Complementary Material - Data Mart Construction based on Semantic Annotation of Scientific Articles: A Case Study for the Prioritization of Drug Targets

Marlon Amaro Coelho Teixeira, Kele Teixeira Belloze, Maria Cláudia Cavalcanti, Floriano Silva-Junior

July 2017

Abstract

Semantic text annotation enables the association of semantic information ontology concepts to text expressions (terms), which are readable by software agents. In the scientific scenario, this is particularly useful because a lot of scientific discoveries are "hidden" within scientific papers. The Biomedical area has more than 300 ontologies, most of them composed of over 500 concepts. These ontologies can be used to annotate these papers and thus, facilitate data extraction. However, in the context of a scientific research, a simple keyword-based query using the interface of a digital scientific texts library can return more than a thousand hits. The analysis of such a large set of texts annotated with such numerous and large ontologies, is not an easy task. Here it is described a scientific scenario, where a corpus of selected papers was annotated using three distinct ontologies, with focus on the research of gene essentiality. The later is a key concept to be considered when searching for genes showing potential as anti-infective drug targets. This work presents how the annotation data was extracted, organized and aggregated into a dimensional schema of a demo Data Mart. Finally, a conclusion is drawn showing some research strategies over these data, and discussing how they can help the scientist to prioritize drug targets in protozoan parasites.

Everyday new discoveries arise in the biomedical area and many of these advances are related to new techniques and new equipments used in high throughput experiments. An increasing volume of structured data has resulted from these experiments and has become available. According to Rigden et al. [28], nowadays, there are about 1,685 online databases for molecular biology, divided into 15 categories and 41 subcategories . The amount and heterogeneity of these databases has become a substantial challenge to the searches and analysis of data. However, not only individual databases are used for searches of data and information. Many biological questions can only be answered by combining data from multiple sources [3]. The large volume of dispersed data hinders the conduction of researches in the biomedical field. Gathering accurate and reliable information from different data sources, in an efficient way, became a complicated activity and often not viable without the development and reuse of a suite of tools to assist in this process.

On the other hand, textual repositories are rich sources from which important information can be extracted by biomedical researchers. One of the most important digital repositories is PubMed [1], which accounts for approximately 26 million scientific texts. Although such repositories contain reports on most of the scientific discoveries, literature review usually takes a lot of time. Based on the reading of selected scientific papers, the scientist may establish what are the topic and scope of his/her research. A typical scientific paper covers topics from distinct domains within the same area.

Digital libraries classify and index large sets of scientific papers according to these topics, facilitating the scientist to find the papers of interest for his/her research interest. However, a simple keyword-based query using the interface of a digital library can return a list of more than a thousand hits. This may complicate the task of ranking the papers according to their relevance and even result in discarding some very relevant papers that appear at the end of the list. Even if a more elaborate query, including all synonyms and different expressions for the same meaning, is built it does not guarantee that the most suitable texts will be found. In addition, keywords may not reveal all domains present in the papers.

Scientific article indexing and annotation has become useful for the biomedical research community [19, 25, 18]. This is because biomedical scientists need to classify articles according to his/her research interest. For instance, it may be the case that a scientist is interested in the use of the shotgun technique to identify peptides. The simple indexation of these two terms would bring this text (and maybe others) to the top of the scientist desk? Perhaps. However, it is usual that a scientist may be interested in many combinations of distinct techniques and proteins.

Besides, other aspects may also be of interest, such as pathways, organisms, etc. Therefore, a simple two-dimension problem becomes a multi-dimensional one.

Text annotation allows the identification of the occurrence of such multidimensional combinations, and therefore makes it possible to rank these articles according to the scientist's interest. The keyword-based tools can also relate different terms by means of AND and OR operators. But these correlations are fragile, because they are possible only with terms explicitly present in articles. If a variation of the term of interest is present in an article, the tool will not be able to identify it.

There are many initiatives on automatic text annotations that are discussed in item III.A. However, none of them handle an analytical view of such annotations, such as the co-occurrence of a set of terms, representing each one a specific aspect of the scientist interest. Moreover, to the best of our knowledge, there are no previous reports on a method to bring an analytical view of a database of scientific text annotations.

Using an analytical view system in scientific research enables to correlate informations about different organisms. Diseases caused by protozoa are considered neglected diseases as they reach the poorest populations of third world countries. For these reasons, experimental data on drug targets from these organisms are still very scarce. The correlation of protozoans information with relevant data from other well studied (model) organisms can direct the researchers' experiments, making the searches less costly and obtaining relevant results in a shorter time [3].

An important concept when trying to identify new drug targets for anti-infective drug discovery is gene essentiality. Genes are considered essential to an organism when its repression, silencing or blockage results in the organism death [13]. Information about essential genes are more likely to be transferable among evolutionarily distant organisms than other types of gene function relationships. The hypothesis is that essential genes remain with a high conservation rate over the evolution of organisms. Thus, the presence of a gene in an organism with an orthology relationship with an essential gene, increases the chances this gene remains essential [6].

The objective of this work is to present a Data Marts design methodology for unstructured data analysis by the means of ontologies. Thus, providing a systematic way to process a large set of scientific articles and support the researcher in better decision making with respect to his/her specific research interests. Decision support approaches and concepts, such as data marts and star schema, detailed further, have been largely used in the business area [8].

We also present a case study on the design and load of a data mart with focus on gene essentiality for the following five protozoa: *Entamoeba histolytica*, *Leishmania major*, *Plasmodium falciparum*, *Trypanosoma brucei* and *Trypanosoma cruzi*. The data used to feed the data mart were extracted from scientific articles through ontology-based text annotation. The process to transform, aggregate, and load these data to the data mart is reported here. Finally, queries are used to illustrate the possible benefits of this approach in the search for new drug targets and demonstrating how the tool can be flexible to needs of users.

This work is organized as follows: Sections I and II present a background. Section III presents an overview of the methodology proposed by this work. Section IV describes the prioritization of drug targets case study. Section V discusses the results of case study. Section VI shows the OLAP interface and the last section presents the conclusions about this work.

I. Semantic Annotation

Semantic annotation is one of the main efforts towards the Semantic Web. According to Berners-Lee et al. [7], the Semantic Web is not a separate Web. Instead, it aims at being an extension of the Web, where text content is associated to its well-defined meaning. Indeed, an annotation should be well defined, not ambiguous and easy to understand by domain specialists, in a way that it could be useful for the information retrieval process [14].

Efforts in this direction are the semantic annotation systems [3, 5, 22], which provide mechanisms to bring semantic to documents through text annotation. It means to handle and associate metadata or ontology concepts with text content. An ontology is a model that represents a domain of reality, i.e., a formal description of concepts and relationships [7]. The use of ontologies is recommended not only to maintain annotations based on a uniform vocabulary, but also to benefit from the richness of the ontological representation. Through ontologies it is possible to make inferences about the annotations, getting information that is not always explicit to the user, and possibly, enriching annotations.

In the last decade, there has been an increasing number of ontologies emerging on biomedical domains. The Open Biological and Biomedical Ontologies (OBO) Foundry [40] and the NCBO BioPortal [41] provide together more than 500 ontologies. On the other hand, taking into account that an ontology is typically designed to cover a single domain, in order to cover a scientific text, typically multi-domain, it is required the use of multiple ontologies while semantically annotating them. The multi-ontology annotation of such texts can be especially valuable for

the biomedical scientist. It enables the identification of the co-occurrence of concepts from different ontologies and allows for ranking papers according to the scientist's interest.

An excerpt of one of these texts is shown in Fig 1. It illustrates possible annotations using three different ontologies. The expression "Drug Target" was annotated with the PHARE ontology, the expression "*Trypanosoma brucei*" was annotated with the NCBI Taxonomy, and finally, the expressions "essential" and "gene knockout" were annotated with the NCI *thesaurus*.

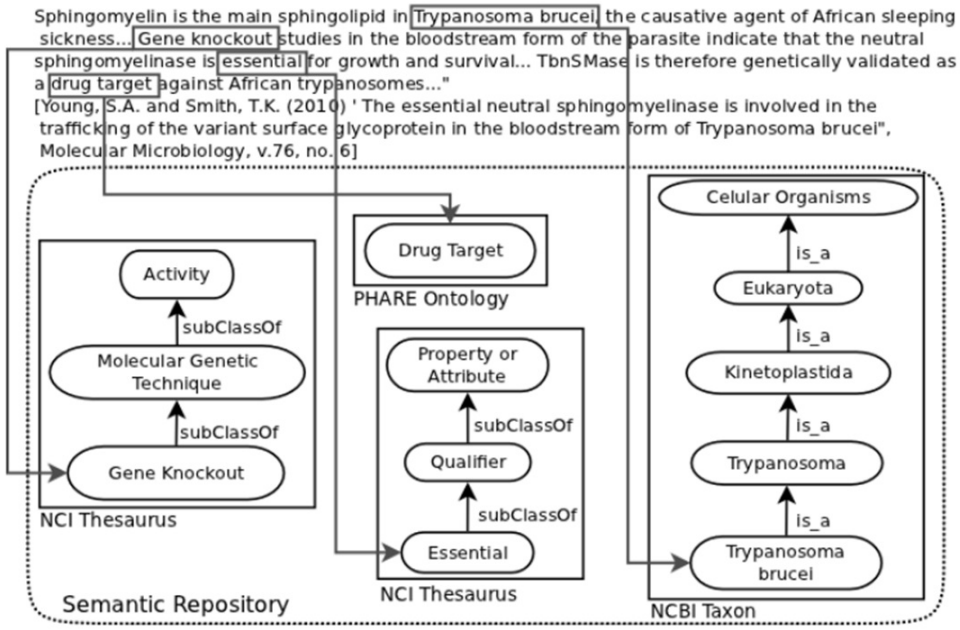

Figure 1: **Associating text expressions to multiple ontology concepts.**

The annotation process can be intrusive or not. It is intrusive when the annotation is inserted inside the document under annotation. It is non-intrusive when the annotation is registered externally (e.g. in a database). In this last case, it is easier to extract annotations and bring them into an analytical database, in order to provide a multidimensional view of a set of annotated texts. In the context of scientific scenarios, such as the biomedical scenario, this multidimensional view can be useful to support scientists on decision making, while conducting their research.

II. Decision Support Systems

Computer technologies developed in order to assist in problem solving and decision making are known as decision support systems (DSS). The studies that substantiate this area emerged in the decades of 50 and 60, but it was in the 70's that those systems were significantly developed and designed to manage complex databases. They allowed to access the information inside and outside the corporations. Despite this complexity, DSS tools should provide user-friendly interfaces, easy handling, enabling users to create interactive queries, besides reporting and graph functions. These tools always focused on how information technology can increase the efficiency and effectiveness of decisions taken [33].

In the 90's a powerful approach for decision support was developed, known as data warehousing (DW). Data Warehouses are non-volatile thematic driven databases, capable of integrating multiple data sources. Different from transactional databases, which are designed to maintain data from day-to-day activities of the corporation at present state, a DW is designed to keep historical data, thus saving all the status of an enterprise over time in order to assist in decision making. Typically, the DW approach is able to handle a large amount of data.

Moreover, besides the historical registry, a DW aims to provide a multidimensional view of the data, enabling DSS tools to answer complex analytical queries. For instance, while a transaction database answers queries like "*which flight goes from city A to B on Saturdays*", a DW should answer more complex queries, such as "*how many flights went from city A to city B during the last 6 months, with a number of passengers under 50% of the plane capacity*". Note that to answer such query, the DW needs to keep information about passengers occupation, for

each flight, time, destiny, origin, and plane.

A DW is modelled according to the dimensional model. It is designed to provide a multidimensional view of the data that facilitates to resolve such complex queries. Such model is formed by a central concept known as fact, which is what needs to be observed. The fact is connected to the aspects or dimensions that describe/characterize each observed fact. In the above example, the fact is the occupation, which is described by a combination of the dimensions: flight, time, destiny, origin, and plane. A typical implementation of the dimensional model is the star schema [24], where the main table represents the facts (central table), which has references to satellite tables, called dimension tables.

One of the most important elements in a dimensional model is the grain. Grain is the definition of the representation of a single tuple in the fact table. The choice of grain impacts directly on the characteristics of the fact and dimension tables. The grain defines the highest level of detail of the information that the DW will provide to the user. For example, if the grain is defined as daily sales of a store, by means of roll-up operation you can get the weekly sales, but it is not possible to perform the drill-down operation to obtain the hourly sales made [24]. The roll-up and drill-down operations will be detailed in the next paragraphs.

In the context of DW arises the concept of Data Mart (DM). A DM is usually defined as a subset of a DW, with a specific focus, or with a reduced number of dimensions. It also aims to help on decision making of a business process. As already mentioned, these approaches consider that maintaining historical data is important, which affects directly the database size. The queries are also more complex because of the manipulation of a wide set of records from multiple tables and the use of common joins and aggregations making the management of critical elements such as performance and response time a challenge [21, 22, 24].

