

**MARS TARGET ENCYCLOPEDIA: INFORMATION EXTRACTION FOR PLANETARY SCIENCE.** K. L. Wagstaff<sup>1</sup>, R. Francis<sup>1</sup>, T. Gowda<sup>1</sup>, Y. Lu<sup>1</sup>, E. Riloff<sup>2</sup>, and K. Singh<sup>1</sup>, <sup>1</sup>Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA 91109, [kiri.wagstaff@jpl.nasa.gov](mailto:kiri.wagstaff@jpl.nasa.gov), <sup>2</sup>University of Utah, Salt Lake City, UT.

**Introduction:** We created a new reference database called the Mars Target Encyclopedia (MTE) that contains compositional information about science targets (usually rocks or soils) on Mars. Users can search for all targets that contain an element (e.g., “calcium”) or mineral (e.g., “hematite”) and see a map view of their spatial locations (see Fig. 1). Clicking on a search result or searching for a specific target of interest (e.g., “Dillinger”) brings up a page that compiles previous findings about its composition (see Fig. 2).

The information that populates the MTE was mined automatically from the planetary science literature using state-of-the-art advances in information extraction. All information in the MTE is linked to the source publication from which it was extracted, so users can easily browse the full context in the original document.

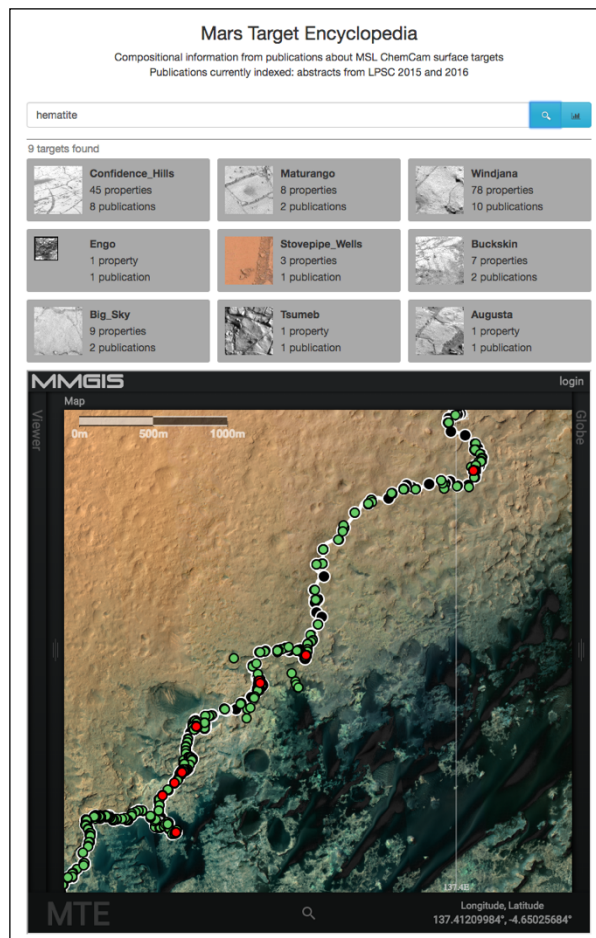


Figure 1. MTE search results for “hematite.”

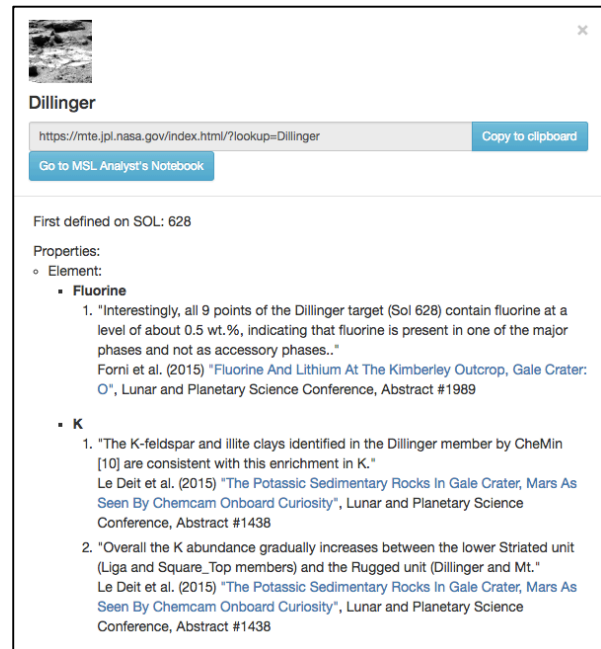


Figure 2. MTE entry for the Dillinger target, with citations.

**Surface Targets on Mars:** Mars rover missions identify new observational targets on a daily basis. Each such rock, soil, or point of interest is given a unique name, often derived from Earth locations (e.g., “Ithaca”, “Staten Island”), Earth people (e.g., “John Klein”), or whimsy (e.g., “Frodo”). The Mars Science Laboratory (MSL) rover has identified more than 7,000 targets in 4.5 years. There are hundreds of publications about these targets.

**Information Extraction methods:** Information extraction (IE) methods have been employed to extract diverse information such as terrorist events in news articles or protein interactions in biomedical documents. We trained an IE system to recognize “named entities” such as elements, minerals, and targets and then identify compositional relations between targets and elements or minerals [1] (see Fig. 3).

