2 Resources

Abstract

The previous chapter presented the importance of text and semantic resources for Health and Life studies. This chapter will describe what kind of text and semantic resources are available, where they can be found, and how they can be accessed and retrieved.

Keywords

Biomedical literature · Programmatic access · UniProt citations service · Semantics · Controlled vocabularies · Ontologies · OWL: Web Ontology Language · URI: Uniform Resource Identifier · DAG: Directed Acyclic Graphs · OBO: Open Biomedical Ontologies

Biomedical Text

Text is still the preferential means of publishing novel knowledge in Health and Life Sciences, and where we can expect to find all the information about the supporting data. Text can be found and explored in multiple types of sources, the main being scientific articles and patents (Krallinger et al. 2017). However, less formal texts are also relevant to explore, such as the ones present nowadays in electronic health records (Blumenthal and Tavenner 2010).

What?

In the biomedical domain, we can find text in different forms, such as:

Statement: a short piece of text, normally containing personal remarks or an evidence about a biomedical phenomenon;

Abstract: a short summary of a larger scientific document;

Full-text: the entire text present in a scientific document including scattered text such as figure labels and footnotes.

Statements contain more syntactic and semantic errors than abstracts, since they normally are not peer-reviewed, but they are normally directly linked to data providing useful details about it. The main advantage of using statements or abstracts is the brief and succinct form on which the information is expressed. In the case of abstracts, there was already an intellectual exercise to present only the main facts and ideas. Nevertheless, a brief description may be insufficient to draw a solid conclusion, that may require some important details not possible to summarize in a short piece of text (Schuemie et al. 2004). These details are normally presented in the form of a full-text document, which contains a complete description of the results obtained. For example, important details are sometimes only present in figure labels (Yeh et al. 2003).

F. M. Couto, *Data and Text Processing for Health and Life Sciences*,
Advances in Experimental Medicine and Biology 1137,
https://doi.org/10.1007/978-3-030-13845-5_2

One major problem of full-text documents is their availability, since their content may have restricted access. In addition, the structure of the full-text and the format on which is available varies according to the journal in where it was published. Having more information does not mean that all of it is beneficial to find what we need. Some of the information may even induce us in error. For example, the relevance of a fact reported in the Results Section may be different if the fact was reported in the Related Work Section. Thus, the usage of full-text may create several problems regarding the quality of information extracted (Shah et al. 2003).

Where?

Access to biomedical literature is normally done using the internet through PubMed[1], an information retrieval system released in 1996 that allows researchers to search and find biomedical texts of relevance to their studies (Canese 2006). PubMed is developed and maintained by the National Center for Biotechnology Information (NCBI), at the U.S. National Library of Medicine (NLM), located at the National Institutes of Health (NIH). Currently, PubMed provides access to more than 28 million citations from MEDLINE, a bibliographic database with references to a comprehensive list of academic journals in Health and Life Sciences[2]. The references include multiple metadata about the documents, such as: title, abstract, authors, journal, publication date. PubMed does not store the full-text documents, but it provides links where we may find the full-text. More recently, biomedical references are also accessible using the European Bioinformatics Institute (EBI) services, such as Europe PMC[3], the Universal Protein Resource (UniProt) with its UniProt citations service[4].

Other generic alternative tools have been also gaining popularity for finding scientific texts, such as Google Scholar[5], Google Patents[6], ResearchGate[7] and Mendeley[8].

More than just text some tools also integrate semantic links. One of the first search engines for biomedical literature to incorporate semantics was GOPubMed[9], that categorized texts according to Gene Ontology terms found in them (Doms and Schroeder 2005). These semantic resources will be described in a following section. A more recent tool is PubTator[10] that provides the text annotated with biological entities generated by state-of-the-art text-mining approaches (Wei et al. 2013).

There is also a movement in the scientific community to produce Open Access Publications, making full-texts freely available with unrestricted use. One of the main free digital archives of free biomedical full-texts is PubMed Central[11] (PMC), currently providing access to more than 5 million documents.