For these reasons, DW and DM are handled through On-Line Analytical Processing (OLAP) tools. The use of such tools allows users to perform operations like aggregation, detailing of hierarchical levels, selection, projection and reorientation of multidimensional view along multiple dimensions, enabling better insight of historical data and heterogeneous sources [10, 32].

Some queries can only be answered if the hierarchy is manipulated. For example, considering a corporation that sells products to the whole world and that desires to find out which city bought more products, in this case the lowest levels of the hierarchy must be analyzed and compared. But, if the corporation wants to find out which country most bought products, consequently, city sales must be summed up.

The most common hierarchy manipulation operations in DM are the roll-up and drill-down operations. The roll-up operation performs aggregations of the information, moving from a lower level to a higher level of the hierarchy. Thus, less detail is presented, but on the other hand, it is possible to obtain a large view of the situation. The drill-down operation is the reverse operation, it goes from the most general element to the most specific one, showing more detailed data [38].

In order to build and maintain data in a DW or DM, a process called ETL (Extract, Transform and Load) should be implemented. Usually, this process also uses a support database, where the data extracted from transactional databases are stored temporarily for transformation operations and before they are loaded into the DW. This database is also called data staging area.

Traditional methods of DM design are well consolidated, these techniques based on structured data sources are generic. Nowadays, a lot of information is available through unstructured data as scientific articles. However, there are just a few studies exploring specific methodologies to build DMs from unstructured data for analysis and decision support.

Some of these approaches use knowledge domains described by ontologies to compose the DM dimension tables, as shown by a previous work [29]. The later considered an article where two terms were mentioned in the same hierarchy but at different levels, one more specific and other one more general. The tool only recorded the presence of the more specific term, thus the fact table only stored references to the same granularity level. This limitation prevents the user from obtaining information about the hierarchical nodes at higher levels, as for example, the reference to protein families.

Other studies used ontologies directly and automatically in the Data Mart design [36, 20, 27], where an ontology was used as input and, at the end of the process, a dimensional diagram was obtained. Nevertheless, the solutions were generic due to the fact that the modeling of biomedical problems handles multiple ontologies according to the different knowledge areas involved.

In the present study, the dimensional elements present in the star schema are identified by the needs of users. The dimension tables represent the description of the aspects involved in these needs and these descriptions are obtained through the multiple ontologies that describe them. The fact table is responsible for registering the co-occurrence of the elements of the dimension tables. The following subsections describe this process in more details.

III. Overview of the Methodology: from Annotations to Data Marts

The proposed method, named TOETL (Text and Ontology ETL), was designed based on the traditional ETL process, to address the scientific texts annotation context. The first step of this method was described in a previous work [3] and summarized in subsection A. The following steps, *Transform* and *Load* steps, which are the contributions of the present work, are detailed in subsections B and C.

Figure 2(ii) provides a brief view of the adaptation of the traditional ETL process, in order to address the specificity of a Scientific text annotation DM. Note that the sources are mainly text-based sources and that the *Extract* step is focused on the Semantic Annotation, using a set of ontologies as input. In this step, the scientist, which is the main user, defines the set of articles of his/her interest, and also the set of ontologies that are more suitable for annotating those articles.

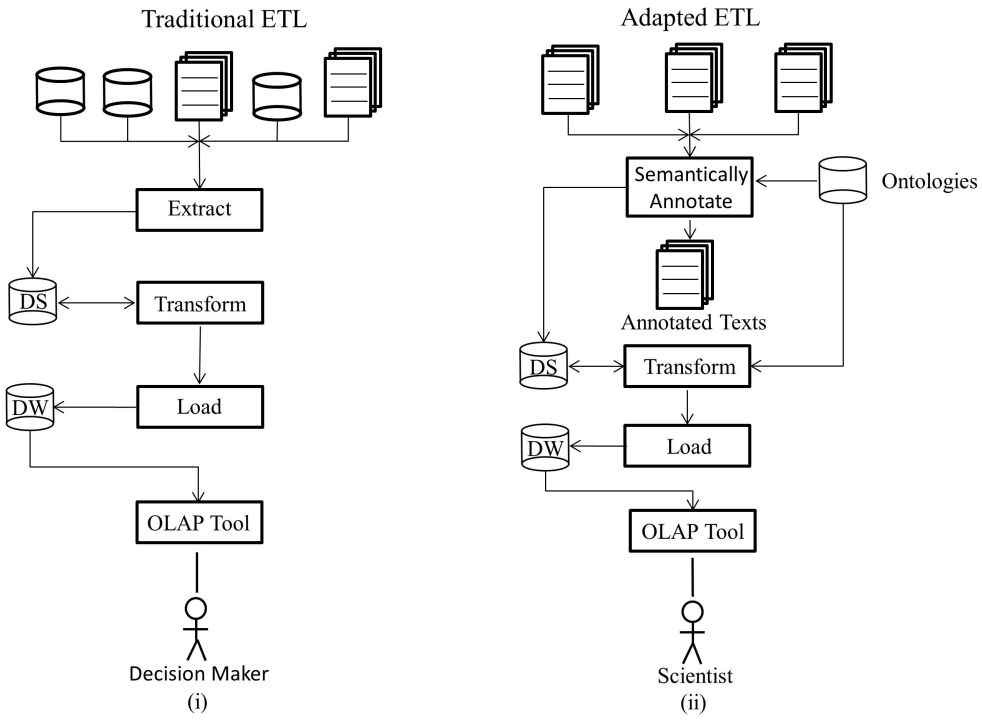

Figure 2: **Traditional and Adapted ETL (TOETL) process.**

A. Annotation Step

Based on a previous work [3], we organized the annotation stage as a set of steps, as shown in Fig 3. The *understanding the research theme* step involves the study of the classic literature in the area, as well as interviews with specialists. A set of terms, expressions and their synonyms are raised in order to identify in the literature a significant amount of scientific texts (corpus). This set should cover all domains across the research theme.

Once defined, this set is then used to proceed to two other steps: the *queries definition* and the *digital library selection* steps. The *queries definition* step uses the terms and expressions to compose keyword search queries on a digital library. Both generic and specific queries should be formed, in order not to miss important literature items while building the corpus. Therefore, the idea is to organize terms according to specific domains, and build logic expressions with the AND operator. First, in pairs of terms, each from a different domain. This will bring a large number of hits, but will guarantee a good recall. Then, more specific queries are formed, by combining more terms using the AND operator. In addition, for each operand in these query expressions, a parenthesis with alternatives or synonyms should be formed, using the OR operator. For instance, for a research theme that involves three domains, the set of terms $S = \{t_{1,1}, t_{1,2}, t_{2,1}, t_{2,2}, t_{2,3}, t_{2,4}, t_{3,1}, t_{3,2}\}$, the following query expressions could be formed:

- $t_{1,1} AND (t_{2,1} OR t_{2,2})$
- $t_{1,2} AND (t_{2,1} OR t_{2,2})$

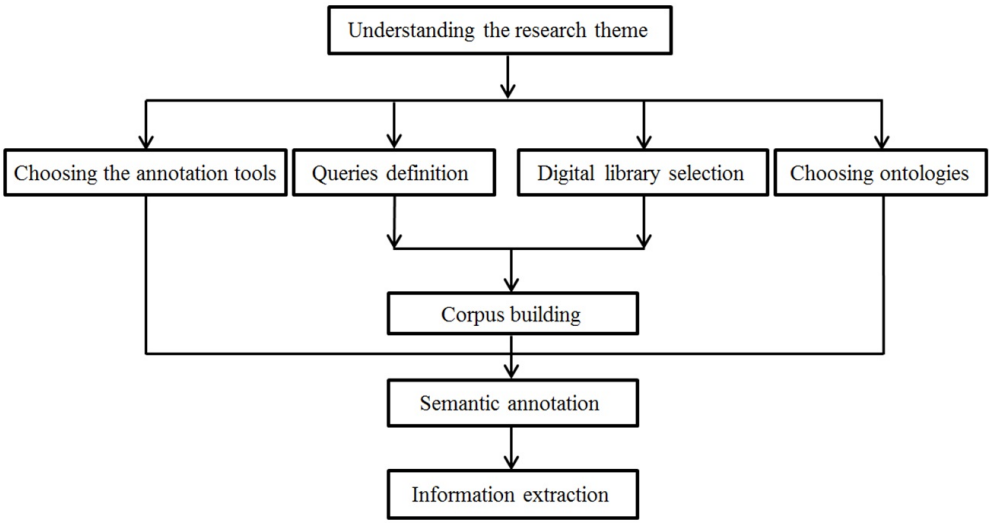

Figure 3: **Flow chart representing the data extraction by means of ontologies.**

- $t_{1,1}AND(t_{2,3}ORt_{2,4})$
- $t_{1,2}AND(t_{2,3}ORt_{2,4})$
- $t_{1,1}AND(t_{2,1}ORt_{2,2})AND(t_{3,1}ORt_{3,2})$
- ...

At the *digital library selection* step, the idea is to choose libraries (one or more) that cover most of the domains involved in the research theme in addition, for a complete analysis that will be performed later on, it is important that most of the texts are fully available (complete texts, not only the abstract). Some articles are available only under payment.

Once the queries expressions are formed and libraries are available, the corpus can be built. At this step, *corpus building*, an automatic tool should be in charge of such repetitive task: issuing each query expression to the library interface and eliminating duplicated hits. Also at this step, fetching, saving and cleaning texts should also be performed.

In parallel, other two steps can be performed: *choosing ontologies* and *choosing annotation tools*. As stated before, the use of ontologies is recommended in order to maintain annotations based on a uniform vocabulary. However, it is not an easy task to choose them. First of all, more than one ontology may be necessary once the research theme usually involves multiple domains. Ontology repositories such as OBO Foundry and NCBO BioPortal cited before, can be used to facilitate the search. In these repositories, one may find many ontologies that cover the same domain, but their coverage may be different, i.e., some may have a deep and narrow coverage, while other may have a large and shallow coverage. Some previous sampling annotations may be used to facilitate the choice. This is done by taking a small and diversified set of texts from the corpus, and submitting them to initial annotations with a first set of selected ontologies. Redundant ontologies should be eliminated, as well as those that present irrelevant annotations or only a small number of relevant ones.

For this methodology, it is not feasible to do manual annotation. Therefore, it is required to choose an automatic annotation tool. Moreover, two other features are required for the annotation tool: (i) It must use any arbitrary ontology for annotation, i.e., the user must be free to choose the ontology of his/her preference; and (ii) the resulting annotated text should be ready for further processing. Some of the available annotation tools do not address both of these requirements, as reported in [5, 4].

A desired but not strictly necessary functionality for annotation tools, is to annotate using hierarchical inference. Some annotation tools provide such functionality [17]. After a term is annotated with an ontology concept, an inference-based annotation occurs when this concept belongs to a branch of concepts. In this case, it may be also annotated with all or some of its parents. This feature guarantees that a scientific text could be associated not only to specific concepts but also to their generalizations.

Once we have the ontologies, the corpus already prepared and the annotation tool selected, then we can proceed to the *semantic annotation* step. This step may be very time consuming. It all depends on the size of the corpus

and of the ontologies. It may take days. Some works on this direction propose solutions that can considerably reduce this processing time [11, 12].

Finally, after the annotation, the *information extraction* is the step that prepares the staging area database for the datamart construction stage. It consists of an automatic processing of the annotated texts. Each annotation in a text will generate a database tuple with (i) the annotated text expression, (ii) the ontology class (id and label) used to annotate it, and (iii) the text id (paper identification).

For annotating the corpus each ontology is used separately. Consequently, a set of annotated articles with the same amount of the corpus files is generated for each ontology. All annotated terms will be stored in the Annotation table. If a term is annotated more than once in an article, it will be inserted more than once. This repetition does not compromise the DM information, because the Annotation table is just a temporary table in the DM populating process. These repetitions will be eliminated at the time of the fact table population, showed in item III.B.4.

Thus, each annotated term is recorded in the Annotation table containing the following columns: id (tuple identifier), article_id (article identifier), label (term annotated), class_id (class annotated identifier). The logical schema for this table could be the following:

$AnnotationOi(id, article_id, label, class_id)$

B. Transform Step

For the second stage, again a sequence of steps are defined based on traditional data mart building methods [22, 24, 37]. However, they were adapted in order to address the scientific texts context.

1. Identification of Analytical Demands. The first step to build the data mart is to identify the analytical questions that must be addressed. At this point, we should count again on interviews with specialists. But now the focus is not on building a corpus of scientific texts. The main idea is to identify which information we can get from the annotation data. This information gathering is of fundamental importance and must be done carefully. Missing some information may lead to an incomplete DM.

Therefore, this step focuses on raising questions about text contents, based on the annotation data. Usually, these questions require the correlation between concepts from different knowledge domains, for example: "*What side effects were most often cited with chagas disease treatment?*". Clearly in this case, the answer to this question depends on the need to correlate three distinct aspects (concepts): articles, side effects and diseases.

