Computer Methods and Programs in Biomedicine 157 (2018) 225–235

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Data mart construction based on semantic annotation of scientific articles: A case study for the prioritization of drug targets

Marlon Amaro Coelho Teixeira[a,d,*], Kele Teixeira Belloze[b], Maria Cláudia Cavalcanti[c,*], Floriano P. Silva-Junior[a,*]

[a] Oswaldo Cruz Institute (IOC), Oswaldo Cruz Foundation (FIOCRUZ), Av. Brasil 4365, Manguinhos, Rio de Janeiro 21040-360, Rio de Janeiro, Brazil
[b] Federal Center for Technological Education Celso Suckow da Fonseca, CEFET/RJ, Av. Maracanã 229, Rio de Janeiro, Rio de Janeiro, Brazil
[c] Computer Engineering Department, Military Institute of Engineering (IME), Praça General Tibúrcio 80, Urca, Rio de Janeiro 20271-064, Rio de Janeiro, Brazil
[d] Acre Federal Institute of Education and Science and Technology (IFAC), Av. Brasil 920, Xavier Maia, Rio Branco 69.903-068, Acre, Brazil

ARTICLE INFO

Article history:
Received 21 July 2017
Revised 21 December 2017
Accepted 10 January 2018

Keywords:
Semantic annotation
Decision support systems
Drug target prioritization

ABSTRACT

Background and objectives: 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 it reveals a lot of scientific discoveries that are hidden within academic articles. The Biomedical area has more than 300 ontologies, most of them composed of over 500 concepts. These ontologies can be used to annotate scientific 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. Therefore, the main objective of this work is to provide a method that could facilitate this task.
Methods: This work describes a method called Text and Ontology ETL (TOETL), to build an analytical view over such texts. First, a corpus of selected papers is semantically annotated using distinct ontologies. Then, the annotation data is extracted, organized and aggregated into the dimensional schema of a data mart.
Results: Besides the TOETL method, this work illustrates its application through the development of the TaP DM (Target Prioritization data mart). This data mart has focus on the research of gene essentiality, a key concept to be considered when searching for genes showing potential as anti-infective drug targets.
Conclusions: This work reveals that the proposed approach is a relevant tool to support decision making in the prioritization of new drug targets, being more efficient than the keyword-based traditional tools.

1. Introduction

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 become available as a result from these experiments. Still, 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. 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, it is usual that a scientist may be interested in many combinations of distinct topics.

Text annotation allows the identification of the occurrence of such multidimensional combination of topics, and therefore makes it possible to rank these articles according to the scientists interest. However, 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 [1]. Semantic annotation is one of the main efforts towards associating text content to its well-defined meaning. There are already many semantic annotation systems [2], which provide mechanisms to bring semantic to documents through text annotation. It means to handle and associate metadata or ontology concepts with text content.

* Corresponding author at: Acre Federal Institute of Education and Science and Technology (IFAC), Av. Brasil, Xavier Maia 69.903-068, Rio Branco, Acre 920, Brazil.
E-mail address: marlon.teixeira@ifac.edu.br (M.A.C. Teixeira).

[1] https://www.ncbi.nlm.nih.gov/pubmed.

https://doi.org/10.1016/j.cmpb.2018.01.010
0169-2607/© 2018 Elsevier B.V. All rights reserved.

An ontology is a model that represents a domain of reality, i.e., a formal description of concepts and relationships [3]. 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.

Nowadays, there are more than 500 ontologies on the biomedical domains [4,5]. 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. Despite many initiatives on semantic annotation [2], 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.

Since the 90s a powerful approach for analytical view and decision support, known as data warehousing (DW), has been largely used. DW are non-volatile thematic driven databases, capable of integrating multiple data sources. In the context of DW arises the concept of data mart (DM). DM is usually defined as a subset of a DW, with a specific focus, or with a reduced number of dimensions [6,7]. There are many initiatives on the development and usage of DW on biomedical data [8,9]. These DW aim to provide a multidimensional view of the data, to answer complex analytical queries, such as "*what is the mortality rate for postmenopausal female patients admitted from the general ward with fever?*" (extracted from a Clinical Decision Support System [10]). Note that to answer such query, the DW needs to keep information about the temperature for each patient, hospital unit, time and patient status. Typically, for maintaining historical data, a DW is a high volume database. Moreover, since complex queries involve the manipulation of a wide set of records from multiple tables and the use of common joins and aggregations, performance and response time is a challenge in DW environments [6,7].

The model behind the DW design is based on facts and dimensions, and aims to address those performance issues. The fact is what needs to be observed, such as the temperature. Each observed fact is described or characterized by aspects or dimensions, such as: patient, hospital unit, time, etc. In order to facilitate the user in expressing such complex queries, the DW database is usually 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 [11]. Using DW and OLAP systems in scientific research can help in the sense that it enables to *observe* scientific texts annotations and to correlate informations about different organisms.

In order to build and maintain data in a DW or DM, a process called ETL (Extract, Transform and Load) should be designed and implemented. Traditional methods of DM design are well consolidated, and usually apply generic techniques for structured data sources. However, there are just a few studies exploring specific methodologies to build DMs from unstructured data for analysis and decision support [12–14]. Some of them [15–18] propose ontology-based approaches, that they claim to be useful for dealing with textual data sources. On the other hand, none of them details how to handle and select large amounts of textual data. As mentioned before, this feature is fundamental to support scientific research, specially in the biomedical field.

Therefore, new approaches are needed to address scientific research, as for example in the search for new drugs in fight against neglected diseases. Neglected diseases are diseases caused by protozoa and 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 [19].

The objective of this work is to present a design methodology of a data mart (DM), usually defined as a smaller DW, for textual data analysis by means of ontologies. Thus, the idea is to provide 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. 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*. At the end, useful queries illustrate the benefits of this approach in the search for new drug targets.

2. Methods

Usually, the existing traditional ETL process [6,7] consists of three steps, as shown in Fig. 1(i). The *Extraction* step receives data from a variety of data sources, including structured data and text data sources, and stores them in an intermediary database, also known as Data Staging area (DS). The next step applies transformations on the collected data and prepares them for the final step, which is responsible to load them according to the dimensional model (dimensions and facts).

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 previous works [19,20] and summarized in Section 2.1. The following steps, *Transform* and *Load* steps, which are the contributions of the present work, are detailed in Sections 2.2 and 2.3.

Fig. 1(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.

2.1. Annotation step (extraction)

The authors have carried out a previous work [19,20], where the semantic annotation is organized as a set of steps (illustrated in Fig. 2) and this step is briefly described here.

It begins with the *understanding the research theme* step, which involves the study of the classic literature in the area, as well as interviews with the users. 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.

