><NAME>extraembryonic structure</NAME><DESCRIPTION>Anatomical
    structure that is contiguous with the embryo and is comprised of
    portions of tissue or cells that will not contribute to the
    embryo.</DESCRIPTION><ID>ZFA:0000020</ID></ITEM>
22 <ITEM><NAME>floor plate</NAME><DESCRIPTION>Multi-tissue structure
    that is the ventral-most aspect of the developing neural tube.
    The floor plate is a specialized glial structure that spans the
    rostral-caudal axis from the midbrain to the tail regions.</
    DESCRIPTION><ID>ZFA:0000022</ID></ITEM>
23 </ZFA>

```

Appendix C

ZP Terms in XML

```
1
2 <ZP\_SET>
3 <ITEM><ZP>ZP:0009054</ZP><ZFA>ZFA:0000598</ZFA><PATO>PATO:0000460</
  PATO></ITEM>
4 <ITEM><ZP>ZP:0009054</ZP><ZFA>ZFA:0000598</ZFA><PATO>PATO:0001623</
  PATO></ITEM>
5 <ITEM><ZP>ZP:0008661</ZP><ZFA>ZFA:0000155</ZFA><PATO>PATO:0000460</
  PATO></ITEM>
6 <ITEM><ZP>ZP:0008661</ZP><ZFA>ZFA:0000155</ZFA><PATO>PATO:0000600</
  PATO></ITEM>
7 <ITEM><ZP>ZP:0008661</ZP><ZFA>ZFA:0001115</ZFA><PATO>PATO:0000460</
  PATO></ITEM>
8 <ITEM><ZP>ZP:0008661</ZP><ZFA>ZFA:0001115</ZFA><PATO>PATO:0000600</
  PATO></ITEM>
9 <ITEM><ZP>ZP:0005097</ZP><ZFA>ZFA:0000547</ZFA><PATO>PATO:0000460</
  PATO></ITEM>
10 <ITEM><ZP>ZP:0005097</ZP><ZFA>ZFA:0000547</ZFA><PATO>PATO:0000610</
  PATO></ITEM>
11 <ITEM><ZP>ZP:0000054</ZP><ZFA>ZFA:0001114</ZFA><PATO>PATO:0000460</
  PATO></ITEM>
12 <ITEM><ZP>ZP:0000054</ZP><ZFA>ZFA:0001114</ZFA><PATO>PATO:0000587</
  PATO></ITEM>
13 <ITEM><ZP>ZP:0001143</ZP><ZFA>ZFA:0001114</ZFA><PATO>PATO:0000460</
  PATO></ITEM>
14 <ITEM><ZP>ZP:0001143</ZP><ZFA>ZFA:0001114</ZFA><PATO>PATO:0000586</
  PATO></ITEM>
15 <ITEM><ZP>ZP:0001772</ZP><ZFA>ZFA:0000152</ZFA><PATO>PATO:0000001</
  PATO></ITEM>
16 <ITEM><ZP>ZP:0001772</ZP><ZFA>ZFA:0000152</ZFA><PATO>PATO:0000460</
  PATO></ITEM>
17 <ITEM><ZP>ZP:0003743</ZP><ZFA>ZFA:0005490</ZFA><PATO>PATO:0000460</
  PATO></ITEM>
18 <ITEM><ZP>ZP:0003743</ZP><ZFA>ZFA:0005490</ZFA><PATO>PATO:0001997</
  PATO></ITEM>
```

