

UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE BIOLOGIA ANIMAL



Validation of Automatic Similarity Measures

Pedro Jorge Mendes dos Santos

Mestrado em Bioinformática e Biologia Computacional

Dissertação orientada por:
Prof. Dr. Francisco José Moreira Couto
Prof. Dr. João Diogo Silva Ferreira

Acknowledgements

First, I would like to thank my advisors Prof Dr. Francisco José Moreira Couto for accepting to take part in the project, and Prof. Dr. João Diogo Silva Ferreira, who always demonstrated motivation in guiding my work and share his knowledge, either related to semantic similarity, web development or bioinformatics in general. I admire his professionalism, his patience every time I did not understand something or had difficulty in expressing my thoughts. It reassured me to know that I could always count on him over the making of my dissertation.

My friends and colleagues who were in the same boat as me, with whom we shared our experiences, laughed, and supported each other when needed.

My parents and brother, without them I would not be where I am, they always motivated me to go after my objectives, no matter how long it took, advised me, and many times simply listened.

Resumo

A capacidade para comparar automaticamente duas entidades biomédicas (p. ex. doenças, vias metabólicas ou artigos científicos) permite que os computadores raciocinem sobre o conhecimento científico. Assim sendo, fazer a validação destas medidas é essencial para garantir que os resultados produzidos por elas reflectam o actual conhecimento colectivo sobre o respectivo domínio.

Uma das estratégias para avaliar se a medida é precisa e funcional é a validação manual por parte de peritos. Contudo, este processo é ineficiente devido a toda a pesquisa secundária necessária para o fazer, o que significa que compilar grandes conjuntos de dados de valores de semelhança atribuídos por humanos é uma tarefa difícil.

“Manual Validation Helper Tool” (MVHT) é uma aplicação web criada com o intuito de acelerar esta validação manual, em conjunto com um formato capaz de acomodar os diversos tipos de dados em forma de anotações, provenientes de diferentes ontologias ou domínios. MVHT foi testada em quatro datasets distintos e um deles foi apresentado a utilizadores piloto para que dessem o seu feedback acerca do que poderia ser melhorado na aplicação, bem como para se obter um gold-standard de semelhança manual. Com o seu auxílio, a ferramenta foi optimizada e encontra-se acessível para ser usada por criadores de medidas de semelhança semântica, que por sua vez podem partilhar os seus datasets de forma prática, os quais peritos podem visitar e rapidamente começar a comparar pares de entidades.

Palavras-chave Ferramenta web, Ontologias biomédicas, Validação manual, Multidisciplinaridade de dados biomédicos.

Abstract

The ability to automatically compare two biomedical entities (e.g. diseases, biochemical pathways, papers) enables the use of computers to reason over scientific knowledge. As such, validating these measures is essential to ensure that the results they produce reflect the current community knowledge on the respective domain.

Manual validation by experts is one of the strategies to assess whether a measure is sound and accurate. However, this is an inefficient process because of the secondary research required to do so, which means that compiling large datasets of human-curated similarity values is difficult.

The “Manual Validation Helper Tool” (MVHT) is a web application created to accelerate this manual validation, coupled to a format that can accommodate different types of data in the form of annotations, from different domains or ontologies. MVHT was tested on four distinct datasets and one of them was given to pilot users so they could provide feedback on the application, as well as to gather a gold-standard of manual similarity. With their help the tool was optimized and is accessible to be used by creators of semantic similarity measures, who can share their datasets in a more practical way via generated URLs, which other people can visit and quickly start comparing pairs of entities.

Keywords Web tool, Biomedical ontologies, Manual validation, Multidisciplinarity of biomedical data.

Resumo Alargado

A quantidade de informação científica tem vindo a aumentar exponencialmente ao longo dos anos. Esta informação continua a ser maioritariamente apresentada sob a forma de linguagem natural em artigos científicos, obtendo-se assim um elevado número de dados não estruturados para os quais é humanamente impossível processar e organizar todos eles. Esforços foram então focados no desenvolvimento de um método mais prático para representar o conhecimento de forma a que fosse manipulável por computadores, levando ao aparecimento das ontologias. Uma ontologia é um conjunto de conceitos referentes a um determinado domínio do conhecimento e das relações estabelecidas entre eles, podendo ser relações hierárquicas, ou não. Estes conceitos podem ser usados para atribuir significado a entidades complexas como proteínas, vias metabólicas, casos clínicos, artigos científicos, etc. Para tal, as entidades são anotadas com conceitos pertencentes a ontologias, num processo chamado anotação semântica, que lhes confere significado que pode ser interpretado computacionalmente. A anotação de entidades remove ambiguidades e facilita a associação com outros recursos, no entanto uma outra aplicação consiste em inferir o nível de semelhança entre entidades comparando os conceitos que as anotam. A esta técnica chama-se semelhança semântica baseada em ontologias. Neste processo é atribuído a um par de conceitos, ou mais frequentemente, de entidades anotada com conceitos, um valor numérico que representa a sua semelhança. Na área da bioinformática, a semelhança semântica tem aplicações em diversas tarefas como prever interações entre proteínas, associação entre genes e doenças, propor novos alvos para fármacos, prever propriedades químicas de metabolitos, entre outros.

Quando uma medida de semelhança semântica é criada, é importante testar a sua eficácia. Esta tarefa, denominada validação ou avaliação, pode ser feita de diversas maneiras e por isso deve ser ponderada qual a opção mais equilibrada, tendo em conta o objectivo da medida, para que não se chegue a uma conclusão precipitada. Uma das formas de validação consiste em comparar a medida de semelhança semântica criada com uma outra já existente (através de um dataset de pares de entidades com valores de semelhança obtidos com a medida existente). É também possível usar um dataset de valores de semelhança manuais, resultante da comparação de pares de entidades por parte de peritos, ao qual se atribui o nome de “gold-standard”. Estes gold-standards são datasets de alta qualidade pois aproximam-se mais fortemente da realidade do que medidas automáticas, contudo esta curação manual é uma tarefa dispendiosa em termos de tempo e não se adapta bem no caso de serem datasets grandes ou que sejam alterados rápida e frequentemente. Seria vantajoso se este passo fosse sistematizado.

Este projecto foi criado com dois objectivos em mente. O objectivo primário seria criar uma ferramenta que pudesse facilitar e acelerar a validação manual e um objectivo secundário, que seria obter um gold-standard. *Manual Validation Helper Tool* (MVHT) é uma ferramenta web desenvolvida a pensar neste problema, permitindo aos criadores de semelhança semântica carregar os seus datasets e partilhá-los com peritos que possam comparar manualmente pares de entidades e assim gerar um gold-standard.

MVHT pode ser dividida em duas partes distintas:

- a) o painel de controle do dono do dataset, contendo informação acerca dos datasets que já foram introduzidos previamente, tais como um identificador, número de entidades, anotações, comparações feitas, URL para ser partilhado com peritos, bem como a opção de apagar o dataset ou de ver os resultados das comparações em maior detalhe.
- b) a vista do curador onde é apresentado um par de entidades lado a lado com anotações agrupadas por ontologia ou domínio, para o qual é esperado que seja atribuído um valor de semelhança. Sempre que seja submetido um valor manual será fornecido um novo par de entidades.

Na vista de painel do dono do dataset, sempre que tenha a intenção de inserir um dataset terá de escolher uma estratégia para fazer o emparelhamento de entidades. De momento a ferramenta disponibiliza três tipos diferentes: a) totalmente aleatório b) ordem fixa e c) ordem fixa por utilizador. Apesar de o emparelhamento ser aleatório em todas as estratégias, cada uma difere na ordem pela qual um par é apresentado ao utilizador, tendo o utilizador a liberdade para decidir dar mais peso a ter mais pares com respostas ou mais respostas por par.

Existe uma grande heterogeneidade na representação do conhecimento biomédico, por exemplo, dados podem ter componentes genómicas, proteómicas, taxonómicas, vias metabólicas, modelos biológicos, casos clínicos, entre outros. Como tal, é necessário testar a ferramenta usando um conjunto de casos que tenham dados provenientes de diferentes fontes. Foram escolhidas quatro fontes de informação para as quais se criaram datasets:

- KEGG Pathways - dataset de vias metabólicas;
- BioModels - dataset de modelos matemáticos de processos bioquímicos;
- CRAFT - dataset de artigos científicos anotados;
- OMIM - dataset de doenças anotadas com factores relacionados.

Outra tarefa necessária para o meu trabalho foi a criação de um formato que se conseguisse adaptar às características de cada um dos tipos de entidades presentes nos datasets acima, ou seja, é importante que o formato seja generalizável dada a multidisciplinaridade dos dados. Foi então criado um esquema de formato em JSON que descreve a informação esperada para cada dataset, cuja documentação se encontra online. Neste formato é obrigatório que esteja presente uma questão para enquadrar os peritos no problema e uma lista das entidades com as respectivas anotações.

A ferramenta foi também testada por utilizadores. Para tal, primeiro foram calculados três parâmetros que ajudassem na caracterização dos datasets (*cobertura, volume e diversidade*). A ideia é que usando estes parâmetros seja possível perceber qual o dataset mais propício a ter pares com mais anotações em comum, a que se chamou *sobreposição esperada*. De todos os datasets, o BioModels foi o que apresentou mais obstáculos na compilação, o que se traduziu no menor valor sobreposição esperada de todos. O dataset com maior sobreposição esperada foi o OMIM, seguido do CRAFT. Visto que o dataset relativo às doenças do OMIM apenas foi criado mais tarde, o dataset referente ao CRAFT foi o escolhido para dar aos utilizadores, a partir do qual se tentou obter um gold-standard. Foram feitas comparações para 67 pares no total e para além disso foram recolhidos comentários acerca do que não era funcional e que poderia ser melhorado na interface. Alguns exemplos de alterações foram a falta de indicadores de que estaria a ser feita uma comparação entre duas coisas, de indicadores que mais informação estaria contida dentro das abas, falta de espaço vertical na página, entre outras. Estes comentários foram usados para melhorar a ferramenta.

Para os pares de entidades que obtiveram respostas procedeu-se ao cálculo da semelhança semântica com uma medida denominada simGIC. A medida simGIC é aplicada em subgrafos definidos a partir dos conceitos que anotam as entidades, incluindo os seus ancestrais, usando uma medida de especificidade para atribuir peso aos nós do grafo. Com os resultados manuais recolhidos calculou-se o coeficiente de correlação de Pearson para ver se havia correlação entre os valores de semelhança manual com os automáticos (simGIC). O valor de correlação obtido foi de 0.132, implicando que não existe correlação entre as duas variáveis. Isto pode ser explicado pelo facto de se ter obtido um dataset de valores de semelhança manuais de fraca qualidade, como tal não foi atingido um gold-standard. No entanto, esta tarefa foi desempenhada sobretudo para demonstrar que é possível recolher facilmente as respostas dentro da ferramenta para um ficheiro e fazer rapidamente o cálculo da correlação com uma medida automática. No futuro, tenciono melhorar as funcionalidades da ferramenta introduzindo mais estratégias de seleção de pares para serem apresentados aos curadores, um validador do esquema JSON e por fim usar o dataset de doenças retiradas do OMIM e apresentá-lo a médicos para que deem a sua

opinião acerca do que pode ser melhorado, bem como para fazerem comparações de entidades e gerar um gold-standard.

Contents

List of Figures	xiv
List of Tables	xvii
Acronyms	xix
1. Introduction.....	1
2. Background.....	3
2.1. Entity annotation.....	3
2.2. Types of Datasets.....	3
2.2.1. Gene products	4
2.2.2. Clinical trials.....	4
2.2.3. Gene-phenotype relations.....	5
2.2.4. Biological pathways	5
2.2.5. Mathematical models	6
2.2.6. Articles.....	6
2.3. Types of semantic similarity.....	6
2.4. The Gold-standard.....	7
2.5. Similar Platforms	8
2.6. Tools/Programming languages used	8
3. Methodology	10
3.1. Format.....	10
3.2. Datasets	10
3.2.1. KEGG	11
3.2.2. BioModels	12
3.2.3. CRAFT.....	14
3.2.4. OMIM.....	14
3.3. Application	16
3.3.1. Architecture	16
3.3.2. Pairing Strategies	17
3.4. Data collection.....	19
3.5. Comparison with automatic measures.....	19
4. Results and Discussion.....	21

4.1. Dataset format	21
4.2. The four case studies	22
4.2.1. KEGG Pathways.....	23
4.2.2. BioModels	23
4.2.3. CRAFT.....	24
4.2.4. OMIM.....	25
4.3. Front-end implementation	26
4.3.1 The current version	26
4.3.2 Interface evolution	28
4.4. Manual comparisons.....	32
4.5. Manual values vs Automatic values.....	33
5. Conclusions.....	35
References	37
Appendices	40
A. List of diseases	40
B. JSON schema	42

List of Figures

Figure 1 - Representation of a pathway entity and all its annotations.....	11
Figure 2 - Extraction of annotations was done via finding the desired property names for each pathway text file, i.e. “DISEASE” and then capture every line containing a KEGG term until reaching the next property, which in this case corresponds to “DRUG”.....	12
Figure 3 - Representation of a biological model entity and all its annotations.	13
Figure 4 - Each species is taken from a different model. The one in the top contains a sboTerm attribute which allows the concept name to be retrieved from SBO. The middle species has no sboTerm so the only option here would be to use the name “BasalACh2” which is not very clear as to what it refers to. In the bottom species there is also no sboTerm attribute and the Id and name attributes are the same “F26DP”.....	13
Figure 5 - Representation of an article entity and its annotations	14
Figure 6 - Representation of a disease entity and all its annotations.....	15
Figure 7 - MVHT architecture and workflow	17
Figure 8 - Representation of the current pairing strategies implemented. Adjacent numbers inside the vertical rectangle represent pairs of entities and each dot on the left with a different color symbolizes a different curator. If a pair has been given a similarity value then a dot, or multiple dots, appears next to it.	18
Figure 9 - “Outer layer” of the schema with the keys question and entities datasets must have.....	21
Figure 10 - Example dataset to demonstrate the type of information that can be processed by the tool. Here the dataset contains two made-up proteins A and B, where it can be seen that for key “functions” there are objects inside them, with properties “name” and “url”, while for key “processes” it consists of a list with annotation names, showing the different ways users may want to express their data.....	22
Figure 11 - View of the dataset owner dashboard after a user logs in. On the left there is a panel with the pairing options, and a place where a dataset can be uploaded, as well as a link to sign out of the account. On the right there is a table with information about each dataset already inserted, and an option to look at the results, as well as a delete button.	26
Figure 12 - Form curators are required to fill before proceeding with the pair comparisons.....	27
Figure 13 – Manual curator view comparing two articles from the CRAFT dataset.	27
Figure 14 - Page containing the results for one of the datasets showing information about each comparison already made.	28
Figure 15 - Default state of the similarity bar when given a new pair of entities.	29
Figure 16 - Submit button appears adjacent to the place where users drag the slider.	29
Figure 17 - Tab containing CHEBI annotations with three Venn diagrams icons, meaning that inside it there are 14 unique annotations for the article in the left, 20 unique annotations for the article in the right and 4 common annotations between them.....	29