The set of terms is then used as input to the *queries definition* step 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. The same set of terms is used at the *digital library selection* step, in order to choose libraries (one or more) that cover most of the domains involved in the research theme. Once the queries expressions are

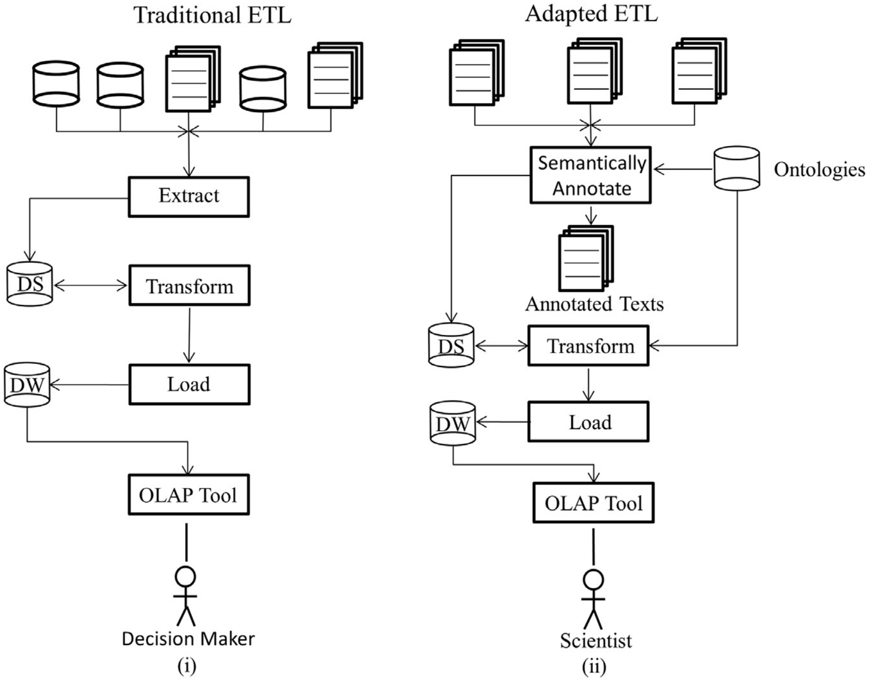


Fig. 1. Traditional and adapted ETL (TOETL) process.

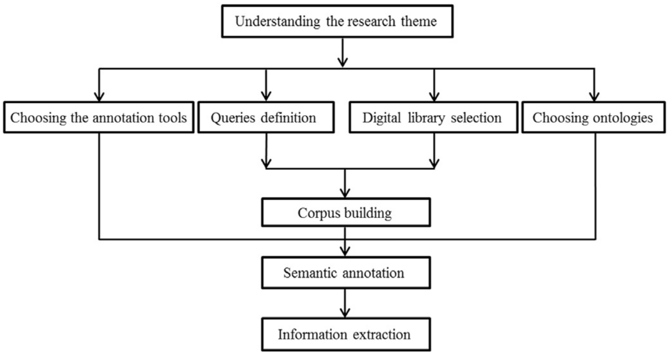


Fig. 2. Annotation step activities (extraction).

formed and libraries are available, the corpus can be built (*corpus building* step). In parallel, ontologies and annotation tools are chosen.

More than one ontology may be necessary once the research theme usually involves multiple domains. Ontology repositories such as OBO Foundry [5] and NCBO BioPortal [4], can be used to facilitate the search.

For the *semantic annotation* step, it is required to choose an automatic annotation tool, that could use any arbitrary ontology (user choice), providing easy access to annotation data for further processing, and possibly, that could provide hierarchical inference while annotating. Based on the analysis of a previous work [21], we recommend the use of automatic annotation tools such as Autometa [22], which uses standard annotation formats, such as W3C RDFa[2], that are easy to extract.

Finally, after the annotation procedure, the *information extraction* step automatically processes the annotated texts, and populates the DS (Annotations database).

According to Fig. 1(ii), after the *Annotation* step, it follows the *Transformation* step, which prepares the data for the DW *Load* step. These steps are the main contribution of this work and are described in the following subsections.

2.2. Transform step

It is important to understand that in the context of a scientific research, although the textual source remains the same, the interest and content of a scientist during a research is dynamic. The focus or his research may change over time. The aspects that are useful for the scientist analysis are defined only after extracting the annotation data. Therefore, differently from the traditional Transformation step, which counts on pre-defined and designed dimensions, in the scientific research scenario, some dimensions have to be defined at this step. In other words, the design of the DM is ongoing during the Transformation step of an ETL process, meaning both processes, ETL and DM design, occur in parallel. During the first ETL execution, a new DM is designed and populated. New extractions may occur, if the scientist maintain the same research focus, to update the already created DM. But, whenever the scientist changes his focus, a new DM will be needed.

The *Transform step* consists of four sequential sub-steps, described as follows. The *Identification of Analytical Demands* is the **first step** and it aims to identify the analytical questions that must be addressed. At this point, we should count again on interviews with users. The main idea is to identify which information is interesting to get from 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 correlation of three distinct aspects: articles, side effects and diseases. After raising such questions, each one 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

[2] https://www.w3.org/TR/rdfa-core/.

Table 1
Ontology based dimensions.

Ontology dimensions
id: BIGINT (PK)
concept_id:BIGINT
concept_label:VARCHAR
category_level:INT
parent_id:BIGINT

classified as related to the diseases domain, and the terms such as fever and body pain can be classified as symptoms.

The **second step** is the *Identification and characterization of the Dimensions and Categories.* 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 [23] 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.

According to the traditional DM design [7], once the dimensions are defined, it is time to characterize them, meaning that we should identify attributes that could describe each dimension instance, including dimension categories to address aggregation demands. Ontologies also help on this step, as their concepts' hierarchies already embeds a categorization. Thus, all ontology based dimensions may be characterized similarly, as shown in Table 1.

The *category_level* attribute is used to label the level of the hierarchy, when it is the case (e.g. species, genus, family, etc.). The *parent_id* attribute represents a recursive relationship, used to allow each tuple to point to the tuple that represents a more generic 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 in the article.

Once defined the ontology based dimensions, the **third step** involves the *cut-off of the ontologies* used to annotate the texts, in order to prepare for the population of those dimensions at the *Load* step. This can be done manually or through the use of ontology segmentation tools such as Locality Module Extractor [24] and SEGMENTATION [25]. As the ontologies can be very broad, where several concepts are described, this step has the objective of removing from the ontology the concepts that are not part of the research interest. Some examples of ontology cut-offs can be found in the supplementary material (Figures 8 and 9).

Usually, a DM includes a *time* dimension, that must also be characterized. In the case of a scientific corpus analysis, it makes sense to analyze scientific articles publication on a yearly or monthly basis, not less than that. Therefore, a time dimension with attributes that characterize each year/month would be sufficient, and may be previously designed and populated.

The **fourth step** is the *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/multiple presence of a term in an article counts one. Therefore, the fact to be observed is the counting for each combination of values of d_1,..., d_n, and month/year, where d_i corresponds to each selected ontology hierarchy (dimension), 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 [26]. In this case, the article itself emerges as a dimension. Then, the fact would observe the counting for each combination of values of d_1,..., d_n, month/year, and article, where d_i corresponds to each selected ontology-based (hierarchies) dimension, 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, but the article must be kept as a dimension.

The article dimension may be implemented with attributes that characterize the publication, such as its title, its identification code assigned at its original repository, a link to its pdf file, etc. It is not necessary to keep data about the publication date of the articles in this dimension, since it is already represented by the MonthYear dimension table.

2.3. Load step

This step consists of two main sub-steps: the population of the dimension tables and the population of the fact table. The population of each dimension table can be automated using as input the corresponding ontologies (or ontology cut-offs) in a structured format, such as RDF/OWL XML file format.

