

A Combined Resource of Biomedical Terminology and its Statistics

Tilia Renate Ellendorff

Institute of Computational Linguistics
University of Zurich
tilia.ellendorff@uzh.ch

Adrian van der Lek

Institute of Computational Linguistics
University of Zurich
adrian.vanderlek@uzh.ch

Lenz Furrer

Institute of Computational Linguistics
University of Zurich
lenz.furrer@uzh.ch

Fabio Rinaldi

Institute of Computational Linguistics
University of Zurich
fabio.rinaldi@uzh.ch

Abstract

In this paper, we present a large biomedical term resource automatically compiled from the terminology of a selection of biomedical databases. The resource has a very simple and intuitive format and therefore can be easily embedded into a system for biomedical text mining and used as a linguistic resource. It is continuously updated and a user interface makes it possible to compile a new term resource according to individual requirements by selecting specific databases to be included. We present statistics for each included biomedical entity type separately as well as in the context of the combined terminology.

1 Introduction

Discovering entities such as genes, chemicals, diseases, species, etc. in the written text of research articles is an important part of biomedical text mining. This task is commonly known as “named entity recognition” (NER). Given a word from a text, it consists in deciding if this word is the name of an entity of interest. However, in the biomedical domain, the main focus of interest are not only the words in the text that refer to specific entities, but also under which database identifiers these entities are registered. The main purpose of using unique database identifiers is to provide unique conceptual referents for entities. Since biomedical entities tend to be highly ambiguous this disambiguation process is crucial. Furthermore, identifiers establish a reference to a database. This can be used either to retrieve additional information

about the specific entity or to add related information to the database.

There are three major approaches for tackling the task of NER. A rule-based approach uses hand-crafted rules capturing structures of the word itself as well as its context. A machine learning approach extracts different lexical and contextual features from annotated corpora and applies a sophisticated statistical analysis. The third approach simply uses a dictionary look-up for discovering entity names which are already known and present in a given database.

Whereas the first and the second approach are possibly able to discover novel entity names which have never been seen in the literature before, the third approach is restricted to the dictionary, which means that only previously seen entity names can be discovered. The third approach, on the other hand, has the big advantage that by applying a dictionary look-up, a reference to a database can be established immediately.

In practice usually a combination of approaches is used. Combined approaches typically include a dictionary look-up as second step after using machine learning or rules for identifying entity mentions in the text. This second step provides database identifiers for the entity mentions discovered in the first step. Therefore, terminological resources are an important component of most systems for biomedical text mining.

There are a range of available biomedical databases containing terminological information for one or more entity types. Most of these databases are not designed to meet the needs of biomedical text mining and are typically available in very different formats. Therefore, when

building a text mining system, it can be time-consuming to extract their terminology and bring it into a format which can be easily used for term look-up.

Furthermore, different biomedical entity types have differences in their lexical properties. For instance genes tend to have a high degree of ambiguity. Chemicals tend to have many synonyms as their official name can be very long and complex, depicting different characteristics of the respective molecule. As a result of this, they are typically replaced by a shorter version in practice.

The aim of this paper is, on the one hand, to present a terminological resource in a very simple format which allows for easy integration as a dictionary in text mining systems, and, on the other hand, to analyze its term content regarding the lexical properties of the different terminologies that have been integrated, as well as the entity types in focus.

In the following sections of this paper we first give a very short introduction to the setting and purpose of biomedical text mining in general. Then we describe known properties of the different entity types. In the main part of the paper, we give a listing of a selection of databases containing terminological information about entities. We describe how we process these databases in order to transfer them into a simple and intuitive format. We consider the characteristics of the different entity types involved, namely genes and proteins, chemicals, diseases, species and cell lines. Finally, we present statistics of the terminology taken from each database on its own, as well as in the context of the whole combined term resource.

2 Biomedical Text Mining

It is essential to keep databases in the domain of life-sciences and biomedicine up to date in order to make knowledge easily accessible to researchers and support them in their daily work flow. New findings in the domain are typically published in the format of scientific articles. In the last decades the task of entering these findings into the database has still mainly relied on human manual work as an expensive and time-intensive process. Nowadays it becomes increasingly impossible for these specially trained curators alone to keep up with the increasing rate of publications in the domain. Automatizing this process, does

not only help save money and time, but with the increasing rate of research and published papers it becomes the only way of coping with the huge inflow of information.

Biomedical text mining can be used to partially automate the process of biomedical literature curation by using computational power for discovering biomedical entities together with interaction and events in which they participate (Rinaldi et al., 2013). A successful biomedical text mining system is typically based on a pipeline which first discovers entities of interest in the text of a scientific article and subsequently looks for interactions between them. As described above, finding the unique database identifiers of the entities in focus is an important step in this process. A dictionary look-up considering all terms in an article is the only way for grounding them to their respective database identifiers. Which database identifiers are used in this process depends largely on the application for which a text mining system is built, or in other words, the database for which the system is designed to extract information.