```

19 <ITEM><ZP>ZP:0001866</ZP><ZFA>ZFA:0000152</ZFA><PATO>PATO:0000460</
    PATO></ITEM>
20 <ITEM><ZP>ZP:0001866</ZP><ZFA>ZFA:0000152</ZFA><PATO>PATO:0001780</
    PATO></ITEM>
21 <ITEM><ZP>ZP:0003916</ZP><ZFA>ZFA:0001283</ZFA><PATO>PATO:0000001</
    PATO></ITEM>
22 <ITEM><ZP>ZP:0003916</ZP><ZFA>ZFA:0001283</ZFA><PATO>PATO:0000460</
    PATO></ITEM>
23 <ITEM><ZP>ZP:0000362</ZP><ZFA>ZFA:0009316</ZFA><PATO>PATO:0000460</
    PATO></ITEM>
24 <ITEM><ZP>ZP:0000362</ZP><ZFA>ZFA:0009316</ZFA><PATO>PATO:0001997</
    PATO></ITEM>
25 <ITEM><ZP>ZP:0000362</ZP><ZFA>ZFA:0000114</ZFA><PATO>PATO:0000460</
    PATO></ITEM>
26 <ITEM><ZP>ZP:0000362</ZP><ZFA>ZFA:0000114</ZFA><PATO>PATO:0001997</
    PATO></ITEM>
27 <ITEM><ZP>ZP:0006147</ZP><ZFA>ZFA:0000105</ZFA><PATO>PATO:0000460</
    PATO></ITEM>
28 <ITEM><ZP>ZP:0006147</ZP><ZFA>ZFA:0000105</ZFA><PATO>PATO:0000591</
    PATO></ITEM>
29 <ITEM><ZP>ZP:0004247</ZP><ZFA>ZFA:0009091</ZFA><PATO>PATO:0000051</
    PATO></ITEM>
30 <ITEM><ZP>ZP:0004247</ZP><ZFA>ZFA:0009091</ZFA><PATO>PATO:0000460</
    PATO></ITEM>
31 <ITEM><ZP>ZP:0004247</ZP><ZFA>ZFA:0000368</ZFA><PATO>PATO:0000051</
    PATO></ITEM>
32 <ITEM><ZP>ZP:0004247</ZP><ZFA>ZFA:0000368</ZFA><PATO>PATO:0000460</
    PATO></ITEM>
33 <ITEM><ZP>ZP:0011205</ZP><ZFA>ZFA:0000100</ZFA><PATO>PATO:0000460</
    PATO></ITEM>
34 <ITEM><ZP>ZP:0011205</ZP><ZFA>ZFA:0000100</ZFA><PATO>PATO:0000645</
    PATO></ITEM>
35 <ITEM><ZP>ZP:0000711</ZP><ZFA>ZFA:0000008</ZFA><PATO>PATO:0000460</
    PATO></ITEM>
36 <ITEM><ZP>ZP:0000711</ZP><ZFA>ZFA:0000008</ZFA><PATO>PATO:0000587</
    PATO></ITEM>
37 <ITEM><ZP>ZP:0001511</ZP><ZFA>ZFA:0005145</ZFA><PATO>PATO:0000460</
    PATO></ITEM>
38 <ITEM><ZP>ZP:0001511</ZP><ZFA>ZFA:0005145</ZFA><PATO>PATO:0001779</
    PATO></ITEM>
39 <ITEM><ZP>ZP:0005185</ZP><ZFA>ZFA:0000123</ZFA><PATO>PATO:0000460</
    PATO></ITEM>
40 <ITEM><ZP>ZP:0005185</ZP><ZFA>ZFA:0000123</ZFA><PATO>PATO:0001624</
    PATO></ITEM>
41 </ZP\_SET>

```

Appendix D

PATO Terms in XML