After raising such questions, each question is analyzed for the identification and classification of terms cited in these questions. We count on the ontologies used for annotation to classify such terms according to their corresponding domains. For example, for the question "*How many articles cite chagas disease, fever and body pain?*", the term chagas disease can be classified as a disease, and the terms fever and body pain can be classified as symptoms. Another example is based on the question "*Which biological processes are most often cited with any organism?*". It is clear that for a consistent response, the occurrence of terms referring to the names of *biological processes* and *organisms* must be correlated and quantified. Therefore, *biological processes* and *organisms* are concepts of interest identified in this example.

2. Identification and characterization of the Dimensions and Categories. The previous step is the base for the dimension identification. Initially, it seems natural to think that each concept of interest would become a dimension and each tuple of the fact table would represent the presence of these terms in an article. But in the biomedical area there is a huge amount of concepts related to organisms and proteins, that could be correlated. Therefore, this approach would not be feasible due to the high dimensionality curse. In other words, it would lead to a very large database with performance issues.

In order to reduce the number of dimensions, once again we can count on the ontologies. Since the main purpose of an ontology is to describe a knowledge domain, organizing it as hierarchies of concepts, thus we can use generic concepts to represent a set of specific concepts identified in the previous step. For instance, concepts such as *Archaea* and *Eukaryota* can be represented by the generic concept named *cellular organisms*. Therefore, it is possible to define a reduced set of dimensions with the help of ontology hierarchies.

The main idea is to use generic concepts to represent a set of specific concepts identified in the previous step. For instance, concepts such as *Archaea* and *Eukaryota* can be represented by the generic concept named cellular organisms. Therefore, it is possible to define a reduced set of dimensions with the help of ontology hierarchies. Usually, ontologies of the Biomedical areas are very large, including concepts of many subdomains. For example, the National Cancer Institute (NCI) thesaurus ontology includes hierarchies of organisms, drugs, chemicals, genes, activities, biological process, etc. In this step the designer should identify and select the hierarchies of interest.

Based on the analytical questions demand, concepts that are combined in the same analytical question should be kept in distinct dimensions. For instance, if the user demands to identify combinations of organisms and chemicals, these hierarchies should correspond to distinct dimensions.

Now that dimensions are defined, it is time to characterize them, meaning that we should identify properties or attributes that could describe each dimension instance. In order to address roll-up analytical demands, dimension categories are defined at this point. On a regular hierarchy, each hierarchical level can be defined as an attribute. For instance, for a product dimension, three hierarchical levels may be defined as attributes: department (e.g. electronics, books, etc.), category (e.g. TVs, computers, etc.) and brand. However, ontologies hierarchies are usually not regular, meaning their hierarchical branches are not balanced (do not have the same height). Therefore, in this case, we can use a generic solution. Instead of creating attributes named after a category, only two attributes are created: a generic attribute that represents the parent, and another generic attribute that represents the category names. For instance, in the case of an organism taxonomy, the *Azorhizobium* term belongs to the genus category level (category name attribute). Its parenthood attribute points to another term, the *Xanthobacteraceae*, which belongs to the family category level.

A DM is by definition a historic database, meaning it maintains historical data. Therefore, there is always a time dimension that must be defined. In the case of a scientific corpus analysis, it makes sense to analyze scientific articles publication at least on a monthly basis, not less. Therefore, a time dimension with attributes that characterize each year/month would be sufficient. Finally, once defined and characterized, dimensions may be implemented as tables and populated. This will be described further below.

3. Cut-off of the ontologies. Usually, ontologies of the Biomedical areas are very large, including concepts of many subdomains. For example, the National Cancer Institute (NCI) *thesaurus* ontology [34] describes various knowledge fields such as *organisms*, *drugs*, *chemicals*, *genes*, *activities*, *biological process*, etc. It is important to separate and select hierarchies, one for each dimension. First, hierarchies that do not include concepts raised at the previous step must be eliminated. Fig 4 shows the selection of the *biological process* hierarchy from the NCI ontology. In addition, for each selected hierarchy, some branches may not be of interest according to the analytical demands. Therefore, these branches may also be eliminated. Fig 5 shows an example where for the *biological process* hierarchy, only the *pathologic process* and the *multicellular process* branches are kept.

For the remaining selected set of hierarchies, based on the analytical questions demand, concepts that are combined in the same analytical question should be kept in distinct dimensions, i.e., each one must constitute a distinct dimension. For instance, if the demand needs to identify combinations of organisms and chemicals, these hierarchies should correspond to distinct dimensions. On the other hand, consider the case where two distinct concepts are used in many combinations, but never combined with each other. If both can be generalized under a single generic concept, then, both can be part of the same dimension. For instance, as shown in Fig 5, if there are analytical questions involving the *fibroplasia* and *necrotic process* branches but without the need to correlate them then these can be kept in the same dimension. On the other hand, if the analytical questions address issues such as "*Which types of fibroplasias are most often cited with some type of Necrotic Process?*", then in these cases, they must be kept on different dimensions, despite belonging to the same branch, as descendants of the pathological processes concept.

In the ontologies handling process, distinct situations may occur. The first situation occurs when the ontology specifically describes the concept of interest. In this case, it is necessary to perform a simple operation to extract the elements from the ontology to the dimension table, the way of visiting the nodes of the ontology hierarchical tree is indifferent.

The second situation happens when the ontology describes more than one concept and it is necessary to extract only the one of interest. If the classes that describe the concepts of interest are concentrated in an ontology branch, it is possible to extract this branch by means of a software. Or based on a bottom-up span of the ontology hierarchy, to analyze the classes identifying which belong to the branch of interest and to copy these elements to another tree.

The third situation is the most delicate. The ontology can describe several concepts and the classes that describe the concept of interest are not concentrated in a specific ontology branch. They are scattered in different branches and in this case, it is necessary to separate them in a module (subset). To create this module, a search in the ontology hierarchy is made with the aim of identifying and separating them from the main ontology. With all the classes identified, a new root node is created and the parent_id attribute of the classes of interest is pointed to the new root node. So, at the end of this process, a module of smaller size and more specific is defined, as shown in Fig 6.

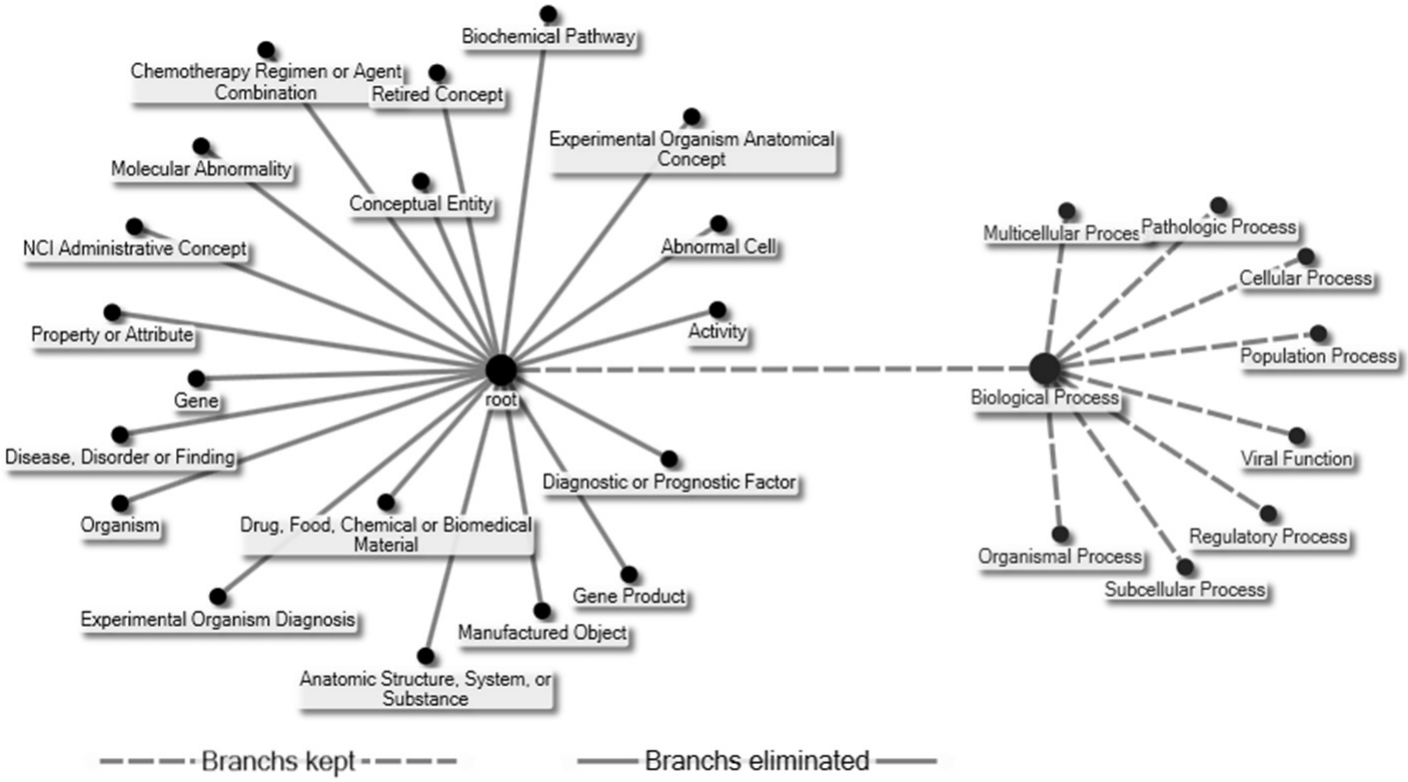

Figure 4: **Selection of the biological process branch.** This figure presents as the concepts present in the ontologies whose are not part of the user's interest must be eliminated.

4. Fact Identification/Definition. The analytical demands for a corpus analysis typically involve finding the number of occurrences of terms throughout the articles in the corpus. Examples of such questions are: (i) "*How many articles mentioned a given term?*", (ii) "*How frequently cited was a specific term throughout articles of a corpus?*", (iii) "*How often two (or more) terms are mentioned together in the same article*", (iv) "*How many occurrences of a term (or two or more terms) are there inside each article*". All these queries aim to observe the number of occurrences of terms, either throughout the corpus, or inside an article. In the first case, a single presence of a term in an article counts. Therefore, the fact would observe the counting for each combination of values of *d1, . . . , dn*, and month/year, where di corresponds to each selected ontology hierarchy, defined in the previous step.

For the second case, for questions like (iv), it is important to count the number of occurrences inside the article. The later would be useful to calculate term relevance with respect to the whole corpus, such as the TF-IDF metric [31]. In this case, the article itself emerges as a dimension. Then, the fact would observe the counting for each combination of values of *d1, . . . , dn*, month/year, and article where di corresponds to each selected ontology hierarchy, defined in the previous step.

A third alternative for the fact identification is to have a Factless situation. Suppose it is necessary to represent the simple yes or no occurrence of a term, meaning this is the only information that matters. In this case, no counting of occurrences is needed. However, it is important to register if a specific combination of terms occurs in a specific article, published on a specific month/year.

Also at this stage, the grain definition is performed. As mentioned in section III.B.3, the grain establishes the representation of a single tuple in the fact table. This definition is important because it will determine which is the maximum level of information detail that the user will view.

C. Load Step

For a relational implementation, the star schema is the usual choice. Each defined dimension may be implemented as a table. The population of each dimension table can be automated using as input the corresponding ontologies

