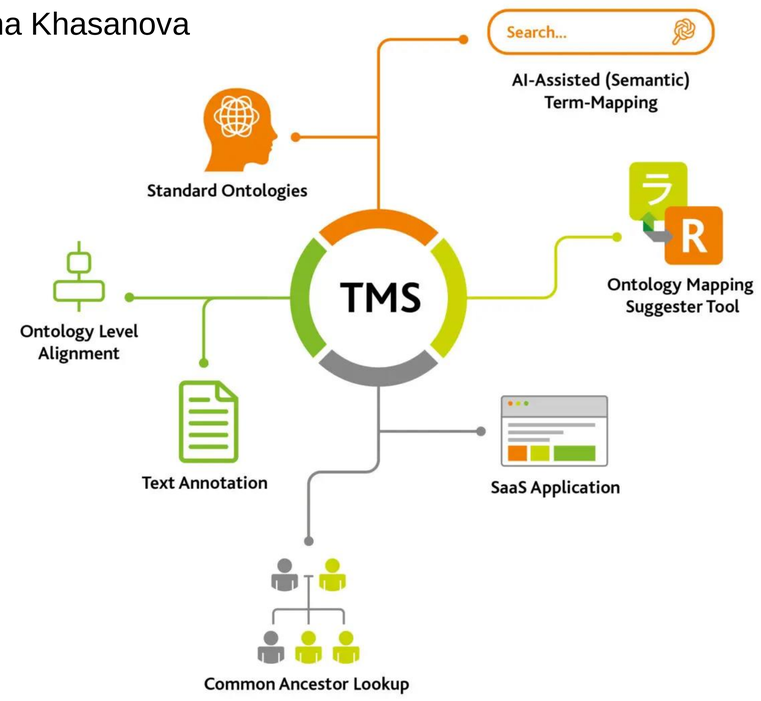
Integrative Approach to Data Harmonization: Empowering Biomedical Research with Rancho Biosciences' Terminology Management Solution (TMS)

Brad Farrell, Hillary Mosso, Alena Fedarovich, Tracy Ballinger, Rob Beetel, Leonya Ivanov, Vishnu Govindaraj, Tatiana Khasanova
Rancho Biosciences, LLC



https://ranchobiosciences.com/rancho-products/#tms

Abstract

In a collaborative effort to enhance data curation in biomedical research, Rancho Biosciences undertook two multifaceted projects involving the integration of diverse knowledge bases and the optimization of curation workflows through advanced AI methodologies. This initiative aimed at addressing the interoperability challenges between various proprietary and standardized ontologies, enhancing the predictive and translational safety assessments in biopharmaceutical research.

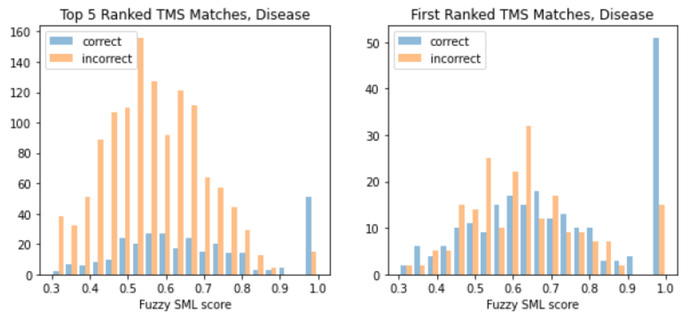
The fundamental component in both projects was the development and deployment of Rancho Biosciences' Terminology Management Solution (TMS), an in-house tool crafted from extensive curation experience. TMS was rigorously evaluated against existing commercial tools across essential tasks such as term harmonization, ontology mapping, and data extraction from unstructured sources. The evaluation focused on not just accuracy and efficiency, but also usability, support, and customization, to provide a comprehensive assessment of each tool's potential in streamlining the curation process.

Additionally, we explored the feasibility of AI-enhanced curation workflows, identifying critical pinch points in the process where AI could significantly boost efficiency. This endeavor encompassed the harmonization of terms within fields across various studies, the mapping of these terms to established ontologies, and the intricate task of aligning terms in raw data with attributes in structured data models.

The results demonstrated the robust capabilities of TMS, particularly its superiority in precision and recall compared to other evaluated tools. TMS not only excelled in accurately mapping a vast array of terms to respective ontologies but also displayed a potential for substantial time-saving in manual curation processes. These outcomes highlight the TMS's role as a pivotal asset in biomedical data curation, promising a significant leap forward in the accuracy and efficiency of data harmonization efforts. The project's success lays a solid foundation for future enhancements and positions Rancho Biosciences at the forefront of innovation in biomedical data management.