Figure 18 - Collapsed tabs with a down arrow icon on the left to signal that there is hidden content if clicked.....	30
Figure 19 - Open tab containing the three columns inside and an upper arrow icon on the left to indicate it can be collapsed.....	30
Figure 20 - If a user clicks on an annotation it will redirect them to the entry “CHEBI:24669” of ChEBI containing more information about “hydroxy carboxylic acid”.	31
Figure 21 - Article titles taken from a CRAFT dataset pair with a “VS” string in the middle to underline there is a comparison being made.....	31
Figure 22 - The header of the page with two KEGG pathways “hsa00260” and “hsa02010” being compared. The description would be open by default using a large portion of the pages space.	31
Figure 23 - Horizontal tabs for each domain in a KEGG’s entity showing the number of unique and common annotations inside them.....	32
Figure 24 - Similarity bar that would fill up to slider with a unique color. Submit button was placed below.....	32
Figure 25 - Histogram with the manual values, with the y-axis being the number of answers and the x-axis being the similarity score, ranging from 0 to 100, assigned to pairs by curators.	33
Figure 26 - Scatter plot and histograms between manual (top histogram) and automatic (right histogram) similarity values for CRAFT pairs.....	34
Figure 27 - JSON schema which MVHT accepts to upload the given dataset.....	42

List of Tables

Table 1 - Statistical metrics coverage, volume and diversity were calculated for each domain in KEGG.	23
Table 2 - Statistical metrics coverage, volume and diversity were calculated for each domain in BioModels.....	24
Table 3 - Statistical metrics coverage, volume and diversity were calculated for each ontology in CRAFT.....	25
Table 4 - Statistical metrics coverage, volume and diversity were calculated for each ontology in OMIM.	25
Table 5 - Expected overlap of common annotations between pairs in each dataset.....	33

Acronyms

BP – Biological Processes (GO)

CC – Cellular Components (GO)

ChEBI – Chemical Entities of Biological Interest

CL – Cell Ontology

CRAFT – Colorado Richly Annotated Full-Text

DOID – Human Disease Ontology

eo – Expected Overlap

GO – Gene Ontology

KEGG – Kyoto Encyclopedia of Genes and Genome

MF – Molecular Functions (GO)

MOP – Molecular Process Ontology

MVHT – Manual Validation Helper Tool

OMIM – Online Mendelian Inheritance in Man

PRO – Protein Ontology

SO – Sequence Ontology

SYMP – Symptom Ontology

UBERON – Uber-anatomy Ontology

1. Introduction

The amount of data and knowledge that gets published every day makes it challenging for researchers to be up to date with all this new information [1]. At the same time, data is published in different formats which also increases the difficulty in aggregating all this information into a common repository [2].

Processing and managing all this information can be achieved by making it machine readable, a challenge that is being explored by using ontologies, which are formal representations of a domain or domains of knowledge that implicitly attribute machine-readable meaning to the concepts of that domain, based on the relationships between them [3]. One application of ontologies is in areas such as information retrieval [4], that depend on a notion of similarity between entities to find whether a certain document or resource is related to a search query. Ontologies do this by being a source of annotations to complex entities. For example, proteins can be regarded as the set of their molecular functions, described in an ontology, which transitively assign machine-readable semantics to the proteins; pathways can be annotated with the biochemical processes they execute, the chemical compounds involved in the reactions, the cellular locations where these reaction occur, etc. This style of annotation with ontology concepts allows for automatic reasoning to be applied to the annotated entities, for example, by comparing the chemical compounds involved in two pathways, provides a means to score the similarity of the pathways. These ideas depend on a notion of ontology-based semantic similarity (named simply “semantic similarity” in the rest of this document), defined as a measure of similarity between concepts, or between entities annotated with concepts, that is computed based on the structure of the ontology.

One important step in developing a similarity measure is a process called validation (or evaluation), which can be done in different manners. Some examples include:

- picking an existing similarity measure, that is known to be appropriate for a specific task, and checking for a correlation [5],
- predicting properties of pairs of entities. For example, one can create a dataset of pairs of proteins that are known to interact and pairs of proteins that are known not to interact, and then test if the semantic similarity measure between them can accurately predict these properties [6].

Manual validation is another approach that consists in giving experts pairs of entities and making them attribute a similarity value to each one [7]–[9]. It is generally accepted that manually assigned similarity values, given by experts in a subject, are the most realistic representation of reality, and result in a high-fidelity dataset to be used when validating. This method is appealing because these experts have extensive knowledge about their fields of work and are more suitable to evaluate entity similarity than other people or automatic measure. However, just as each type of validation contains its strengths and its limitations, for manual validation the issue lies in the substantial amount of time necessary to carry this task, where it is necessary for these experts to understand what is being asked, to look for more information in other places and for dataset owners to collect the data from them, making it only practical to use this validation method for smaller datasets [10].

The Manual Validation Helper Tool (MVHT) was created with this idea in mind. It is a web application that is accessible to anyone, that tries to simplify the manual validation process, where the only requirement is for dataset owners to supply their dataset in a designed JSON format, share the generated links with experts or curators, and grab the results once available, having access to real-time analytics such as how many experts have already answered, their fields of expertise, the number of answers, and for what pairs. The application also benefits experts, since they can do this task wherever

and whenever they want, with other sources of information associated with the entities and their annotations more at hand, as well as having an accessible way to submit a similarity score and navigate through the pairs of entities.

This document is organized into 5 chapters. This first chapter introduces and motivates the project. The second chapter explains the notions related to the problem that are necessary to understand it, similar works published previously, and the tools that were used in my work. The third chapter describes the methodology of this project (the datasets that were compiled to test the application and the design and implementation of the application). The fourth chapter delves into the type of information each dataset contained, explains the dataset format defined to be used by the application, the improvements made to the tool over time, and finally, presents a comparison between a collected gold-standard and an automatic similarity measure. The fifth chapter presents some conclusions, some limitations to my contributions and potential future work.

2. Background

2.1. Entity annotation

Over the years there has been an exponential increase in the amount of documented research in the biomedical field; however, this information tends to be presented in an unstructured manner, particularly in natural text within scientific literature [2]. As the amount of information grows, extracting the knowledge and information from this unstructured content in an efficient way becomes more difficult, as the man-power needed to properly read, analyse and extrapolate from the text-encoded knowledge exceeds the possibilities within the scientific community. One way to assist in the processing of this unstructured information is a process called semantic annotation, which consists in identifying concepts in those documents, and annotating them with an identifier that refers to that concept defined in an ontology [11]. An ontology is a representation of a domain of knowledge, in a hierarchical or tree-like shape, made of concepts and the relationships between these concepts, whose purpose is to provide machine readable semantic meaning to those concepts, granting the ability to perform automatic (computer-assisted) operations with the data. The concepts are represented in the form of classes, where related concepts that precede a specific concept are superclasses (also known as hypernyms); in contrast, related concepts that succeed a specific concept are subclasses (hyponyms). Concepts can also be related to one another by relationships other than this class-subclass type. For example, an ontology of anatomy can relate the concepts “aorta” and “heart” by means of the relationship “adjacent-to”; an ontology of biochemical processes may define that a process “regulates” another one.

Ontology concepts are small units that can be used to ascribe a broader meaning to complex entities, such as medical notes, proteins, biological pathways, scientific articles, etc. These entities can be annotated with concepts from ontologies of appropriate domains, improving their machine-readable meaning. Let us use as an example an article that contains the word “Turkey”. If this word is annotated with a concept from a geospatial ontology, it would refer to the country, and not the animal, thus making the formal meaning of the word explicit. Another example using the UniProt database could be elastin, the protein present in connective tissue which confers elasticity properties to tissues making them return to a default state (identifier [Uniprot:P15502](#)). This protein is annotated as having the molecular function “extracellular matrix binding”, biological process “animal organ morphogenesis” and cellular compartment “extracellular matrix”. Both these examples consist of entities that upon enrichment with these annotations become objective and unambiguous regarding their formal meaning. The end goal is for these documents or other pieces of content, after being annotated semantically, to become reusable and interoperable sources of information, achieved by linking them to other already existent data repositories.

2.2. Types of Datasets

In the biological field, knowledge can be extracted from very diverse sources of data. Certain kinds of “wet lab experiments” are notorious for the high volume of generated data (like spectrometry, DNA sequencing, etc), scientific literature is constantly being published, etc. This information can be stored in different formats, for example in the form of the datasets of annotations mentioned in the previous section, which can be stored in some sort of repository, such as files or databases. These repositories are often interlinked with other databases that contain an entry referring to the same concept.

Datasets diverge from each other depending on the domains used to annotate the entities that are part of them, as well as the way they are structured. This format heterogeneity, consequently, reduces the flexibility of the systems since it may be necessary to adopt a data format or use specific tools which are more suited for the dataset's intrinsic structure [12]. Even though this heterogeneity represents a challenge in the analysis and processing of datasets, with the help of ontologies this problem can be mitigated, by disambiguating annotations.

Here I present a set of digital datasets related to areas of life sciences, showing some common characteristics in this field: gene products, clinical trials, gene-phenotype relations, biological pathways, mathematical models and articles.

2.2.1. Gene products

Gene product studies have been one of the most popular topics in biological research. For protein datasets, it is relevant to know their functions, places where they can be found in the cell, other proteins that interact with them, families they belong to, 3D structure, diseases associated with them, etc. In terms of resources available to search for this data, UniProtKB is the “go to” Knowledgebase, rich in protein sequence and functional information. It consists of two sections called UniProtKB/TrEMBL and UniProtKB/Swiss-Prot, differing in the way annotations are made. For UniProtKB/TrEMBL, protein annotations are generated computationally, meaning that they are unreviewed, unlike UniProtKB/Swiss-Prot, where annotations are provided manually, meaning their information comes from literature and is reviewed by human curators [13]. In UniProtKB, there is a controlled vocabulary developed for the entries according to their content, called “keywords”, which in other words are the annotations, and each keyword can be categorized into one of 10 categories, e.g. “ligand” or “cellular component”. Categorization is done in order to help indexing entries. These keywords are then manually associated with another independent controlled vocabulary provided by the Gene Ontology Project (GO), called GO terms. GO is the most widely known and used knowledge source of gene functions. All annotations are mainly divided into three different categories or domains, organized into hierarchical vocabulary sets:

- Molecular Functions which describe activities executed at the molecular level by gene products, such as “transporter activity”;
- Biological Processes, a series of molecular functions carried out by one or more gene products, like “organelle organization”;
- Cellular Components which are the cellular locations where those actions take place [14].

These categories are, in fact, three different ontologies with no class-subclass relations between terms of distinct domains. Even though they are disjoint ontologies in this sense, there are still some type of relations present, such as molecular function term “enzyme regulator activity” ([GO:0030234](#)) having a *part_of* relation to the biological process term “regulation of catalytic activity” ([GO:0050790](#)), or “regulation of water channel activity” ([GO:1902427](#)) which is a term from the biological process branch, having a *regulates* relation to “water channel activity” ([GO:0015250](#)), a term from the branch molecular function.

2.2.2. Clinical trials

Data from clinical trials can be taken from [ClinicalTrials.gov](#), which is a registry of research studies done with the assistance of human volunteers who are asked to participate in interventions of some sort. These clinical studies are meant to evaluate the changes in participants who were submitted to interventions, contributing to the amount of medical knowledge. For clinical trials, factors of interest

can be the condition or disease being studied; the focus of the research (what is the question trying to be solved by researchers); the type of strategies, design and methods practiced; starting and ending date of the study; number of people that participated; the criteria for selecting who is apt to participate in the study; the treatment effects, expected and not expected; the type of study, which could be interventional, meaning that some type of treatment is administered (such as vaccines, drugs, medical devices or other procedures, although the methods can also be non-invasive, i.e. alterations in diet or exercise), or the study could be merely observational, where diagnosis or other types of interventions can still be received, but no treatment is given, i.e. collecting patients' medical condition information [15].

2.2.3. Gene-phenotype relations

There are also datasets focused on gene-phenotype relationships. This requires the gathering of the locus/genes where the mutation occurs, the type of inheritance (if it is autosomal or somatic, recessive, or dominant), the diseases derived from it, etc. A renowned resource with this type of information is the Online Mendelian Inheritance in Man (OMIM), a knowledgebase of genes and genetic disorders, containing a team of curators who continuously review biomedical literature and update the compendium. At the moment of writing this thesis, OMIM contained a total of 25,516 entries. Each entry is assigned a unique six-digit number and is categorized in terms of what type of information they describe (genes, phenotype, or both). OMIM entries contain links that redirect the user to other resources, like in the case of clinical synopsis, containing anatomically organized clinical characteristics present in the disorder, linked to ontologies such as Human Phenotype Ontology (HPO) or Disease Ontology (DO). The key feature of OMIM is the review of literature about genetic mutations, genes and diseases associated with the OMIM entry, written in textual form, compiled from scientific sources and reviewed by the curators. Every paragraph contains a citation and a link to the original text corpus [16].

2.2.4. Biological pathways

Data from biological pathways, either from signalling, regulatory or metabolic pathways can also be represented and stored in databases such as KEGG or Reactome. The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a collection of databases containing information for understanding functions and utilities of biological systems, from genomic and molecular information. It consists of 18 databases, manually created based on published literature, categorized into genomic information, chemical information, health information and systems information. Database entries are biological entities from molecular to higher levels called KEGG objects. Each object contains an identifier, usually with the form of a prefix followed by a five-digit number, which is unique for all databases.

The genomic information category contains organisms with complete and catalogued genomes, taken mostly from GenBank, and is present in the GENOME and GENES databases. The chemical information category contains chemical substances and reactions in COMPOUND, REACTION, ENZYME and GLYCAN databases. The health information category contains information about drugs, groups of functionally identical drugs, diseases, and disease networks, stored inside databases like DRUG and DISEASE. In the systems information category (PATHWAY, BRITE and MODULE databases) there are diagrams representing (a) molecular interactions and reaction networks (PATHWAY); (b) functional hierarchies of genes, proteins, drugs, compounds, diseases, relations and organisms (BRITE); (c) functional units of gene sets in metabolic pathways and phenotypic features (MODULE). For KEGG PATHWAY, as of September 2020, is comprised of 538 reference pathways, with manually drawn maps and kegg objects stored in other KEGG databases such as: genes and proteins; diseases associated with the pathway; drugs that target gene products in it; chemical compounds substances and chemical reactions [17].

2.2.5. Mathematical models

An example of a dataset describing biological models can be made using BioModels, a repository containing structured information about mathematical models describing dynamic interactions [18]. The increase in mathematical models of biological processes resulted in an increase of divergences in terms of the way models were presented, which consequently decreases the reusability of published models [19]. To address this necessity, standard and machine-readable formats were created, and Systems Biology Markup Language (SBML) became the most prominent format in systems biology. SBML keeps being updated and every major edition in the composition and structure is defined as a *Level*. A Level can suffer smaller alterations and is released as a new Version for that Level. Currently the latest Level available is Level 3 Version 2¹. BioModels was created having these necessities in mind, presenting a repository where models could be deposited, and accessed by everyone so that they could be expanded and improved, using SBML as its core format. Models stored in this repository can either be described in literature or generated automatically from pathway resources such as KEGG, resulting in a collection of models called Path2Models [20].