Other relevant source of biomedical texts is the electronic health records stored in health institutions, but the texts they contain are normally directly linked to patients and therefore their access is restricted due to ethical and privacy issues. As example, the THYME corpus[12] includes more than one thousand de-identified clinical notes from the Mayo Clinic, but is only available for text processing research under a data use agreement (DUA) with Mayo Clinic (Styler IV et al. 2014).

From generic texts we can also sometimes find relevant biomedical information. For example, some recent biomedical studies have been processing the texts in social networks to identify new trends and insights about a disease, such as processing tweets to predict flu outbreaks (Aramaki et al. 2011).

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

[2] https://www.nlm.nih.gov/bsd/medline.html

[3] http://europepmc.org/

[4] https://www.uniprot.org/citations/

[5] http://scholar.google.com/

[6] http://www.google.com/patents

[7] https://www.researchgate.net/

[8] https://www.mendeley.com/

[9] https://gopubmed.org/

[10] http://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/PubTator/

[11] https://www.ncbi.nlm.nih.gov/pmc/

[12] http://thyme.healthnlp.org/

How?

To automatically process text, we need programmatic access to it, this means that from the previous biomedical data repositories we can only use the ones that allow this kind of access. These limitations are imposed because many biomedical documents have copyright restrictions hold by their publishers. And some restrictions may define that only manual access is granted, and no programmatic access is allowed. These restrictions are normally detailed in the terms of service of each repository. However, when browsing the repository if we face a CAPTCHA challenge to determine whether we are humans or not, probably means that some access restrictions are in place.

Fortunately, NCBI[13] and EBI[14] online services, such as PubMed, Europe PMC, or UniProt Citations, allow programmatic access (Li et al. 2015). Both institutions provide Web APIs[15] that fully document how web services can be programmatically invoked. Some resources can inclusively be accessed using RESTful web services[16] that are characterized by a simple uniform interface that make any Uniform Resource Locator (URL) almost self-explanatory (Richardson and Ruby 2008). The same URL shown by our web browser is the only thing we need to know to retrieve the data using a command line tool.

For example, if we search for *caffeine* using the UniProt Citations service[17], select the first two entries, and click on download, the browser will show information about those two documents using a tabular format.

```
PubMed ID Title Authors/Groups
    Abstract/Summary
27702941 Genome-wide association
    ...
22333316 Modeling caffeine
    concentrations ...
```

More important is to check the URL that is now being used:

```
https://www.uniprot.org/
    citations/?sort=score&desc=&
    compress=no&query=id
    :27702941%20OR%20id:22333316&
    format=tab&columns=id
```

We can check that the URL has three main components: the scheme (`https`), the hostname (`www.uniprot.org`), the service (`citations`) and the data parameters. The scheme represents the type of web connection to get the data, and usually is one of these protocols: Hypertext Transfer Protocol (HTTP) or HTTP Secure (HTTPS)[18]. The hostname represents the physical site where the service is available. The list of parameters depends on the data available from the different services. We can change any value of the parameters (arguments) to get different results. For example, we can replace the two PubMed identifiers by the following one 29029291[19], and our browser will now display the information about this new document:

```
PubMed ID Title Authors/Groups
    Abstract/Summary
29029291 Nutrition Influences...
```

The good news is that we can use this link with a command line tool and automatize the retrieval of the data, including extracting the abstract to process its text.