3 Known Properties of Biomedical Entities

Compared to NER for other text genres, biomedical NER is known to be especially challenging. This is mainly due to the lexical properties of biomedical entities. The most prominent of these properties is the variability of entity names. Despite the existence of terminology recommendations by nomenclature organizations, authors are still free to use whatever variant of an entity name they prefer. Therefore, every author of the domain tends to use his or her own variant, often due to orthographic variations (Krauthammer and Nenadic, 2004). Furthermore, since many standard entity names are long and complicated, a very large range of abbreviations are continuously invented. Even though this is a characteristic shared by most entity types of the biomedical domain, each of them still has their own typical properties.

The biomedical entity type that has been most investigated in the past are genes and gene products, i.e. proteins. Usually, genes and proteins are treated as one group in biomedical NER, as their names are frequently used in place of each other. It is characteristic of genes/proteins to have a high level of ambiguity. One reason for this is that

genes are typically treated as different concepts, depending on the species for which the gene is described. Even though the concept is understood as different, often the same entity name is used. For example the gene *p53* (tumor suppressor gene) is found in a range of species, among them humans, mice and rats, and for all of them the same gene name is used. Another reason why gene names are so ambiguous is the fact that gene names can be very arbitrary, either because the discovering researcher chose an unusual name or because they are named according to their function, position on a gene or relation to other genes¹. Furthermore, they can be built up of letters, numbers, punctuation, stopwords or non-alphabetical characters and frequently they are multi-word units (e.g. *daughters against decapentaplegic*, *short stop*, *cheap date*).

Chemicals have gained the reputation to belong to the most challenging entity names of the biomedical domain. They are highly heterogeneous as they can include generic names (e.g. *water*, *alcohol* or *cigarette smoke*), brand names (e.g. *Aspirin*), IUPAC (International Union of Pure and Applied Chemistry) names (*2-(Acetyloxy)benzoic Acid*) to name just a few (Rocktäschel et al., 2012). Chemical formulas are also used as entity names (e.g. *Al2(SO4)3*), some of which consist only of one letter. Apart from all these, partially highly ambiguous variants, authors introduce even more variants by using their own abbreviations, e.g. *NF-κB* instead of *NF-kappa-B inhibitor*. Furthermore, many standardized names, such as those from IUPAC, contain a lot of separating non-alphanumeric characters, such as slashes or commas, which can vary according to what the author prefers. On the other hand, if and where a name contains brackets, determines its meaning in many cases (Rocktäschel et al., 2012). For all these reasons, chemicals typically have a very large number of synonyms which can consist in completely different strings of letters, numbers and non-alphabetical characters, some of these as multi-word units. Because the synonyms have such different surface forms, usual normalization steps as for example fuzzy matching (checking for words similar to those in the dictionary), do not reach very good results for chemical entity names.

From all biomedical named entity types, dis-

ease names, as for example *HIV*, *Back Pain* or *Breast Cancer*, are possibly presenting the least difficulties to NER. In the past they have been shown to have less variability than named entities of genes and chemicals. This is mainly due to the reason that disease names are highly standardized throughout the literature (Jimeno-Yepes et al., 2008). One effect of this is that using a dictionary look-up on its own without further normalization can reach reasonable results as long as the dictionary is complete, containing all disease entity names of interest. Maybe even more than for chemicals and genes/proteins, it is very typical for disease names to consist in multi-word units.

In biomedical NER, names of organisms and species are usually treated together under the type “species”. Species names can also have a high level of ambiguity, often the same species name is used to refer to several different entities (*C. elegans* can be used to refer to up to 41 different species in the NCBI taxonomy (Gerner et al., 2010)). Similar to gene entity names, species entity names can contain common English words which are part of the name and they share a range of acronyms with other entity types, like genes (Gerner et al., 2010). All of this is prone to introducing a high rate of false positives. Another characteristic that species names share with gene and chemical names is the high variability, introduced through the usage by the authors and sometimes even by misspellings.

An overview of cell line nomenclature has been given by Sarntivijai et al. (2008). Cell lines names share the characteristics of other biomedical entity names: there is no obligation to use standardized names and authors are free to use whatever variants they like. One more reason for ambiguity of cell lines is the scientific experimental setting: cell lines mutate or become subject to contamination, which also bring along a change of concept (Sarntivijai et al., 2008).