```
1
2 <PATO>
3 <ITEM><NAME>quality</NAME><DESCRIPTION>A dependent entity that
   inheres in a bearer by virtue of how the bearer is related to
   other entities</DESCRIPTION><ID>PATO:0000001</ID></ITEM>
4 <ITEM><NAME>mobility</NAME><DESCRIPTION>A quality of inhering in a
   bearer by virtue of the bearer's disposition to move freely.</
   DESCRIPTION><ID>PATO:0000004</ID></ITEM>
5 <ITEM><NAME>speed</NAME><DESCRIPTION>A physical quality inhering in
   a bearer by virtue of the bearer's scalar absolute value of the
   rate of change of the bearer's position.</DESCRIPTION><ID>
   PATO:0000008</ID></ITEM>
6 <ITEM><NAME>age</NAME><DESCRIPTION>A time quality inhering in a
   bearer by virtue of how long the bearer has existed.</
   DESCRIPTION><ID>PATO:0000011</ID></ITEM>
7 <ITEM><NAME>color</NAME><DESCRIPTION>A composite chromatic quality
   composed of hue, saturation and intensity parts.</DESCRIPTION><
   ID>PATO:0000014</ID></ITEM>
8 <ITEM><NAME>color hue</NAME><DESCRIPTION>A chromatic scalar-
   circular quality inhering in an object that manifests in an
   observer by virtue of the dominant wavelength of the visible
   light; may be subject to fiat divisions, typically into 7 or 8
   spectra.</DESCRIPTION><ID>PATO:0000015</ID></ITEM>
9 <ITEM><NAME>color brightness</NAME><DESCRIPTION>A scalar optical
   property that is the intensity, value or amount of perceived
   light.</DESCRIPTION><ID>PATO:0000016</ID></ITEM>
10 <ITEM><NAME>color saturation</NAME><DESCRIPTION>A scalar chromatic
   property that is the degree of purity of perceived light.</
   DESCRIPTION><ID>PATO:0000017</ID></ITEM>
11 <ITEM><NAME>fluorescence</NAME><DESCRIPTION>A luminous flux quality
   inhering in a bearer by virtue of the bearer's emitting longer
   wavelength light following the absorption of shorter wavelength
   radiation; fluorescence is common with aromatic compounds with
   several rings joined together.</DESCRIPTION><ID>PATO:0000018</ID>
```

```

    ></ITEM>
12 <ITEM><NAME>color pattern</NAME><DESCRIPTION>A chromatic property
    that is the relative position of different hues or degrees of
    saturation.</DESCRIPTION><ID>PATO:0000019</ID></ITEM>
13 <ITEM><NAME>compatibility</NAME><DESCRIPTION>A quality inhering in
    a bearer by virtue of the bearer's disposition to harmonious
    coexistence.</DESCRIPTION><ID>PATO:0000021</ID></ITEM>
14 <ITEM><NAME>composition</NAME><DESCRIPTION>A single physical entity
    inhering in an bearer by virtue of the bearer's quantities or
    relative ratios of subparts.</DESCRIPTION><ID>PATO:0000025</ID><
    /ITEM>
15 <ITEM><NAME>concentration of</NAME><DESCRIPTION>A quality inhering
    in a substance by virtue of the amount of the bearer's there is
    mixed with another substance.</DESCRIPTION><ID>PATO:0000033</ID
    ></ITEM>
16 <ITEM><NAME>consistency</NAME><DESCRIPTION>A physical quality
    inhering in a bearer by virtue of the bearer's density, firmness
    , or viscosity.</DESCRIPTION><ID>PATO:0000037</ID></ITEM>
17 <ITEM><NAME>direction</NAME><DESCRIPTION>A physical quality
    inhering in a bearer by virtue of the bearer's orientation in
    space.</DESCRIPTION><ID>PATO:0000039</ID></ITEM>
18 <ITEM><NAME>distance</NAME><DESCRIPTION>A quality that is the
    extent of space between two entities.</DESCRIPTION><ID>
    PATO:0000040</ID></ITEM>
19 <ITEM><NAME>flavor</NAME><DESCRIPTION>A quality of a physical
    entity inhering in a bearer by virtue of whether the bearer's
    molecules are being perceived by a taste and odorant receptors.<
    /DESCRIPTION><ID>PATO:0000043</ID></ITEM>
20 <ITEM><NAME>frequency</NAME><DESCRIPTION>A physical quality which
    inheres in a bearer by virtue of the number of the bearer's
    repetitive actions in a particular time.</DESCRIPTION><ID>
    PATO:0000044</ID></ITEM>
21 <ITEM><NAME>biological sex</NAME><DESCRIPTION>An organismal quality
    inhering in a bearer by virtue of the bearer's ability to
    undergo sexual reproduction in order to differentiate the
    individuals or types involved.</DESCRIPTION><ID>PATO:0000047</ID
    ></ITEM>
22 <ITEM><NAME>hardness</NAME><DESCRIPTION>A physical quality inhering
    in a bearer by virtue of the bearer's disposition to being
    turned, bowed, or twisted without breaking.</DESCRIPTION><ID>
    PATO:0000048</ID></ITEM>
23 <ITEM><NAME>intensity</NAME><DESCRIPTION>A quality inhering in a
    bearer by virtue of the bearer's possessing or displaying a
    distinctive feature in type or degree or effect or force.</
    DESCRIPTION><ID>PATO:0000049</ID></ITEM>
24 </PATO>

```


Appendix E

Result in SDD