2.2.6. Articles

Scientific literature, the most traditional form of documenting research can also be used as a basis to create a dataset. The Colorado Richly Annotated Full-Text (CRAFT) corpus consists of 97 full text journal articles from PubMed that have been manually annotated regarding the concepts mentioned in the text of the papers. This annotation project was done in order to show that article bodies can contain important information that is missed just by directing attention to the abstract. The papers in this collection have all been used by the Mouse Genome Informatics group². The procedure was getting all textual references to terms from an existing ontology, resulting in seven different ontologies used, and more than 100,000 concepts. The annotation process was divided in named entity annotation, linguistic annotation and coreference annotation. Named entity annotation or annotation of concepts was executed by PhD students and PhDs in the biological sciences. Linguistic annotation was done by linguist graduate students where the text structure, part-of-speech tags and sections were annotated. Coreference annotation was done by combining PhD students and linguistics and checking for disagreements [21].

2.3. Types of semantic similarity

The act of annotating entities allows us to index concepts, makes them easier to link to other resources, and resolves ambiguities in text, but that is not the only thing semantic annotation enables. By comparing two entities that are annotated with concepts from an ontology, we can measure how similar the two entities are to one another. This is called ontology-based semantic similarity, or simply semantic similarity. These are algorithms that make use of the meaning behind the concepts being compared and assign to the pair of concepts a numeric value of similarity, representing how close the two entities are [22]. Semantic similarity can be used to compare two ontology concepts or, more generally, two entities annotated with ontology concepts.

In the field of bioinformatics, semantic similarity measures have been used in tasks such as finding interactions between proteins based on their functions [23]; finding gene functions in articles using information extraction methods [24]; finding new uses for drugs [25]; predicting chemical properties in metabolites [26], among others.

¹ <http://sbml.org/Documents/Specifications>

² <http://www.informatics.jax.org/>

There have been many different algorithms proposed to measure semantic similarity, each tailored to a specific application, specific goal scenario, or specific ontology. Edge-based measures look into the edge distances in the hierarchy of concepts in the ontology. In other words, they count the minimum number of relations (edges) that must be travelled in order to go from one concept to the other, so the fewer the number of edges between them, the higher the level of similarity. Node-based approaches use the nodes (concepts), instead of edges, as the main data sources to measure semantic similarity. They usually start by assigning to each concept a measure of its specificity called Information Content (IC), which depends on the frequency with which that concept appears in an annotation corpus. Concepts that are more specific (i.e. pneumothorax), as in, carry more information, have higher IC values, whereas concepts that are less specific (i.e. respiratory disease) have a lower IC value since they are more abstract.

When comparing two entities, semantic similarity measures that can compare sets of concepts are necessary. Typical approaches can be divided into pairwise and groupwise. Pairwise approaches first calculate semantic similarity for every pair of concepts between two annotated entities, then they combine the various numeric values into a global score using a combination method, such as the average, maximum or sum. Groupwise approaches are algorithms devised to deal directly with sets of annotations and calculate semantic similarity directly using set, vector or graph methods. In set methods only direct annotations are considered and similarity is calculated using set similarity measures. In vector methods entities are represented in a vector space, where each concept corresponds to a dimension, then a similarity is calculated using a vector similarity measure. In graph methods, entities are represented as subgraphs of an ontology containing all their annotations. Similarity can be calculated by using graph matching techniques or by defining a subgraph for an entity, containing the concepts that annotate said entity, including its ancestors, and then doing the same for the other entity [27].

Finally, there have been semantic similarity measures developed to use information from multiple ontologies, which enables their use in comparing entities that are annotated with concepts from different domains (such as the ones mentioned in the previous section) [28].

2.4. The Gold-standard

Similarity measures must go through a validation step, where they are tested to determine their performance, taking into account their original objective. There are various possible validation strategies to opt for, therefore, picking which method to use should be a careful process because it could have an effect on the level of success for the measure, as in, an unfit validation method could result in an erroneous acceptance of its capacities and, consequently, lead to incorrect knowledge.

There is no validation strategy that is universally useful for all purposes, and developers of semantic similarity measures either reuse a validation measure developed in the past or create a new one based on their goal. Therefore, a standardization of the process could help instil some guidelines not only for measure developers but also for users of these measures and publishers.

There has been an attempt [22] to group these strategies into four different types (all GO-based semantic similarity):

- **Classification strategies:** Through machine learning a model is trained to predict properties related to the entities being compared, using the semantic similarity measure as part of the machine-learning approach.
- **Comparison strategies:** The semantic similarity measure being tested is compared to a previous one, which has been validated before.
- **Theoretical validation:** The semantic similarity measure is evaluated using just its own mathematical properties.

- **Contextual validation:** Uses statistical methods to show that the measure is able to detect a statistically significant difference between the similarity of two related entities and the similarity of two unrelated entities.

Among the possible “comparison strategies”, there is a specific method relevant to my work called *Manual similarity*. Here, humans are given pairs of concepts and they give a score that captures how similar they think the pair is, based on their expertise in the domain. Then looking at the manual and automatic values, correlation can be computed. In principle, a higher correlation means that the automatic measure is more qualified to reflect the human perception of similarity. This idea assumes that similarity values given by humans are trustworthy since experts' cognition is a close representation of reality. When manual semantic similarity values are collected, what is obtained is a “gold-standard”. The “gold-standard” is a benchmark that can be used with confidence as a means of validating new similarity measures, at least for the specific type of data present on the dataset.

2.5. Similar Platforms

The idea of using manual similarity values for semantic similarity validation is not new, and there has been a set of previous works that used the idea of asking experts about the similarity between pairs of concepts or ideas. Rong [7] calculated the semantic similarity of GO terms using Resnik’s algorithm, then asked 10 biologists to compare 25 GO term pairs containing a mixture of high, intermediate, and low similarity pairs. Pedersen et al [8] performed similarly, computing similarity of biomedical concepts and asking 13 people to assign a similarity score to 120 pairs. Soğancıoğlu [9] validated measures with the help of 5 experts that manually annotated 100 sentence pairs.

To the best of my knowledge, there are no software or web applications that attempt to help validate automatic measures by creating a gold-standard. On July 21st 2020, searches were made in PubMed Central and in the academic search engine Google Scholar querying for “Semantic similarity tool”, “Semantic similarity evaluation tool”, “Manual evaluation/validation tool”. The same queries were made but instead of “tool” using other related words such as “web application”, “software” and “platform”. In fact, there are some related platforms, which also help validate semantic similarity measures, but based on the idea that the platform validates a user defined measure itself, usually through automatic means, and not on the idea of producing a manual gold standard. Two such examples include CESSM, (Collaborative Evaluation of Semantic Similarity Measures) whose purpose is to evaluate GO-based semantic similarity measures, but in this case, a dataset containing properly characterized protein pairs is used as means of comparison for other measures [29]. Another one is Sematch, a framework used to compute semantic similarity scores of concepts, words, and entities for Knowledge Graphs. In Sematch, evaluation of a semantic similarity measure is made by picking one dataset, out of a group of datasets, containing manual similarity scores and calculating the Pearson or Spearman correlation coefficient [30]. In both these tools, datasets are installed *a priori* into it, so users do not have the power to provide new ones, they simply test how their measures perform, contrary to MVHT which does not evaluate measures directly, instead it allows the creation of gold-standards of manually assigned similarity scores which can be used to do the validation.

2.6. Tools/Programming languages used

Several programming tools were used for different aspects of this work. Client side processes were made using Javascript. Server side was made with PHP, one of the oldest, and once, the most used language for server-side development. There were other adequate alternatives that could have been opted for instead, like Node.js, Java or Python’s Django framework, each containing its strengths and weaknesses.

Although PHP nowadays is not necessarily the cutting edge of server-side languages, it remains a good foundation, and improved performance-wise with the release of PHP 7. Secondary scripts made in order to transform and prepare the data were mostly done using Python programming language. In the following sections, I briefly explain the characteristics of these technologies and their relevance for this project.

PHP is a server-side scripting language that is popular for enabling dynamic web pages to be written and is used by established sites such as Facebook, Wikipedia and Yahoo. Scripts are interpreted on a server containing a PHP module, meaning that computer clients making requests from the server do not need to have PHP installed, only a web browser is necessary to obtain request resources from the server. It contains built-in functions that are meant to easily interoperate with MySQL and just like in JavaScript, PHP can be embedded into HTML pages. The majority of MVHT was built using PHP.

HTML/CSS are two important technologies for the creation of web pages, usually complemented by JavaScript. They are not true programming languages, such as JavaScript or PHP, due to the fact that they are simply instructions parsed by the browser. HTML, which stands for Hypertext Markup Language, has a role in defining the structure of a web page, if it will contain buttons, hyperlinks, headers, forms, photos, lists, etc. Cascade Style Sheet (CSS) is a style language responsible for the presentation of an HTML document. It determines the size, colors, layout, among other things of its elements and the style can be shared among different pages.

JavaScript is a programming language essential for any interactive web application alongside with HTML and CSS, it is used in the client-side components for the majority of websites although JavaScript is not limited to client-side, it can also be used in the server-side or in software. Where it truly shines is by turning static web pages into interactive ones, i.e. hiding or showing elements on the page when a user clicks on them; change HTML attribute values; creating animations that move or transform certain elements; autocomplete; dropdown menus; etc.

SQL (Structured Query Language) is a declarative programming language designed to work with relational databases. The maintenance of these relational databases is usually done via relational database management systems (RDBMS). MySQL is an example of an RDBM and it includes an SQL server, a programming interface, client programs to access the server and administrative tools.

Python is a general purpose programming language with an object oriented design, an easy to learn syntax, and furthermore, its methods and functions resemble natural English language, which increase readability. Python offers a plethora of packages, so chances are that when developing something, it is possible to find some online free resource capable of helping with that task. Lxml is a library specialized in processing XML and HTML in python, based on the ElementTree API. Another library for handling large datasets efficiently is Pandas, even though the syntax is very distinct when compared to the python code, upon surpassing the initial learning slope, it can become a powerful tool for customizing and making data more flexible.

JSON (Javascript Object Notation) is an interchange data format derived from JavaScript programming language. It became increasingly more popular due to the simplicity in its grammar, when compared to XML, which allowed JSON to be highly interoperable. Data is structured in either ordered lists of values or collections of attribute-value pairs, in other words, they are presented in the form of arrays or objects.

3. Methodology

3.1. Format

One of the goals of the present tool is to enable researchers to include their datasets into its back-end database. These datasets must abide by the rules of the format and as such I needed to create a format that conforms to different types of datasets and that is flexible enough to faithfully represent the information future dataset owners may require. Some examples:

- a dataset of phenotypes related to hereditary diseases and their respective annotations, like the genes and diseases associated with them.
- a dataset of proteins with annotations for their family, 3D structure, number of amino acids, method of discovery, etc.
- a dataset of animal species, annotated with their taxonomic classification, physical characteristics, habitat, geographic range, type of reproduction, behaviour, food habits and their negative and positive impacts on the environment.

To support all these use cases, and because a generalizable format makes the application more appealing to a larger set of users, I defined the format taking into account the characteristics of the four datasets used as case studies (KEGG Pathways, BioModels, OMIM and CRAFT -- see section [3.2](#)).

My focus was making the format as simple as possible while at the same time having all the information deemed important in the datasets (their entities and annotations) and cross-references to other sources. I selected JSON to be the base format, because it is a text-based format that is more compact with a relatively simple syntax, that also has the advantage of being language independent, which made it a good fit for the purpose.

MVHT looks for an optimized way to collect results from as many curators as possible, meaning it is important to try and explain to these curators what the study is about and what they should take into consideration before comparing entities. To accommodate for this, dataset owners must provide a question that curators will see at the time of making manual comparisons. For all the case studies in my work, the manual validation question introduced into the application was “How similar are these two X”, (where X is “pathways” for KEGG Pathways, “models” for BioModels, “diseases” for OMIM and “articles” for CRAFT). These questions are somewhat vague. For example, in CRAFT, by simply asking how similar the two articles are, some users may give more weight to the type of methodology used, while others focus more on the field of the experiment and what type of organisms or entities they are dealing with. Dataset owners should be more mindful of the question they include in the dataset, as to decrease the ambiguity and improve inter-curator coherence. Another possible question for a dataset, which depends on the ultimate goal of the dataset owner, could have been, “Do these diseases reveal a common pathophysiology?” for a dataset of diseases with annotations of genes as well as biological processes that are associated with or cause them.

3.2. Datasets

To evaluate the tool that was developed in this project, it was necessary to experiment on how it would perform in displaying information from different sources. From the high number of biological datasets that include annotated entities, I selected four datasets as my use cases:

- KEGG PATHWAYS
- BioModels
- CRAFT annotated papers
- Diseases (OMIM)

The general idea was to pick data repositories that had distinct types of annotations to see if the tool could be proven effective in displaying that data in a coherent manner. The datasets are explained further below.

3.2.1. KEGG

The first case study was made with data taken from the KEGG PATHWAY repository, using its API. In this dataset, entities correspond to human pathways and each entity is annotated with concepts from KEGG BRITE ontology database. Each entity contains a name, description, URL and lists of annotations from different types (diseases, drugs, compounds and genes) as it is represented in Figure 1.

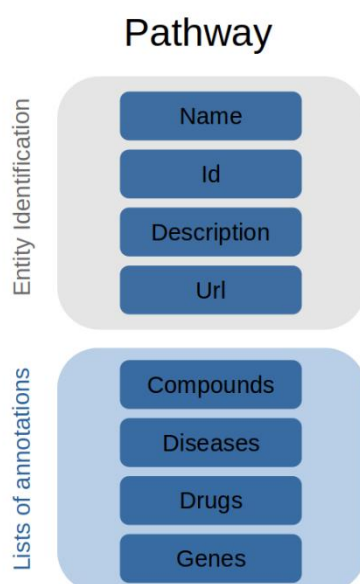


Figure 1 - Representation of a pathway entity and all its annotations

I used the web service at <http://rest.kegg.jp/list/pathway/hsa> so I could save a list with all the human pathways. Then for all entries I used the API command http://rest.kegg.jp/get/<pathway_id> and saved the contents into text files, one for each pathway. From these text files I picked the KEGG terms associated with *disease*, *gene*, *drug* and *compound* as the annotation terms. In order to do that I searched for those words in the text file which would appear at the beginning of the line, and then extracted all the terms until a new section was reached in the file (as illustrated in Figure 2). IDs and URLs for these KEGG terms were extracted as well. In the end I obtained a dataset with 337 entities, one for each pathway, each associated with a name, an identifier (KEGG ID), and a URL to the KEGG pathway entry page source. Not all pathways had a description for it, so some entities had the field *description* set to null (in the tool this section would appear empty). Some of these pathways had only annotations of IDs, with missing names for the compounds, so in this case an extra step was necessary in order to match the ID with its entry in the KEGG COMPOUND database.