4 Combining Terminology from Different Databases

With the aim of building a terminological resource for biomedical text mining, we decided to focus on a selection of commonly used databases, each of which contains terminological information about one or more of the most typical biomedical entity types. In this section, we first describe the struc-

¹<http://www.curioustaxonomy.net/gene/fly.html>

ture of the resulting combined terminological resource before we give details about the databases from which we take the terminological information, which files we use, how we process them and the general architecture that we apply.

4.1 Structure of the Combined Terminological Resource

The terminological resource which we compile using terminological data extracted from the selected databases is contained in one single file. This file has the very simple format of comma separated values (csv). The fieldnames, defining the contents of each of the six columns are the following: 'oid', 'resource', 'original_id', 'term', 'preferred_term', 'entity_type' (Table 1).

For each ID in each original database, an internal Base36 identifier 'oid' is generated. Base36 is a binary-to-text encoding scheme², using digits and all letters (in this case capital) of the English alphabet. A five digit sequence may e.g. encode over 60 million decimal values, while taking up only five bytes as opposed to eight. Synonyms are assigned the same oid as the main term. As such, the oid is not a unique identifier. Hence, the primary row key of the output file is a combination of the oid and the (synonymous) term.

The contents of the term field are matched in the text by the dictionary look up. The preferred term is the most standard term for a concept, which is preferred over other term variants.

Finally, the entity type field contains the type of entity, in this case normalized to the following entity types: gene/protein, chemical, disease, species and cell line.

By restricting the database to these fields, we focused on the most important information from the selected databases. The intention is to exclude any redundant information from the term resource and keep it as lightweight as possible by only focusing on information that is absolutely necessary for its application in a biomedical text mining system.

4.2 Included Resources by Entity Types

We decided to include terminology from the databases described in the current section, sorted by entity types. This selection is only made for illustrative purpose and to cover a sample of the

most commonly used databases. However, as mentioned before, the format of the terminological resource is flexible and allows for easy integration of terminology from further databases.

Genes and Proteins

NCBI Gene NCBI Gene (Brown et al., 2015) ("Entrez Gene") is the gene database of the National Center Biotechnology Information (NCBI)³. It contains gene data from a wide range of species. Entrez Gene uses its own unique gene identifiers to track gene records. NCBI provides a downloadable file⁴ which is updated on a daily basis. This file contains one gene identifier per line together with the gene symbol and synonyms, among other information.

UniProtKB/SwissProt The UniProt Knowledgebase (UniProtKB) of the Universal Protein Resource provides various functional information on proteins. Only the section "UniProtKB/SwissProt" is considered, because it is manually annotated and reviewed, whereas the (much larger) UniProtKB/TreMBL resource is automatically curated and not manually reviewed. UniProt uses a mnemonic identifier and one or more accession numbers. Both identifiers and accession numbers are considered in our work. The identifier has the quality of being a human-readable mnemonic code, unlike the accession number⁵. In order to reduce redundancy, only the first accession number is considered, although a separate mapping is maintained for accession numbers that refer to the same gene.

Diseases

MeSH diseases MeSH (Medical Subject Headings)⁶ is a controlled vocabulary maintained by the United States National Library of Medicine (NLM) and updated annually. It contains keywords used to manually annotate PubMed abstracts and NLM's book database with the aim of facilitating search by providing the subjects of a text. These so called subject headings are available in the format of descriptors which are hierarchically sorted. A connected tree number defines the position in the hierarchy and, at the same time,

³<http://www.ncbi.nlm.nih.gov/gene>

⁴ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/GENE_INFO/All_Data.gene_info.gz

⁵UniProt KB User Manual

⁶<https://www.nlm.nih.gov/mesh/>

²<https://en.wikipedia.org/wiki/Base36>

Table 1: Overview of the fields in the combined terminology.

oid	<i>The internal identifier created by our system</i>
resource	<i>The original database from which the terminology of the term was extracted</i>
original_id	<i>The original unique identifier from the database of origin</i>
term	<i>The term itself which is searched in the text during the process of a dictionary look-up</i>
preferred_term	<i>This is the standard term for a concept (if available in the database of origin)</i>
entity_type	<i>The entity type of the term (e.g. gene/protein, chemical, disease, species or cell line)</i>

provides information about the entity type described by a descriptor. The tree branches that we considered recursively with all their subbranches are those for chemicals and drugs (branch D), diseases (branch C) and organisms (branch B). Apart from MeSH descriptors we also included MeSH supplementary records. Supplementary records (SCRs) are updated weekly and contain additional terms which might occur in the literature as concept data (MeSH, 2015). SCRs exist mainly for chemical substances and rare diseases and are mapped to descriptors. This mapping defines their position in the hierarchical structure of MeSH according to which we determined their entity type. MeSH descriptors as well as supplementary records use their own unique identifiers by which information about specific database entries can be retrieved from the database.