The population of the Article dimension is based on the set of selected annotated texts. Their *ids* and *titles* are loaded into the Article dimension table. In addition, the link to electronic version of the article is also stored. This information facilitates the access to the article contents, which is requested quite often.

Finally, the population of the Fact table is done as follows: For each article stored in the Article table, a list with all the distinct annotated terms is created. This list contains terms of all dimensions annotated in the article. Then, this list is separated in a sublist of terms for each dimension. Each fact tuple comes from the combination of all the elements present in these lists. For example, if we consider three dimensions and an article has two terms in each dimension (i.e, each dimension list has two elements), eight tuples will be needed to represent the facts.

It may occur cases in which there are no annotation of terms from a specific dimension. To address this anomaly, a special tuple labeled NA (not annotated), which is present in all ontology-based dimensions is referenced.

3. Results

The following subsections present the results that were achieved based on the proposed method. First, in Section 3.1. a real case study describes how the TOETL method helped on building a specific DM, with focus on drug target prioritization. Then, Section 3.2 describes query strategies for the selection of new targets for a specific organism. Finally, Section 3.3 presents the TaP Olap tool, and shows its flexibility and availability for querying the DM.

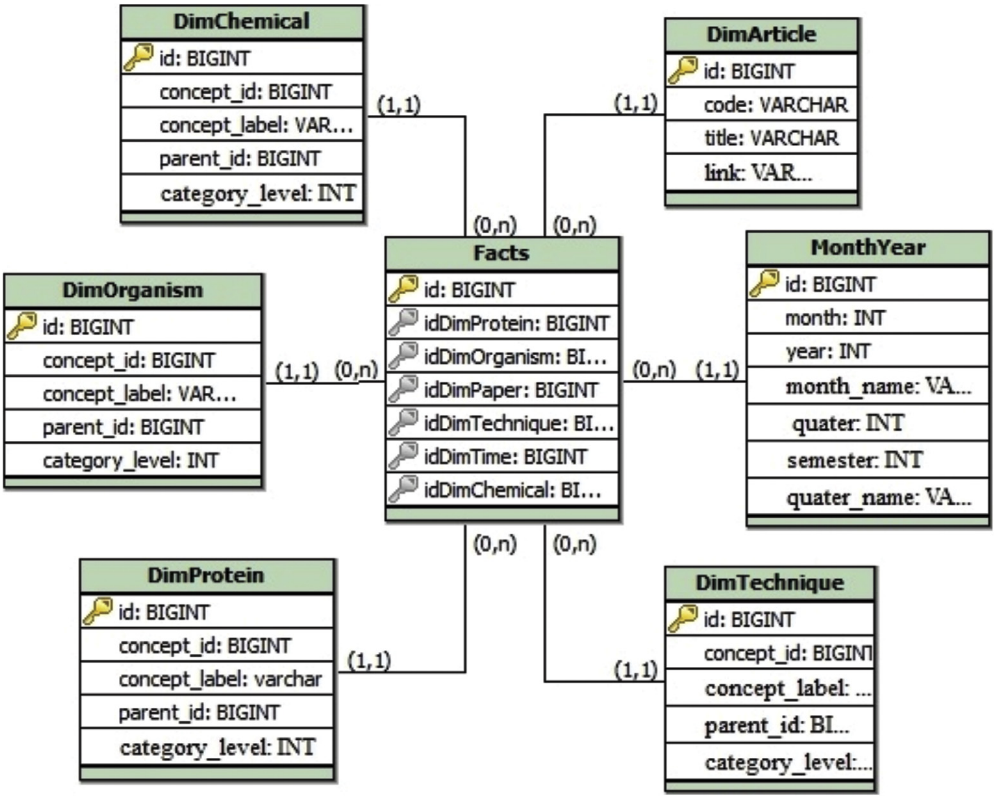


Fig. 3. Data mart design. This is the final design of TaP DM that was populated using data extracted from articles.

Table 2
Result of extractingdata.

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

3.1. Application to the prioritization of drug targets

Based on the methodology described previously, this section describes the TaP (Target Prioritization) DM construction and population. This DM is built as a case study, to support the decision on the prioritization of new drug targets, focusing especially on the gene essentiality techniques.

In order to generate a sample of articles that will serve as a data source for the data mart, terms involving organisms of interest and techniques of essentiality were identified. With the help of end users (researchers working on the drug discovery field), a set of articles were selected, a set of ontologies were identified and those texts were annotated. This generated a set of annotated texts[3], as described and discussed in previous works [19,20].

These annotated texts were processed and a table was generated with the following schema: article identifier, the annotated term, the ontology class identifier and the ontology class name, as shown in Table 2. The next section describes how the *Transform* and *Load* steps were followed during the TaP DM construction.

3.1.1. Transform step

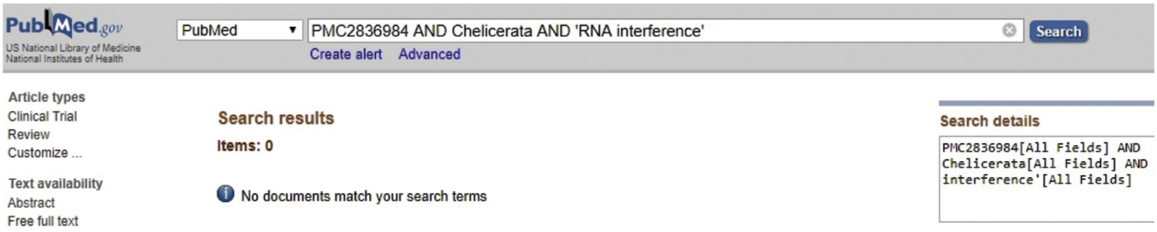
As mentioned in Section 2.2, the first step in the design and construction of the tool is the *Identification of Analytical Demands*, where the questions that the DM must be able to answer are raised. Specialists of the Computational and Experimental Biochemistry of Drugs Laboratory (LaBECFar) at IOC/Fiocruz, raised 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. Based on these questions, the next step (*Identification and characterization of Dimensions and Categories*) identifies 6 concepts as candidates for *TaP* DM dimensions: protein, organism, gene essentiality techniques, chemical, article and time. Note that most of them are mentioned in those queries.

For each of these terms a corresponding dimension table was created for the TaP DM: DimProtein, DimChemical, DimOrganism, DimTechnique, DimArticle and DimMonthYear. Fig. 3 shows that DimProtein, DimChemical, DimOrganism and DimTechnique were designed to represent ontology terms and, consequently, to address ontology hierarchies.

In the third step, for each dimension, a subset of the ontology should be generated. This is not an easy task, since some ontologies are very large. There are tools that facilitate this process [27,28]. In this work the Protégé suite was used, it allows to visualize graphically the hierarchies of the ontologies and it facilitates to eliminate, manually, the branches that are not of interest. There are works [29] that propose more agile solutions for this task.

The Molecule Role ontology was edited to generate two cut-offs, one that included chemical related terms and another for protein and protein family terms. It is important to note that the dimension table must be populated only with terms that correspond to the meaning of what the dimension represents. For instance, the

[3] Available at https://github.com/TapDM/AnnotatedTexts.