Semantics

Lack of use of standard nomenclatures across biological text makes text processing a non-trivial task. Often, we can find different labels (synonyms, acronyms) for the same biomedical entities, or, even more problematic, different entities sharing the same label (homonyms) (Rebholz-Schuhmann et al. 2005). Sense disambiguation to select the correct meaning of an expression in

[13] https://www.ncbi.nlm.nih.gov/home/develop/api/

[14] https://www.ebi.ac.uk/seqdb/confluence/display/JDSAT/

[15] https://en.wikipedia.org/wiki/Web_API

[16] https://www.ebi.ac.uk/seqdb/confluence/pages/viewpage.action?pageId=68165098

[17] https://www.uniprot.org/citations/

[18] https://en.wikipedia.org/wiki/Hypertext_Transfer_Protocol

[19] https://www.uniprot.org/citations/?sort=score&desc=&compress=no&query=id:29029291&format=tab&columns=id

a given piece of text is therefore a crucial issue. For example, if we find the disease acronym *ATS* in a text, we may have to figure out if it representing the *Andersen-Tawil syndrome*[20] or the *X-linked Alport syndrome*[21]. Further in the book, we will address this issue by using ontologies and semantic similarity between their classes (Couto and Lamurias 2019).

What?

In 1993, Gruber (1993) proposed a short but comprehensive definition of ontology as an:

> an explicit specification of a conceptualization

In 1997 and 1998, Borst and Borst (1997) and Studer et al. (1998) refined this definition to:

> a formal, explicit specification of a shared conceptualization

A conceptualization is an abstract view of the concepts and the relationships of a given domain. A shared conceptualization means that a group of individuals agree on that view, normally established by a common agreement among the members of a community. The specification is a representation of that conceptualization using a given language. The language needs to be formal and explicit, so computers can deal with it.

Languages

The Web Ontology Language (OWL)[22] is nowadays becoming one of the most common languages to specify biomedical ontologies (McGuinness et al. 2004). Another popular alternative is the Open Biomedical Ontology (OBO)[23] format developed by the OBO foundry. OBO established a set of principles to ensure high quality, formal rigor and interoperability between other OBO ontologies (Smith et al. 2007). One important principle is that OBO ontologies need to be open and available without any constraint other than acknowledging their origin.

Concepts are defined as OWL classes that may include multiple properties. For text processing important properties include the labels that may be used to mention that class. The labels may include the official name, acronyms, exact synonyms, and even related terms. For example, a class defining the disease *malignant hyperthermia* may include as synonym *anesthesia related hyperthermia*. Two distinct classes may share the same label, such as *Andersen-Tawil syndrome* and *X-linked Alport syndrome* that have *ATS* as an exact synonym.

Formality

The representation of classes and the relationships may use different levels of formality, such as controlled vocabularies, taxonomies and thesaurus, that even may include logical axioms.

Controlled vocabularies are list of terms without specifying any relation between them. Taxonomies are controlled vocabularies that include subsumption relations, for example specifying that *malignant hyperthermia* is a *muscle tissue disease*. This *is-a* or subclass relations are normally the backbone of ontologies. We should note that some ontologies may include multiple inheritance, i.e. the same concept may be a specialization of two different concepts. Therefore, many ontologies are organized as a directed acyclic graphs (DAG) and not as hierarchical trees, as the one represented in Fig. 2.1. A thesaurus includes other types of relations besides subsumption, for example specifying that *caffeine* has role *mutagen*.

Gold Related Documents

The importance of these relations can be easily understood by considering the domain modeled by the ontology in Fig. 2.1, and the need to find texts related to *gold*. Assume a corpus with one distinct document mentioning each metal, except for *gold* that no document mentions. So, which documents should we read first?

The document mentioning *silver* is probably the most related since it shares with *gold* two parents, *precious* and *coinage*. However, choos-