Chemicals

MeSH Chemicals and Drugs The chemical and drug branch (branch D) of MeSH. Its structure and method applied are the same as described above for diseases.

ChEBI ontology ChEBI (Chemical Entities of Biological Interest)⁷ is a “freely available dictionary of molecular entities focused on ‘small’ chemical compounds”. ChEBI’s terminology follows IUPAC (International Union of Pure and Applied Chemistry) and NC-IUBMB (Nomenclature Committee of the International Union of Biochemistry and Molecular Biology) but establishes its own unique and stable identifiers. Data is not only manually curated by ChEBI curators but also integrated from different sources, such as IntEnz, KEGG and PDBeChem.

Organisms and Species

NCBI Taxonomy The NCBI Taxonomy is a “curated classification and nomenclature”⁸ of species referenced in other Entrez databases. At the time of writing this paper, the coverage of the taxonomy is 10% of all described species of life (NCBI, 2015). From the various files that constitute the NCBI taxonomy, only the mapping from taxonomy IDs to names, synonyms and other properties is processed when creating the resource. During preprocessing, the node file, describing parent-child relationships between nodes in the taxonomy, is used to filter all names that are not assigned to leaf nodes.

MeSH Organisms The organisms branch (branch C) of MeSH (description above).

Cell Lines

Cellosaurus Cellosaurus⁹ is a thesaurus of cell lines. It is described as a “controlled vocabulary of cell lines” that is used in biomedical research. The resource is freely available under the Creative Commons Attribution-No Derivs License 3.0.

4.3 Implementation

All components of the program were implemented in CPython 2.7.

Automatic downloading and preprocessing

All resources are downloaded automatically using a standalone downloader. The script queries a text file specifying source URLs and, if present, loads timestamps from previous downloads from a log file. For each FTP or HTTP URL specified, the script attempts to query the modification timestamp from the server. If it deviates from the timestamp recorded in the log file, the file is downloaded and, if compressed, extracted from its archive. Finally, the updated log file is written.

⁷<https://www.ebi.ac.uk/chebi/>

⁸<http://www.ncbi.nlm.nih.gov/taxonomy>

⁹<http://web.expasy.org/cellosaurus/>

For resources that are provided in a form not ready for automatic processing, a preprocessing step is triggered. In the Entrez Gene gene info file, the header row is reformatted to correspond to the tab-separated format and only columns relevant to term resource are kept. In the NCBI taxonomy names listing, only names for records with the taxonomic rank *species* are kept.

Parsing

For each resource, a parser was implemented, tailored to extract the identifiers and terms from each original database. For each term in the database, the parser produces a row hash with name-value pairs for each field. If a record specifies synonyms or multiple IDs, additional hashes are created for each synonymous term or additional ID. The main term is specified as the preferred term for each synonym.

In order to keep the memory footprint to a minimum, most parsers process resources in an iterative way, row by row. An exception is the MeSH resource, which depends on two separate, inter-linked files, hence, the entire resource is loaded into memory.

Finally, a resource builder iterates over all specified resource parser objects and writes each row hash to a row in the output TSV.

Uniprot and Cellosaurus parser Based on a parser described on *Mannheimia goes programming*¹⁰ Uniprot sequence entries and Cellosaurus cell lines are specified as lists of space-separated key value pairs, separated by delimiter lines. Keys may be mandatory or optional and can occur once or multiple times. Multiple occurrences of the species key are concatenated, as these constitute continuations of previous values, not additional values (see Uniprot knowledgebase). Relevant information (terms and accession numbers) are mapped to the row hashes mentioned above. Some keys may have multiple values, each separated by a semicolon. For the Uniprot resource, a copy of the row hash is created for each accession number indicated in a sequence entry.

NCBI Taxonomy parser The names file is provided in a pipe-separated TSV-like format. For each id, a list of terms is provided, each paired

with a name class, specifying the type of term. The entry with the name class “scientific term” is used as preferred name. All name classes are considered with the exception of “authority”, as these terms specify not only terms but authorship and publication dates. For obscure reasons, entries of the name class “synonym” also occasionally cite authorship. Regular expressions are used to detect and remove citations and additional information in parenthesis, as well as to strip double quotes from these entries. A small subset of entries specify a unique name. For these cases, additional row hashes are generated.

Entrez Gene gene info parser The gene info format is a standard TSV file. A preprocessor converts the non-standard headers and generates a reduced file containing only relevant columns (containing the gene ID, symbol and all synonyms). Synonyms are specified as pipe-separated sequences. For each synonym, an additional row hash is generated.