Rancho TMS: Mapping Algorithm - Fuzzy

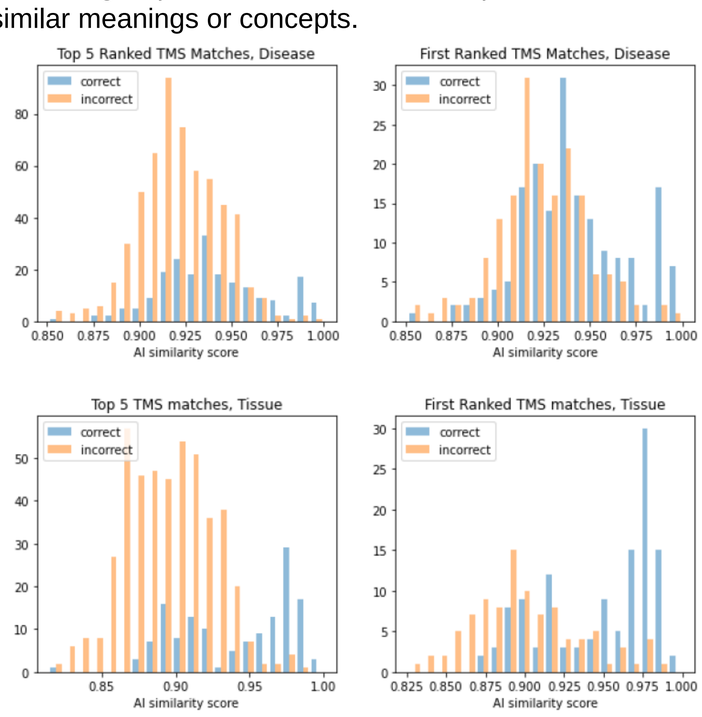
Fuzzy or phonetic mapping refers to the technique of associating words or terms based on their sound or approximate spelling, enabling the identification of similar-sounding words across different datasets or search queries, even in the presence of minor errors or variations in spelling. Fuzzy algorithm allows fast phonetic mapping with similarity score outputs.



Similarity (SML) scores are shown for correct and incorrect matches of disease and tissue terms, for all top 5 ranking matches on the left and for only the best matches on the right.

Rancho TMS: Mapping Algorithm - AI

Semantic (using AI) term mapping involves the process of linking and translating terms between different vocabularies or databases, facilitating the understanding and integration of diverse data sources by establishing equivalences or relationships between terms that have similar meanings or concepts.



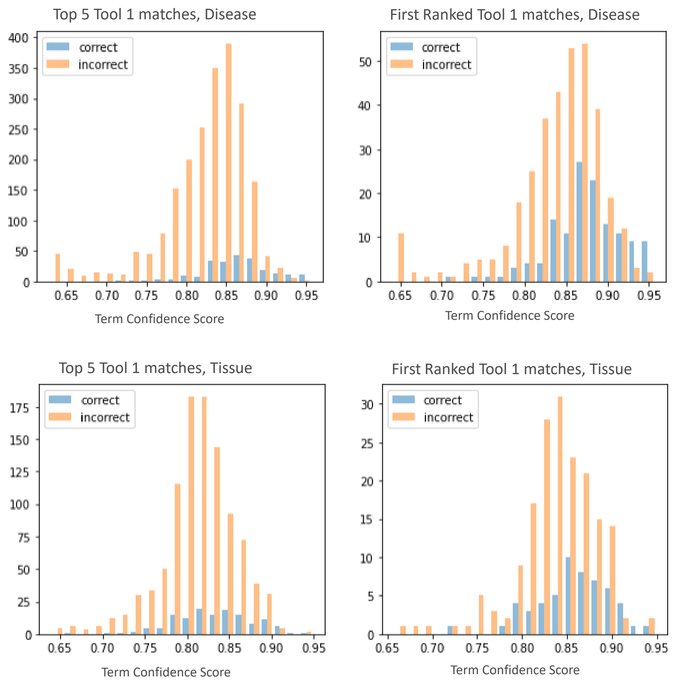
AI cosine similarity scores are shown for correct and incorrect matches of disease and tissue terms, for all top 5 ranking matches on the left and for only the best matches on the right.
The cosine similarity score shows a better separation between correct and incorrect mappings than TMS Fuzzy SML score or commercial Tool 1 confidence scores.

How does TMS Compare

vs. Commercial Tool 1

Both TMS algorithms (Fuzzy and AI) were compared to a commercially available ontology term mapping tool, referred to as Tool 1.