PATHWAY_MAP	hsa00010	Glycolysis / Gluconeogenesis
MODULE	hsa_M00001	Glycolysis (Embden-Meyerhof pathway), glucose => pyruvate [
	hsa_M00002	Glycolysis, core module involving three-carbon compounds [P
	hsa_M00003	Gluconeogenesis, oxaloacetate => fructose-6P [PATH:hsa00010
	hsa_M00307	Pyruvate oxidation, pyruvate => acetyl-CoA [PATH:hsa00010]
NETWORK	nt06017	Glycogen metabolism
ELEMENT	N00731	Glycolysis
	N00733	LDHA deficiency in glycolysis
	N00735	EN03 deficiency in glycolysis
	N00737	PGAM2 deficiency in glycolysis
	N00739	ALDOA deficiency in glycolysis
	N00740	PFKM deficiency in glycolysis
	H00069	Glycogen storage disease
	H00071	Hereditary fructose intolerance
	H00072	Pyruvate dehydrogenase complex deficiency
	H00114	Fructose-1,6-bisphosphatase deficiency
	H00664	Anemia due to disorders of glycolytic enzymes
	H01071	Acute alcohol sensitivity
	H01096	Pyruvate kinase deficiency
	H01267	Familial hyperinsulinemic hypoglycemia
	H01760	Hepatic glycogen storage disease
	H01762	Muscle glycogen storage disease
	H01939	Glycogen storage disease type I
	H01945	Glycogen storage disease type VII
	H01997	Pyruvate dehydrogenase E1-alpha deficiency
	H01998	Pyruvate dehydrogenase E1-beta deficiency
	H01999	Pyruvate dehydrogenase E2 deficiency
	H02000	Dihydropyridine dehydrogenase deficiency
	D00123	Cyanamide (JP17)
	D00131	Disulfiram (JP17/USP/INN)
	D07257	Lonidamine (INN)
	D08970	Piragliatin (USAN)
	D11342	Dorzagliatin (USAN)

Figure 2 - Extraction of annotations was done via finding the desired property names for each pathway text file, i.e. "DISEASE" and then capture every line containing a KEGG term until reaching the next property, which in this case corresponds to "DRUG".

3.2.2. BioModels

The second case study was BioModels data (I used the data from the latest version, which is dated from 2017³). 618 was the number of curated models used to retrieve the information. The entities in this dataset are models of biological or biomedical systems and annotations are taken from several ontologies including the Systems Biology Ontology (SBO). Each entity has a name, ID, description, URL, and lists of "Species" (chemical compounds), "Reactions" and cellular "Compartments" (see Figure 3). The models also contain elements such as mathematical functions and values for the parameters in said functions, however I decided not to include this information. The reason for this decision was because models with multiple reactions in the tool would each need to be separated into two sections, one for the mathematical equation and the other for the parameters with the respective values, however to curators it would mean that they would see a list of parameter values, which they had to associate to the variables in those equations, in an unorganized manner. Models are in XML format and files can be in one of three different namespaces, meaning element names and the overall file structure differed between them. To do this procedure I used the python package lxml from the ElementTree API which specializes in XML processing. Adding to the varying XML namespaces, each author who develops the model can define the names for the variables relative to the species, reactions and compartments in a different and subjective way, making this process of data-mining difficult (see Figure 4).

³ ftp://ftp.ebi.ac.uk/pub/databases/biomodels/weekly_archives/2017/

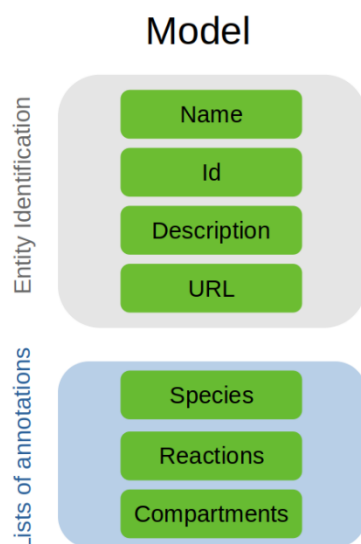


Figure 3 - Representation of a biological model entity and all its annotations.

Several parallel approaches were used to extract as much information as possible in this dataset: I first collected only the value associated with the XML “name” attribute, but some annotations did not have a “name” attribute, only an “ID” attribute, while in others the names were acronyms or the same as the ID. Another approach was to make use of an attribute called `sboTerm` that could be linked to the corresponding entry in SBO, however, there were a great number of models that also did not have a `sboTerm`. A third approach was to use links contained in the XML files that would redirect to an entry in some type of repository, i.e. InterPro, UniProtKB, GO, Reactome, ChEBI, etc. The problem was that certain authors did not provide any link to other resources, while other times links were relative to entries that were either obsolete or non-existent. Also, even if the links were present and functional, most times they redirected to entries that did not give too much information that could differentiate the annotations. For example, in the model with the identifier “BIOMD0000000001”, most species are linked to the entry named “nicotinic acetylcholine receptor”, from InterPro ([IPR002394](https://www.ebi.ac.uk/interpro/entry/interpro:IPR002394)). If one was to use this to name annotations in the dataset, then multiple entries would have the name of this entry making its identification difficult to extract. Since most models had no name attribute or `sboTerm` but contained links to other resources, the third approach was selected as a priority to extract the annotation names. When not available, `sboTerm` attribute was used and if this attribute was also non-existent, the name attribute on the XML tags was used instead.

```
<species id="BLL" name="BasalACh2" metaid="_000003" sboTerm="SBO:0000297" compartment="comp1">
<species id="BLL" name="BasalACh2" metaid="_986142" compartment="comp1">
<species id="F26DP" name="F26DP" constant="false">
```

Figure 4 - Each species is taken from a different model. The one in the top contains a `sboTerm` attribute which allows the concept name to be retrieved from SBO. The middle species has no `sboTerm` so the only option here would be to use the name “BasalACh2” which is not very clear as to what it refers to. In the bottom species there is also no `sboTerm` attribute and the `Id` and `name` attributes are the same “F26DP”.

3.2.3. CRAFT

CRAFT⁴ is the third case study for the application used to test the tool. This repository is composed of 97 scientific papers about mouse genomics, and annotations for multiple ontologies were extracted for each one of them. Annotation group names were “Chemical entities” (ChEBI), “Cells” (CL), “Biological processes” (BP), “Cellular Components” (CC), “Molecular Functions” (MF), “Molecular Processes” (MOP), “Anatomical entities” (UBERON), “Proteins” (PRO) and “Genetic terms” (SO). NCBITaxon was another ontology present in CRAFT annotations, but I decided not to use it in my work since for the vast majority of articles, the only annotations present were “mouse”, “human” and its plural forms, which would not contribute anything to the comparison process. A representation of this entity can be seen in Figure 5. Just like in BioModels, CRAFT annotations were taken from XML files, one for each article, but here the schema was constant for all of them, facilitating the extraction process. XML files were organized in different folders, one for each ontology. For the *description* I used the papers’ abstract, for the *name* I used the title of the article and for the *URL* I used the pubmed link where the article can be found. Also, in the process of annotating the articles, words in the singular and plural forms were annotated (ex. gene and genes), even though they correspond to the same concept, so for the creation of this dataset only unique identifiers were considered, excluding plural forms.

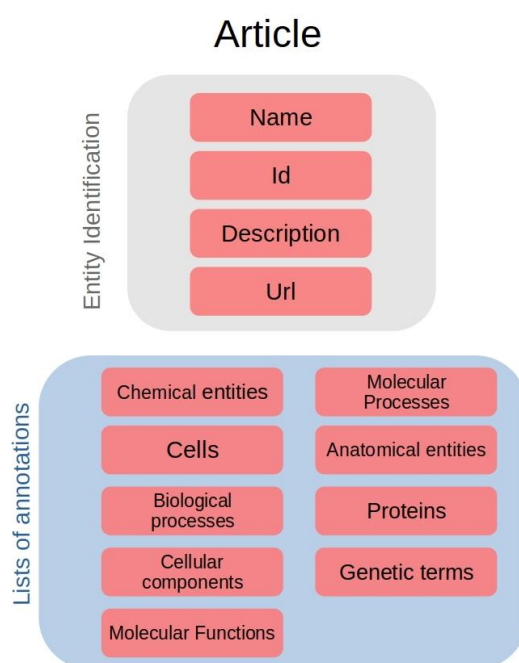


Figure 5 - Representation of an article entity and its annotations

3.2.4. OMIM

The last dataset constructed, containing diseases using OMIM as its data source, was meant to be given to physicians before the covid-19 pandemic occurred. With diseases, information like its clinical features, diagnosis, pathogenesis, biochemical features, molecular genetics, etc is relevant. First I selected 20 relatively well known and common diseases (full list of diseases can be found in [Appendices-A. List of diseases](#)), that could be found as an entry in OMIM⁵, extracted all textual information relative

⁴ <https://github.com/UCDenver-ccp/CRAFT>

⁵ <https://www.omim.org/>

to said entry using OMIM's API and stored it in text files. Then I used an API endpoint from BioPortal called *annotator* on each of the files created from the previous step. What annotator does is examine the text and find classes that match a certain word or expression on the text, for an ontology, or multiple ontologies. DOID (Human Disease Ontology), SYMP (Symptoms Ontology), ChEBI, CL (Cell Ontology), PRO (Protein Ontology), UBERON and GO were the ontologies used to grab the annotations (see Figure 6). An issue that rises from this selection is that certain ontologies "share" the same concepts, for example, the concept "protein" is defined either in PR, CHEBI, or GO; the concept "cell" also exists as a class in CL or in GO. This can make it so that curators see some common annotations between different tabs on the manual curator view (see section 4.3.1).

Since this is a natural language processing (text-mining) step, there are certain occurrences in the text that do not necessarily reflect a strong association between the OMIM disease and the ontology concept, because sometimes there are negative associations. As an example, when looking into the OMIM's entry page for Alzheimer's disease ([#104300](#)), there is a segment in the text explaining that studies were made that found no association between patients with Down syndrome and an excessive number of dementia (suggestive of Alzheimer's disease) cases in their families, which results in the concept "Down syndrome" being grabbed by the annotator API for the DOID ontology, and making it seem as if it is a disease associated, or caused by Alzheimer's disease, even though that is not the case. Similarly, for hemophilia A (entry [#306700](#)) the differences in symptoms between this disease and von Willebrand disease (entry [#193400](#)) are highlighted, meaning that the API will grab not only "von Willebrand disease" as a related disease (they are in fact both diseases that affect blood coagulation) but it will also grab its symptoms from SYMP ontology and compile them together with the other symptom annotations, as well as annotations from other domains.

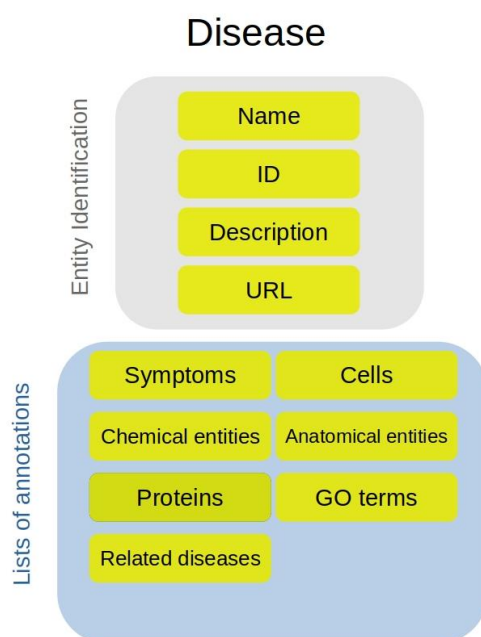


Figure 6 - Representation of a disease entity and all its annotations

3.3. Application

3.3.1. Architecture

MVHT has two types of users:

- **Dataset owners**, the primary target of the tool, are researchers that develop automatic similarity measures between entities, and want to test their measure against curated similarity values.
- **Curators**, who may be experts in a certain field (for example researchers of that field), or even people not related to biomedical science, depending on the technical level of the dataset's subject.

Dataset owners must prepare and upload their own datasets. The application will then process and store them in an internal database, generate shareable links for dataset owners to share with curators, form entity pairs (see section [3.3.2](#)) and deal with content display. Multiple datasets can be inserted into the application, each one belonging to a different dataset owner and managed by him.

Curators, the other type of MVHT users, are the experts that will use the tool in order to help dataset creators (they gain nothing from the experience except the notion of aiding dataset owners in attaining their objectives). Accomplishing this will enable the creation of “gold-standards” (see section [2.4](#)), which dataset owners can use to validate their measures of similarity.

The application can be divided into two main parts, one for each type of user, which communicate with each other through a central application layer (Figure 7):

- The **dataset owner dashboard**, allows users to (i) insert a dataset into the application (ii) inspect the state of each dataset (its size, the number of pairs that have been compared by manual curators, etc.) and (iii) download the results.
- The **manual curator view**, which asks users for a few personal details (name and expertise) and presents them with pairs of entities so that they can compare them and thus contribute to building a gold-standard of manual comparison values.

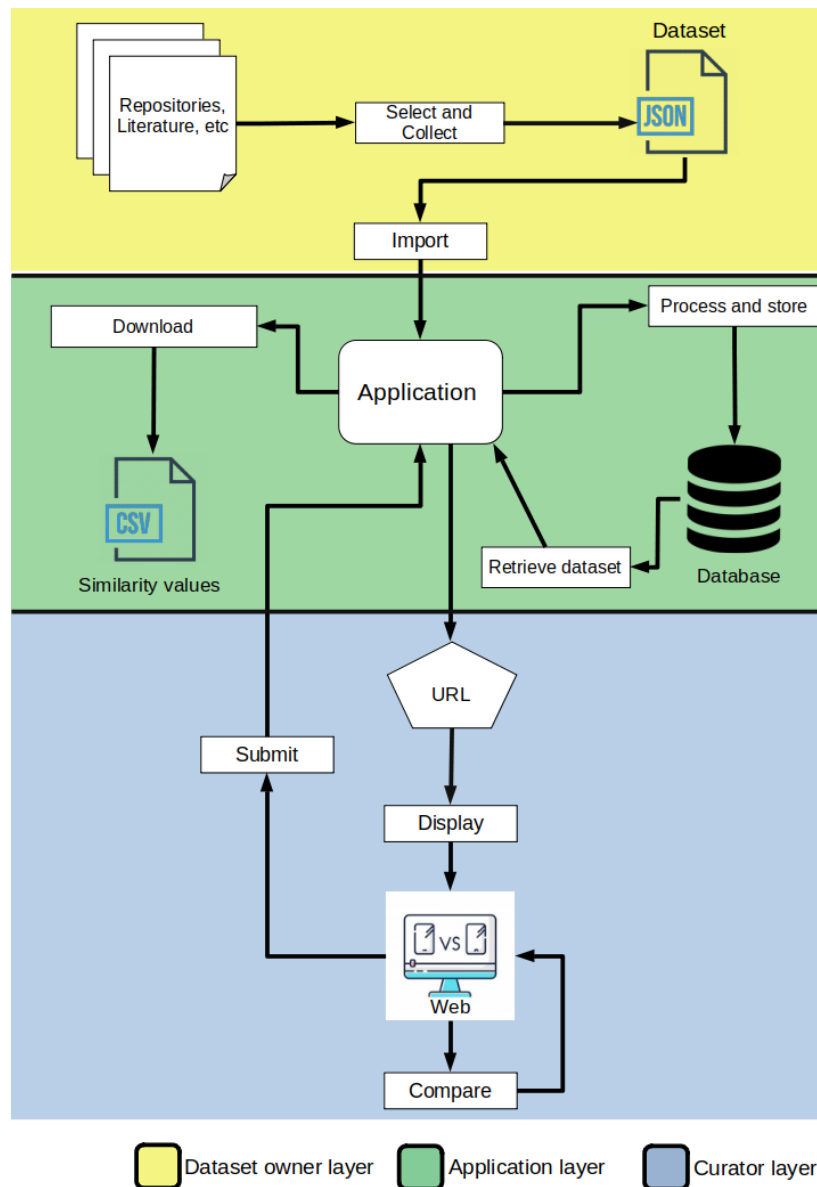


Figure 7 - MVHT architecture and workflow

3.3.2. Pairing Strategies

Dataset owners are presented with the option of picking how pairs of entities are shown to curators. This is important because it allows them to test how their measures perform when subjected to different pairing factors. Also, it may guide other researchers into picking a semantic similarity measure that has been validated with a pairing strategy that better corresponds to their needs.