MeSH XML parser MeSH is provided in two separate files, a descriptor record set and an associated supplement record set. First, both files are parsed into memory. Only the tree structures *Organisms* [B], *Diseases* [C] and *Chemicals and Drugs* [D] are considered. Additionally, a look-up table is generated from the descriptor record set, mapping the ID of a descriptor record to IDs of all trees, in which it occurs. Each supplement is mapped to its descriptor record using the look-up table. Finally, for each descriptor and supplement record, a row hash is generated.

CHEBI OBO parser The CHEBI OBO parser wraps the OBO parser from the *Orange Bioinformatics* add-on for the open-source data mining utility *Orange*¹¹. A row hash is generated for each term and, if present, for each synonym. Placeholder synonyms (containing only periods as terms) are discarded.

Web interface

The combined resource is generated and accessed through a web interface¹². The interface is controlled by a Python Common Gateway Interface (CGI) script, which allows direct control of the creation pipeline. Visitors to the website

¹⁰<http://mannheimiagoesprogramming.blogspot.ch/2012/04/uniprot-keylist-file-parser-in-python.html>

¹¹<http://orange.birolab.si/>

¹²pub.cl.uzh.ch/purl/biodb

can select the desired resources through checkboxes. Additionally, the user may provide pattern/replacement pairs for changing how the resources and their entity types are labeled. The reason for this lies in the extensive labeling of certain resources. This is particularly true for MeSH, where labels like *mesh desc(Anatomy)*, *mesh desc(Chemicals and Drugs)* etc. give a detailed description of origin. By replacing the regular expression `mesh desc.*` with MeSH, for example, the number of resource labels can be reduced.

After submitting the request, the resource creation process is started in the background. As the creation process may take several minutes, the user is provided an individual link that allows downloading the resource file as soon as it is ready.

5 Term Resource Statistics

Graphs depicting frequency distributions for the terms in the term resource, for each entity type as well as for the whole term resource, can be found in the appendix¹³. Each graph shows the ambiguity of a term (how many IDs per each term), and the reverse property, i.e. how many terms are available for each ID. For example, while cell lines and organisms are mostly unambiguous (the vast majority of them shows a 1:1 correspondence between IDs and terms), chemicals show a much higher degree of ambiguity. Since common matching strategies in dictionary look-up are to lowercase the terms, or to strip them of non-alphanumeric characters with the aim of increasing recall, we also show the additional ambiguity generated by this process.

6 Related Work

Recently, the focus of research on biomedical named entity recognition has rather been on machine learning approaches than on dictionary based approaches, however, assigning unique database identifiers is often a necessity and is frequently used as a second step after the application of a machine learning system.

The task of retrieving database IDs for named entities is promoted by shared tasks, such as the BioCreative workshop which includes normalization sub-tasks by asking participants to provide

database identifiers for the named entities in the text. BioCreative II included a gene normalization task encouraging the development of systems which are able to assign an Entrez Gene identifier to genes found in PubMed articles (Morgan et al., 2008).

Thompson et al. (2011) have compiled a very large biomedical dictionary called BioLexicon which, similar to our work, brings together terms from different resources. But additionally, BioLexicon contains terms extracted from text as well as further linguistic information, such as grammatical information and semantic verb roles, all of which is located in a relational database.

Kaljurand et al. (2009) have also compiled a term resource from different database terminologies but focuses mainly on the identification of protein mentions. Furthermore, the current work considers a richer set of terminology, and presents more detailed statistics.

One system, that has successfully used a dictionary look up for gene and protein NER has been described by Hanisch et al. (2005). The lexical properties of the gene ontology (which we have not used in our work so far) have been explored by McCray et al. (2002).

Compared to these related resources, the combined term resource presented in this paper has several advantages. It is continuously updated with the newest version of all included databases, which keeps it up to date. A user can produce a customized file with selected databases, according to specific needs of a project. Apart from this, it can be easily extended with further resources, which we are planning to do in the future. Last but not least the format is very simple and therefore allows for easy integration into any kind of text mining system. Furthermore, also due to its simplicity, the format is more lightweight, taking up less memory than other formats commonly used for terminological databases.