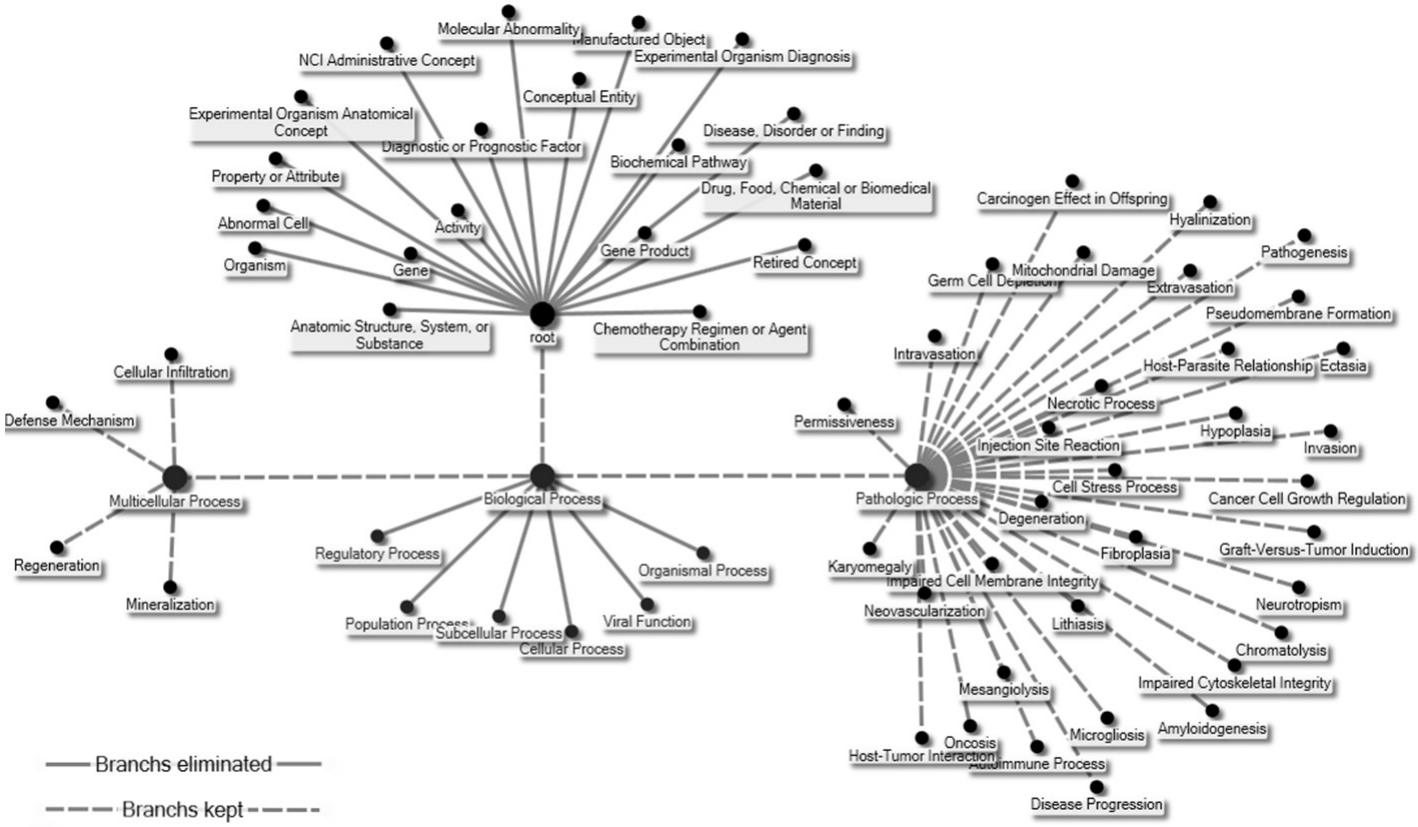

Figure 5: **Selection of the pathological process and multicellular process branches.** This image represents how analytical issues influence the choices of the branches of the ontologies. How much more the questions involved the deeper classes of the ontology, the tendency is that more branches to be eliminated.

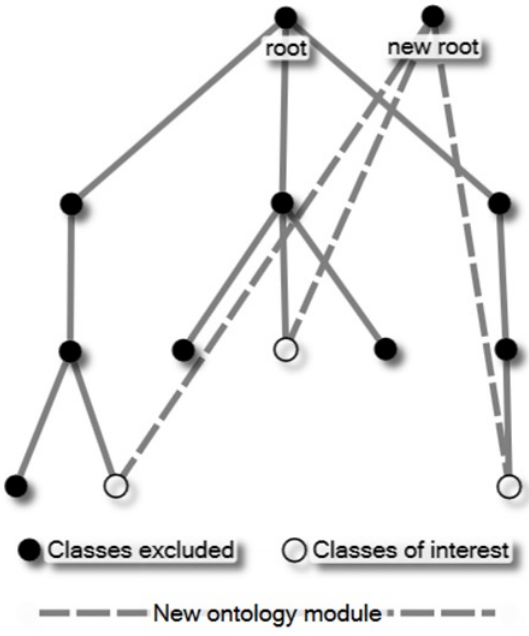

Figure 6: **Ontology module definition.** This figure presents the identification process of the classes of an ontology that describe the terms of interest. Thus, It is possible to extract them, generating a module of smaller size and focused on the concept of interest.

in a structured format, such as RDF/OWL XML file format. For each dimension, the following relational schema is used, and may be extended with additional attributes.

DimensionI(id, concept_id, concept_label, parent_id, category_level), concept_id is unique and not null, parent_id references DimensionI(id).

According to this schema, each concept of the hierarchy is identified by a surrogate key (id) and the id of the concept used by the ontology representation (concept_id). Both are candidate keys, but the surrogate key is chosen as the primary key. In addition, each dimension tuple may point to another tuple in the same dimension, to maintain the hierarchical information level.

The arise of multiple inheritance in the hierarchical ontology tree is another reason not to choose the field concept_id as key. This kind of relationship allows a node to have more than one parent node. To represent these relations, the dimension tables record a new tuple for each parent node, keeping the same value of the field concept_id, invalidating the choice of this field as key.

Keeping information that allows representing the hierarchical levels of ontologies is of vital importance. Thereby, all the wealth of information contained in the hierarchies is transferred to the DM. This greatly increases the ability to handle information from the most specific to the more generic levels of a concept. It is by means of this hierarchical information that the DM is able to perform citation counts of more generic terms even when they are not explicitly cited.

Differently, the article dimension may be implemented with attributes that characterize the publication, such as title, link for the pdf file, etc. A simple and basic schema for this dimension could be the following:

Article(id, title, link).

The population of the Article dimension should aim to keep information to facilitate the access to the document, because searches with purpose to find the articles of interest are quite common. It is not necessary to keep data about publication dates of articles in this dimension, these data are of responsibility of the MonthYear dimension table.

Despite the fact that the publication date (day, month and year) is an attribute that characterizes an article, to represent it as a separate dimension allows to characterize the whole fact and not only the article. Thus, queries such as "*Which proteins have been cited in a specific year and that had never been cited previously?*" are possible to be executed. For the month/year dimension, a typical schema, based on usual time dimension implementations is the following:

MonthYear(id, month, year, month_name, quarter, semester, quarter_name).

The fact is also implemented as a table. The schema for the fact table of the scientific annotation data mart is the following:

Fact(id, dim1_id, dim2_id, dim 3_id, dim4_id, mmyy_id, article_id), dim1_id references Dimension1(id), dim2_id references Dimension2(id), dim3_id references Dimension3(id), dim4_id references Dimension4(id), mmyy_id references MonthYear(id), article_id references Article(id).

However, the population of Fact table is not an easy task. Each annotation table stored at the staging area should be read in parallel in order to build the fact tuples. For each article stored in Article table, a list with all the annotated terms is created. This is accomplished through the article_id field of the annotation table. The terms are added to this list only once, no matter how many times they are present in the annotation table.

This is necessary because in the fact table is not registered the frequency of occurrence of the terms in an article, but only if they are or not cited. In this list are present the annotated terms of all dimensions contained in the article. To start the fact table population, it is necessary to classify the terms by dimension. Then for each dimension a list is created and through the class_id field the terms are classified and grouped in each list.

The representation of the facts in an article is the combination of all the elements present in the lists. For example, if we consider three dimensions and an article has two terms in each dimension (i.e, the lists of each dimension have two elements), eight tuples will be needed to represent the facts, basically is performed the cartesian product among the elements of these lists. Fig 7 shows the overview of the identification facts process in an article.

When a certain list is empty, it means that no term of a specific dimension has been annotated. For these cases, it is not advisable to leave the null value in the tuple column [24]. Thus, for each dimension, a default value is created to represent these cases. When there is no citation of a certain dimension in an article, this default value is referenced in the corresponding dimension reference field of the fact tuple. At the end of this process, the design (Fig 8) and population of the data mart are completed.

Fig 9 shows the overview and flow between the steps of the TOETL methodology.

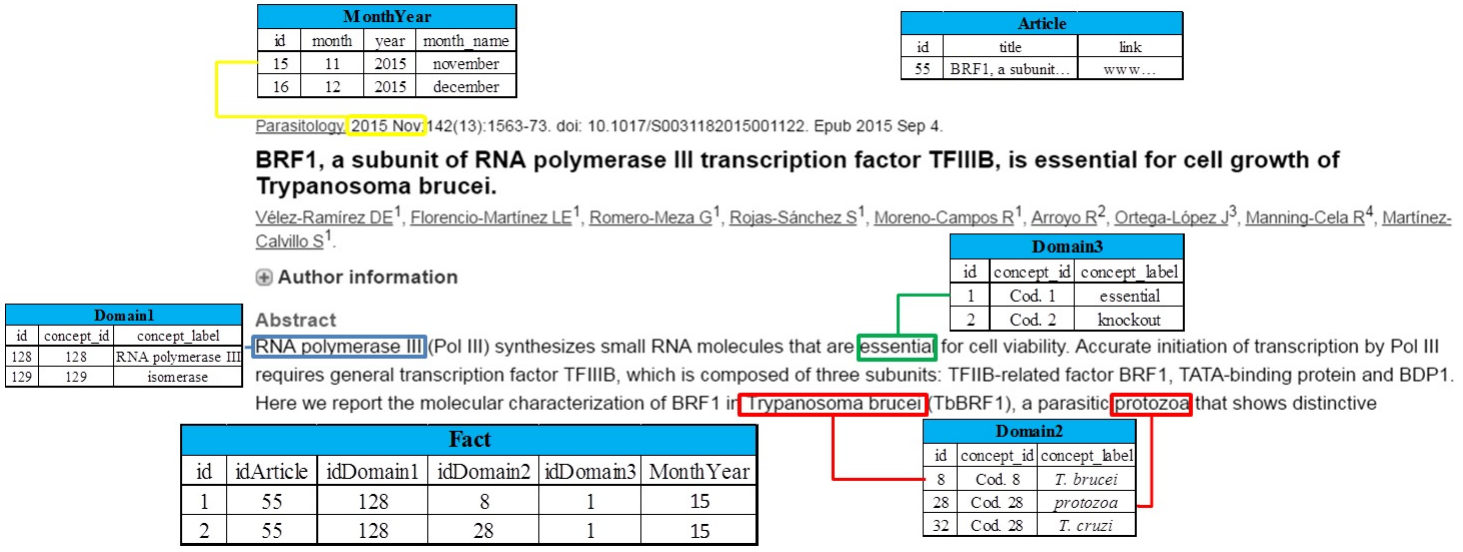

Figure 7: **Fact table generation** Note that the fact table is a factless tables where references to the dimensions table are stored. Another peculiarity of the fact table is the combination among all the terms of each dimension present in the article.

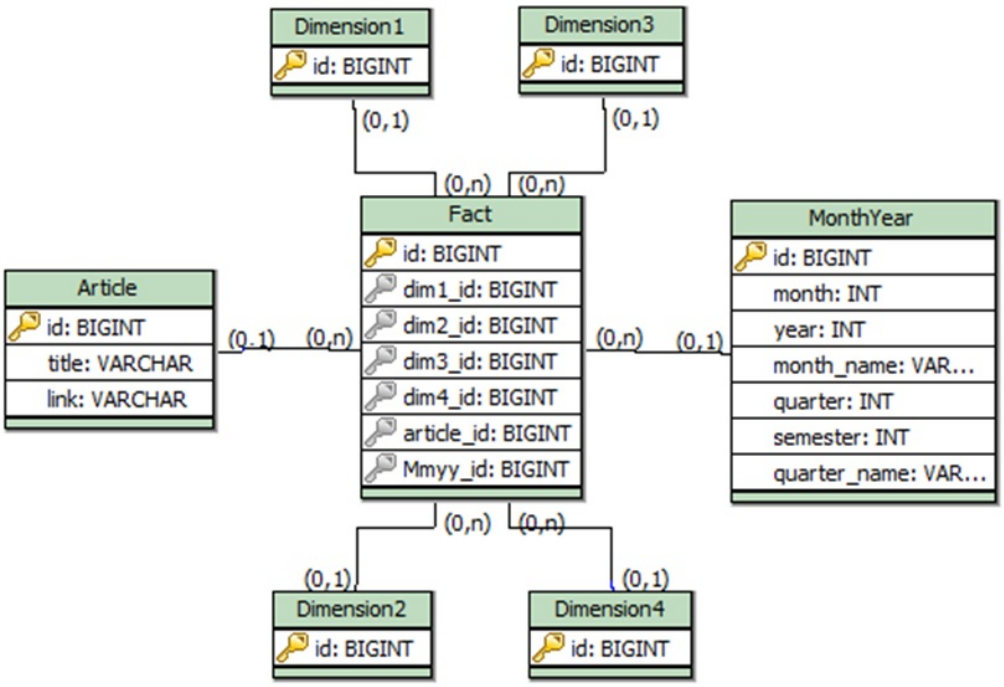

Figure 8: **Star schema.** DM design based on star schema where a fact table (factless table) stores references to dimension tables.

IV. Prioritization of Drug Targets Case Study

Based on the methodology described in the previous section, the following subsections describe a case study that applies such methodology with focus on the research of new drug targets, and their prioritization. Both stages of the proposed methodology are described. In the first one, we built an annotation database, which was the staging area database for the second stage, i.e., the Tap DM construction and population.

