Rancho Term Mapping Solution

(Fuzzy Tool)

Rancho BioSciences

November 2021

Problem Statement

In their trend-setting paper, "The FAIR Guiding Principles for scientific data management and stewardship", M. Wilkinson, *et al*., state:

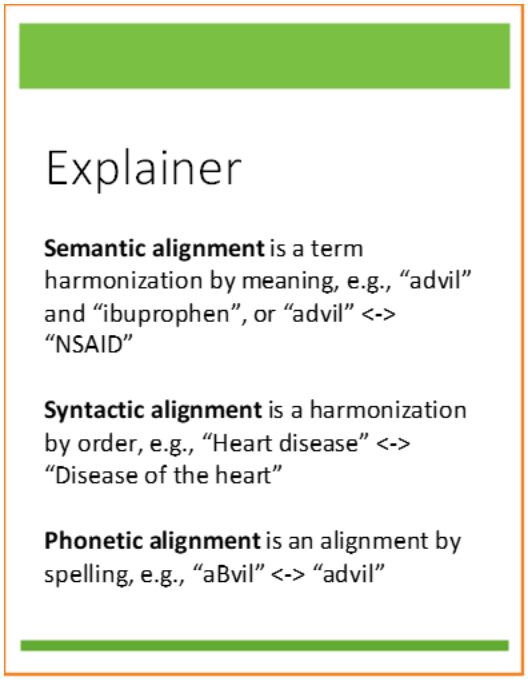
> *"Good data management is not a goal in itself, but rather is the key conduit leading to knowledge discovery and innovation, and to* ***subsequent data and knowledge integration and reuse by the community*** *after the data publication process."* [1]

For data to be findable, interoperable, and reusable, it first needs to be normalized (so that data from different sources can be aligned) and most importantly, it needs to be cleaned up, so it is free from original human and machine errors.

For both tasks, it is a standard practice to align data to well-established standard ontologies[2] and controlled vocabularies and to curate it, both manually and digitally. While there is no automated solution that can guarantee clean and well-aligned data, an efficient semi-automated solution can do all preliminary work, leaving curators with fewer, more complex cases.

One of the main goals of automated curation is to harmonize data semantically, syntactically, and phonetically, so it can be discovered and shared between studies and domains. *This paper is focused mostly on Rancho BioSciences' solution for rapid, practical data harmonization based on phonetic alignment.*

Before mapping the terms, one needs to select the best vocabulary source for alignment, preferably as domain specific as possible[3]. For example, if one considers a collection of chemical compounds captured from an Electronic Lab Notebook, one may want to select ChEMBL[4] as a vocabulary source. The selected source then needs to be pre-processed, loaded and pre-indexed. For the term mapping task, the goal is to choose the best (fastest and more precise) algorithm for calculating the distance between standard and misspelled terms, presenting mapping sorted by similarity score (**SML**).



[1] M Wilkinson, M Dumontier, I Aalbersberg, *et al*., The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3,** 160018 (2016). https://doi.org/10.1038/sdata.2016.18.

[2] C Tao, J Pathak, HR Solbrig, WQ Wei, CG Chute, Terminology representation guidelines for biomedical ontologies in the semantic web notations. *J Biomed Inform* **46**(1), 128-138 (2013). https://doi.org/10.1016/j.jbi.2012.09.003.

[3] J Malone, R Stevens, S Jupp, T Hancocks, H Parkinson, C Brooksbank, Ten Simple Rules for Selecting a Bio-ontology. *PLoS Comput Biol* **12**(2), e1004743 (2016). https://doi.org/10.1371/journal.pcbi.1004743.

[4] A Gaulton, A Hersey, M Nowotka, AP Bento, J Chambers, D Mendez, P Mutowo, F Atkinson, LJ Bellis, E Cibrián-Uhalte, M Davies, N Dedman, A Karlsson, MP Magariños, JP Overington, G Papadatos, I Smit, AR Leach, The ChEMBL database in 2017. *Nucleic Acids Res* **45**(D1), D945-D954 (2017). https://doi.org/10.1093/nar/gkw1074.