7 Conclusion

In this paper we presented a new terminological resource for biomedical text mining. The resource is compiled of terminology extracted from a selection of biomedical databases. Its very simple format facilitates integration into any system for biomedical text mining in order to perform a dictionary look-up. The novelty of our approach

¹³The graphs were generated using *plotly* (<https://plot.ly/>)

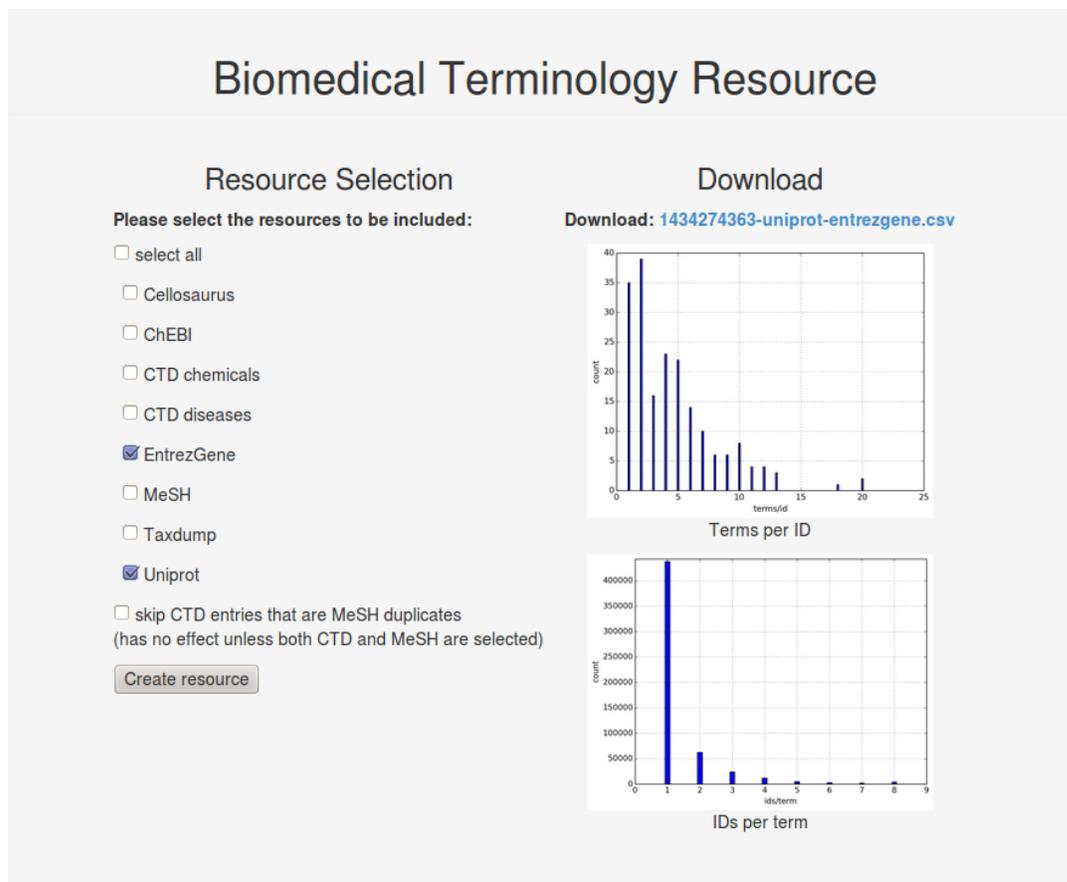


Figure 1: Web interface.

is that the resource is continuously updated with new terms found in the respective databases. Furthermore, a user interface provides a user with a choice of different databases and entity types so that an individual resource can be compiled. Various statistics for the compiled termfile give insight to the lexical properties of the terms contained in the resource. This sheds light on differences between the lexical properties of different entity types. Additional terminological resources will be included in future versions of the system.

References

- Garth R. Brown, Vichet Hem, Kenneth S. Katz, Michael Ovetsky, Craig Wallin, Olga Ermolaeva, Igor Tolstoy, Tatiana Tatusova, Kim D. Pruitt, Donna R. Maglott, and Terence D. Murphy. 2015. Gene: a gene-centered information resource at ncbi. *Nucleic Acids Research*, 43(D1):D36–D42.
- Martin Gerner, Goran Nenadic, and Casey M. Bergman. 2010. Linnaeus: A species name identification system for biomedical literature. *BMC Bioinformatics*, 11:85.
- Daniel Hanisch, Katrin Fundel, Heinz-Theodor Mevissen, Ralf Zimmer, and Juliane Fluck. 2005. Prominer: rule-based protein and gene entity recognition. *BMC Bioinformatics*, pages –1–1.
- Antonio Jimeno-Yepes, Ernesto Jiménez-Ruiz, Vivian Lee, Sylvain Gaudan, Rafael Berlanga Llavori, and Dietrich Rebholz-Schuhmann. 2008. Assessment of disease named entity recognition on a corpus of annotated sentences. *BMC Bioinformatics*, 9(S-3).
- Kaarel Kaljurand, Fabio Rinaldi, Thomas Kappeler, and Gerold Schneider. 2009. Using existing biomedical resources to detect and ground terms in biomedical literature. In Carlo Combi, Yuval Shahar, and Ameen Abu-Hanna, editors, *AIME*, volume 5651 of *Lecture Notes in Computer Science*, pages 225–234.
- Michael Krauthammer and Goran Nenadic. 2004. Term identification in the biomedical literature. *Journal of Biomedical Informatics*, 37(6):512 – 526. Named Entity Recognition in Biomedicine.
- Alexa T McCray, Allen C Browne, and Olivier Bodenreider. 2002. The lexical properties of the gene ontology. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, pages 504–508.
- MeSH. 2015. Introduction to MeSH – 2015.