```
1
2 <Datasets>
3   <Dataset xml:lang="en">
4     <Characters>
5       <CategoricalCharacter id="ZFA:0000001">
6         <Representation>
7           <Label>Kupffer's vesicle</Label>
8           <Detail>Small but distinctive epithelial sac containing
              fluid, located midventrally posterior to the yolk cell
              or its extension, and transiently present during most
              of the segmentation period. Kupffer's vesicle has
              been compared to the mouse embryonic node.</Detail>
9         <States>
10           <StateDefinition id="ZFA:0000001;PATO:0000460">
11             <Representation>
12               <Label>abnormal</Label>
13               <Detail>A quality inhering in a bearer by virtue of
                  the bearer's deviation from normal or average
                  .</Detail>
14             </Representation>
15           </StateDefinition>
16           <StateDefinition id="ZFA:0000001;PATO:0000586">
17             <Representation>
18               <Label>increased size</Label>
19               <Detail>A size quality which is relatively high.</
                  Detail>
20             </Representation>
21           </StateDefinition>
22           <StateDefinition id="ZFA:0000001;PATO:0000646">
23             <Representation>
24               <Label>malformed</Label>
25               <Detail>A morphological quality inhering in a
                  bearer by virtue of the bearer's being distorted
                  during formation.</Detail>
```

```

26         </Representation>
27     </StateDefinition>
28     <StateDefinition id="ZFA:0000001;PAT0:0001241">
29         <Representation>
30             <Label>physical object quality</Label>
31             <Detail>A quality which inheres in a continuant.</Detail>
32         </Representation>
33     </StateDefinition>
34     <StateDefinition id="ZFA:0000001;PAT0:0001997">
35         <Representation>
36             <Label>decreased amount</Label>
37             <Detail>An amount which is relatively low.</Detail>
38         </Representation>
39     </StateDefinition>
40     <StateDefinition id="ZFA:0000001;PAT0:0002001">
41         <Representation>
42             <Label>has fewer parts of type</Label>
43             <Detail>The bearer of this quality has_part < n
44                 AND has_part > 0 of the indicated entity type
45                 , where n is the normal amount for a comparable
46                 organism. Note that the bearer of the quality is
47                 the whole, not the part. Formally: If a bearer
48                 entity e has fewer parts of type X at time t,
49                 then the number of instances x of X at t such
50                 that x part_of e is < n, where n is either
51                 the normal number for comparable entities, or n
52                 is stated explicitly.</Detail>
53         </Representation>
54     </StateDefinition>
55     <StateDefinition id="ZFA:0000001;PAT0:0000596">
56         <Representation>
57             <Label>decreased volume</Label>
58             <Detail>A volume which is relatively low.</Detail>
59         </Representation>
60     </StateDefinition>
61     <StateDefinition id="ZFA:0000001;PAT0:0001896">
62         <Representation>
63             <Label>unlumenized</Label>
64             <Detail>A structure quality inhering in a bearer by
65                 virtue of the bearer's lacking of a three
66                 dimensional space surrounded by one or more
67                 anatomical structures and containing one or more
68                 anatomical substances.</Detail>
69         </Representation>
70     </StateDefinition>
71 </States>

```

```
59         </Representation>  
60     </CategoricalCharacter>  
61 </Characters>  
62 </Dataset>  
63 </Datasets>
```