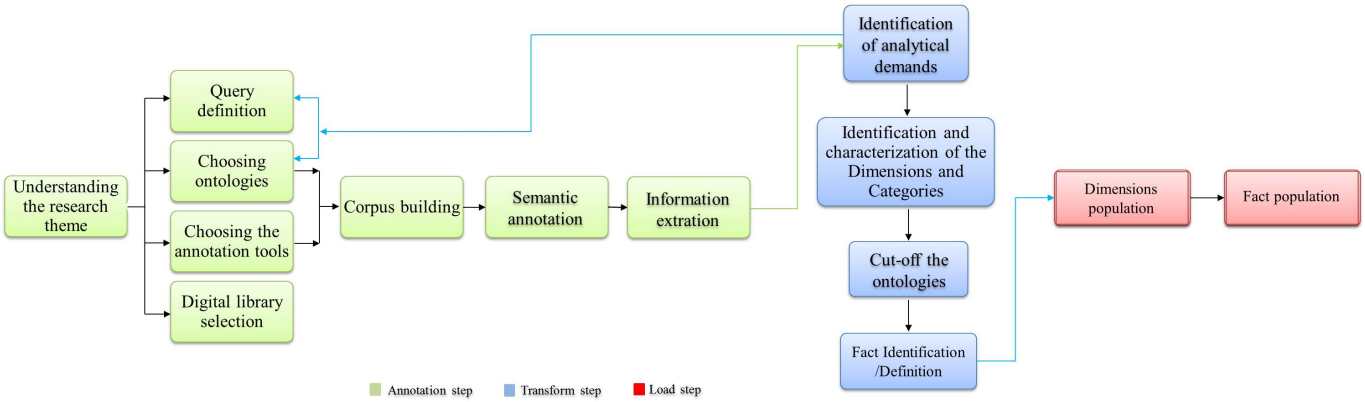

Figure 9: **TOETL overview.**

Semantic Annotation

A. Annotation Stage: Building Tap Staging Area Database.

In order to generate a sample of articles that will serve as a data source for the datamart, terms involving organisms of interest and techniques of essentiality were identified. This step, known as understanding the research theme, was fulfilled through meetings with experts and resulted in the following list of terms: Gene Essentiality: gene, protein, essential, essentiality, reverse genetic, knockout, knockdown, RNA interference, RNAi, lethal phenotype, survival, null mutants. Organisms(protozoan targets and models): *Entamoeba histolytica*, *Leishmania major*, *Plasmodium falciparum*, *Trypanosoma brucei*, *Trypanosoma cruzi*, *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Danio rerio*, *zebrafish*, *Mus musculus*, *Saccharomyces cerevisiae*, *baker's yeast* and *Escherichia coli*.

From the combination of these identified terms were built 21 queries (query definition step) that were then submitted to a repository of scientific articles. In this scenario, PubMed Central (PMC) was chosen as a textual base since it concentrates thousands of citations and abstracts about themes in health and biomedical areas (*digital library selection* step). The queries produced a total of 1383 articles (*corpus building* step).

In order to perform the semantic annotation, it was necessary to define the set of related ontologies (*ontology selection* step). At the present time, there are many ontologies on the biomedical domain that can be found at The Open Biological and Biomedical Ontologies (OBO) Foundry and at The National Center for Biomedical Ontology (NCBO) BioPortal. Besides being a repository, the OBO Foundry [35] plays the role of a reference organization, which reviews and certifies a set of ontologies on the biomedical domain. The NCBO BioPortal [39] is another ontology repository that provides more than 500 ontologies, and most of them have more than 500 classes [5], characterizing them as medium to large size ontologies.

To select the ontologies it was used the Annotator tool [23]. It is able to annotate with all ontologies present in the NCBO repository, simultaneously, although it only annotates texts with up to 500 words. Thus, fragments of a set of 44 articles were annotated. Based on the annotation results, three out of the five top ontologies in terms of annotation coverage were selected:

- Molecule Role [41]: represents a controlled vocabulary of proteins, protein families and chemicals;
- NCBI Organismal Classification (NCBI Taxon) [15]: represents the taxonomic classification of live organisms;
- NCI *thesaurus* (NCIt) [34]: contains a vocabulary to represent medical care, basic and translational research, information to the public and administrative activities. It includes terms related to gene essentiality techniques.

With the set of articles and ontologies defined, the *semantic annotation* step was ready to start. But before, it was necessary to submit the articles to a cleaning process, where figures, references and stop words were removed. Then using the annotation tool Autômeta [16], the articles were annotated intrusively, i.e., the terms cited in the text that belong to ontology were marked by an RDFa (Resource Description Framework in Attributes) tag.

After the *semantic annotation* step, the extraction of the annotated data (*information extraction* step) was performed through the RDFa API and stored in a database table. Each tuple on this table is composed of the article

identifier, the annotated term, the ontology class identifier and the ontology class name, as shown in Table 1. The next section describes the Data Mart construction Stage, which includes the identification of the main dimensional elements (dimension and fact tables) that will compose the star schema modeling.

Table 1: **Result of extracting data.**

Annotation				
id	idArticle	term	idClass	class_name
1	5	Protein	IMR_0000001	MoleculeRole
2	2	small GTPase	IMR_0000914	GTP-binding protein
3	38	transferase	IMR_0000207	enzyme

B. Transform Step: TaP Data Mart Modeling and Population.

1) Identification of Tap Analytical Demands. As a case study, this paper presents the construction of a data mart involving the prioritization of new drug targets, focusing especially on the gene essentiality techniques. As mentioned in section III.B.1, the first step in the design and construction of the tool is the lifting of the questions that it must answer. It was accomplished by means of interviews with experts in the knowledge area, where questions such as "*What technique was most often cited with some specific organism?*","*What organisms were most often cited with some specific protein?*,"*Which articles cites an organism and leastwise 2 gene essentiality techniques?*" and "*What chemical was most often cited with some gene essentiality techniques?*" were identified. They served as information source for this first stage of the data mart design process. So the terms present in the mentioned analytical questions were organized and classified in 6 concepts identified: protein, organism, gene essentiality techniques, chemical, article and time.

2) Identification and characterization of Tap Dimensions and Categories. As mentioned in the item III.B.2, the dimensions store the ontologies that describe the concepts involved in the DM. All the data and hierarchical organization must be stored by the dimensions. With the identification of 6 concepts involved in DM and the ontologies that best describe them, the identification of the dimension tables is clearer. So in this case study, 6 dimension tables were identified: DimProtein, DimChemical, DimOrganism, DimTechnique, DimArticle and DimMonthYear.

The protein and chemical dimension tables are designed to store information and the hierarchical classification structure of proteins families/subfamilies and chemicals respectively. This information was extracted from the Molecule Role Ontology (OWL representation file), available from the OBO Foundry site. Using the software Protégé [30] it was possible to separate the two (protein and chemical) ontology branches, storing them in different files and tables. The protein dimension (DimProtein) and the chemical dimension (DimChemical) tables have the following attributes: ID (identifier), concept_id(OBO identifier), concept_label (protein/chemical name), parent_id (element parent id) and category_level (element hierarchical level), as shown in Table 2 and Table 3:

Table 2: **Protein dimension table.**

DimProtein				
id	concept_id	concept_label	parent_id	category_level
1	IMR_0000002	ligant	5	2
2	IMR_0000003	Wnt	6	3
3	IMR_0000004	hedgehog	8	6

The organism dimension table is accountable to store structured information on organisms and their taxonomic classification. The information was extracted from NCBI Taxon ontology provided by the National Center for Biotechnology Information (NCBI). The organism dimension (DimOrganism) table has the ID (identifier), concept_id (NCBI identifier), concept_label (organism name), parent_id (element parent id) and category_level (element hierarchical level) fields, as shown in Table 4.

The technique dimension (DimTechnique) table is designed to store the terms, its variations and synonyms associated to the gene essentiality techniques. These terms were presented at the section II. Each tuple is composed

Table 3: **Chemical dimension tables.**

DimChemical				
id	concept_id	concept_label	parent_id	category_level
1	IMR_0000947	chemical	90	2
2	IMR_0001349	nucleotide	50	3
3	IMR_0001351	ATP	80	6

Table 4: **Organism dimension table.**

DimChemical				
id	concept_id	concept_label	parent_id	category_level
1	2	bacteria	5	13
2	6	azorhizobium	1254	15
3	13	dictyoglomus	658	8

of the following attributes: ID (identifier), concept_id(NCI *thesaurus* identifier), parent_id (element parent id), concept_label (technique name) and parent_id (element parent id) and category_level (element hierarchical level) as shown in Table 5.

Table 5: **Technique dimension table.**

DimTechnique				
id	concept_id	concept_label	parent_id	category_level
1	C22491	knockout	1	1
2	C20153	RNA Interference	1	1
3	C71578	essential	1	1

The paper dimension (DimArticle) table is responsible for storing the information of the scientifics articles that were selected for data extraction. This table has the following attributes: id (identifier), title (article title) and link (article path), as shown in Table 6.

Table 6: **Article dimension tables.**

DimArticle		
id	title	link
1	The Proteins...	www...
2	The cell...	www...
3	Aproaches...	www...

The time dimension (MonthYear) stores data about the article publication dates. It allows the data mart to exhibit the data over time. This is fundamental for the researcher to do the analysis of historical information, contributing to future decisions. This table has id (identifier), month (article publication month), year (article publication year), month_name (month name), quarter (quarter of the year), semester (semester of the year) and quarter_name (quater name) attributes, as shown in Table 7.

3) Cut-off of the ontologies. The DimOrganism table stores the informations contained in NCBI Taxon ontology. The DimTechnique table stores the informations contained NCI *thesaurus* ontology. But analyzing the data in DimTechnique table, we concluded that information about the hierarchy of elements would not be useful for users. Performing roll-up and drill-down operations in this dimension adds nothing to the data analysis.

Furthermore, the NCI *thesaurus* ontology is very broad and describes several concepts from different areas of knowledge. The classes that describe the essentiality study techniques represent a very small part of the ontology

Table 7: **MonthYear dimension table.**

MonthYear						
id	month	year	month_name	quarter	semester	quarter_name
1	1	2010	january	1	1	first quarter
2	2	2010	february	1	1	first quarter
3	3	2010	march	1	1	first quarter

and in addition, these classes are dispersed in several ontology branches. Considering these characteristics, the DimTechnique table does not store all the NCI *thesaurus* ontology, but a specific module. This module contains all classes of interest, in this study case, these classes are responsible for describing the terms raised about essentiality (item IV.A). The building this module followed as described in item III.B.4.

The DimProtein and DimChemical tables store information contained in the Molecule Role ontology. The later has two main branches (proteins and chemicals) connected to the root node of the hierarchy. With the purpose of allowing the DM to correlate these two concepts, these two branches were stored in different dimensions.

Another factor must be observed in this situation, even without the need to correlate between these two concepts, storing the Molecule Role ontology completely in just one dimension could cause some kind of inconsistency in the DM response. For example, the query *How many proteins are cited in a particular article?*, if all Molecule Role ontology is stored in the DimProtein table, the DM would consider the chemicals present in the ontology such as proteins. For this reason, the answer to this question would not be correct. It is very important to select the branches that will be stored by the dimension tables. The implementation of the DimArticle and DimMonthYear dimension tables was performed as described in the item III.B.4.

4) Identification of Tap Facts. In this case study, the fact does not include observing term occurrences inside the article. However, at first it was important to keep track of the id of the articles. Therefore, we chose the factless approach, i.e., each fact represent the occurrence of a given combination of three or more terms, from the defined dimensions, in a specific article, published in a specific month/year.

By consequence, using a factless table, the occurrences of the dimensions terms are made by the recording of its references. Therefore, each tuple of the fact table stores the identifiers of terms of each dimension present in an article. Thereby, in this study case, each tuple of the fact table consists of 7 fields, its key field and the 6 fields connecting to the 6 dimension tables. Considering a hypothetical case in which an article is annotated with one term of each dimension, one tuple of the fact table would be enough to represent the presence of the dimension tables in this article.

But in most cases, in a single article occurs the presence of more than one term of one dimension. Thus, all these elements are combined with the cited elements of others dimensions, as mentioned in item III.B.3. Cases where there is no annotation of terms from a specific dimension, was adopted the reference 99999999 pointing to the NA(not annotated) term of the dimension. In this case study, an article is recorded in the fact table, if it cites at least 3 dimension tables, excluding the MonthYear and Article tables, since every article already contains them. This was a choice of implementation, once for the researchers, registering articles where only one or two dimensions occur isn't meaningful. Table 8 shows examples of facts.

Table 8: **Fact table.**

Fact					
id	idDimArticle	idDimChemical	idDimProtein	idDimTechnique	idDimOrganism
1	20	5	10	3	6
2	21	2	15	3	7
3	21	2	15	4	7
4	22	8	12	99999999	5

It is important to emphasize the definition of grain in the case study. The grain represents, in an article, the presence of at least 3 dimensions in a certain month of the year. The month information is taken from the date of publication of the article. It was established as the smallest unit, because there is no need to perform analyzes for

smaller temporal spaces (hour, day, etc..). Starting of the month unit and by means of the drill-up operations, it is possible to show more general analyzes, as for example bimonthly, quarterly and yearly analyzes.

At the end of this process, the data mart is designed, built and populated, as shown in Fig 10. The MySQL database management system (DBMS) was used to deploy the tool, enabling consultation and to check if the tool is able to correlate data from different knowledge fields. Answering experts' queries in item V of the next section are some examples.