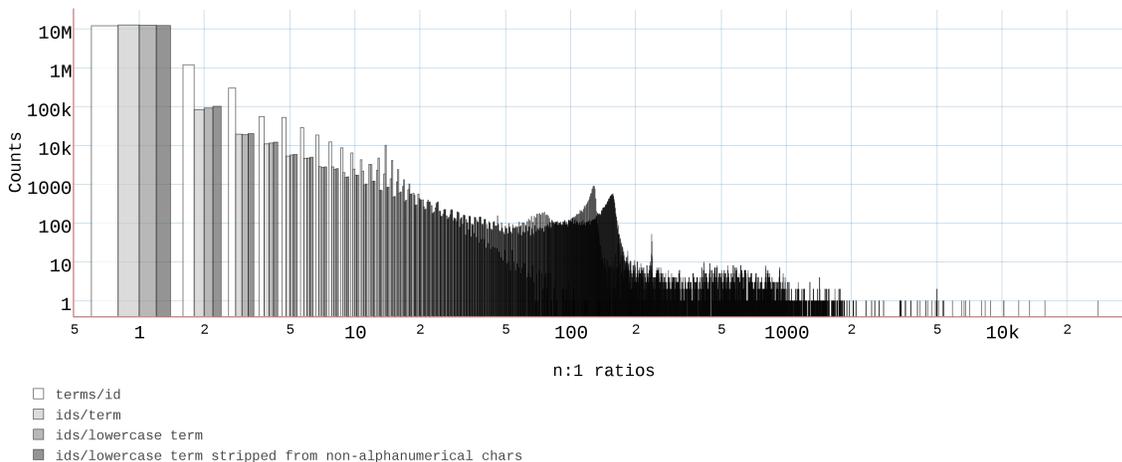
- <https://www.nlm.nih.gov/mesh/introduction.html>. Accessed: 2015-06-16.
- Alexander Morgan, Zhiyong Lu, Xinglong Wang, Aaron Cohen, Juliane Fluck, Patrick Ruch, Anna Divoli, Katrin Fundel, Robert Leaman, Jorg Hakenberg, Chengjie Sun, Heng-hui Liu, Rafael Torres, Michael Krauthammer, William Lau, Hongfang Liu, Chun-Nan Hsu, Martijn Schuemie, K Bretonnel Cohen, and Lynette Hirschman. 2008. Overview of biocreative ii gene normalization. *Genome Biology*, 9(Suppl 2):S3.
- NCBI. 2015. NCBI Taxonomy - Frontpage. <http://www.ncbi.nlm.nih.gov/taxonomy>. Accessed: 2015-07-02.
- Fabio Rinaldi, Allan Peter Davis, Christopher Southan, Simon Clematide, Tilia Renate Ellendorff, and Gerold Schneider. 2013. Odin: a customizable literature curation tool. In *Fourth BioCreative Challenge Evaluation Workshop*, volume 1, pages 219–223. Biocreative, October.
- Tim Rocktäschel, Michael Weidlich, and Ulf Leser. 2012. Chemsport: a hybrid system for chemical named entity recognition. *Bioinformatics*, 28(12):1633–1640.
- Sirarat Sarntivijai, Alexander S. Ade, Brian D. Athey, and David J. States. 2008. A bioinformatics analysis of the cell line nomenclature. *Bioinformatics*, 24(23):2760–2766.
- P. Thompson, J. McNaught, S. Montemagni, N. Calzolari, R. del Gratta, V. Lee, S. Marchi, M. Monachini, P. Pezik, V. Quochi, C. J. Rupp, Y. Sasaki, G. Venturi, D. Rebholz-Schuhmann, and S. Ananiadou. 2011. The biolexicon: a large-scale terminological resource for biomedical text mining. *BMC Bioinformatics*, 12:397.