A



B

Algorithm	Tool 1	TMS Fuzzy	TMS AI
Disease Rank 1	132 (38%)	219 (46%)	179 (38%)
Disease Top 5	225 (47%)	296 (62%)	207 (43%)
Tissue Rank 1	55 (24%)	137 (59%)	123 (53%)
Tissue Top 5	134 (58%)	171 (74%)	142 (61%)

(A) Tool 1 confidence scores for correct and incorrect matches of disease and tissue terms for all top 5 ranking matches on the left and for only the best matches on the right.

(B) Accuracy of Rancho TMS with two scoring algorithms shown. Each tool was tested against a set of 476 dirty disease terms, which were mapped to DOID, and 232 dirty tissue terms, which were mapped to Uberon. The top five ranked results were returned for each tool, and the number of correct matches were counted both within the top match and within the top five matches. TMS using Fuzzy scoring performs the best, but TMS AI methods also do well, with up to 74% and 61% respectively, of terms correctly mapped.

vs. Commercial Tool 2

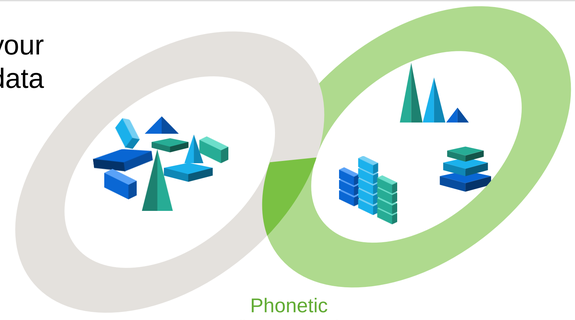
TMS Fuzzy was also compared with another automated mapping tool, called Tool 2 here. We ran 770 terms against DOID using both tools. TMS performed slightly better on the recall. Although precision was the same for both tools.

Automatic matching results review	Tool 2	TMS Fuzzy
True positive (TP)	491	513
False positive (FP)	219	234
False negative (FN)	42	5
Not matched terms by either algorithm	18	18
All	770	770
Recall (TP/(TP+FN))	0.92	0.99
Precision (TP/(TP + FP))	0.69	0.69

Why Use TMS

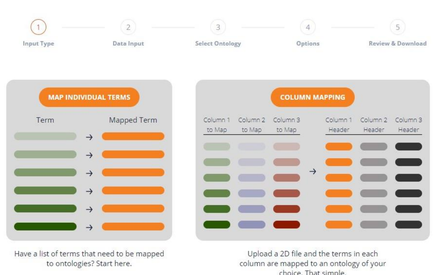
Rancho's TMS empowers you to align term lists or datasets with your preferred terminology standard, ensuring a streamlined and consistent data ecosystem.

- Can use public and custom standards or ontologies.
- Simplifies the alignment process, putting the power of precise data representation at your fingertips
- Accessible via a **user-friendly interface** or powerful RESTful APIs

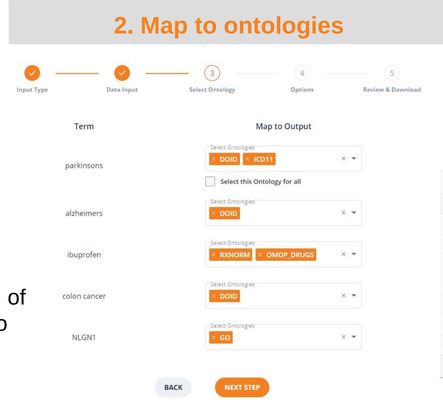


Phonetic (by spelling)
Raw Terms Across Different Domains → Mapping → Machine-readable Standard Terms
Semantic (AI-assisted)

1. Select mapping type

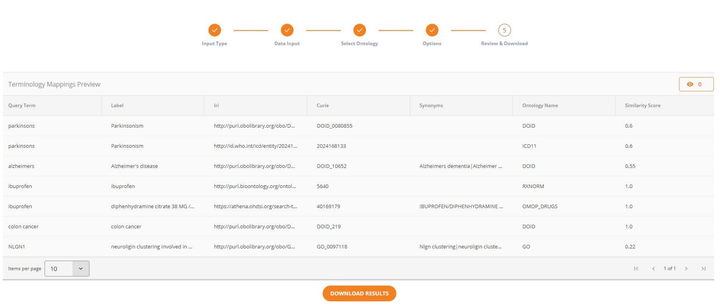


2. Map to ontologies



2-dimensional mapping feature allows upload of a 2d csv file and map terms in each column to ontologies of their choice and is for harmonization of sample-level metadata.

3. Preview and download results (csv)



Rancho biosciences