Solution

Prior Art: Existing algorithms for fuzzy string matching

The most used methods for fuzzy string matching[5] can be grouped in five main categories:

Soundex (metaphone, double-metaphone) – assigns same key to similar sounding English-based phonemes, thus accounting for possible misspellings.

List Method – lists all possible misspellings of a given term, and then finds best match between these terms and a given "dirty" term.

Edit Distance Method[6] **(Levenshtein, Jaro-Winkler, Jaccard**[7]**)** – calculates how many transformations it takes to get from the standard term to its "dirty" variant.

Statistical Similarity Method – trains the model to recognize similar terms based on a large training set of similar pairs.

Hybrid method(s)[8] – first uses a common key (like metaphone) method for high recall and then uses a statistical method to achieve high precision.

Rancho Solution and Methods

When building our term-mapping solution, we found that most of the existing methods are either too restrictive (**statistical methods**) or too permissive (**soundex method**). Furthermore, the better the method, the more expensive it is computationally. Our solution is two-pronged: first, instead of looking for all possible language misspellings, to select the best ontologies to map the data to, and second, to pre-index the standard ontology terms to accelerate the mapping process.

Our solution towards efficient term mapping has two components. First, instead of looking for all possible language misspellings we select the best ontologies, such as those described by at the Open Biology and Biomedical Ontology (OBO) Foundry[9], to which the data should be mapped to, and second, to accelerate the mapping process, we pre-index the standard ontology terms.

---

[5] DH Kraft, G Bordogna, G Pasi, An extended fuzzy linguistic approach to generalize boolean information retrieval, *Information Sciences-Applications*, **2**(3):119-134, (1994). https://doi.org/10.1016/1069-0115(94)90032-9.

[6] H Tissot, R Dobson, Combining string and phonetic similarity matching to identify misspelt names of drugs in medical records written in Portuguese. *Journal of Biomedical Semantics*, **10** (Suppl 1), Article 17 (2019). https://doi.org/10.1186/s13326-019-0216-2.

[7] G Navarro, A guided tour to approximate string matching, *ACM Computing Surveys (CSUR)*, **33**(1), 31-88 (2001). https://doi.org/10.1145/375360.375365.

[8] GV Bard, Spelling-error tolerant, order-independent pass-phrases via the damerau-levenshtein string-edit distance metric. *ACSW '07 Proceedings of the fifth Australasian symposium on ACSW frontiers*, **68**, 117-124 (2007). https://dl.acm.org/doi/10.5555/1274531.1274545.

[9] R Jackson, N Matentzoglu, JA Overton, R Vita, JP Balhoff, PL Buttigieg, S Carbon, M Courtot, AD Diehl, DM Dooley, WD Duncan, NL Harris, MA Haendel, SE Lewis, DA Natale, D Osumi-Sutherland, A Ruttenberg, LM Schriml, B Smith, CJ Stoeckert Jr., NA Vasilevsky, RL Walls, J Zheng, CJ Mungall, B Peters, OBO Foundry in 2021: operationalizing open data principles to evaluate ontologies, *Database*, **2021**, baab069 (2021), https://doi.org/10.1093/database/baab069.

To help with selecting the ontologies, we have developed an *ontology mapping suggester* tool (http://scigraphplus.rbsdb.net:8080/fuzzy/suggest/ - currently in beta). This tool allows users to paste their curated text or term list, and as its name may imply, it then suggests the best ontologies, simultaneously using Rancho, BioPortal (https://bioportal.bioontology.org/)[10], and EBI OLS suggester[11] tools.

Once the desired ontology or ontologies are selected by the curator, we encounter the first technical problem that needs to be solved – slow performance. As anyone who has used services such as BioPortal can attest, existing ontology stores are often exceptionally large and quite slow because they store hierarchies and relationships as well as additional information.