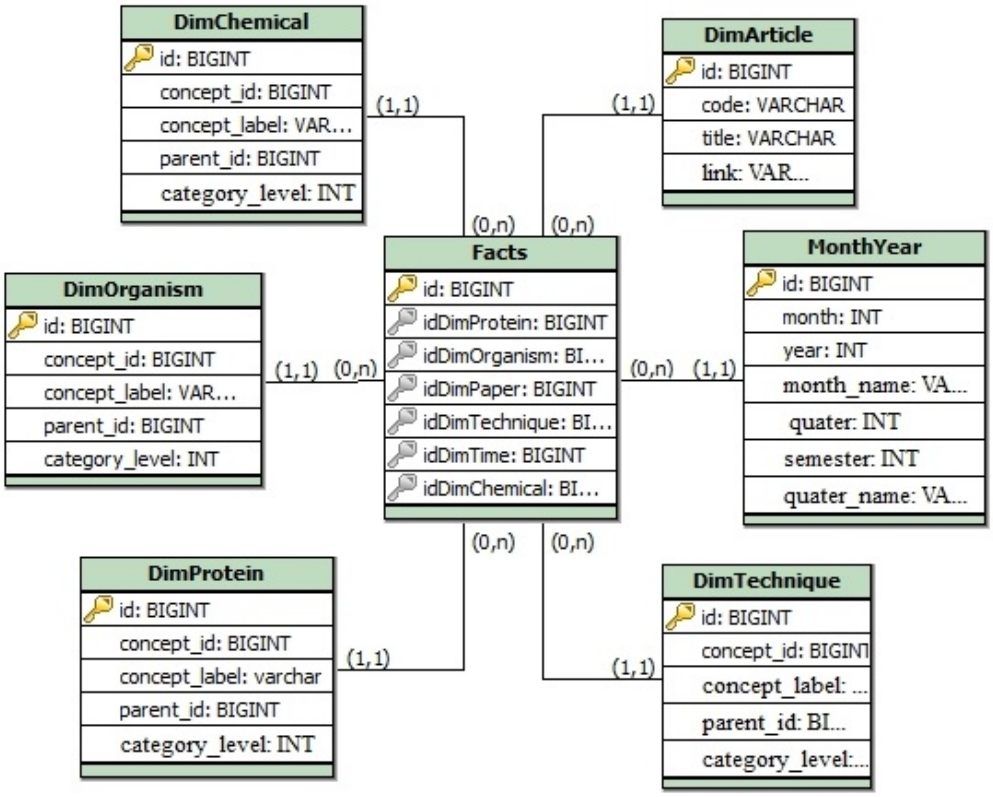

Figure 10: **Data Mart design.** This is the final design of Tap DM that was populated using data extracted from articles.

V. Case Study Results and Discussion

In this section we present some examples of queries submitted to the *Tap* DM. They have the aim to show how a decision support tool allows more flexibility in the search for information. Another evident characteristic of the results, is the easiness of correlating several concepts, besides capturing information that are not explicitly found in the articles. Two queries are presented: (i) one that can be answered directly through the OLAP interface, and (ii) another one that is more complex and demands a sequence of queries to be answered. It is worth to mention that, although (i) and (ii) are specific questions, other similar questions may be formulated in the same way.

(i) What is the citation history of the NMT1 protein with *Saccharomyces cerevisiae* organism? This query shows that it is possible to combine different concepts and to study the other elements over time. Table 9 shows the query result. Protein NMT1 has been systematically cited throughout the years (in average, every 2 years) and it appears with several chemicals in 9 articles. Analyzing these articles (PMC1460106, PMC3818516, PMC2643609, PMC1694798, PMC1061621, PMC181578, PMC1461765, PMC25529 and PMC2139860) where the NMT1 and *Saccharomyces cerevisiae* concepts manifest, in none of them these terms are directly connected, in other words, the articles do not report a relation between the two concepts. But the constant coexistence of concepts over time could indicate a possible relationship and it was confirmed by the UniProt[1] database, NMT1 belongs to *S. cerevisiae* organism.

[1] http://www.uniprot.org/uniprot/P14743

Table 9: **NMT1 protein and *Saccharomyces cerevisiae* organism historical data.**

Year	Amount of article	Amount of chemical
2013	1	10
2008	1	7
2006	1	11
2004	1	2
2003	1	6
2001	1	16
1999	1	6
1997	2	17

These results are an evidence that the DM is able to show data trends. Exploring its flexibility, it is possible to establish parameters of interest to each user and to obtain specific answers for each situation. In this query, with the combination of the protein and organism concepts, the trend of time, article and chemical concepts could be observed.

It is important to note that a correlation between two terms found at the Tap DM does NOT mean that there is a biological relationship between them. For instance, even if a protein is not expressed by an organism, it can be correlated to it when they are cited together in an article.

For this reason, the definition of the time-interval of the selected articles (corpus), is crucial. The larger the time-interval is, the richer the results of such time-based queries can be. This allows the identification of co-ocurrences of the same concepts along several time points, which increases the likelihood of these concepts to have a biological relationship. In other words, this ability is important to reveal hidden relationships between concepts that still have not been explicitly established, expanding new possibilities for users.

This example is just one among the various analytical queries that can be submitted to the DM. In contrast, keyword-based traditional tools, although more commonly used than DM, do not allow this kind of refinement.

(ii) What are the best possible new targets for a particular organism? To address the (ii) question, a complex procedure, which involves a set of queries should be performed. It is detailed as follows, using Relational Algebra expressions. But before, let us assume the following definitions to facilitate the understanding of the query expressions:

- P : proteins set (represented by the DimProtein table)
- O : organisms set (represented by the DimOrganism table)
- F : fact set (represented by the fact table)

In order to find new targets, then the candidate proteins are proteins that have never been cited previously with that particular organism. Queries 1 and 2 may be used to get those proteins. The first query retrieves all Proteins that were cited with that particular organism, while the second one uses the result of the first query to retrieve all other proteins that were not cited with that organism (NPC).

$$PC \leftarrow \Pi_{concept\_id,concept\_label}\left((\sigma_{concept\_label='particular\_organism'}(O)) * F * P\right) \tag{1}$$

$$NPC \leftarrow \Pi_{concept\_id,concept\_label}(P - PC) \tag{2}$$

According to the main premise of this case study, for these new proteins to have a higher likelihood of being new targets for the organism of interest, they must be cited with the techniques of essentiality. To flexibilize this requirement, the $nTec$ parameter was used in the query and can be changed according to user needs. It determines the amount of distinct techniques of essentiality that must be cited with the protein for it to be considered a possible new target. Thus, increasing the $nTec$ parameter value, the probability that the protein has an important role in the essentiality study also increases. Queries 3 and 4 may be used to retrieve, from the NPC set, the set of the proteins that are cited with more then $nTec$ essentiality techniques. Thus, the NTP is the new set of target proteins.

$$Q_0 \leftarrow_{concept\_id,concept\_label} \xi_{countdistinct(idDimTechnique)}\left(NPC * F\right) \tag{3}$$

$$NTP \leftarrow \Pi_{id,label}\left(\sigma_{tottech \geq nTec}(\rho_{Q_0(id,label,tottech)}\left(Q_0\right))\right) \tag{4}$$

Once defined the *NTP* set, how should we correlate its proteins to the organism of interest since they were never mentioned together? The strategy adopted was to correlate NTP proteins to the PC proteins that have an important role in the gene essentiality study for the organism of interest. In other words, if they are cited together in most of the articles, probably the NTP protein may have a high probability of being essential.

In order to do that we assumed that proteins have an important role in the essentiality of a particular organism, when they have been cited with that organism and the techniques of essentiality throughout a given time. In this context, the *nYear* parameter was created to determine the number of years that the protein must be cited with the techniques of essentiality, for it to be considered a protein that plays an important role in the organism. Similar to the *nTec* parameter, increasing the *nYear* parameter value, the probability that the protein has an important role in the essentiality study also increases. Queries (5-7) retrieve the set of the essentiality related proteins (*EP*).

$$Q_1 \leftarrow_{concept\_id,concept\_label} \xi_{countdistinct(idMonthYear),countdistinct(idDimTechnique)}\left(PC \star F\right), \tag{5}$$

$$Q_2 \leftarrow \sigma_{cYear \geq nYear}\left(\rho_{Q_1(id,label,cYear,cTec)}(Q_1)\right), \tag{6}$$

$$EP \leftarrow \Pi_{id,label}\left(\sigma_{cTec \geq nTec}\left(Q_2\right)\right) \tag{7}$$

Figure 11 gives an overview of the approach and illustrates the relationships among the organism of interest, the proteins with more probability to have an important role in the study of its essentiality (*EP* set) and the proteins never mentioned with it (*NTP* set). Note that these are disjoint sets.

After defining the *EP* and *NTP* sets, it is necessary to calculate the rate between their elements. As mentioned before, the idea is to correlate NTP proteins and EP proteins, by identifying those EP proteins that are essential to some organism and that are cited with some NTP protein. This can be done by calculating for each pair of proteins (P_{in}, P_{jn}) that belongs to the set $\{EP \times NTP\}$, the rate that they co-occur throughout the articles. In (1), $cNTP_{jn}$ is the number of articles in which P_{jn} was cited and $cNTP_{jn}EP_{in}$ is the number of articles that P_{jn} was cited with P_{in}. The more they co-occur, the higher is the TR_{jnin} rate. If it is close to 100, it means that P_{jn} always occurs with P_{in}, and according to our assumption, that it has a high probability of being essential to the organism of interest.

$$TR_{jnin} = \left(\frac{cNTP_{jn} * 100}{cNTP_{jn}EP_{in}}\right) \tag{8}$$

In order to study this strategy applied to the Data Mart, the *T. brucei* was selected as the organism of interest. Queries (3) and (4) were performed, obtaining 1159 and 23 proteins, for *NTP* and *EP* sets, respectively. The *nTec* and *nYear* parameters were set to 3 and 7, respectively. Thus, for a protein to belong to the *NTP* set, in addition to have never been mentioned with the organism of interest, it must be cited in all articles where it appears at least essentiality-related techniques/terms. To belong to the EP set, a protein term must achieve this standard by at least 7 years, jointly citing the organism of interest. The values of the nTec and nYear parameters were defined with the objective of obtaining proteins with characteristics belonging to the NTP and EP sets. Obviously, these values can be redefined in order to restrict or flex these properties.

Then, query (6) was submitted to the TaP DM and the TR rate was calculated for each pair of proteins belonging to the set $\{EP \times NTP\}$. The pairs of proteins with *TR* equals to 100.0 (selection criteria) were selected and resulted in 302 (P_{jn}) proteins. These 302 proteins belong to NTP set and, in other words, they have never been cited with the *T. brucei* and may have no relation to it. To select only proteins that have a relationship with the organism of interest, it is necessary to validate these proteins using another data source (database validation). With this purpose in mind, the UniProt [2] and TdrTargets [26] databases were used to identify the *T. brucei* proteins. The proteins involving the subspecies of *T. brucei* were also considered. At the end of this process 91 proteins were selected (Supplementary Material), and they are the most likely to be potential new targets for *T. brucei*.

Some of those proteins were validated by just one of those databases due to the use of different terms for the same protein. For instance, the *PI3-kinase* (synonym of O18683_DROME) protein term is present in the TdrTargets and TaP DM, but the Uniprot database uses the term *Phosphatidylinositol 3-kinase, putative*.

To validate the resulting 91 proteins, the TdrTarget database was used and 39 proteins were positively confirmed as targets. But the same problem mentioned early, about the different terms adopted by databases, happens in this case. Therefore, the 52 proteins that were pointed out by the Data Mart as possible targets, but were not

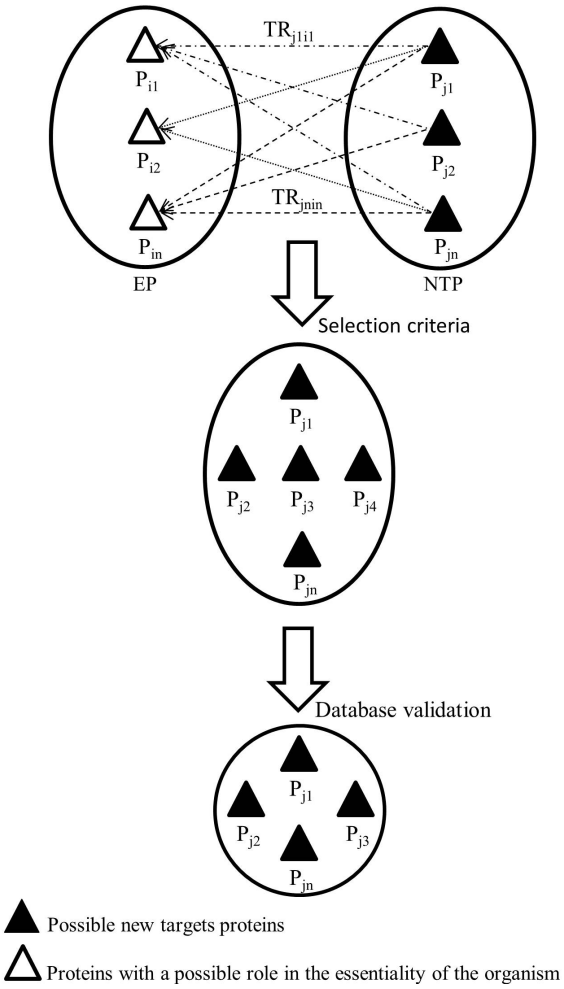

Figure 11: **Overview of the search strategy about essentiality.** The TaP DM allows to relate data creating strategies to focus on research goals, it combines elements to infer relations

confirmed by TdrTargets, may still prove to be validated targets. As an example let us consider the protein *phosphofructokinase*. It is among the 52 proteins lacking validation because the TdrTargets employs the term *ATP-dependent phosphofructokinase* instead. Thence, the DM is correct in selecting the protein *phosphofructokinase* as a possible target. Figure 12 demonstrates with greater detail how the process of data analysis was done to identify possible new targets for the organism *T. brucei*.