PMC2836984 AND Chelicerata AND 'RNA interference'

Search results

Items: 0

No documents match your search terms

Search details

```
PMC2836984[All Fields] AND
Chelicerata[All Fields] AND
interference'[All Fields]
```

Fig. 4. Search for articles that cite the terms *Chelicerata* and RNA interference in the PubMed interface.

Table 3
Fact table.

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

cut-off prepared for the protein dimension population, should not include chemical terms. A wrong cut-off of the ontology may lead to wrong analytical results. On the other hand, terms that are not yet classified as proteins, but that are newborn proteins, may be included in the cut-off, in order to facilitate the analyses. The NCI *thesaurus* had to be browsed to find out the terms about essentiality, thus at the end of the process, a module containing all classes of interest was obtained.

The next step is the Fact Identification/Definition. 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 dimensions in a specific article, published in a specific month/year. Each tuple of the fact table consists of 7 fields: its key field and the 6 fields which point to (foreign keys) tuples in the corresponding 6 dimension tables. The complete star schema can be seen in Fig. 3.

3.1.2. Load step

The ontologies used to annotate the selected texts, were also used to populate the dimension tables. The DimOrganism table was populated based on the entire NCBI Taxon ontology, provided by the National Center for Biotechnology Information (NCBI), which embeds a deep hierarchy representation of the organisms.

Similarly, the DimProtein and DimChemical tables were also designed to store information extracted from a specific ontology, the Molecule Role ontology. However, they did not use the whole ontology. Each table used an ontology cut-off that corresponds to one of its main branches.

The DimTechnique table was populated based on a cut-off of selected concepts from the NCI *thesaurus* ontology. In this case, terms were not categorized, because their hierarchy is shallow and of no utility from the analytical point of view.

Differently, the population of the DimArticle and DimMonthYear dimension tables was not based on ontologies. Instead, it was performed by means of extracting data from the set of selected and annotated articles. The DimArticle table was populated with data (title, link, etc.) extracted from the PubMed articles, allowing them to be retrieved in its original form. The DimMonthYear was populated with the months of the time interval that covered the whole set of selected articles.

Finally, the population of the fact table counted on the already populated dimensions, and on the DS (Annotations Database). As described before, for each article id, it gets the set of related tuples in the annotation tables, identifies the corresponding tuples in the ontology based dimension tables, and combines their references to form a fact tuple. Table 3 shows a few tuples on the generated Fact table. Note that tuple4 points to the NA tuple in the DimTechnique table (value 9999999). This means that article 22 does not cite terms with respect to techniques.

At the end of this process, the data mart is designed, built and populated. The MySQL[4] database management system (DBMS) was used to host the *TaP* DM and to connect to the decision support tool (OLAP tool), which will enable end users to correlate data. Next section presents how these correlations can guide the scientist on his research choices.

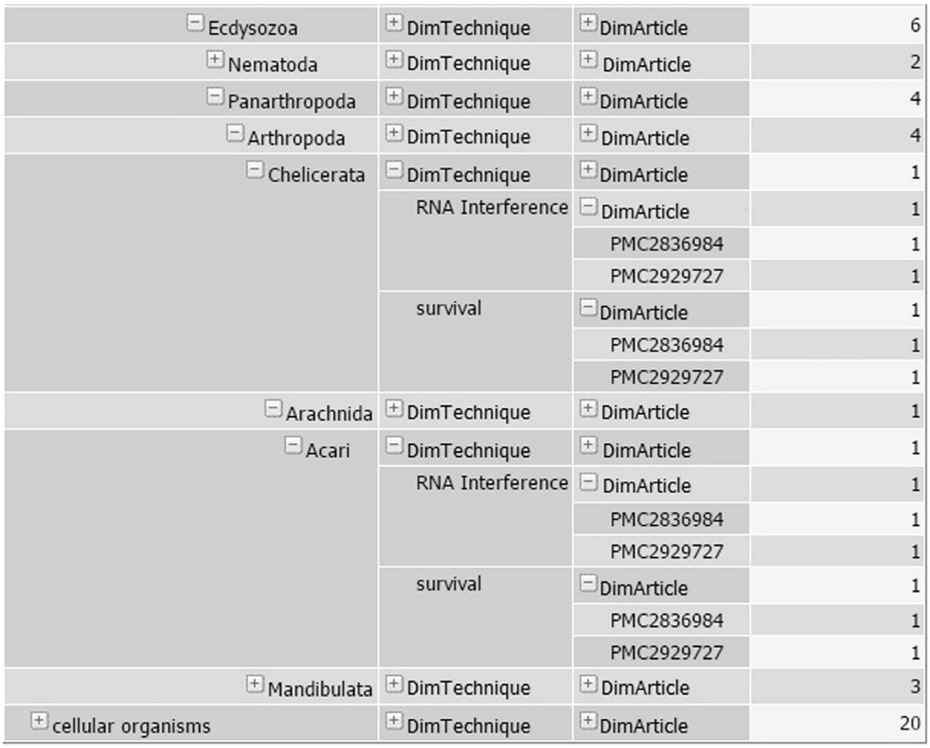
3.2. Examples of queries

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) Which articles contain the Chelicerata *subphylum and the RNA interference essentiality technique?.* Using the PubMed interface, one could construct the query *Chelicerata* AND 'RNA interference'. Among the indexed articles, the article PMC2836984 would not be present, because he does not have the term *Chelicerata* (Fig. 4).

On the other hand, performing a similar search in TaP DM returns the presence of the terms *Chelicerata* and RNA interference in article PMC2836984 (Fig. 5). This occurs because in this article, the *Acari* term is present, and it relates to an organism of the *Chelicerata* subphylum. If this article contains important information for the researcher, he/she could not have found it using the Pubmed interface. Even considering that the researcher had occasionally accessed this article, during its reading he could not realize the relation between *Acari* and *Chelicerata* or this detail could go unnoticed.

[4] https://www.mysql.com/.



⊟Ecdysozoa	⊞DimTechnique	⊞DimArticle	6
⊞Nematoda	⊞DimTechnique	⊞DimArticle	2
⊟Panarthropoda	⊞DimTechnique	⊞DimArticle	4
⊟Arthropoda	⊞DimTechnique	⊞DimArticle	4
⊟Chelicerata	⊟DimTechnique	⊞DimArticle	1
	RNA Interference	⊟DimArticle	1
		PMC2836984	1
		PMC2929727	1
	survival	⊟DimArticle	1
		PMC2836984	1
		PMC2929727	1
⊟Arachnida	⊞DimTechnique	⊞DimArticle	1
⊟Acari	⊟DimTechnique	⊞DimArticle	1
	RNA Interference	⊟DimArticle	1
		PMC2836984	1
		PMC2929727	1
	survival	⊟DimArticle	1
		PMC2836984	1
		PMC2929727	1
⊞Mandibulata	⊞DimTechnique	⊞DimArticle	3
⊞cellular organisms	⊞DimTechnique	⊞DimArticle	20

Fig. 5. Search for articles that cite the terms *Chelicerata* and RNA interference in the TaP DM interface.

Another point of comparison is the time to obtain the desired answers. The time spent on DM design and ETL steps (TOETL approach) is directly related to the computational power of the machines involved in these processes. This is due to the use of large ontologies in the semantic annotation of large corpora. Some techniques such as modularization of ontologies can be applied, drastically reducing the time spent in ETL processes [19].