However, for mapping purposes, we only need just a few fields, namely *label*, *synonyms*, and *CURIEs (See Glossary)*.

Thus, we can simply extract this information and store it in a fast (indexed) database, while maintaining CURIEs as a back-link to the full ontology store. We achieve through the development of ETL (Extract, Transform, and Load) scripts that take standard ontology formats (e.g., OWL, OBO, TTL) as input and convert them into the CSV format for the ingestion into the Rancho ontology store.

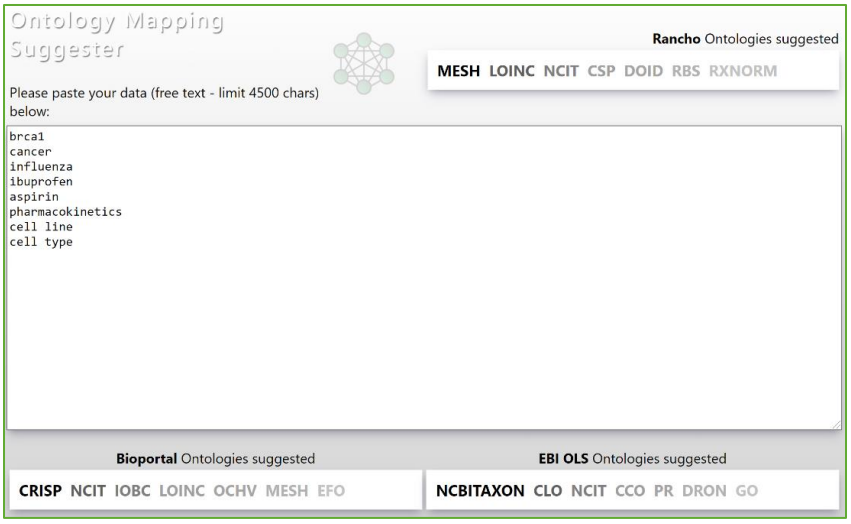


Figure 1: Sample Rancho Ontology Mapping Suggester output.

[10] PL Whetzel, NF Noy, NH Shah, PR Alexander, C Nyulas, T Tudorache, MA 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**(Web Server issue):W541-5. (2011). https://doi.org/10.1093/nar/gkr469.

[11] S Jupp, T Burdett, J Malone, C Leroy, M Pearce, J. McMurry, H Parkinson, A new Ontology Lookup Service at EMBL-EBI. In: Malone, J. et al. (eds.) *Proceedings of SWAT4LS International Conference* (2015). http://ceur-ws.org/Vol-1546/paper_29.pdf.

For the ontology store, we use **SciGraph** (https://github.com/SciGraph/SciGraph), an open-source tool, running on **Neo4j**. We have built several *Java* applications that interface with SciGraph to perform ontology-related tasks.

There are a few established mapping methods that allow indexing. The method we found most efficient is the **trigram method** implemented as a *PostgreSQL* extension (refer to e.g., https://www.postgresonline.com/journal/archives/169-Fuzzy-string-matching-with-Trigram-and-Trigraphs.html for detailed information and https://ii.nlm.nih.gov/MTI/Details/trigram.shtml for applying to **MeSH** subject headers).

While indexing with *trigrams* takes some time (about four hours[12] for our entire ontology store; it is comparable with or better than previously reported methods), the actual similarity score calculation is fast enough to use in live applications. This means that once background indexing has completed, our implementation allows curators to interact with an ontology with sub-second speed.

We align raw data to standard ontologies that may or may not be combined with customer-specific vocabularies. This approach allows us to apply domain knowledge across different datasets. In addition, upon mapping, standard ontologies enrich the source term with synonyms and abbreviations, as well as hierarchical and other types of relationships to other terms and concepts.