VI. Tap OLAP Tool

OLAP tools are able to handle large amounts of data, providing an intuitive user interface to build queries interactively, without the need to learn a new technology or language. Among several commercial tools available, for this study it was chosen the Pentaho suite. Pentaho is an open source tool. It includes an OLAP for Web interface called Mondrian [9].

Figure 13 shows Mondrian interface, where in (a) the user is able to have an overview on the dimensions that the DM covers. Figure 13(b) shows the number of organisms cited together with specific chemicals and the years in which this occurred. For instance, the *inositol phosphate* chemical was cited with 4 organisms in 2008 and 21 organisms in 2015. Figure 13(c) presents the number of essentiality techniques cited with specific proteins.

All the examples presented in (a), (b) and (c) are obtained simply by manipulating the hierarchies (clicks on the "+" and "-" icons), thus, by means of drill down/up operations, the users can browse the whole hierarchies. These examples show data about only some dimensions, but it is easy to add/combine all dimensions and obtain other interesting analytical views.

Different from traditional keyword-based tools, the TOETL approach is able to retrieve articles based on ontology

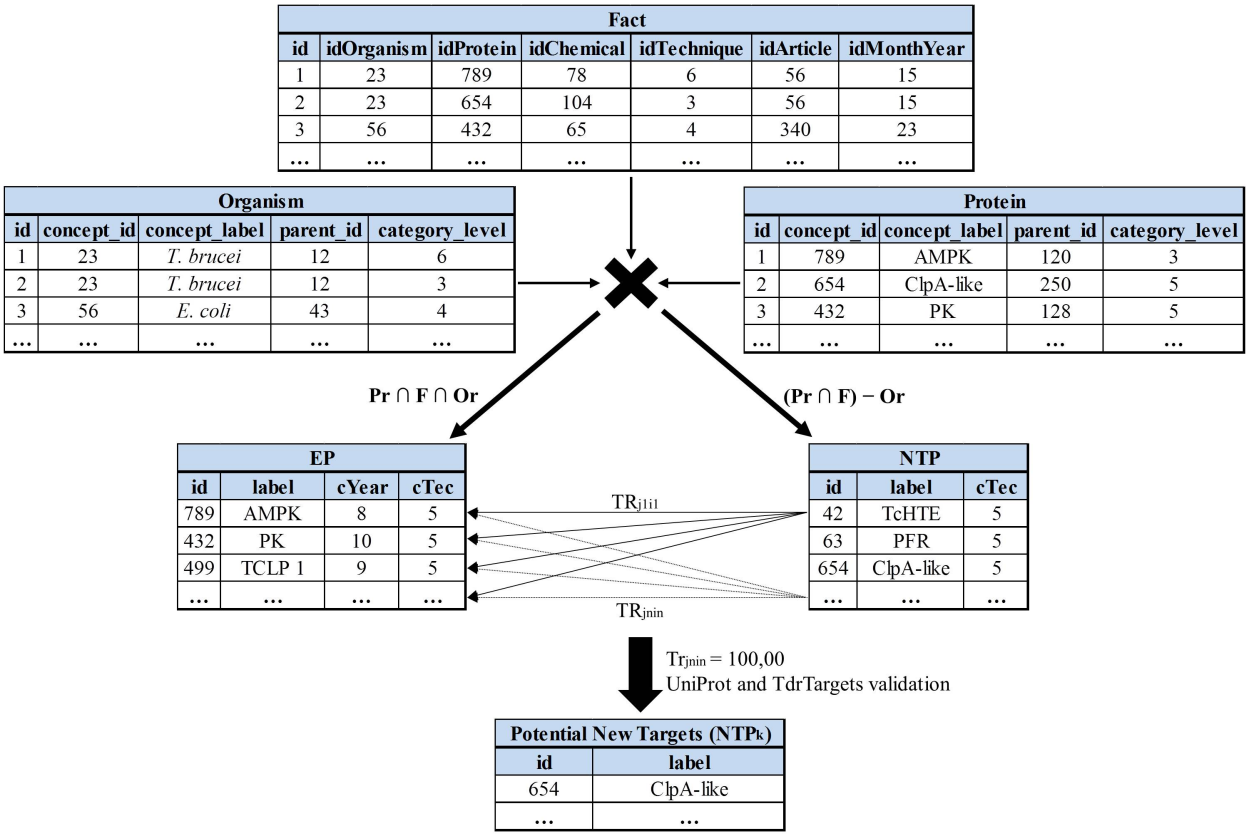

Figure 12: **Fragment of the tables used in the search strategy of the essentiality in *T. brucei* organism.**

concepts. It benefits from the ontology structure, which provides context and rich relations between concepts. Through the use of ontologies, even simple queries, such as to find articles that cite a certain term, will return articles with the exact term, but also articles that mention more specific terms, or synonyms. Using a keyword-based tool, many relevant articles may not be retrieved because it does not consider, for instance, hierarchical relations between terms. On the other hand, the TOETL approach increases the possibility of important informations being duly found. A comparison showing the advantage of using the TaP OLAP interface instead of a keyword-based interface, can be found in the supplementary materials.

The TaP OLAP DM is available as a demo[2], populated with the data of the reported case study. More details about the TaP OLAP DM are available at its manual[3].

Conclusion

This work presented a scientific scenario where the scientist aims at prioritizing drug targets, where the challenge is to filter useful information from a huge amount of scientific texts supporting decision making researcher. This work shows that with the use of ontologies and data mart it is possible to extract information and store it in a suitable form to help the researchers taking decisions about complex problems.

Searches performed directly on a data mart demonstrate that this approach can assist in survey of pointers to direct or prioritize research of new drug targets or any other research field or subject. Our approach also helped to start filling a gap in research about methods of extracting information through semantic annotation to support decision making in scientific research.

Other highlights are the differences between the Tap approach and keyword-based traditional tools. With the use of ontologies, even in simple queries, as to find articles that cite a certain term, the comprehensiveness of responses is greater. Using the keyword-based tools, the user needs to inform exactly how the term is cited in the text. If an

[2]http://157.86.114.152:8086/pentaho; for login use "Joe(admin)", the password will be filled automatically (if it does not happen, the password is "password")

[3]http://157.86.114.152:8086/manual/manual.pdf

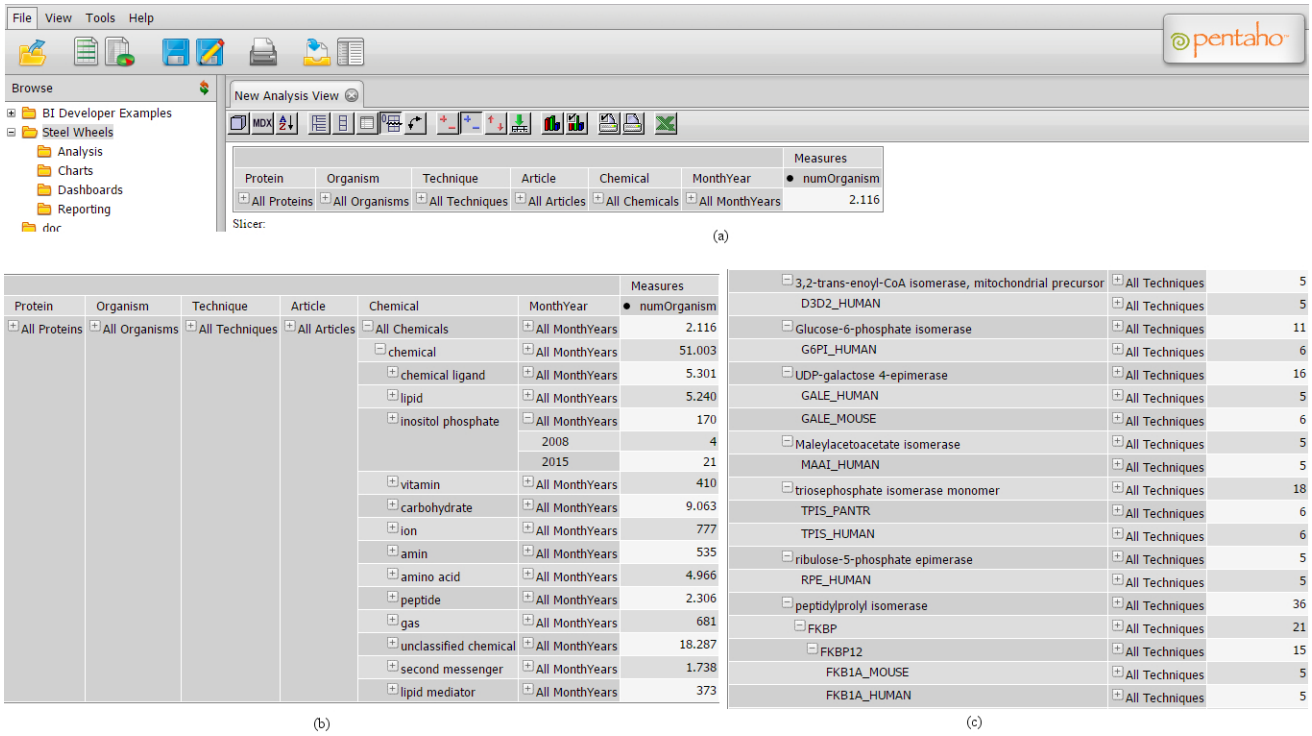

Figure 13: **Mondrian OLAP tool.** (a) shows the complete interface with all dimensions. (b) Presents the possibilities of manipulation of the hierarchies. (c) presents a screen patch of the query (ii)(3), for each protein, the number of essentiality techniques is presented.

important article uses a synonym of a term of interest, the user may not have access to valuable information. This can happen by simple fact of the user does not know the synonym term or there are so many synonyms for the term that it makes impossible to consult all possibilities. This situation also applies to hierarchical informations. A huge amount of information is not accessed because the traditional tools do not consider hierarchical relations between terms.

On the other hand, the Tap approach increases the possibility of important informations being duly found. The description of the concepts and its diversity of terms present in the ontology, it allows to the DM to identify the presence of a concept in an article more efficiently than keyword-based tools.

Future work includes the integration of data from the homology analysis of infectious agents and model organisms' genes. The idea is to identify druggable targets on infectious agents. The integration of these data with the annotation data will enrich the information contained in the data mart, increasing the chances of finding unexplored relationships.

1 Supplementary material

Table 10: Proteins resulting from the approach 4.2.ii

TaP DM term	**UniProt id**	**Tdr Target term**	**Confirmed target**
HDAC3_MOUSE	Q8T6T6;	-	false
PTPA_HUMAN	Q57TS2; D6XM75;	-	false
CRM1	Q7Z2C0;	exportin 1	true
PTN11_HUMAN	Q584E1; D6XFI2;	-	false
TRAF2_MOUSE	Q586L8; D7SGH1;	-	false
AAPK2_HUMAN	Q57YQ0; D6XM77;	-	false
NMT1_MOUSE	Q388H8;	EC 2.3.1.97	true