But what we should take into account is that regardless of the time spent on the DM development and loading procedures, it is still less than the time of a manual analysis. Conducting manual reviews on a set of hundreds of articles to answer questions such as “How often is the kinase protein family cited with phylum Nematoda?” would be absolutely infeasible. To get this answer, the researcher must know all the protein terms (also all variations of these terms) belonging to kinase family and all terms of the classifications and organisms belonging to phylum Nematoda, only this way would he/she get this answer.

In a rough estimate, consider that if a person could carefully analyze 3 articles per day, it would take 461 days to cover the 1383 articles. To deploy a DM using TOETL, the time spent in the Annotation (Extraction), Transformation, and Load steps are approximately 30 [29], 15 and 5 days respectively, totaling 50 days. It is clear that these estimates may vary depending on the case, but the difference between the time spent of the two approaches tends to be very large.

Another problem that makes the manual analysis of these issues impracticable for decision making, is the fact that for each new question or adding/removing of the elements of interest, a new analysis of all articles should be done. DM are so flexible that it makes possible to answer a large set of questions, besides allowing the visualization of the problem from different perspectives. Once the dimensions and facts are loaded, new questions or adding/removing of the elements of interest is done by means of simple operations and without involving the rerun of the ETL process.

(ii) What are the best possible new targets for a particular organism?. To address this question, a procedure which involves a set of queries should be performed. Using the organism *T. brucei* as an example, two proteins sets (EP and NTP) were selected:

- EP : proteins that have an important role in the essentiality of a particular organism. The criteria are: proteins that have been cited with *T. brucei* and with all techniques of essentiality throughout a given time (as an example, we used seven years minimum)
- NTP : the candidate target proteins. They are proteins that have never been cited previously with *T. brucei* organism and they must be cited with all techniques of essentiality.

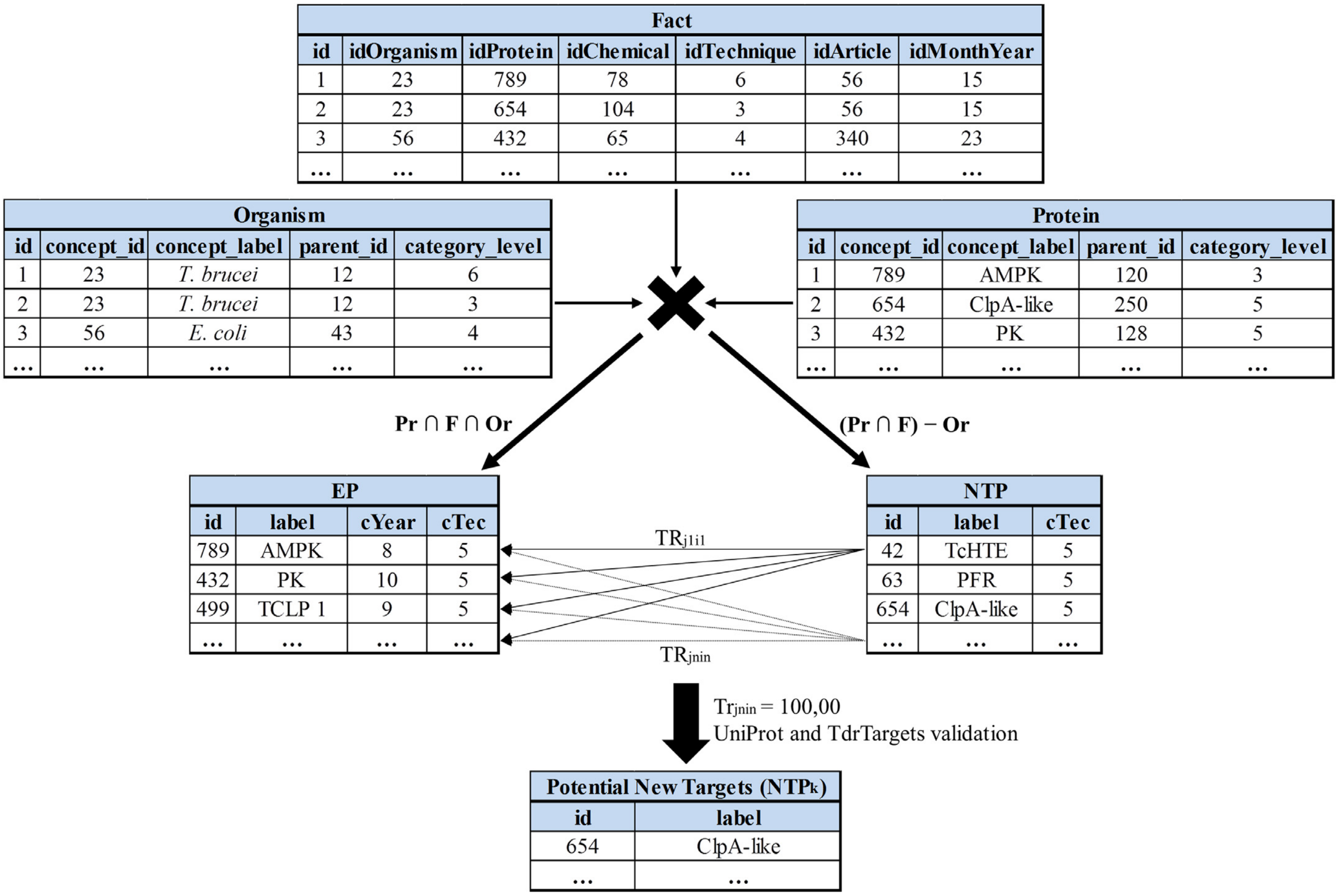
After defining the *EP* and *NTP* sets, it is necessary to calculate the rate between their elements. The idea is to correlate NTP proteins and EP proteins, by identifying those EP proteins that are essential to some organism and that are cited with some NTP protein. This can be done by calculating for each pair of proteins (P_{in}, P_{jn}) that belongs to the set $\{EP \times NTP\}$, the TR_{jnin} rate that they co-occur throughout the articles. For this purpose, it’s defined:

- $cNTP_{jn}$ is the number of articles in which P_{jn} was cited;
- $cNTP_{jn}EP_{in}$ is the number of articles that P_{jn} was cited with P_{in}.

Thus, the TR_{jnin} rate can be calculate, as shown in Eq. (1). 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.

$$TR_{jnin} = \left(\frac{cNTP_{jn} * 100}{cNTP_{jn}EP_{in}} \right) \tag{1}$$

Ultimately, 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



Fact						
id	idOrganism	idProtein	idChemical	idTechnique	idArticle	idMonthYear
1	23	789	78	6	56	15
2	23	654	104	3	56	15
3	56	432	65	4	340	23
…	…	…	…	…	…	…

Organism				
id	concept_id	concept_label	parent_id	category_level
1	23	*T. brucei*	12	6
2	23	*T. brucei*	12	3
3	56	*E. coli*	43	4
…	…	…	…	…

Protein				
id	concept_id	concept_label	parent_id	category_level
1	789	AMPK	120	3
2	654	ClpA-like	250	5
3	432	PK	128	5
…	…	…	…	…

EP			
id	label	cYear	cTec
789	AMPK	8	5
432	PK	10	5
499	TCLP 1	9	5
…	…	…	…

NTP		
id	label	cTec
42	TcHTE	5
63	PFR	5
654	ClpA-like	5
…	…	…

Potential New Targets (NTPk)	
id	label
654	ClpA-like
…	…

Fig. 6. Fragment of the tables used in the search strategy of the essentiality in *T. brucei* organism. EP: proteins that have been cited with *T. brucei* and with all techniques of essentiality throughout seven years minimum. NTP: proteins that have never been cited previously with *T. brucei*, but that have been cited with all techniques of essentiality. TR_{jnin}: relationship rate between the proteins of the EP and NTP sets.

organism of interest, it is necessary to validate these proteins using another data source (database validation). With this purpose in mind, the UniProt [30] and TdrTargets [31] 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 (Table 4 of the 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 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.