The resultant indexed **Postgres Trigram** database allows fuzzy term mapping yet preserves a link (via *CURIE*) to the ontology server that is used to perform hierarchical tasks, such as *common ancestor queries, term level alignment,* and *annotation services*.

label	iri	curie	synonyms	onto_name	sml
ibuprofen	http://purl.obolibrary.org/obo/CHEBI_5855	CHEBI_5855		EFO	1
ibuprofen	http://purl.bioontology.org/ontology/CSP/2270-2077	2270-2077		CRISP	1
ibuprofen	http://purl.obolibrary.org/obo/CHEBI_5855	CHEBI_5855	|2-[4-(2-methylpropyl)phenyl]propanoic acid|Ibuprofen|	CHEBI	1
Ibuprofen	http://purl.bioontology.org/ontology/MESH/D007052	C0887195	|Ibuprofen, (+-)-Isomer|Ibuprofen, Potassium Salt|Ibuprofen-Zinc|Ibuprofen, Sodium Salt|Ibuprofen, Copper (2+) Salt|Ibuprofen, Zinc Salt|Ibuprofen Zinc|Ibumetin|Benzeneacetic acid, alpha-methyl-4-(2-methylpropyl)-|Ibuprofen, Magnesium Salt|Ibuprofen, Calcium Salt|Motrin|Ibuprofen, (R)-Isomer|Ibuprofen, Aluminum Salt|Ibuprofen, (S)-Isomer|Trauma-Dolgit Gel|Rufen|Salprofen|Trauma Dolgit Gel|Nuprin|alpha-Methyl-4-(2-methylpropyl)benzeneacetic Acid|Brufen|	MESH	1
Ibuprofen	http://snomed.info/387207008	387207008	|p-Isobutylhydratropic acid|	SNOMEDCT	1
Ibuprofen	http://purl.bioontology.org/ontology/MESH/D007052	D007052	|IP 82|Motrin|Nuprin|Ibuprofen, (S)-Isomer|Aluminum Salt Ibuprofen|Rufen|Sodium Salt Ibuprofen|Ibuprofen, (+-)-Isomer|Benzeneacetic acid, alpha-methyl-4-(2-methylpropyl)-|Ibuprofen Zinc|Ibuprofen, Magnesium Salt|Ibuprofen, Aluminum Salt|Calcium Salt Ibuprofen|Ibuprofen, Sodium Salt|Ibuprofen, Copper (2+) Salt|Ibuprofen, (R)-Isomer|Ibuprofen, Zinc Salt|Potassium Salt Ibuprofen|IP82|Brufen|I.V. Solution, Ibuprofen|	MESHF	1
Ibuprofen	http://purl.bioontology.org/ontology/RXNORM/5640	5640		RXNORM	1
Ibuprofen	http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C561	C561		NCIT	1
ibuprofen(1-)	http://purl.obolibrary.org/obo/CHEBI_132922	CHEBI_132922	|2-[4-(2-methylpropyl)phenyl]propanoate|	CHEBI	0.833333
4-(5-thioxo-5H-1,2-dithiol-3-yl)-phenyl 2-(4-isobutyphenyl)propanoate	http://purl.bioontology.org/ontology/MESH/C571295	C571295	|HS-IBU|HS-ibuprofen|	MESHF	0.714286
4-(5-thioxo-5H-1,2-dithiol-3-yl)-phenyl 2-(4-isobutyphenyl)propanoate	http://purl.bioontology.org/ontology/MESH/C571295	C3491417	|HS-IBU|HS-ibuprofen|	MESH	0.714286
Ibuprofen Pill	http://purl.bioontology.org/ontology/RXNORM/1156278	1156278		RXNORM	0.666667
Ibuprofen 50 MG	http://purl.bioontology.org/ontology/RXNORM/316076	316076		RXNORM	0.625

Figure 2: Example of a Rancho Term Mapping Endpoint (similarity score SML is between 0 and 1, 1 being the exact match).