For now, there are three pairing options (see Figure 8):

1. **Totally random** - Two entities will be selected randomly from the dataset, the only condition being they must be different from each other.
2. **Fixed order** - This method focuses on having distinct pairs with similarity values assigned, so only pairs without any similarity values are shown to users. If all pairs have already been compared, then it resets and repeats the process until all of them have been compared again.
3. **Fixed order per user** - Every user will be given the same entity pairs in a specific order. This method focuses on having multiple comparisons between the same entities.

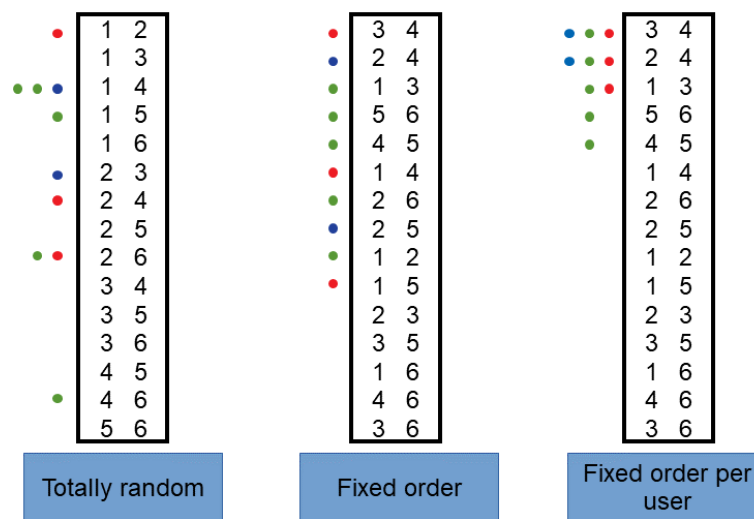


Figure 8 - Representation of the current pairing strategies implemented. Adjacent numbers inside the vertical rectangle represent pairs of entities and each dot on the left with a different color symbolizes a different curator. If a pair has been given a similarity value then a dot, or multiple dots, appears next to it.

For both *fixed order* and *fixed order per user* options, the tool will grab entities provided in the dataset to make all possible combination pairs and shuffle them. The tool will present pairs in the order generated, so although the pairing is random, the order is not. What differs one method from the other is how the pairs of entities will be displayed. If the dataset owner chose the *fixed order* option, entity pairs with no comparison values will be given priority, meaning that if a pair of entities has been given a value of similarity it will not appear anymore until all other pairs have been given answers. What this method tries to achieve is assigning a similarity value to all pairs of entities, so that in the end the amount of similarity values for each pair is approximately equal among all of them. When the last pair of the randomly generated pairs is given a similarity value, then it resets and goes back to the top of the list, repeating the process until all of them have now two answers.

In case a dataset owner chooses *fixed order per user*, pairs will be given to curators in a predefined order just like before, but now each curator runs through the list independently of each other, meaning that every time a new curator accesses the link, he will start from the beginning of the chain, unlike the *fixed order* option where new users would start in the pair subjacent of the last pair with a similarity value given. The objective here is enforcing multiple answers for the same pair, where dataset owners can then obtain a mean value for that pair when retrieving the results, as well as compute a measure of coherence between the curators.

In the future, this step can be improved by adding more diverse and complex pairing strategies. Therefore, the tool accommodates for this and is built in a way that allows for the addition of more options without compromising already existing ones or changing too much of its architecture.

3.4. Data collection

Two types of results are expected from my work: the development of MVHT as a tool (including getting feedback from pilot users and improving the tool accordingly), and the generation of a gold standard with one of the datasets collected for the case studies.

Initially, the idea for this project was that once the tool was practically finished and working, doctors would be given a link to make manual comparisons using a dataset of diseases (see section 3.2.4). However, due to the pandemic of the new COVID-19 virus, this was deemed untimely (I still intend to release the dataset and to ask physicians for their input, just not before the end of this master's thesis). Instead, colleagues from biological fields provided their assistance in the dataset of scientific papers (CRAFT). There were two stages of pilot users asked to give feedback: on the first stage 5 people simply experimented and looked into what the tool offered and gave their opinions on what they thought could be improved visually and in terms of user interaction. Then later, 6 people (the same 5 people from the first stage and 1 new one) were asked to use the tool and make comparisons for a dataset with annotations of articles taken from CRAFT repository. These manual values were later downloaded from the tool as a CSV file.

3.5. Comparison with automatic measures

After obtaining manual similarity values from curators, the next step would be to compare them to automatic values generated computationally and check for a correlation. This step is primarily meant to infer if a dataset owner can quickly test for a correlation after downloading the results from the tool. As a case study, in this work I selected the *simGIC* measure [31], a graph-based measure that is an extension of the *simUI* measure [32].

In *simUI*, given two articles, represented as entities A and B, each annotated with a set of concepts, an induced graph can be formed, containing all direct annotations in the entry as well as the indirect ancestors up until the root node of the ontology, which I will call $\alpha(A)$ and $\alpha(B)$. Then, semantic similarity is calculated by using the number of terms intersecting in the two graphs and then divide them with the number of terms they have together:

$$simUI(A, B) = \frac{|\alpha(A) \cap \alpha(B)|}{|\alpha(A) \cup \alpha(B)|} \quad (3.1)$$

SimGIC is similar, but additionally it weights each concept by its information content (IC), instead of just counting the number of concepts in each set:

$$simGIC = \frac{\sum_{t \in \alpha(A) \cap \alpha(B)} IC(t)}{\sum_{t \in \alpha(A) \cup \alpha(B)} IC(t)} \quad (3.2)$$

simGIC was only calculated for entities that were compared by curators in the CRAFT dataset, resulting in 67 pairs. Finally, Pearson's *r* correlation coefficient was used to check for linear correlations between the manual values and the computed values. Pearson's correlation is a correlation coefficient meaning it is a statistic that quantifies the relationship between two sets of values. It can range from -1 to 1, where -1 indicates a perfect negative relationship, 0 indicates there is no relationship between the data and 1 indicates a perfect positive relationship. Pairs had *simGIC* calculated for all 9 ontologies and the mean of those values was used as a final automatic similarity score. Pearson's *r* was then calculated using these means against the human values.

4. Results and Discussion

In this section, I present the main results obtained in this work. The section first describes the dataset format developed to be used by the tool. The second section describes the major characteristics of the four collected datasets, whose collection process was instrumental in defining the format. The third section describes the details of the tool implementation. The fourth section describes the results obtained by using the tool with one of the four collected datasets. The fifth section presents the results of correlation between those collected results and the automatic measures of similarity presented previously.

4.1. Dataset format

To insert a dataset into the application, the dataset owner must upload a JSON file according to a certain schema (documentation about the schema can be found on the “Format Guide” page of the website⁶ and in [Appendices- B. JSON schema](#)).

The JSON file representing a dataset must contain an object with two attributes: the key “question” associated with a string of a question that dataset owners want human curators to have in mind when making comparisons; and a key “entities” associated with the list of entities of the dataset (Figure 9).

```
{
  "question": "How similar are these proteins?",
  "entities": [ ...
]
```

Figure 9 - “Outer layer” of the schema with the keys question and entities datasets must have.

Every entity inside the “entities” list is an object that can have the following attributes: “name”, “id”, “description” and “url”. Apart from “name” and “id”, other attributes are optional. Besides these first four keys, the user can decide which other keys to include. These must be associated with lists of annotations and serve as a way to organize the annotations in sections, within the entity.

This is the part of the format that provides generalization since it allows the user to insert things like diseases, metabolites, species, symptoms, etc., which is why the format can be used to describe datasets about such disparate things like pathways, mathematical models of biological systems, articles, and even other non life-science related entities, such as movies.

These custom attributes define the categories of annotation for an entity and are always lists. Each item in the list can either be a string, which represents the name of the concept annotating the entity within that category, or an object with two attributes (“name” and “url”). In this case, the “name” property is mandatory, while “url” is optional (Figure 10).

⁶ <http://mvht.lasige.di.fc.ul.pt/format.html>

```

{
  "question": "How similar are these proteins?",
  "entities": [
    {
      "name": "Protein A",
      "id": "001",
      "url": "http://examplesite/001",
      "description": "Protein involved in cellular respiration reactions",
      "functions": [
        {
          "name": "function A",
          "url": "http://examplesite/f001"
        },
        {
          "name": "function B",
          "url": "http://examplesite/f002"
        }
      ],
      "processes": [
        "process A", "process B", "process C"
      ]
    },
    {
      "name": "Protein B",
      "id": "002",
      "description": null,
      "functions": [
        {
          "name": "function A",
          "url": "http://examplesite/f001"
        },
        "function C"
      ],
      "processes": [
        "process D", "process E", "process F"
      ]
    }
  ]
}

```

Figure 10 - Example dataset to demonstrate the type of information that can be processed by the tool. Here the dataset contains two made-up proteins A and B, where it can be seen that for key “functions” there are objects inside them, with properties “name” and “url”, while for key “processes” it consists of a list with annotation names, showing the different ways users may want to express their data.

4.2. The four case studies

I collected four datasets as part of this project which are available to download online⁷. This collection was done mainly with two purposes:

- to study the characteristics of annotated datasets that may be relevant in deciding the features of the tool;
- to test the tool with real human users in order to validate its usefulness.

In this section, I will describe the four case study datasets, talk about their limitations, strengths and how well they shape themselves into the tool. The datasets are also statistically described in terms of three measures, which capture their multidisciplinary [33]: the fraction of entities with annotations in each domain (Coverage), the average number of annotations in these entities for a domain (Volume)

⁷ <https://github.com/pedrojmds/MVHT>

and the number of distinct concepts in that domain used to annotate the entities (Diversity). The idea is that these measures can help to better understand the structure of a dataset. For example, if a domain contains low coverage, volume and diversity, then it shows that the domain is not very relevant to the dataset as it gives no new information to the entities, meaning that it could be removed from the dataset and not be used for the semantic similarity calculation (see section 3.5). Alternatively, higher coverage and volume, but low diversity could mean that there is a substantial number of common annotations between entities, relative to that domain.

4.2.1. KEGG Pathways

The average number of annotations in each entity was 138. This dataset contained a high number of terms in some domains, such as “Gene” and “Drug”, for many of the entities; but there were very few pathways with many terms in common, which curators could find it difficult to discern if the pathways are similar or not without previous background knowledge on them. Also, not all pathways contained annotations for the same custom properties, i.e. the Mitophagy pathway⁸ contains only the list of genes, but it is lacking annotations for “Disease”, “Drug” and “Compound”. Since some pathways had no KEGG terms annotated for certain domains, just like in the example above, this dataset presented too many pairs with zero matching annotations, even though, upon closer inspection, the entities do have drugs associated with them, just not represented in the KEGG database. This is the case for the Prion disease pathway, with no drug annotations, but in fact there are drugs known to have been tested and used in treating the condition [34], [35].

Looking at Table 1, “Drug” is the domain that is less represented out of all pathways. The domain with the highest average number of annotations was “Gene”, with a value more than 3 times larger than the second one (“Drug”), which means that this is a domain highly represented in the dataset, and almost 16 times larger than the domain “Disease”. In terms of diversity, results showed a large number of distinct annotations for all domains, which relative to the volume values and the high number of 56616 possible pairs between all 337 entities, make it so that common annotations do not appear often, and if they do, it is on very low numbers. This factor contributed for KEGG’s dataset to not be given to pilot users, since with a totally random pairing strategy (see section 3.3.2) it would be difficult to obtain similar pathways, even using fixed order and fixed order per user.

Table 1 - Statistical metrics coverage, volume and diversity were calculated for each domain in KEGG.

Domain	Coverage	Volume	Diversity
Disease	0.86	16.81	1043
Drug	0.64	79.51	4239
Gene	0.97	266.21	7978
Compound	0.82	52.21	3281

4.2.2. BioModels

Due to a lack of consensus, as different encoders use different acronyms to represent the same entities, this dataset did not show a very “clean” set of names. When using the names provided by model authors,

⁸ https://www.genome.jp/dbget-bin/www_bget?hsa04137

reactions would often be defined as “React0”, “React1”, etc, while if using the `sboTerm` attribute contained in the XML tags it would often lead to nonspecific names such as “Material entity” or “Macromolecules” for the entire set of species. Also, due to differences in the schema versions of the XML files, the data contained in specific elements varied, where in certain files some elements contained attributes that others did not (as shown in Fig. 3). Other times no description was given for the model, and only the abstract from the article where the model was created was present. All these details in conjunction to a lacking nomenclature present in some models, made the extraction of data for many entities difficult and vague, therefore making BioModels a nonoptimal dataset to insert in the tool.

The inconsistencies in the way XML files stored the annotations (see section 3.2.2), made it so that the extraction of some annotations was not possible, so the statistical measures in Table 2 are not a faithful representation of the real values, for example, none of the domains covered the whole dataset even though the reality is that every model had annotations in each domain. There was a large number of unique annotations in both Species and Reaction domains, however the average number of annotations per entity was relatively small, approximately 27 and 21, respectively. This high diversity together with lower volume values decrease the chance of finding common annotations in pairs of entities. Models of biological systems usually take place in a specific location, either in a cell, a component of a cell, in the extracellular space, etc. This justifies the volume of only 2 annotations on average per entity for the Compartment domain. Each entity had an average of 53 annotations.

Table 2 - Statistical metrics coverage, volume and diversity were calculated for each domain in BioModels.

Domain	Coverage	Volume	Diversity
Species	0.93	26.84	7173
Reaction	0.77	21.33	8037
Compartment	0.82	2.3	149

4.2.3. CRAFT

Common annotations between these entities can be mostly found in the *genetic terms* category (SO) due to the fact that more generic concepts appear so often, such as “genes”, “genome” and “allele”.