PSN2_MOUSE	Q38F54;	-	false
AAPK1_HUMAN	Q57YQ0; D6XM77;	-	false
RAF1_HUMAN	Q57YQ0; D6XM77;	-	false
KS6A3_HUMAN	Q57YQ0; D6XM77;	-	false
PAK1_MOUSE	Q57YQ0; D6XM77;	-	false
ERCC2_MOUSE	Q581V4; D6XF63;	-	false
RAF1_RAT	Q57YQ0; D6XM77;	-	false
ERCC3_MOUSE	Q581V4; D6XF63;	-	false
Glyceraldehyde 3-phosphate dehydrogenase, liver	P10097;	-	false
Glucose-6-phosphate isomerase	Q6RZY0;	glucose-6-phosphate isomerase	true
PSN2_HUMAN	Q38F54;	-	false
TOR1_SCHPO	Q57YM2; D6XM05;	-	false
PSN2_RAT	Q38F54;	-	false
RPCY	Q584K3; D6XEQ9;	-	false
TOR2_SCHPO	Q57YM2; D6XM05;	-	false
DNA-directed RNA polymerase III subunit 22.9 kDa polypeptide	Q584K3; D6XEQ9;	-	false
AKT1_RAT	Q57YQ0; D6XM77;	-	false
FKB1A_MOUSE	Q57TS2; D6XM75;	rotamase	true
FKB1A_HUMAN	Q57TS2; D6XM75;	PPIase	true
NMT1	Q388H8;	EC 2.3.1.97	true
RS27_HUMAN	false	40S ribosomal protein S27	true
AGM1_HUMAN	Q57XH7; D6XMR6;	N-acetylglucosamine-phosphate mutase	true
IDI1_HUMAN	Q38E87;	isopentenyl-diphosphate delta-isomerase	true
TRAF6_HUMAN	Q586L8; D7SGH1;	-	false
GSK3_CAEEL	Q388M1;	-	false
MK03_HUMAN	Q580X5; D6XL90;	mitogen-activated protein kinase 1	true
bamacan	false	structural maintenance of chromosome 3	true
cytochrome P450	false	cytochrome P450	true
DHSA_MOUSE	Q57TV5; D6XM42;	-	false
ARF6_RAT	false	ADP-ribosylation factor 6	true
PGS1_HUMAN	P13377;	-	false
RL32_HUMAN	Q38CW9;	60S ribosomal protein L32	true
RL32_MOUSE	Q38CW9;	60S ribosomal protein L32	true
ERK1	false	mitogen-activated protein kinase 3	true
DUT_HUMAN	Q57ZH3; D6XJ47;	deoxyuridine triphosphatase	true
RL32_RAT	Q38CW9;	60S ribosomal protein L32	true
peptidylprolyl isomerase	Q57UJ5; D6XLQ7;	PPIase	true
protein tyrosine phosphatase	false	protein tyrosine phosphatase	true
METK2_HUMAN	Q586G3; D6XH67;	-	false
PMM1_HUMAN	F4NCC2;	-	false
UBE2N_MOUSE	Q57XC5; D6XL01;	-	false
DNA-directed RNA polymerase III largest subunit	P08968;	DNA-directed RNA polymerase iii largest subunit	true
SPEE_HUMAN	Q38EH6;	spermidine synthase	true
PRI2_HUMAN	Q38BT5;	DNA primase large subunit	true

KPYR_HUMAN	Q388I8;	pyruvate kinase 1	true
ADHX_MOUSE	Q385N2;	-	false
XPO1_MOUSE	Q7Z2C0;	-	false
GALE_HUMAN	Q8T8E9;	UDP-galactose 4-epimerase	true
KPYM_HUMAN	Q388I8;	EC 2.7.1.40	true
MK08_MOUSE	Q580X5; D6XL90;	-	false
XPO1_RAT	Q7Z2C0;	-	false
MAAI_HUMAN	Q386Y7;	-	false
PLK1_HUMAN	Q57VI0; D6XJG2;	-	false
GDIA_HUMAN	false	RAB GDP dissociation inhibitor alpha	true
XRCC5_HUMAN	false	KU80 protein	true
HHAT_MOUSE	Q582B4; D6XGI9;	-	false
UBE3A_HUMAN	Q586L8; D7SGH1;	-	false
HPRT_HUMAN	Q07010;	hypoxanthine-guanine phosphoribosyltransferase	true
CG21_YEAST	P41179;	-	false
PDPK1_DROME	Q57YQ0; D6XM77;	-	false
O18683_DROME	false	PI3-kinase	true
KS6B1_HUMAN	Q57YQ0; D6XM77;	-	false
KS6B1_MOUSE	Q57YQ0; D6XM77;	-	false
NDKA_RAT	Q581Q9; D6XFA8;	-	false
Rad21	false	double-strand-break repair protein rad21 homolog	true
F16P1_HUMAN	Q38EA4;	-	false
HHAT_HUMAN	Q582B4; D6XGI9;	-	false
ADK_MOUSE	O61069;	adenosine kinase	true
ENOA_HUMAN	Q9NDH8;	EC 4.2.1.11	true
DPOLB_HUMAN	Q6V1L6;	-	false
EP300_HUMAN	Q57YY2; D6XL66;	-	false
IF4E_RAT	false	eukaryotic translation initiation factor 4e	true
DNLI4_HUMAN	Q587E4; D6XH61;	-	false
GNPI1_HUMAN	Q381A9;	-	false
PTEN_HUMAN	Q57V97; D6XGA4;	-	false
phosphofructokinase	D6XDN4;	-	false
Q9V3L4_DROME	Q57YS3; D6XMA0;	-	false
kinesin-like protein	D6XFK9;	kinesin-like protein	true
MDM2_DANRE	Q586L8; D7SGH1;	-	false
PDXK_HUMAN	D6XI07;	pyridoxal kinase	true
SGK1_CAEEL	Q57YQ0; D6XM77;	-	false
ESPL1_HUMAN	Q4GYR1;	separase	true
ESP1_YEAST	Q4GYR1;	separase	true
CUT1_SCHPO	Q4GYR1;	separase	true

References

[1] PubMed us national library of medicine national institutes of health.

[2] Rolf Apweiler, Amos Bairoch, Cathy H Wu, Winona C Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, Hongzhan Huang, Rodrigo Lopez, Michele Magrane, et al. Uniprot: the universal protein knowledgebase. *Nucleic acids research*, 32(suppl 1):D115–D119, 2004.

[3] Kele T. Belloze. Priorização de alvos para fármacos no combate a doenças tropicais negligenciadas causadas por protozoários, 2013.

[4] Kele T. Belloze, Daniel Igor S. B. Monteiro, Tulio F. Lima, Floriano P. Silva Jr., and Maria Cláudia Cavalcanti. Analyzing tools for biomedical text annotation with multiple ontologies. In *Proceedings of the 3rd International Conference on Biomedical Ontology (ICBO 2012), KR-MED Series, Graz, Austria, July 21-25, 2012*, 2012.

[5] Kele T. Belloze, Daniel Igor S. B. Monteiro, Tulio F. Lima, Floriano P. Silva, and Maria Cláudia Reis Cavalcanti. An evaluation of annotation tools for biomedical texts. In *ONTOBRAS-MOST*, 2012.

[6] Tobias Bergmiller, Martin Ackermann, and Olin K. Silander. Patterns of evolutionary conservation of essential genes correlate with their compensability. *PLOS Genetics*, 8(6):1–13, 06 2012.

[7] Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web. *Scientific American*, 284(5):34–43, May 2001.

[8] Ranjit Bose and Vijayan Sugumaran. Application of intelligent agent technology for managerial data analysis and mining. *SIGMIS Database*, 30(1):77–94, January 1999.

[9] Matt Casters, Roland Bouman, and Jos Van Dongen. *Pentaho Kettle solutions: building open source ETL solutions with Pentaho Data Integration*. John Wiley & Sons, 2010.

[10] Surajit Chaudhuri and Umeshwar Dayal. An overview of data warehousing and olap technology. *SIGMOD Rec.*, 26(1):65–74, March 1997.

[11] Marcus Albert Alves da Silva and Maria Cláudia Cavalcanti. Combining ontology modules for scientific text annotation. *JIDM*, 5(3):238–251, 2014.

[12] Marcus Albert Alves da Silva, Maria Cláudia Reis Cavalcanti, Kele Teixeira Belloze, and Floriano Silva-Junior. Agile semantic annotation of scientific texts at the biomedical scenario. In *10th IEEE International Conference on e-Science, eScience 2014, Sao Paulo, Brazil, October 20-24, 2014*, pages 100–107, 2014.

[13] Michael A. D'Elia, Mark P. Pereira, and Eric D. Brown. Are essential genes really essential? *Trends in Microbiology*, 17(10):433 – 438, 2009.

[14] David W. Embley, Yihong Ding, Stephen W. Liddle, and Mark Vickers. Automatic creation and simplified querying of semantic web content: An approach based on information-extraction ontologies. In *In Proceedings of the first Asian Semantic Web Conference (ASWC 2006) LNCS 4185*, pages 400–414, 2006.

[15] Scott Federhen. The ncbi taxonomy database. *Nucleic acids research*, 40(D1):D136–D143, 2012.

[16] CA Fontes. Explorando inferência em um sistema de anotaçao semântica. *Master's thesis, Dept. of Computer Science, Military Inst. of Engineering, Rio de Janeiro, Brazil*, 2011.

[17] Celso Araujo Fontes, Maria Cláudia Cavalcanti, and Ana Maria de Carvalho Moura. An ontology-based reasoning approach for document annotation. In *2013 IEEE Seventh International Conference on Semantic Computing, Irvine, CA, USA, September 16-18, 2013*, pages 160–167, 2013.

[18] Thomas R. Gruber. A translation approach to portable ontology specifications. *Knowl. Acquis.*, 5(2):199–220, June 1993.

[19] Thomas R. Gruber. Toward principles for the design of ontologies used for knowledge sharing? *International Journal of Human-Computer Studies*, 43(5):907 – 928, 1995.

[20] M. Gulić. Transformation of owl ontology sources into data warehouse. In *2013 36th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 1143–1148, May 2013.

[21] W. H. Inmon. The data warehouse and data mining. *Commun. ACM*, 39(11):49–50, November 1996.

[22] W. H. Inmon. *Building the Data Warehouse (4Th Ed.)*. John Wiley & Sons, Inc., New York, NY, USA, 2005.

[23] Clement Jonquet, Nigam Shah, and Mark Musen. The open biomedical annotator. In *AMIA summit on translational bioinformatics*, pages 56–60, 2009.

[24] Ralph Kimball and Margy Ross. *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling.* Wiley Publishing, 3rd edition, 2013.

[25] Obitko M., Snasel V., and Smid J. Ontology design with formal concept analysis. *CLA 2004*, pages 111–119, 2004.

[26] María P Magariños, Santiago J Carmona, Gregory J Crowther, Stuart A Ralph, David S Roos, Dhanasekaran Shanmugam, Wesley C Van Voorhis, and Fernán Agüero. Tdr targets: a chemogenomics resource for neglected diseases. *Nucleic acids research*, 40(D1):D1118–D1127, 2012.

[27] Jesús Pardillo and Jose-Norberto Mazón. Using ontologies for the design of data warehouses. *CoRR*, abs/1106.0304, 2011.

[28] Daniel J. Rigden, Xosé M. Fernández-Suárez, and Michael Y. Galperin. The 2016 database issue of *Nucleic Acids Research* and an updated molecular biology database collection. *Nucleic Acids Research*, 44(Database-Issue):1–6, 2016.

[29] Oscar Romero and Alberto Abelló. Automating multidimensional design from ontologies. In *Proceedings of the ACM Tenth International Workshop on Data Warehousing and OLAP*, DOLAP '07, pages 1–8, New York, NY, USA, 2007. ACM.

[30] Daniel L Rubin, Natalya F Noy, and Mark A Musen. Protege: a tool for managing and using terminology in radiology applications. *Journal of Digital Imaging*, 20(1):34–46, 2007.

[31] Gerard Salton and Chris Buckley. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, 24(5):513–523, 1988.

[32] Gayatri Sathe and Sunita Sarawagi. Intelligent rollups in multidimensional OLAP data. In *VLDB 2001, Proceedings of 27th International Conference on Very Large Data Bases, September 11-14, 2001, Roma, Italy*, pages 531–540, 2001.

[33] Jung P. Shim, Merrill Warkentin, James F. Courtney, Daniel J. Power, Ramesh Sharda, and Christer Carlsson. Past, present, and future of decision support technology. *Decision Support Systems*, 33(2):111–126, 2002.

[34] Nicholas Sioutos, Sherri de Coronado, Margaret W. Haber, Frank W. Hartel, Wen-Ling Shaiu, and Lawrence W. Wright. NCI thesaurus: A semantic model integrating cancer-related clinical and molecular information. *Journal of Biomedical Informatics*, 40(1):30–43, 2007.

[35] Barry Smith, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J Goldberg, Karen Eilbeck, Amelia Ireland, Christopher J Mungall, et al. The obo foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, 25(11):1251–1255, 2007.

[36] M. Thenmozhi and K. Vivekanandan. A tool for data warehouse multidimensional schema design using ontology. page 161–168, 2013.

[37] Manole Velicanu and Gheorghe Matei. Building a data warehouse step by step. *Economic Informatics, Forthcoming*, 2007.

[38] Min Wang and Bala Iyer. Efficient roll-up and drill-down analysis in relational databases. In *In 1997 SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery*, pages 39–43, 1997.

[39] Patricia L. Whetzel, Natalya F. Noy, Nigam H. Shah, Paul R. Alexander, Csongor Nyulas, Tania Tudorache, and Mark A. Musen. Bioportal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. *Nucleic Acids Research*, 39(suppl 2):W541–W545, 2011.

[40] Patricia L. Whetzel, Natalya Fridman Noy, Nigam H. Shah, Paul R. Alexander, Csongor Nyulas, Tania Tudorache, and Mark A. Musen. Bioportal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. *Nucleic Acids Research*, 39(Web-Server-Issue):541–545, 2011.

[41] Satoko Yamamoto, Takao Asanuma, Toshihisa Takagi, and Ken Ichiro Fukuda. The molecule role ontology: An ontology for annotation of signal transduction pathway molecules in the scientific literature. *Comparative and Functional Genomics*, 5:528 – 536, 2004.