Our dual **Ontology-Database Solution** has several endpoints that allow us to feed term mapping results to upstream applications and bulk mapping scripts.

[12] CPU: Intel(R) Xeon(R) Gold 6132 CPU, Frequency: 2.60GHz, Processor count: 2, Core count (per processor): 14, Total core count: 28, RAM: 192 GB.

In addition to scripts and simple UI applications, we have developed a **Microsoft Excel plugin** that allows users to annotate their terms directly within Excel.

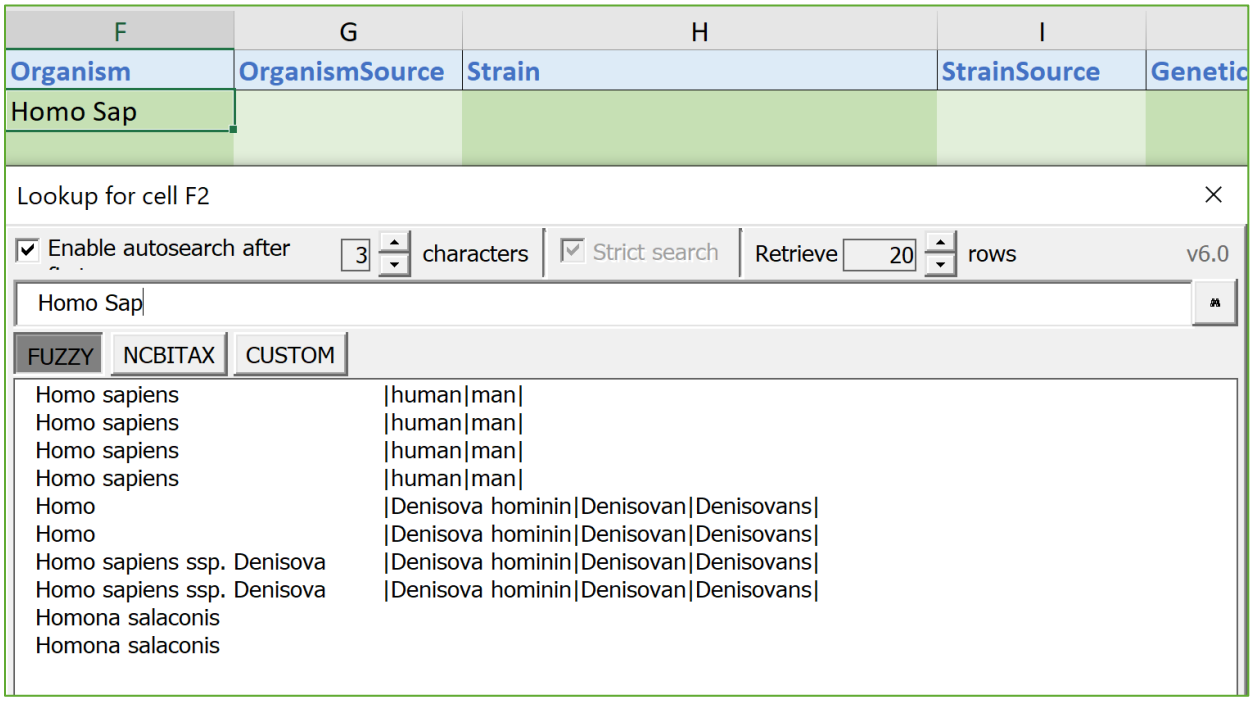


Figure 3: Example of Microsoft Excel Plugin Mapping Results.

For faster delivery and deployment, we containerize our solution, so we can quickly create custom versions with task-specific ontologies.

Discussion

We approach the term-mapping task with an understanding that real data harmonization cannot be fully automatic; instead, the question that must first be asked is whether it is more cost effective to manually curate the data. The answer to this question depends on two parameters: data size and mapping success rate. Based on our internal benchmarks for larger datasets, automatic curation is inefficient if its success rate is less than **65%**. Importantly, regardless of success rate, a rigorous manual QC is always required post-automation.