Table 3 shows that most articles contained at least one annotation from every ontology, only the Molecular Functions ontology from GO (GO_MF) and Molecular Processes Ontology (MOP) contained a lower coverage of 66% and 68% respectively. Additionally, both these ontologies presented very low values of volume and diversity, meaning that these ontologies could be removed from the dataset without a lot of information being left out. The ontology with the highest number of annotations throughout the articles was Uberon, an ontology that represents anatomical structures of animals, with an average of 95 annotations per article, followed by GO Biological Processes (70.14) and ChEBI (64.64). Protein Ontology (PRO) had the highest number of unique concepts in the ontology, with 1231 terms, followed by Uberon with 1041 unique concepts. The average number of annotations per entity was 143, the highest of all datasets.

Table 3 - Statistical metrics coverage, volume and diversity were calculated for each ontology in CRAFT.

Ontology	Coverage	Volume	Diversity
PRO	1.00	53.94	1231
Uberon	0.99	95.51	1041
GO_CC	0.99	27.49	248
GO_MF	0.66	2.10	5
GO_BP	1.00	70.14	721
ChEBI	1.00	64.64	587
CL	0.97	25.79	285
SO	1.00	54.43	198
MOP	0.68	3.62	18

4.2.4. OMIM

This dataset, just like CRAFT, is multiple-ontology, however unlike CRAFT whose annotations were selected manually by a team [21], OMIM's dataset annotations were grabbed using an API endpoint, by providing 7 different ontologies in its parameters. The average number of annotations per entity was 102.

Looking at Table 4, only ChEBI ontology had annotations for every disease on the dataset, however all ontologies had high coverage values (≥ 0.90). The average number of annotations was relatively similar between all ontologies, only CL had a slightly lower value of 12.15. This is not unexpected, as the cellular terms are used to describe a disease less often than drugs, symptoms or proteins. In terms of diversity ChEBI presented the highest value with 351 distinct annotations, followed by DOID with 270 and GO with 202.

Table 4 - Statistical metrics coverage, volume and diversity were calculated for each ontology in OMIM.

Ontology	Coverage	Volume	Diversity
ChEBI	1.00	22.75	351
CL	0.90	12.15	47
DOID	0.95	22.1	270
GO	0.95	22.1	202
PR	0.95	20.5	126

SYMP	0.95	18.0	98
------	------	------	----

4.3. Front-end implementation

In this section, I will fully describe MVHT, its functionality and the way the users interact with it. I start by describing the current version of the tool, and then explore some of the interface details that changed throughout the project based on user feedback.

4.3.1 The current version

The **dataset owner dashboard** can be accessed by visiting the link <http://mvht.lasige.di.fc.ul.pt/>, which shows a description of the tool, its purpose, what other pages there are and the expected dataset format for data to be inputted. The application has a “sign on”/”sign in” form that handles authentication and allows users to register with a dataset owner account. Once logged in, the dataset owner enters the dashboard, where he can upload a dataset and pick what type of pairing strategy he desires from the available options (see section 3.3.2). Additionally, the dashboard contains a table of previously uploaded datasets, with analytics for them: generated dataset code, name, the number of entities it possesses, the total number of annotations, how many pairs of entities have been manually compared, a unique shareable URL to give to curators, and a button to delete the dataset (see Figure 1). Another necessary aspect of the tool is the ability to show the results from answers given by manual curators, so from the dashboard table a dataset owner can click on a dataset and go to a page that displays the results given by curators. There is also an option to download the results as a CSV file.

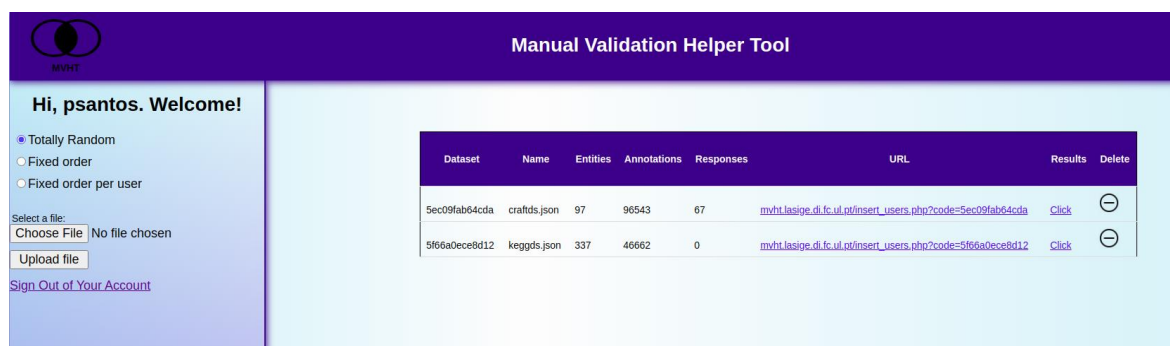


Figure 11 - View of the dataset owner dashboard after a user logs in. On the left there is a panel with the pairing options, and a place where a dataset can be uploaded, as well as a link to sign out of the account. On the right there is a table with information about each dataset already inserted, and an option to look at the results, as well as a delete button.

The unique dataset URL can be given to manual curators so that they can contribute to building the gold-standard of similarity. This URL is a practical mechanism through which dataset owners can quickly gather answers from manual curators, as they can distribute the URL to their connections and thus ask for participation with the intent of obtaining a gold standard. Upon entering the application through this link, the curator must first insert a name and a field of expertise (see Figure 12). For this type of user, no form of login or account creation was implemented because in a realistic scenario people given the link will most likely only access the site once, and maybe a small portion will only do a certain number of comparisons, leave and then decide to continue later, so having them create a password and username would not be optimal. What I decided to do was ask for a name and a field of expertise as soon as the link is visited, and automatically register a new user into the database.

Figure 12 - Form curators are required to fill before proceeding with the pair comparisons.

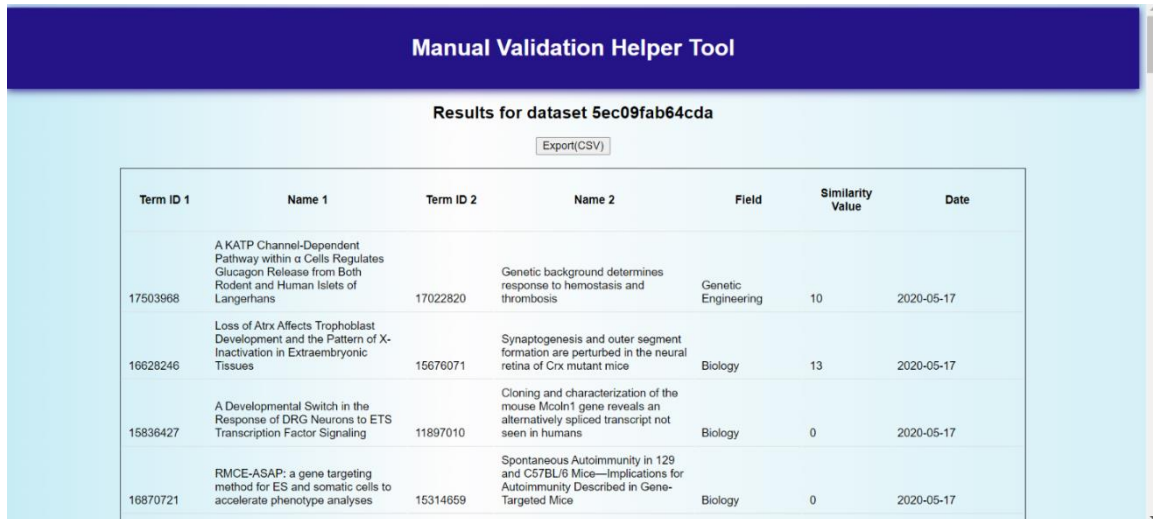
This creates the issue that if an already registered curator decides to come back later, and visits the link, he will be asked to insert the same information as before and a new user is created, possibly resulting in duplicates. This scenario could be mitigated by the use of client-side data storage mechanisms, like HTTP cookies, web storage, or web cache, but they were regarded as not ultimately needed for the application at this stage, since the expected number of returning curator users is low, and the benefits of being able to trace the same user across similarity value assignments is not clear.

After this step curators will be asked to compare two random entities from the dataset in the **manual curator view** (see Figure 13), a page containing a random pair of entities specific to that dataset (computed based on the randomization strategy selected by the dataset owner). In this page there will be a pair of entities being compared with a description of the entities and all the groups of annotations from diverse domains, as they were provided by the dataset owner. All annotation groups are divided into different panels, all clickable so users can collapse or expand the information as they please. To make the comparisons there will be a range bar with a slider that users may click when they decide on a similarity value for that pair. After submitting, a new pair will be shown. Another feature of MVHT is the ability to click either entity names at the top of the page, or any of the annotations inside the horizontal tabs and be redirected to a source page with more information about that annotation or entity so users can learn about them. This navigation aspect is fully supported by the dataset owner, which is responsible for including proper functioning URLs in the file they upload.

Figure 13 – Manual curator view comparing two articles from the CRAFT dataset.

When a curator submits a similarity value, the tool stores which two entities were compared, the value of similarity attributed, ranging from 0 to 100, the user who did the comparison, its field of expertise and the date, which the dataset owner can view in the results page (Figure 14). The only

information about the curators that a dataset owner will have is the field of expertise. For datasets with a substantial number of answers from different people, the field of expertise data may provide some type of insight in future studies. As an example, when asking workers of the medical field and biologists to compare genetic diseases, there may be some differences in the answers between the two, as a doctor may focus more in the symptoms of the disease, whereas biologists may turn their attention to the type of mutations and where in the transcriptional machinery lies the problem.



Manual Validation Helper Tool

Results for dataset 5ec09fab64cda

[Export\(CSV\)](#)

Term ID 1	Name 1	Term ID 2	Name 2	Field	Similarity Value	Date
17503968	A KATP Channel-Dependent Pathway within α Cells Regulates Glucagon Release from Both Rodent and Human Islets of Langerhans	17022820	Genetic background determines response to hemostasis and thrombosis	Genetic Engineering	10	2020-05-17
16628246	Loss of Atrx Affects Trophoblast Development and the Pattern of X-Inactivation in Extraembryonic Tissues	15676071	Synaptogenesis and outer segment formation are perturbed in the neural retina of Crx mutant mice	Biology	13	2020-05-17
15836427	A Developmental Switch in the Response of DRG Neurons to ETS Transcription Factor Signaling	11897010	Cloning and characterization of the mouse <i>Micr1</i> gene reveals an alternatively spliced transcript not seen in humans	Biology	0	2020-05-17
16870721	RMCE-ASAP: a gene targeting method for ES and somatic cells to accelerate phenotype analyses	15314659	Spontaneous Autoimmunity in 129 and C57BL/6 Mice—Implications for Autoimmunity Described in Gene-Targeted Mice	Biology	0	2020-05-17

Figure 14 - Page containing the results for one of the datasets showing information about each comparison already made.

4.3.2 Interface evolution

MVHT was subject to changes over time, to accommodate aesthetic and functional feedback. After experimentation by pilot users, they were asked to express their opinions on what could be improved on the page intended to display the entities being compared. I will proceed to enumerate the most relevant.

Similarity slider

For the similarity slider, it was decided early on that MVHT would not show a numerical value, for users not to be too indecisive and picky between possible values when comparing entities. The issue with this decision was the difficulty to assert if the value they wanted to give in the current comparison was higher or lower than the last one made, because the range bar was filled with a single color that would match the position of the slider, so there was no visual landmark except for their perception of length. Instead, a colored gradient ranging from red (very different), to white (somewhat similar), to green (very similar) was implemented, as well as the words “Different” and “Equivalent” on the left and right side of the bar, respectively. With the color gradient users would not be disoriented as to what was the general position on the range bar of a previous pair they curated and may want to use as a guideline for the present pair, but at the same time not having too much indecision if, to them, a pair was more similar by just a couple numbers, that a numerical indicator could instigate.

The range slider was also positioned at the bottom of the page, following the scroll at all times so it can be quickly reached with the mouse and given a value to submit. The question dataset owners give in their datasets (see section 4.1) was also moved to be just above the similarity bar, in order for curators to remember and take the question into consideration at the time of deciding on a similarity value (see Figure 15). Similarly, in order to accelerate the process of curation, the submit button was made to appear at a position near where users click on the slider, so that they can quickly click it and move on to the next pair of entities (see Figure 16).



Figure 15 - Default state of the similarity bar when given a new pair of entities.



Figure 16 - Submit button appears adjacent to the place where users drag the slider.

Venn diagrams

Venn diagram icons were added to each of the custom annotation tabs (see Figure 17). The Venn diagrams help in visually understanding that the three columns of annotation refer to annotations for the left most entity, to both entities, or to the right-most entity. So if two entities A and B (which will be our sets), are being compared, after a curator clicks on a tab and sees the three columns, it becomes clear that the panel on the left contains only the unique annotations from the entity on the left:

$$B^c \cap A = A \setminus B \quad (4.1)$$

The column on the right contains only the unique annotations from the entity on the right:

$$A^c \cap B = B \setminus A \quad (4.2)$$

And the column in the middle represents the intersection between A and B, this results in only annotations which are common between them showing up:

$$A \cap B \quad (4.3)$$

Alongside the icons, annotation counters are present so that users can have an idea about the number of annotations for each entity, as well as if there are any common annotations between them, without even needing to click on the collapsible tab.

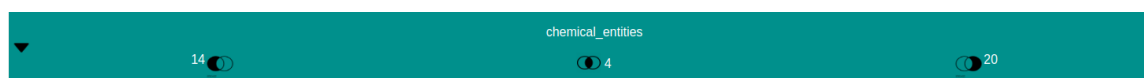


Figure 17 - Tab containing CHEBI annotations with three Venn diagrams icons, meaning that inside it there are 14 unique annotations for the article in the left, 20 unique annotations for the article in the right and 4 common annotations between them.

Vertical space

With a limited vertical space on the page, deciding on what elements to appear at all times was necessary. As it was referred above, the similarity bar was made to be fixed to the bottom, even when scrolling, and by making the submit button appear only when a user clicks the similarity bar, there is one less element filling the page constantly. Entity titles were fixed to the top alongside the description tab, however, this occupied a large portion of the space so not only was the description content made to be collapsible and hidden by default, but it was made to scroll along with the rest of the page, disappearing when the user scrolls down. These changes saved a little bit of space on the page to see the rest of the

information and also made it a less strenuous job for users to need to scroll up to refresh their memory on which two entities were being compared, or scroll to the very bottom of the page to reach the submit button every time a similarity value was given.

Collapsible panels

With the tool's potential for having any number of custom tabs as dataset owners want, page space would be occupied very quickly if all tabs and respective columns were visible as the page loaded and the user experience would be degraded. To mitigate this, all tabs, including the entity description tab were made collapsible and hidden by default, with changing arrow icons to indicate that there is more information hidden in them. This way when a user accesses the page for the first time, he will see the various custom annotation domains and decide which tab, or multiple tabs, to open and close at a time (see Figure 18 and Figure 19).