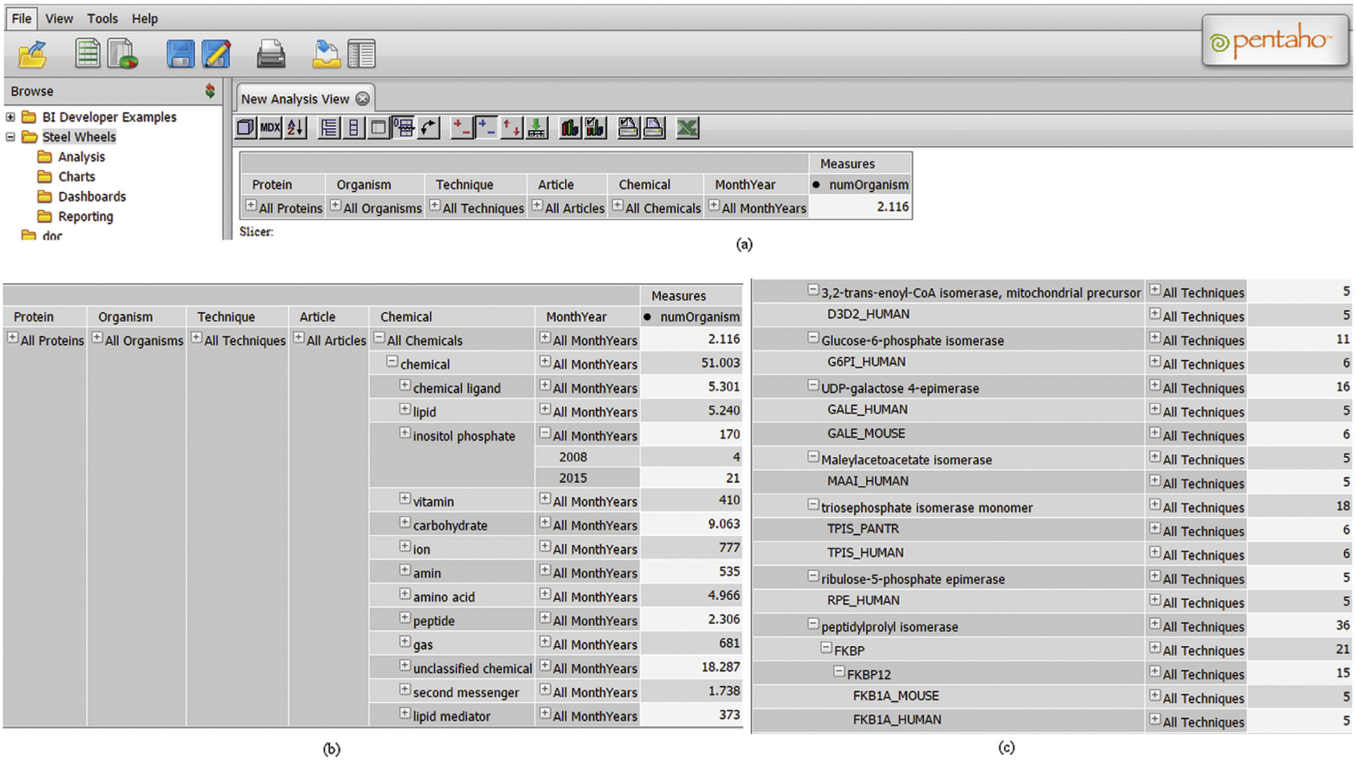
Fig. 6 gives an overview of the approach and illustrates the relationships among the organism of interest, the proteins with more probability to have an important role in the study of its essentiality (*EP* set) and the proteins never mentioned with it (*NTP* set). Note that these are disjoint sets, in other words, to establish relationships by reading the articles, among the proteins of the NTP set and the organism *T. brucei* would be very difficult, since they were never mentioned together.

3.3. 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 [32].

Fig. 7 shows Mondrian interface, where in (a) the user is able to have an overview on the dimensions that the DM covers. Fig. 7(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. Fig. 7(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.



Protein	Organism	Technique	Article	Chemical	MonthYear	Measures
						● numOrganism
⊞All Proteins	⊞All Organisms	⊞All Techniques	⊞All Articles	⊞All Chemicals	⊞All MonthYears	2.116

Protein	Organism	Technique	Article	Chemical	MonthYear	Measures
						● numOrganism
⊞All Proteins	⊞All Organisms	⊞All Techniques	⊞All Articles	⊟All Chemicals	⊞All MonthYears	2.116
				⊟chemical	⊞All MonthYears	51.003
				⊞chemical ligand	⊞All MonthYears	5.301
				⊞lipid	⊞All MonthYears	5.240
				⊞inositol phosphate	⊟All MonthYears	170
					2008	4
					2015	21
				⊞vitamin	⊞All MonthYears	410
				⊞carbohydrate	⊞All MonthYears	9.063
				⊞ion	⊞All MonthYears	777
				⊞amin	⊞All MonthYears	535
				⊞amino acid	⊞All MonthYears	4.966
				⊞peptide	⊞All MonthYears	2.306
				⊞gas	⊞All MonthYears	681
				⊞unclassified chemical	⊞All MonthYears	18.287
				⊞second messenger	⊞All MonthYears	1.738
				⊞lipid mediator	⊞All MonthYears	373

⊟3,2-trans-enoyl-CoA isomerase, mitochondrial precursor	⊞All Techniques	5
D3D2_HUMAN	⊞All Techniques	5
⊟Glucose-6-phosphate isomerase	⊞All Techniques	11
G6PI_HUMAN	⊞All Techniques	6
⊟UDP-galactose 4-epimerase	⊞All Techniques	16
GALE_HUMAN	⊞All Techniques	5
GALE_MOUSE	⊞All Techniques	6
⊟Maleylacetoacetate isomerase	⊞All Techniques	5
MAAI_HUMAN	⊞All Techniques	5
⊟triosephosphate isomerase monomer	⊞All Techniques	18
TPIS_PANTR	⊞All Techniques	6
TPIS_HUMAN	⊞All Techniques	6
⊟ribulose-5-phosphate epimerase	⊞All Techniques	5
RPE_HUMAN	⊞All Techniques	5
⊟peptidylprolyl isomerase	⊞All Techniques	36
⊟FKBP	⊞All Techniques	21
⊟FKBP12	⊞All Techniques	15
FKB1A_MOUSE	⊞All Techniques	5
FKB1A_HUMAN	⊞All Techniques	5

Fig. 7. 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.