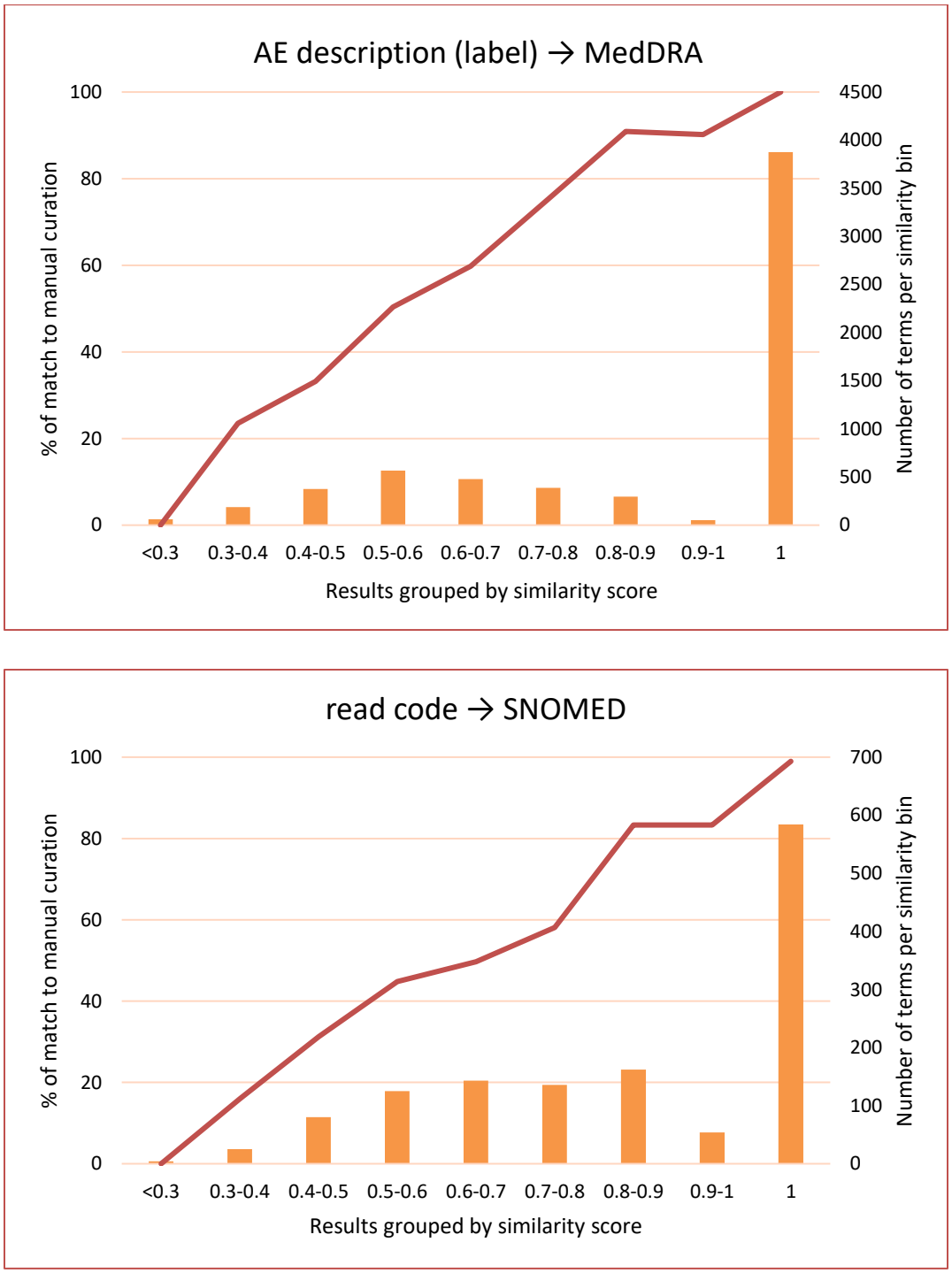


Figure 4: Success rate of fuzzy tool as a function of similarity score provided by the tool for the tasks of mapping adverse events description to MedDRA terminology (top diagram) and mapping between read code (medical coding system used in the UK) and SNOMED (bottom diagram). Green bars represent total number of terms with a specified similarity score.