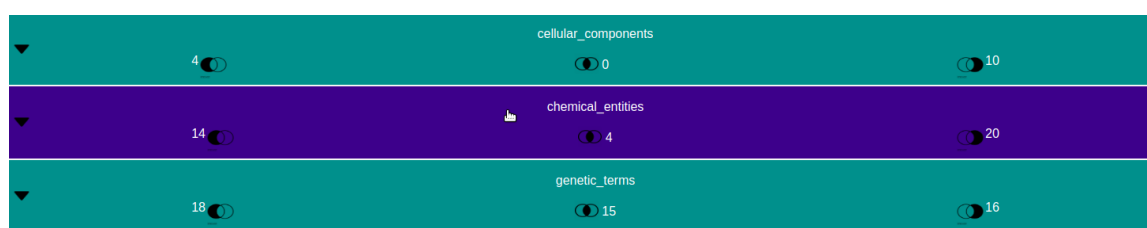


Figure 18 - Collapsed tabs with a down arrow icon on the left to signal that there is hidden content if clicked.

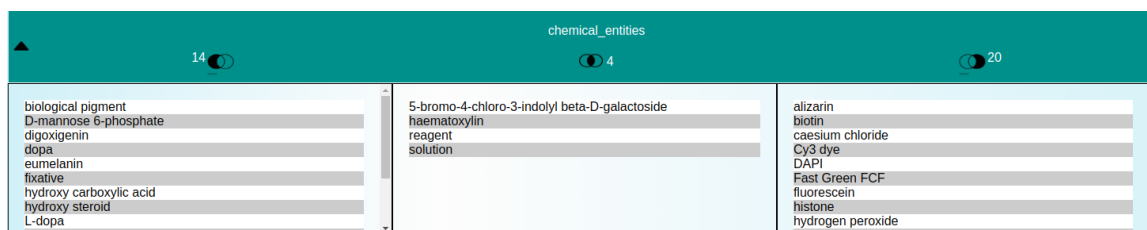


Figure 19 - Open tab containing the three columns inside and an upper arrow icon on the left to indicate it can be collapsed.

Actionable annotations

Annotation names were initially displayed alongside IDs of the concepts in their respective ontologies, however this type of information was not helpful to users, so instead, the concept IDs were removed and annotations were made clickable, so that users could be redirected to other sources that gave a better definition of what the concept represented (see Figure 20). This is particularly useful when the annotation concept is an ontology term, since the link can redirect the user to that concept's definition.

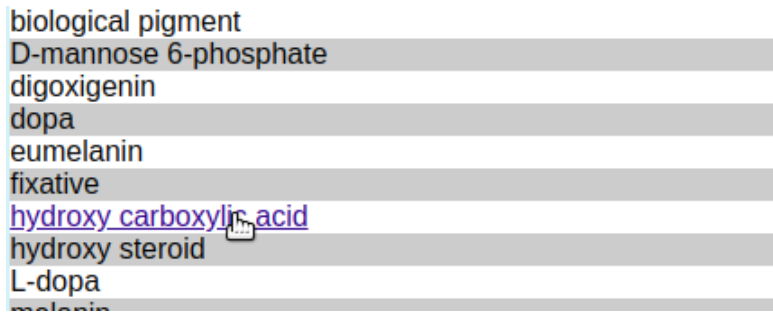


Figure 20 - If a user clicks on an annotation it will redirect them to the entry “CHEBI:24669” of ChEBI containing more information about “hydroxy carboxylic acid”.

Improved comparison symbology

To make the notion that two things were being compared to one another more intuitive, the two entities were positioned side by side, with a column for each entity inside the tabs, a middle column for overlapping annotations, and the symbology of the Venn diagrams discussed above. Also, a “VS” string (for versus) was added on the top banner, between the two entity names (see Figure 21).



Figure 21 - Article titles taken from a CRAFT dataset pair with a “VS” string in the middle to underline there is a comparison being made.

The early stages

As examples of what changed throughout the evolution of the project, I present in Figures 22 to 24 aspects of the original interface, so that they can be compared to the polished version presented above. Figure 22 shows that the header was not suggestive that two things were being compared and the description of each entity occupied a substantial part of the webpage’s space. Entities had an identifier right next to the title that users could click to go to another source of information.

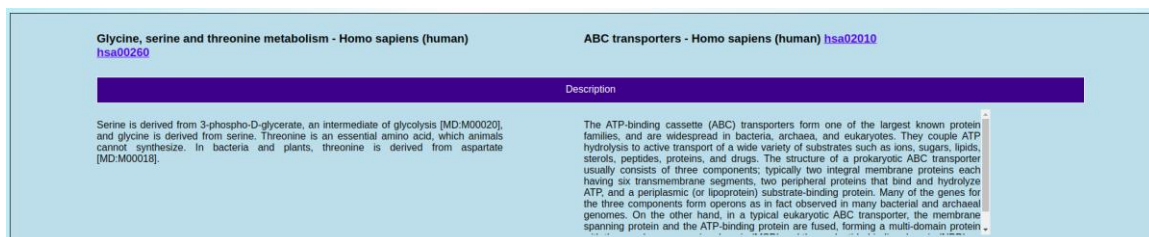


Figure 22 - The header of the page with two KEGG pathways “hsa00260” and “hsa02010” being compared. The description would be open by default using a large portion of the pages space.

A second example, in Figure 23, is that domain tabs were collapsible but had no visual indicator to tell users there was more information hidden inside it unless they clicked on it. Tabs had counters for the number of annotations in each column, however it was again not suggestive for users as to what those numerical values represented, unless previously explained to them:

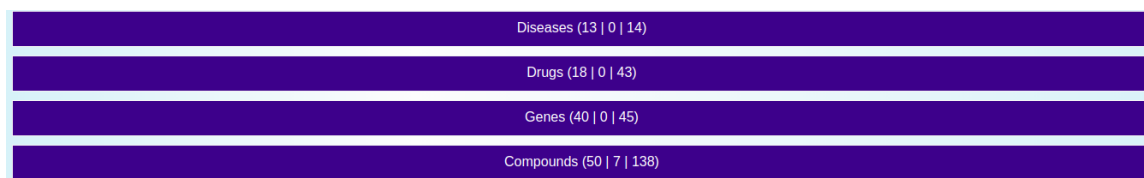


Figure 23 - Horizontal tabs for each domain in a KEGG's entity showing the number of unique and common annotations inside them.

Finally, the last example is the original similarity bar, in Figure 24, which was monocolored and would fill up to the point where the slider was dragged to, which made it difficult for users to remember the position of previous pairs. The submit button was permanently positioned at the bottom center:

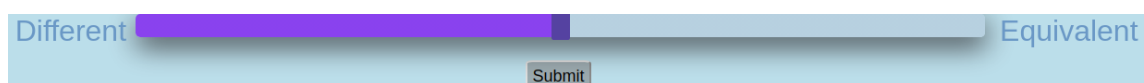


Figure 24 - Similarity bar that would fill up to slider with a unique color. Submit button was placed below.

4.4. Manual comparisons

The criteria for selecting which dataset to present to curators was based on the metrics introduced in section 4.2. For a domain, by multiplying the coverage with the total number of entities, the number of entities containing annotations in said domain is obtained. Additionally, if this number is multiplied by the average number of annotations in an entity, specific for the domain (volume), then an expected total number of annotations is obtained, however the real value is smaller due to the fact that some annotations can be repeated between entities. If we subtract the diversity, we get a value that reflects the amount of duplicate annotations in the dataset, regarding that domain. Finally, this value is distributed between the number of distinct pairs of entities. Using these metrics, it is possible to calculate the value of how expected it is for a pair of entities to contain common annotations between them in a domain, defined as “expected overlap” (eo). The expected overlap was calculated using the following formula:

$$eo = \frac{n * c * v - d}{C(n, 2)} \quad (4.4)$$

Where n is the number of entities, c is the coverage, v is the volume, d is the diversity and $C(n,2)$ is the number of pairs. The expected overlap was calculated for every domain in each dataset, then an average of those values was also calculated for the full dataset (see Table 5). OMIM's dataset was made after the pilot test, so the decision of which dataset to provide took only KEGG Pathways, CRAFT and BioModels into consideration. CRAFT was the dataset with the highest overlap overall, almost 2 times larger than KEGG Pathway (although the domain “Gene” had overlap values matching the highest ones obtained in CRAFT) and 40 times larger than BioModels, which had the lowest overlap. Due to this, CRAFT was the dataset selected to be used in the second pilot test. Expected overlap was also calculated for the diseases dataset and it showed the highest eo out of all datasets, which is related to the fact that it was more carefully handpicked.

Table 5 - Expected overlap of common annotations between pairs in each dataset.

Dataset	Expected Overlap
KEGG Pathways	0.475
BioModels	0.02
CRAFT	0.809
OMIM	1.032

All the participants were master's graduates whose fields of expertise, as provided by themselves, were *marine biology (2)*, *bioinformatics (1)*, *ecology (1)*, *mitochondria biology and neuroscience (1)* and *molecular biology (1)*. They were selected due to their background in biology and capacity to understand the concepts in the articles, although none of them were real experts in the field of mouse genomics, which was the subject of all the papers in CRAFT. They also had no previous experience in curating tasks. The total number of answers was 67, with the highest similarity value being 74 and the minimum being 0. Only 4 pairs had a score above 70 and the vast majority of scores were below 30 (see Figure 25). The similarity average was 17.44, meaning that the overall consensus was that articles were very different from each other.

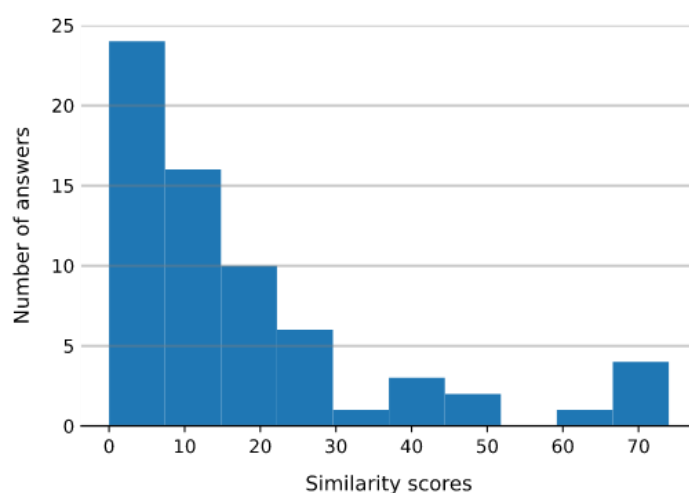


Figure 25 - Histogram with the manual values, with the y-axis being the number of answers and the x-axis being the similarity score, ranging from 0 to 100, assigned to pairs by curators.

4.5. Manual values vs Automatic values

Pearson's correlation coefficient of manual values against values obtained computationally using simGIC were very low, with a value of 0.132 (corresponding to a p -value of 0.287). Additionally, by looking at the distribution in Figure 26 it is clear that there is no resemblance of a correlation line, and in fact some of the highest values obtained from the automatic measure correspond to some of the lowest from the manual answers, and vice-versa. The scatter plot coupled with the calculation of Pearson's r seem to infer that there is no correlation between automatic values and the values given by humans. This

issue is explained by the fact that a) the experiment I ran had a low number of testers (only 6 people), and even though their academic background was related to biology, none of them specialized in genomics, which was the area related to every article in the CRAFT dataset; b) the low number of answers (67) as well as a lack of multiple answers for a single pair to see if there was inter-curator coherence, increases the probability of skewing the results since some of the highest automatic values correspond to some of the lowest manual values c) the dataset picked to be tested by the users was not given a pairing strategy that guaranteed the same rate of appearance for similar, intermediate and different pairs of entities.

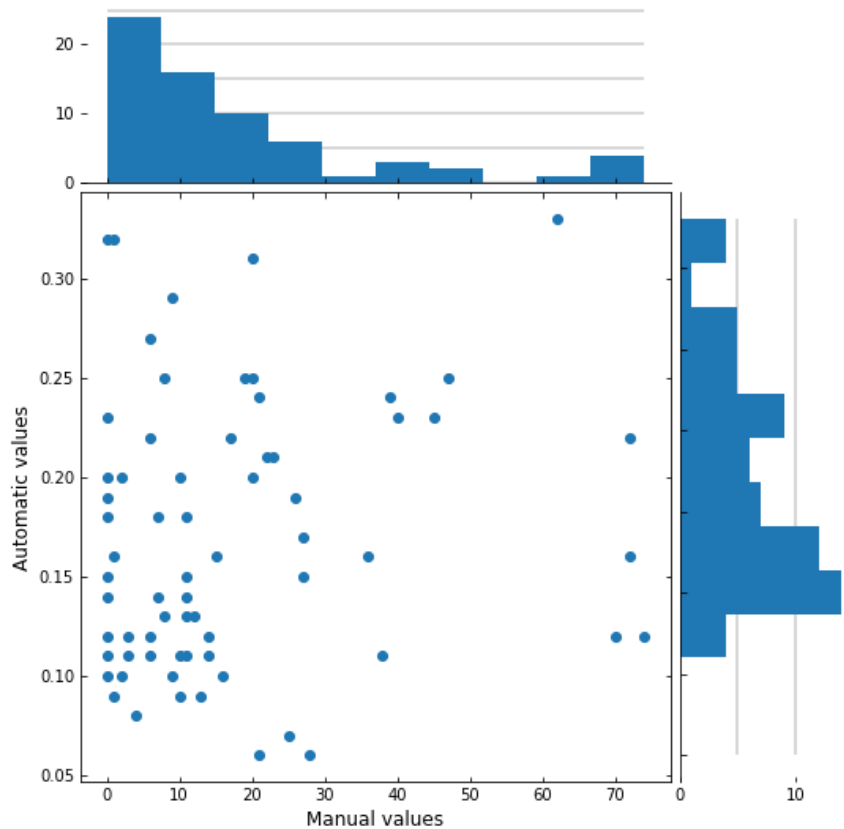


Figure 26 - Scatter plot and histograms between manual (top histogram) and automatic (right histogram) similarity values for CRAFT pairs.

5. Conclusions

This project proposed the creation of an online tool that facilitates the creation of gold-standards of annotated entity similarity values. The tool allows developers of automatic similarity measures to upload a dataset of annotated entities, providing a URL for that user to distribute to experts, who will then assign similarity values to random pairs within that dataset. The original user can then collect all the answers and thus create the required gold-standard.

The project successfully achieved that main goal, resulting in the application currently available at <http://mvht.lasige.di.fc.ul.pt>.

One of the minor goals was to develop a dataset format to be used in the upload phase. It seems to be capable of representing heterogeneous datasets, having been used with four datasets from different backgrounds (pathways, scientific articles, mathematical models of biological systems and diseases), and is adapted to work with multi-domain similarity measures, since it can represent entities with annotations from different domains. Even though my work was directed at the biomedical domain, there is potential to apply the tool to other areas as well. For example, it should be possible to develop a dataset containing movies, using IMDb as the source of data, where each movie is annotated with the genre, director, writers, cast, movie duration, date of release, budget, etc.

For the duration of the project, I faced some challenges, some of which resulted in limitations to the application, as well as limitations of the project as a whole.