[20] http://purl.obolibrary.org/obo/DOID_0050434

[21] http://purl.obolibrary.org/obo/DOID_0110034

[22] https://en.wikipedia.org/wiki/Web_Ontology_Language

[23] https://en.wikipedia.org/wiki/Open_Biomedical_Ontologies

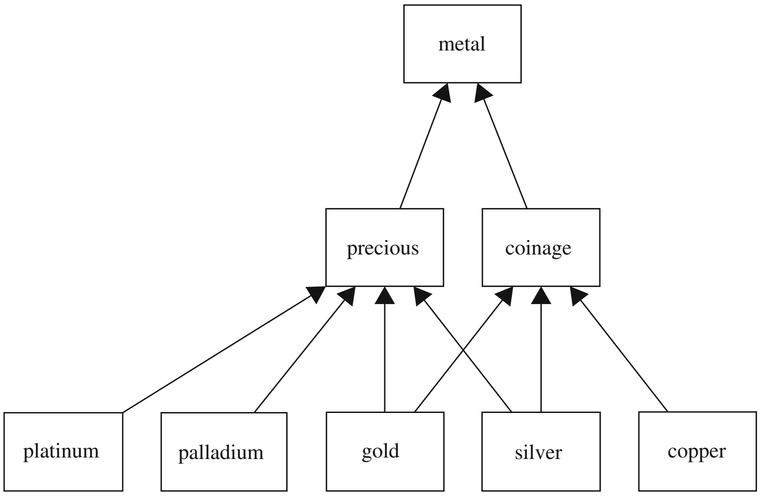

Fig. 2.1 A DAG representing a classification of metals with multiple inheritance, since *gold* and *silver* are considered both precious and coinage metals (All the links represent *is-a* relations)

ing between the documents mentioning *platinum* or *palladium* or the document mentioning *copper* depends on our information need. This information can be obtained by our previous searches or reads. For example, assuming that our last searches included the word *coinage*, then document mentioning *copper* is probably the second-most related. The importance of these semantic resources is evidenced by the development of the knowledge graph[24] by Google to enhance their search engine (Singhal 2012).

Where?

Most of the biomedical ontologies are available through BioPortal[25]. In December of 2018, BioPortal provided access to more than 750 ontologies representing more than 9 million classes. BioPortal allows us to search for an ontology or a specific class. For example, if we search for *caffeine*, we will be able to see the large list of ontologies that define it. Each of these classes represent conceptualizations of *caffeine* in different domains and using alternative perspectives. To improve interoperability some ontologies include class properties with a link to similar classes in other ontologies. One of the main goals of the OBO initiative was precisely to tackle this somehow disorderly spread of definitions for the same concepts. Each OBO ontology covers a clearly specified scope that is clearly identified.

OBO Ontologies

A major example of success of OBO ontologies is the Gene Ontology (GO) that has been widely and consistently used to describe the molecular function, biological process and cellular component of gene-products, in a uniform way across different species (Ashburner et al. 2000). Another OBO ontology is the Disease Ontology (DO) that provides human disease terms, phenotype characteristics and related medical vocabulary disease concepts (Schriml et al. 2018). Another OBO ontology is the Chemical Entities of Biological Interest (ChEBI) that provides a classification of molecular entities with biological interest with a focus on small chemical compounds (Degtyarenko et al. 2007).

Popular Controlled Vocabularies

Besides OBO ontologies, other popular controlled vocabularies also exist. One of them is the International Classification of Diseases (ICD)[26], maintained by the World Health Organization (WHO). This vocabulary contains a list of

[24] https://en.wikipedia.org/wiki/Knowledge_Graph

[25] http://bioportal.bioontology.org/

[26] https://www.who.int/classifications/icd/en/

generic clinical terms mainly arranged and classified according to anatomy or etiology. Another example is the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT)[27], currently maintained and distributed by the International Health Terminology Standards Development Organization (IHTSDO). The SNOMED CT is a highly comprehensive and detailed set of clinical terms used in many biomedical systems. The Medical Subject Headings (MeSH)[28] is a comprehensive controlled vocabulary maintained by the National Library of Medicine (NLM) for classifying biomedical and health-related information and documents. Both MeSH and SNOMED CT are included in the Metathesaurus of the Unified Medical Language System (UMLS)[29], maintained by the U.S National Library of Medicine. This is a large resource that integrates most of the available biomedical vocabularies. The 2015AB release covered more than three million concepts.

Another alternative to BioPortal is Ontobee[30], a repository of ontologies used by most OBO ontologies, but it also includes many non-OBO ontologies. In December 2018, Ontobee provided access to 187 ontologies (Ong et al. 2016).