The TaP OLAP DM is available as a demo[5], populated with the data of the reported case study. More details about the TaP OLAP DM are available at its manual[6].

4. Discussions

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 the query of example (i), with the combination of the protein and organism concepts, non-explicit relations can be found, in contrast, keyword-based traditional tools, although more commonly used than DM, do not allow this kind of refinement.

The query in example (ii) shows that the researcher can define search strategies to find out data trends. In 1383 articles, where 4187 protein terms were cited, the data mart allowed to select a few dozen terms that have the researcher's interest profile. It is important to mention that these few selected proteins were never mentioned with *T. brucei*, on this account, they would hardly be considered likely targets, even reading all 1383 articles.

Using a keyword-based tool, many relevant articles may not be retrieved because it does not consider, for instance, hierarchical relations between terms. Moreover, the TOETL approach increases the possibility of important information being duly found.

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-occurrences 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.

Lastly, the TOETL methodology is designed to be flexible and of intuitive deployment, allowing it to be applied in other scenarios according to the demands of the researchers. Applying the approach in a scenario where researchers want to map, through scientific articles, the incidence of diseases in the world, it would be necessary to annotate the corpus of articles with ontologies that describe diseases and regions of the world.

With this, the concepts (terms) and their relationships contained in the ontologies are then stored in the dimensions, enabling the generation of the facts. Thereat, it would be possible to identify the countries mentioned with each disease, allowing to navigate in hierarchies such as continents, hemispheres and so on. It would be possible to obtain indications of the incidence of the diseases, for example: Which continents a particular disease occurs more?, Which diseases are present in a specific region? or Which regions never recorded a particular disease?.

Although there are some initiatives on building data marts based on ontologies [15–18], differently from the TOETL approach, they do not focus on exploring textual data sources. One of them [16] reinforces the need to use ontologies on the DW design to overcome important shortcomings. In [15], the authors use ontologies to guide the DW design, analyzing each data source and the analysis requirements.

[5] 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").

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

Other studies used ontologies directly and automatically in the data mart design, where an ontology was used as input and, at the end of the process, a dimensional diagram was obtained [17,18]. Nevertheless, they focus only on one ontology. Due to the fact that the modeling of biomedical problems needs to use multiple ontologies, according to the different sub-domains involved, these are not suitable approaches. Moreover, in our approach, we focus on a specific analysis interest and identify cut-offs of a set of selected ontologies. Differently, their idea is to explore completely a single ontology. As mentioned before, biomedicine ontologies are quite large, and exploring the whole ontology may lead to too large or too many dimensions.

Besides, it is important to emphasize that those related initiatives discuss or propose the use of ontologies on the data mart design, and not on the ETL process. In our approach, ontologies are used first to annotate texts extracted from selected data sources (ETL first step), and then to guide the DM design. In the context of scientific textual data sources, it is necessary to identify the focus of the research in order to proceed to the DM design. Therefore, our approach integrates both the ETL and DM design processes.

Some authors [33,34] used text mining techniques applied to scientific texts to extract hidden relationships between terms. These links are difficult to obtain only by reading the articles. The GOPubMed initiative uses the Gene Ontology (GO) applied to the abstracts of the articles, making the search more efficient [35] than keyword-based interfaces. Other initiatives involving mining and the use of ontologies are Bio2RDF [36] and EBI RDF [37], but their focus is different from this work, because they propose integrating structured data from different sources and formats. Moreover, finding links between concepts may not be sufficient to take a better decision. Decision makers (researchers) need to analyze the facts and get answers to their specific questions [38].

Unlike these cited works that use text mining together (or not) with ontologies, the TOETL methodology is designed to use multiple ontologies and DW techniques. It is able to associate data present in the articles from different domains of knowledge, broadening the perspective of the problems under analysis. Data warehouses is a mature approach which is highly explored by the corporate world, able to answer issues involving who, what, when, where and other questions [39], providing analysis over the time.

Another important work [40] addresses the construction of DW from unstructured data, detailing an 11-step ETL process. But this work proposes a NLP-based approach to the ETL, and applies techniques, such as phrase recognition, stop word filtering, and synonym replacement. Differently, it does not use ontology-based annotations to define dimensions, nor use the ontologies to enrich dimension hierarchies. Therefore, to the best of our knowledge, there is no work similar to the present work.

Finally, the idea to explore the PubMed database to identify drug targets, for treatment of neglected diseases is not new. In [41] they report on the use of text mining steps to extract terms related to protein names not yet explored as targets for those diseases. However, they do not focus on building a DM for further analysis, nor they use ontologies to classify or organize terms according to their context.

5. Conclusions

This work presented a scientific scenario where the scientist aims at prioritize drug targets. In such scientific scenarios, the challenge is to filter useful information from a huge amount of scientific texts, in an agile way so it can support the researcher decision making. This work shows that with the use of ontologies and semantic annotation techniques, it is possible to extract information from a corpus of scientific texts, and store it in such a way that it helps the scientists on making decisions with respect to his research directions.

By means of searches on TaP OLAP DM, this work reveals that the proposed approach can aid in the prioritization of new drug targets, finding 91 proteins out of hundreds to start with. It innovates by proposing the TOETL method for extracting information from a corpus, through semantic annotation, to support decision making. This approach can be applied on any other research field or subject.

It is also worth to mention that the TOETL approach is more efficient than keyword-based traditional tools since it is able to index articles through generic and synonym terms. Moreover, articles can be ranked and filtered through analytical queries according to the user research focus.

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.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cmpb.2018.01.010.

References

[1] D.W. Embley, Y. Ding, S.W. Liddle, M. Vickers, Automatic creation and simplified querying of semantic web content: an approach based on information-extraction ontologies, in: Proceedings of the First Asian Semantic Web Conference (ASWC), in: LNCS, 4185, 2006, pp. 400–414.

[2] K. T., D.I.S.B. Monteiro, T.F. Lima, F.P.S. Jr., M.C. Cavalcanti, Analyzing tools for biomedical text annotation with multiple ontologies, in: Proceedings of the 3rd International Conference on Biomedical Ontology (ICBO), KR-MED Series, Graz, Austria, 2012.

[3] T. Berners-Lee, J. Hendler, O. Lassila, The semantic web, Sci. Am. 284 (5) (2001) 34–43.

[4] P.L. Whetzel, N.F. Noy, N.H. Shah, P.R. Alexander, C. Nyulas, T. Tudorache, M.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 Res. 39 (2011) 541–545, doi:10.1093/nar/gkr469.

[5] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L.J. Goldberg, K. Eilbeck, A. Ireland, C.J. Mungall, et al., The obo foundry: coordinated evolution of ontologies to support biomedical data integration, Nat. Biotechnol. 25 (11) (2007) 1251–1255.

[6] W.H. Inmon, Building the Data Warehouse, fourth ed., John Wiley & Sons, Inc., New York, NY, USA, 2005.

[7] R. Kimball, M. Ross, The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling, third ed., Wiley Publishing, 2013.

[8] B. Haarbrandt, E. Tute, M. Marschollek, Automated population of an i2b2 clinical data warehouse from an openehr-based data repository, J. Biomed. Inform. 63 (2016) 277–294, doi:10.1016/j.jbi.2016.08.007.