Regarding the collection of the datasets, KEGG PATHWAYS dataset contained too many unrelated pathways in the pairings. Frequently entity pairs had only one or two common annotations between them. This was possibly due to the fact that all human related pathways were used (337) resulting in a mixture of pathways related to diseases, cellular process pathways, organismal systems, metabolism pathways, etc. For this type of entity, a smaller and meticulous selection of pathways should provide a better dataset, for example, using a dataset containing only pathways related to human metabolism. The BioModels dataset was difficult to extract due to the way models are stored, some in different *levels* (see section [3.2.2](#)) which alters the structure of the files; even within files with the same *level* there were different *versions* with no common consensus in the way XML attributes were provided, resulting in lack of annotations in some entities, for certain custom annotations, mainly for the chemical compound annotations. The CRAFT dataset used articles as entities with multidisciplinary concepts for its annotations, where every entity contained no lack of information in terms of descriptions for the articles, links to other sources with more information, a substantial number of annotations related to each ontology whose custom properties were based on, and overall a good dataset to present to users. Finally, for the disease dataset from OMIM, expected overlap was the highest of all datasets, mainly due the methodical selection of diseases known to be similar and others known to be different, although limitations in natural language processing tasks resulted in the integration of false positive annotations into the dataset, which can be misleading to curators.

Another limitation of my work is that Pearson's correlation coefficient resulted in an almost null correlation value, which was a consequence of a low-quality manual similarity dataset. Therefore a “gold-standard” was not achieved at all with the current state of the work. However, the focus of this part of my work was to show that it is possible to quickly reason over the results collected by the tool and obtain an R value, which was achieved.

I intend that in the future, I can use a more robust dataset containing diseases, using OMIM as the data source and present it to doctors (as it was originally intended before the events of Covid-19) and document the results in a paper. I also intend to upgrade MVHT in terms of implementing a dataset validator, using the developed JSON schema (see section [4.1](#)) as a way to let dataset owners know if their dataset formats are in concordance with it, and implement more pairing strategies between entities.

For example, a “balanced strategy” can be created that starts by running one or more semantic similarity measures on the pairs of entities, sorting them by similarity in bins (like in a histogram). The curators are then given pairs in an order that assures that the number of pairs compared from each bin is approximately the same, thus balancing the gold standard in terms of the range of similarity values it contains. Another example could be to take pairs where there is a lot of inter-curator discordance (some curators think the pair is similar while others think it is different) and select that specific pair to be shown to more people, so that the mean value of those answers can be used as some type of agreement between curators.

References

- [1] N. Shadbolt, T. Berners-Lee, and W. Hall, “The semantic web revisited,” *IEEE Intell. Syst.*, vol. 21, no. 3, pp. 96–101, 2006.
- [2] A. Ruttenberg *et al.*, “Advancing translational research with the Semantic Web,” *BMC Bioinformatics*, vol. 8, no. 3, pp. 1–16, 2007.
- [3] J. D. Ferreira, D. C. Teixeira, and C. Pesquita, “Biomedical Ontologies: Coverage, Access and Use,” 2020.
- [4] M.-F. Sy, S. Ranwez, J. Montmain, A. Regnault, M. Crampes, and V. Ranwez, “User centered and ontology based information retrieval system for life sciences,” *BMC Bioinformatics*, vol. 13, no. 1, pp. 1–12, 2012.
- [5] H. Yang, T. Nepusz, and A. Paccanaro, “Improving GO semantic similarity measures by exploring the ontology beneath the terms and modelling uncertainty,” *Bioinformatics*, vol. 28, no. 10, pp. 1383–1389, 2012.
- [6] F. Azuaje and O. Bodenreider, “Incorporating ontology-driven similarity knowledge into functional genomics: An exploratory study,” in *Proceedings. Fourth IEEE Symposium on Bioinformatics and Bioengineering*, pp. 317–324, 2004.
- [7] L. Rong *et al.*, “A measure of semantic similarity between gene ontology terms based on semantic pathway covering,” *Prog. Nat. Sci.*, vol. 16, no. 7, pp. 721–726, 2006.
- [8] T. Pedersen, S. V. S. Pakhomov, S. Patwardhan, and C. G. Chute, “Measures of semantic similarity and relatedness in the biomedical domain,” *J. Biomed. Inform.*, vol. 40, no. 3, pp. 288–299, 2007.
- [9] G. Soğancıoğlu, H. Öztürk, and A. Özgür, “BIOSSES: a semantic sentence similarity estimation system for the biomedical domain,” *Bioinformatics*, vol. 33, no. 14, pp. i49–i58, 2017.
- [10] S. V. S. Pakhomov, G. Finley, R. McEwan, Y. Wang, and G. B. Melton, “Corpus domain effects on distributional semantic modeling of medical terms,” *Bioinformatics*, vol. 32, no. 23, pp. 3635–3644, 2016.
- [11] J. Jovanović and E. Bagheri, “Semantic annotation in biomedicine: The current landscape,” *J. Biomed. Semantics*, vol. 8, no. 1, pp. 1–18, 2017.
- [12] C. M. Machado, D. Rebholz-Schuhmann, A. T. Freitas, and F. M. Couto, “The semantic web in translational medicine: current applications and future directions,” *Brief. Bioinform.*, vol. 16, no. 1, pp. 89–103, 2015.
- [13] U. Consortium, “UniProt: a worldwide hub of protein knowledge,” *Nucleic Acids Res.*, vol. 47, no. D1, pp. D506–D515, 2019.
- [14] M. Ashburner *et al.*, “Gene ontology: tool for the unification of biology,” *Nat. Genet.*, vol. 25, no. 1, pp. 25–29, 2000.
- [15] D. A. Zarin, T. Tse, R. J. Williams, R. M. Califf, and N. C. Ide, “The ClinicalTrials.gov results database—update and key issues,” *N. Engl. J. Med.*, vol. 364, no. 9, pp. 852–860, 2011.
- [16] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick, “Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders,” *Nucleic Acids Res.*, vol. 33, no. DATABASE ISS., pp. 514–517, 2005.

- [17] M. Kanehisa, “Toward understanding the origin and evolution of cellular organisms,” *Protein Sci.*, vol. 28, no. 11, pp. 1947–1951, 2019.
- [18] C. Li *et al.*, “BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models,” *BMC Syst. Biol.*, vol. 4, no. 1, p. 92, 2010.
- [19] V. Chelliah *et al.*, “BioModels: ten-year anniversary,” *Nucleic Acids Res.*, vol. 43, no. D1, pp. D542–D548, 2014.
- [20] R. S. Malik-Sheriff *et al.*, “BioModels—15 years of sharing computational models in life science,” *Nucleic Acids Res.*, vol. 48, no. D1, pp. D407–D415, 2020.
- [21] K. B. Cohen *et al.*, “The Colorado Richly Annotated Full Text (CRAFT) corpus: Multi-model annotation in the biomedical domain,” in *Handbook of Linguistic Annotation*, Springer, pp. 1379–1394, 2017.
- [22] J. D. S. Ferreira, “Semantic Similarity Across Biomedical Ontologies.” Universidade de Lisboa (Portugal), 2016.
- [23] P. Dutta, S. Basu, and M. Kundu, “Assessment of semantic similarity between proteins using information content and topological properties of the gene ontology graph,” *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 15, no. 3, pp. 839–849, 2017.
- [24] E. Emadzadeh, A. Nikfarjam, R. E. Ginn, and G. Gonzalez, “Unsupervised gene function extraction using semantic vectors,” *Database*, vol. 2014, 2014.
- [25] A. Kastrin, P. Ferik, and B. Leskošek, “Predicting potential drug-drug interactions on topological and semantic similarity features using statistical learning,” *PLoS One*, vol. 13, no. 5, p. e0196865, 2018.
- [26] J. D. Ferreira and F. M. Couto, “Semantic similarity for automatic classification of chemical compounds,” *PLoS Comput Biol*, vol. 6, no. 9, p. e1000937, 2010.
- [27] C. Pesquita, D. Faria, A. O. Falcão, P. Lord, and F. M. Couto, “Semantic similarity in biomedical ontologies,” *PLoS Comput. Biol.*, vol. 5, no. 7, 2009.
- [28] J. D. Ferreira and F. M. Couto, “Multi-domain semantic similarity in biomedical research,” *BMC Bioinformatics*, vol. 20, no. 10, pp. 23–31, 2019.
- [29] C. Pesquita, D. Pessoa, D. Faria, and F. M. Couto, “CESSM: Collaborative Evaluation of Semantic Similarity Measures,” *JB2009 Challenges Bioinforma.*, vol. 157, no. November 2016, p. 190, 2009.
- [30] G. Zhu and C. A. Iglesias, “Sematch: Semantic similarity framework for knowledge graphs,” *Knowledge-Based Syst.*, vol. 130, pp. 30–32, 2017.
- [31] C. Pesquita, D. Faria, H. Bastos, A. E. N. Ferreira, A. O. Falcão, and F. M. Couto, “Metrics for GO based protein semantic similarity: A systematic evaluation,” *BMC Bioinformatics*, vol. 9, no. SUPPL. 5, pp. 1–16, 2008.
- [32] R. Gentleman, “Visualizing and distances using GO,” URL <http://www.bioconductor.org/docs/vignettes.html>, vol. 38, 2005.
- [33] J. D. Ferreira and F. M. Couto, “Multi-domain semantic similarity in biomedical research,” *BMC Bioinformatics*, vol. 20, no. 10, pp. 23–31, 2019.
- [34] J. Collinge *et al.*, “Safety and efficacy of quinacrine in human prion disease (PRION-1 study): a patient-preference trial,” *Lancet Neurol.*, vol. 8, no. 4, pp. 334–344, 2009.

- [35] S. Haik *et al.*, “Doxycycline in Creutzfeldt-Jakob disease: a phase 2, randomised, double-blind, placebo-controlled trial,” *Lancet Neurol.*, vol. 13, no. 2, pp. 150–158, 2014.

Appendices

A. List of diseases

The selection process consisted in searching on google for “common genetic disorders” and picking a sample of 20 diseases, while trying to integrate diseases that can be grouped in a similar category but at the same time maintaining some diversity, e.g. respiratory diseases, oncological diseases, neurodegenerative diseases. OMIM entries contain an identifier using “#” (the hashtag symbol means it is a phenotype description with a known molecular basis) followed by a six-digit number. The list of diseases was as follows:

Alzheimer’s Disease (#104300) is the most common form of dementia that progressively gets worse as time advances, affecting the elderly. It is a neurodegenerative disorder characterized by the presence of intracellular neurofibrillary tangles and accumulation of amyloid plaques in the brain.

Parkinson’s Disease (#168600) is the second most common neurodegenerative disorder that manifests in the form of resting tremor, muscular rigidity, bradykinesia and postural instability.

Down’s Syndrome (#190685), also called trisomy 21 caused by the presence of a third copy, or a partial copy, of chromosome 21. Its symptoms include mental retardation and characteristic facies.

Cystic fibrosis (#219700) is a progressive disease that causes lung infections and limits the ability to breathe over time due to the inability of the CFTR protein to help move chloride to the cell surface. Chloride is responsible for preventing mucus to be thick and sticky.

Sickle cell anemia (#603903) is an inherited blood cell disorder. It is most commonly caused by a variant in hemoglobin S which makes red blood cells rigid, misshapen and with the possibility of occurring vaso-occlusion.

Phenylketonuria (#261600) results from a deficiency in the enzyme phenylalanine hydroxylase, responsible for catabolizing the hydroxylation of phenylalanine to tyrosine.

Angelman’s Syndrome (#105830) is a disorder characterized by mental retardation, movement or balance disorders, limitations in speech and language and sometimes seizures. It is usually caused by complications with a gene located in chromosome 15.

Achromatopsia (#216900) also called “total color blindness” is a disorder characterized by photophobia, reduced visual acuity, and the inability to discriminate between colors. Currently there are five different genes that are known to cause achromatopsia.

Hemophilia A (#306700) can manifest with different levels of severity depending on the plasma levels of coagulation factor VIII, which becomes deficient for people with this disorder. The symptoms are excessive bleeding after trauma or surgery.

Hemophilia B (#306900) is caused by a deficiency in factor IX and is difficult to distinguish from hemophilia A in terms of phenotype.

Alpha-thalassemia ([#604131](#)) is caused by mutations in the alpha-globin genes HBA1 and HBA2 and result in an impaired production of hemoglobin.

Huntington's Disease ([#143100](#)) is a progressive neurodegenerative disorder whose symptoms include involuntary movements, incoordination, cognitive decline and behavioural difficulties due to the death of cell brains. It is caused by a higher than normal number of heterozygous trinucleotide repeats in the huntingtin gene.

Duchenne type muscular dystrophy ([#310200](#)) is caused by a mutation in the gene encoding dystrophin and is characterized by a gradual weakening of muscles.

Von Willebrand Disease ([#193400](#)) is the most common inherited bleeding disorder, caused by a heterozygous mutation in the gene encoding von Willebrand factor, a protein required for platelet adhesion. There are three types of the disease, with type 1 being the most common variation of the disorder characterized by mucocutaneous hemorrhage.

Tourette Syndrome ([#137580](#)) is a neurobehavioral disorder that manifests in the form of vocal and motor tics and is associated with behavioural abnormalities. It is caused by a combination of genetic and environmental factors that are yet not very known.

Lung cancer ([#211980](#)), also called lung carcinoma is an uncontrolled cell growth that happens in the tissues of the lung, that usually manifests as a cough (sometimes together with blood), shortness of breath and chest pain.

Colorectal cancer ([#114500](#)) is a type of cancer that appears in parts of the large intestine and is caused mostly by lifestyle and environmental factors, but gene defects can contribute to the appearance of the disease. Symptoms may include blood in the stool, fatigue and weight loss among others.

Fragile X Syndrome ([#300624](#)) is characterized by mental retardation, distinct facial features and abnormal size of the testicles. It is caused by mutations in the FMR1 gene present in the X chromosome, where the trinucleotide CGC repeat is expanded.

Rheumatoid Arthritis ([#180300](#)) is an inflammatory disease that affects primarily the joints, with autoimmune responses and multiple genetic factors can influence the susceptibility to the disease.

Type 2 Diabetes Mellitus ([#125853](#)), also known as noninsulin-dependent diabetes mellitus, is characterized by high blood sugar and lack of insulin, along with other long term complications. It is usually associated with obesity and certain genetic variations can increase the susceptibility for it.

B. JSON schema

```
{
  "$schema": "http://json-schema.org/draft/2019-09/schema#",
  "title": "Dataset",
  "type": "object",
  "definitions": {
    "annotationString": {
      "type": "string"
    },
    "annotationObject": {
      "type": "object",
      "properties": {
        "name": {"type": "string"},
        "url": {"type": "string"}
      },
      "required": [
        "name"
      ]
    }
  },
  "properties": {
    "question": {"type": "string"},
    "entities": {
      "type": "array",
      "items": [
        {
          "type": "object",
          "properties": {
            "name": {"type": "string"},
            "id": {"type": "string"},
            "description": {"type": "string"},
            "url": {"type": "string"},
            "selected annotations": {
              "type": "array",
              "items": {
                "anyOf": [
                  {"$ref": "#/definitions/annotationString"},
                  {"$ref": "#/definitions/annotationObject"}
                ]
              }
            }
          }
        }
      ],
      "required": [
        "name",
        "id"
      ]
    }
  ],
  "required": [
    "question",
    "entities"
  ]
}
```

Figure 27 - JSON schema which MVHT accepts to upload the given dataset.