A Termfile Statistics

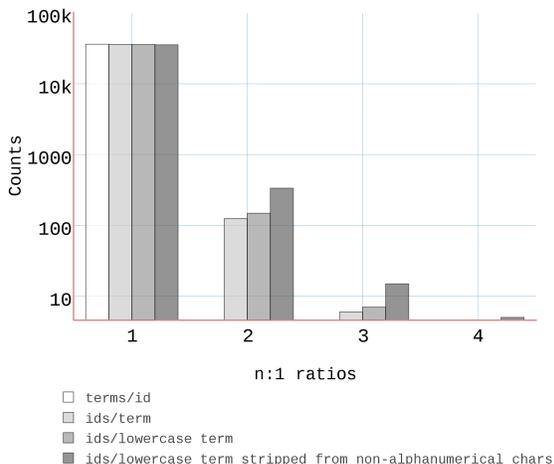
Table 2: Overview of termfile statistics.

	genes/proteins	chemicals	diseases	species	cell lines	all entity types (whole resource)
Number of Terms in Resource	10,429,162	979,418	67,614	1,333,903	36,249	12,846,346
Average of Term Length (number letters)	11.73	37.49	26.98	22.87	7.611	14.92
Average of terms per original IDs	1.1455	3.545	6.018	1.326	1.000	1.328
Average of original IDs per term	1.371	1.049	1.000	1.003	1.004	1.306
Average of original IDs per term (case insensitive)	1.383	1.062	1.000	1.003	1.004	1.317
Average of original IDs per term (case insensitive, non-alphanumeric characters removed)	1.387	1.086	1.000	1.006	1.010	1.324

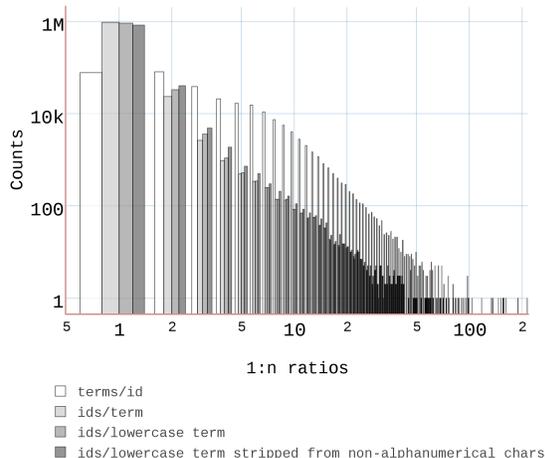
Total



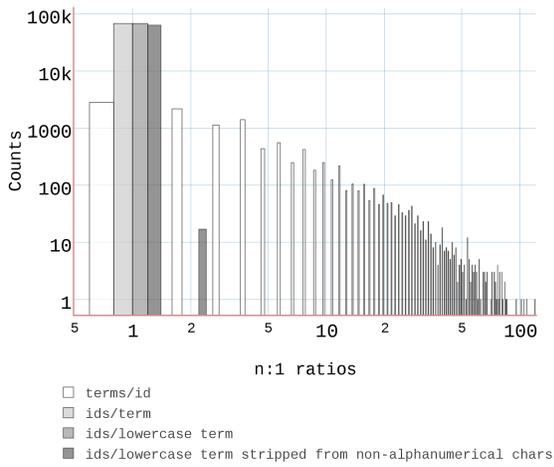
Cell lines



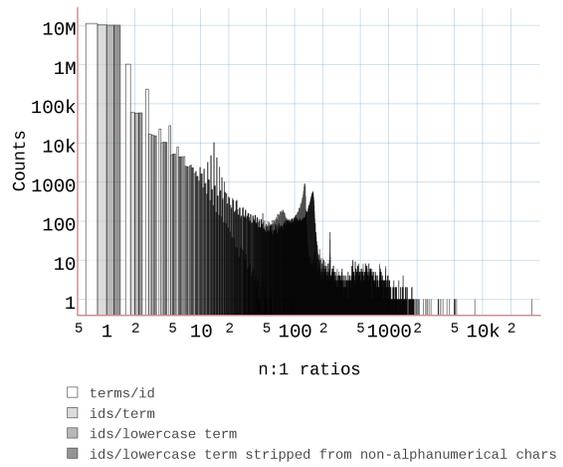
Chemicals



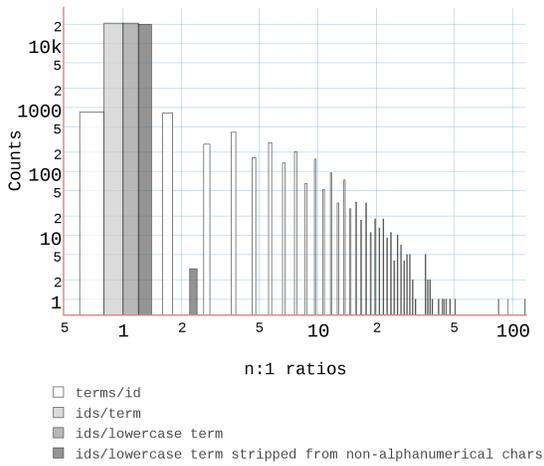
Diseases



Genes & Proteins



Organisms



Species