One of the typical issues with large ontologies is that they represent ontology-specific knowledge, yet they also contain data from other ontologies. For example, the **EFO** (https://www.ebi.ac.uk/efo/) ontology contains many more terms pulled from other ontologies, such as **UO** (https://bioportal.bioontology.org/ontologies/UO), **ChEBI** (https://www.ebi.ac.uk/chebi/), and **Uberon** (https://uberon.github.io/), rather than its own original terms. This presents an interesting issue: if the solution mapped a term to a UO concept, is it different from the "full" UO ontology mapping? Furthermore, as most ontologies are not updated regularly, term mapping performed today may become partially obsolete in future years.

To solve these issues, Rancho is developing an *Ontology Versioning System* that will keep track of term provenance, perform regular ontology updates from sources of record, and keep track of the ontology versions.

Algorithm Improvements

One of the common issues with term mapping is that it tends to be biased towards word length. Thus, most mapping algorithms will give *"adil tablet"* ↔ *"aspirin tablet"* mapping higher score than the *"adil tablet"* ↔ *"advil"* mapping because the word *"tablet"* is longer than the word *"advil"*. To account for this, we implemented a ***Fuzzy+*** version of the solution that uses the **TF*IDF (term frequency * inverse document frequency, stemmed, stop words omitted)** algorithm that adjusts mapping scores according to the frequency of terms in the ontology. While this algorithm strongly improves the mapping quality, it comes at a cost, as in its current iteration, it is computationally quite expensive. It is thus only used for specific, large volume mapping tasks.

One potential area for further improvement in our Fuzzy Mapping tool is to migrate it from Excel to a stand-alone application, as the *Excel Plugin* tends to be slower than the direct mapping (e.g., using a custom script). In addition, the *Excel Plugin* needs to be supported to account for new and platform-specific versions of Excel. To solve that, we plan to develop a web curation UI that can interface quickly and directly with the **Rancho Fuzzy Mapping Tool**, while preserving the basic functionality of Excel.

Glossary

Controlled vocabulary is a list of terms which a community has agreed upon. For example: Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, and Sunday are the days of the week.

Concepts are the units of thought —ideas, meanings, or (categories of) objects and events. Concepts exist in the mind as abstract entities which are independent of the terms used to label them. Concepts are identified by URIs. To group concepts, we use *concept schemes* and *collections.*

CURIEs are Compact URIs of the form [resource:id], e.g. [wikipedia.en:leonya].

IRI is an Internationalized Resource Identifier. This is an internationalized extension of the URL.

Ontology describes what types of things exist in the domain and how they are related. A vocabulary is composed of terms with clear definitions controlled by some internal or external authority. For example, the ontology triple *ex:dog skos:broader ex:mammal* states that dog is part of the broader concept mammal.

Taxonomy is a controlled vocabulary organized in a hierarchy. For example, we can have the terms Computer, Tablet and Laptop and the concepts Tablet and Laptop are subclasses of Computer because a Tablet and Laptop are types of Computers.

Thesaurus is a taxonomy with information about each concept including preferred and alternative terms ("Computer" in English, "Computador" or "Ordenador" in Spanish). Thesaurus may contain relationships to related concepts. For example, the concepts "Computer" and "Software" have some type of relationship.

URI is a Uniform Resource Identifier. URIs include URL (locator), URN (name, e.g., doi) etc., in the format:

scheme:[//authority]path[?query][#fragment]