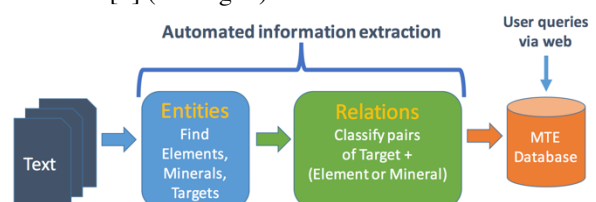


Figure 3. Information extraction process for the MTE.

We trained and evaluated the MTE using two-page abstracts from the Lunar and Planetary Science Conference. First, we extracted text from the PDF abstracts using Tika [2] and stripped out the “References” section in each document to avoid spurious detections (author initials are easily mistaken for element abbreviations). Next, we used the brat tool (<http://brat.nlplab.org/>) to hand-label entities and relations in each document that mentioned the MSL ChemCam instrument from LPSC 2015 (n=63) and 2016 (n=55). We trained the system on the hand-labeled documents from LPSC 2015 plus an additional 1069 documents from LPSC 2014 and 2015 that were automatically annotated using lists of known elements, minerals, and targets and then manually reviewed and corrected as necessary. We evaluated the system on 35 hand-labeled documents from LPSC 2016 (the remaining 20 documents were used for development).

**Named entity recognition (NER).** We created a custom named entity recognizer using known lists of elements, minerals, and targets. We compared the list-based NER system to a machine learning approach that used the Stanford CoreNLP system [3] to train a classifier to recognize elements, minerals, and targets. The CoreNLP NER classifier uses local context, entity type frequency, spelling, and “word shape” (patterns of vowels and consonants) to identify the class of each word (entity). Performance was high overall (nearly 0.90; see Fig. 4). Both methods performed about the same for the Element class, but the list-based method performed better than the CoreNLP NER for Minerals and Targets.

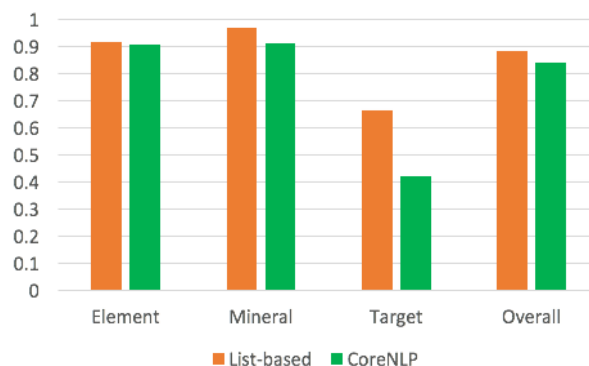


Figure 4. NER performance on LPSC 2016 documents.

**Relation extraction.** We used the jSRE package [4] to train a classifier to decide whether a compositional relation exists for a candidate (Target, Component) pair in the text (e.g., “Target Epworth contains calcium”). A Component is any Element or Mineral. The classifier uses a “bag-of-words” representation (i.e., ignores the order of words) of the sentence containing the target and

component and knowledge about the position of the target and component within the sentence. We compiled training and test sets consisting of all candidate (Target, Component) pairs that were automatically extracted by the NER system. The number of candidates for each Component type are shown in Table 1.

Corpus	Element	Mineral
LPSC 2015 (train)	273	151
LPSC 2016 (test)	34	9

Table 1. Number of candidates for relation extraction.

Performance on relation extraction is shown in Fig. 5. We compared the trained classifier to a simple baseline method that classifies all Target-Component pairs as showing a compositional relationship (“All-yes”). This baseline performs quite well, indicating that whenever the system finds a Target-Component pair within a sentence, there is a high probability that they are in a compositional relationship. However, using machine learning to refine this decision process (“Classifier”) improved performance for the Target-Mineral pairs.

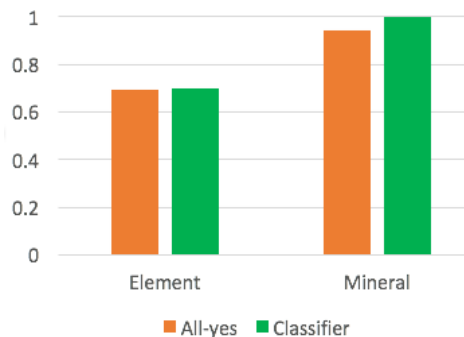


Figure 5. Relation extraction performance on LPSC 2016.

**Future Work:** We plan to extend the MTE to encompass longer, peer-reviewed journal articles. We will also experiment with ways to identify relations that cross sentence boundaries, which requires a deeper processing of the document to resolve pronouns and other ambiguous terms and connect them with specific targets and components.

**Acknowledgments:** This work was funded by the AMMOS program and the PDS and was carried out in part at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. Government sponsorship acknowledged.

**References:** [1] Wagstaff, K. L. et al. (2015) *AAAI Workshop on Knowledge Extraction from Text*. [2] Mattmann, C. & Zitting, J. (2011) *Tika in Action*. [3] Finkel, J. R. et al. (2005) *ACL*, 363-370. [4] Giuliano, C. et al. (2006) *EACL*.