Other alternatives outside the biomedical domain include the list of vocabularies gathered by the W3C SWEO Linking Open Data community project[31], and by the W3C Library Linked Data Incubator Group[32].

How?

After finding the ontologies that cover our domain of interest in the previous catalogs, a good idea is to find their home page and download the files from there. This way, we will be sure that we get the most recent release in the original format and select the subset of the ontology that really matter for our work. For example, ChEBI provides three versions: LITE, CORE and FULL[33]. Since we are interested in using the ontology just for text processing, we are probably not interested in chemical data and structures that is available in CORE. Thus, LITE is probably the best solution, and it will be the one we will use in this book. However, we may be missing synonyms that are only included in the FULL version.

OWL

The OWL language is the prevailing language to represent ontologies, and for that reason will be the format we will use in this book. OWL extends RDF Schema (RDFS) with more complex statements using description logic. RDFS is an extension of RDF with additional statements, such as class-subclass or property-subproperty relationships. RDF is a data model that stores information in statements represented as triples of the form subject, predicate and object. Originally, W3C recommended RDF data to be encoded using Extensible Markup Language (XML) syntax, also named RDF/XML. XML is a self-descriptive mark-up language composed of data elements.

For example, the following example represents an XML file specifying that *caffeine* is a drug that may treat the condition of sleepiness, but without being an official treatment:

```
<treatment category="non-
   official">
 <drug>caffeine</drug>
 <condition>sleepiness</
    condition>
</treatment>
```

The information is organized in an hierarchical structure of data elements. `treatment` is the parent element of `drug` and `condition`. The character < means that a new data element is being specified, and the characters </ means

[27] https://digital.nhs.uk/services/terminology-and-classifications/snomed-ct

[28] https://www.nlm.nih.gov/mesh/

[29] https://www.nlm.nih.gov/research/umls/

[30] http://www.ontobee.org/

[31] http://www.w3.org/wiki/TaskForces/CommunityProjects/LinkingOpenData/CommonVocabularies

[32] http://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset-20111025

[33] https://www.ebi.ac.uk/chebi/downloadsForward.do

that a specification of data element will end. The `treatment` element has a property named `category` with the value `non-official`. The `drug` and `condition` elements have as values `caffeine` and `sleepiness`, respectively. This is a very simple XML example, but large XML files are almost unreadable by humans.

To address this issue other encoding languages for RDF are now being used, such as N3[34] and Turtle[35]. Nevertheless, most biomedical ontologies are available in OWL using XML encoding.

URI

The Uniform Resource Identifier (URI) was defined as the standard global identifier of classes in an ontology. For example, the class `caffeine` in ChEBI is identified by the following URI:

```
http://purl.obolibrary.org/obo/
    CHEBI_27732
```

If a URI represents a link to a retrievable resource is considered a Uniform Resource Locator, or URL. In other words, a URI is a URL if we open it in a web browser and obtain a resource describing that class.

Sometimes, ontologies are also available as database dumps. These dumps are normally SQL files that need to be fed to a DataBase Management System (DBMS)[36]. If for any reason we must deal with these files, we can use the simple command line tool named `sqlite3`. The tool has the option to execute the SQL commands to import the data into a database (`.read` command), and to export the data into a CSV file (`.mode` command) (Allen and Owens 2011).

Further Reading

One important read if we need to know more about biomedical resources is the Arthur Lesk's book about bioinformatics (Lesk 2014). The book has entire chapters dedicated to where data and text can be found, providing a comprehensive overview of the type of biomedical information available, nowadays.

A more pragmatic approach is to explore the vast number of manuals, tutorials, seminars and courses provided by the EBI[37] and NCBI[38].

[34] https://en.wikipedia.org/wiki/Notation3

[35] https://en.wikipedia.org/wiki/Turtle_(syntax)

[36] https://en.wikipedia.org/wiki/Database#Database_management_system

[37] https://www.ebi.ac.uk/training

[38] https://www.ncbi.nlm.nih.gov/home/learn/