[9] A.-S. Jannot, E. Zapletal, P. Avillach, M.-F. Mamzer, A. Burgun, P. Degoulet, The georges pompidou university hospital clinical data warehouse: a 8-years follow-up experience, Int. J. Med. Inform. 102 (2017) 21–28, doi:10.1016/j.ijmedinf.2017.02.006.

[10] M. De Mul, P. Alons, P. Van der Velde, I. Konings, J. Bakker, J. Hazelzet, Development of a clinical data warehouse from an intensive care clinical information system, Comput. Methods Progr. Biomed. 105 (1) (2012) 22–30.

[11] G. Sathe, S. Sarawagi, Intelligent rollups in multidimensional OLAP data, in: Proceedings of the 27th International Conference on Very Large Data Bases (VLDB), Roma, Italy, 2001, pp. 531–540.

[12] J.L. Moreira, K. de Faria Cordeiro, M.L.M. Campos, Jointolap–sistema de informação para exploração conjunta de dados estruturados e textuais: Um estudo de caso no setor elétrico (2013).

[13] P. Wongthongtham, B. Abu-Salih, Ontology and trust based data warehouse in new generation of business intelligence: State-of-the-art, challenges, and opportunities, in: Proceedings of the IEEE 13th International Conference on Industrial Informatics (INDIN), IEEE, 2015, pp. 476–483.

[14] M.F. Abdullah, K. Ahmad, Business intelligence model for unstructured data management, in: Proceedings of the International Conference on Electrical Engineering and Informatics (ICEEI), IEEE, 2015, pp. 473–477.

[15] M. Thenmozhi, K. Vivekanandan, A tool for data warehouse multidimensional schema design using ontology, Int. J. Comput. Sci. Issues 10 (2) (2013) 161–168.

[16] J. Pardillo, J. Mazón, Using ontologies for the design of data warehouses, J. Database Manag 3 (2) (2011) 73–87.

[17] M. Gulić, Transformation of owl ontology sources into data warehouse, in: Proceedings of the 36th International Convection on Information and Communication Technology Electronics and Microelectronics (MIPRO), 2013, pp. 1143–1148.

[18] O. Romero, A. Abelló, Automating multidimensional design from ontologies, in: Proceedings of the ACM Tenth International Workshop on Data Warehousing and OLAP, DOLAP, ACM, New York, NY, USA, 2007, pp. 1–8, doi:10.1145/1317331.1317333.

[19] M.A.A. da Silva, M.C.R. Cavalcanti, K.T. Belloze, F. Silva-Junior, Agile semantic annotation of scientific texts at the biomedical scenario, in: Proceedings of the 10th IEEE International Conference on e-Science, (eScience), Sao Paulo, Brazil, 2014, pp. 100–107, doi:10.1109/eScience.2014.46.

[20] K.T. Belloze, Priorização de alvos para fármacos no combate a doenças tropicais negligenciadas causadas por protozoários (*in portuguese*), 2013.

[21] K.T. Belloze, D.I.S. Monteiro, T.F. Lima, F.P. Silva Jr, M.C.R. Cavalcanti, An evaluation of annotation tools for biomedical texts., ONTOBRAS-MOST 108 (2012) 119.

[22] C.A. Fontes, M.C. Cavalcanti, A.M. de Carvalho Moura, An ontology-based reasoning approach for document annotation, in: Proceedings of the IEEE 7th International Conference on Semantic Computing, Irvine, CA, USA, 2013, pp. 160–167, doi:10.1109/ICSC.2013.37.

[23] N. Sioutos, S. de Coronado, M.W. Haber, F.W. Hartel, W. Shaiu, L.W. Wright, NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information, J. Biomed. Inform. 40 (1) (2007) 30–43, doi:10.1016/j.jbi.2006.02.013.

[24] B.C. Grau, I. Horrocks, Y. Kazakov, U. Sattler, Modular reuse of ontologies: theory and practice, J. Art. Int. Res. 31 (2008) 273–318.

[25] J. Seidenberg, Web ontology segmentation: extraction, transformation, evaluation, in: Modular Ontologies, 2009, pp. 211–243.

[26] G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, Inf. Process. Manage. 24 (5) (1988) 513–523, doi:10.1016/0306-4573(88)90021-0.

[27] P. Hasse, H. Lewen, R. Studer and M. Erdmann, The NeOn Ontology Engineering Toolkit, 2008. http://watson.kmi.open.ac.uk/Downloads%20and%20Publications_files/neon-toolkit.pdf.

[28] R. Liepinš, M. Grasmanis, U. Bojars, Owlgred ontology visualizer, in: Proceedings of the 2014 International Conference on Developers, vol. 1268, CEUR-WS.org, 2014, pp. 37–42.

[29] M.A.A. da Silva, M.C. Cavalcanti, Combining ontology modules for scientific text annotation, JIDM 5 (3) (2014) 238–251.

[30] R. Apweiler, A. Bairoch, C.H. Wu, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, et al., Uniprot: the universal protein knowledgebase, Nucleic Acids Res. 32 (2004) D115–D119.

[31] M.P. Magariños, S.J. Carmona, G.J. Crowther, S.A. Ralph, D.S. Roos, D. Shanmugam, W.C. Van Voorhis, F. Agüero, TDR targets: a chemogenomics resource for neglected diseases, Nucleic Acids Res. 40 (D1) (2012) D1118–D1127.

[32] M. Casters, R. Bouman, J. Van Dongen, Pentaho Kettle Solutions: Building Open Source ETL Solutions with Pentaho Data Integration, John Wiley & Sons, 2010.

[33] K. Kasemsap, Text mining: current trends and applications 338 (2016).

[34] W.W. Fleuren, W. Alkema, Application of text mining in the biomedical domain, Methods 74 (2015) 97–106.

[35] A. Doms, M. Schroeder, Gopubmed: exploring pubmed with the gene ontology, Nucleic Acids Res. 33 (2005) W783–W786.

[36] A. Callahan, J. Cruz-Toledo, P. Ansell, M. Dumontier, Bio2RDF release 2: improved coverage, interoperability and provenance of life science linked data, in: Extended Semantic Web Conference, Springer, 2013, pp. 200–212.

[37] S. Jupp, J. Malone, J. Bolleman, M. Brandizi, M. Davies, L. Garcia, A. Gaulton, S. Gehant, C. Laibe, N. Redaschi, et al., The EBI RDF platform: linked open data for the life sciences, Bioinformatics 30 (9) (2014) 1338–1339.

[38] K. Prasad, S. Ramakrishna, Text analytics to data warehousing, Int. J. Comput. Sci. Eng. 2 (6) (2010) 2201–2207.

[39] L. Gao, E. Chang, S. Han, Powerful tool to expand business intelligence: text mining, in: Proceedings of the Transactions on Enformatika, Systems Sciences and Engineering(ESSE), vol. 8, International Academy of Sciences, Budapest, 2005, pp. 110–115.

[40] B. Inmon, K. Krishnan, Building the Unstructured Data Warehouse: Architecture, Analysis, and Design, Technics Publications, 2011.

[41] E. Barçante, M. Jezuz, F. Duval, E. Caffarena, O.G. Cruz, F. Silva, Identifying drug repositioning targets using text mining, in: Proceedings of the International Conference on Knowledge Discovery and Information Retrieval (KDIR), 2014, pp. 348–353, doi:10.5220/0005134903480353.