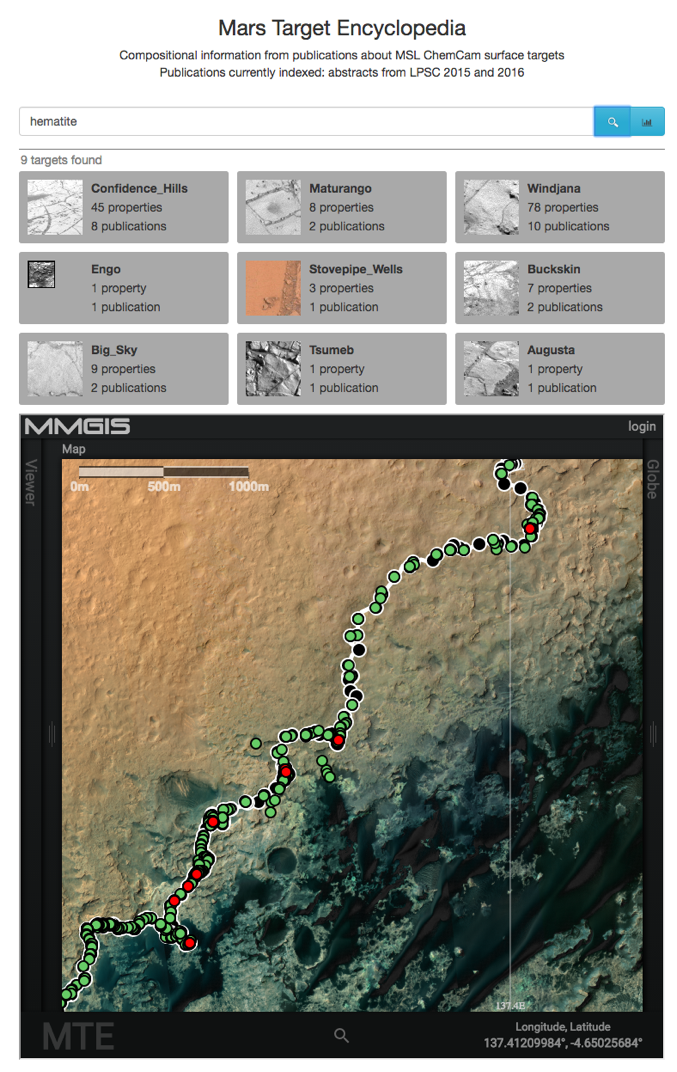
**MARS TARGET ENCYLOPEDIA: INFORMATION EXTRACTION FOR PLANETARY SCIENCE.** K. L. Wagstaff1, R. Francis1, T. Gowda1, Y. Lu1, E. Riloff2, and K. Singh1, 1Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA 91109, kiri.wagstaff@jpl.nasa.gov, 2University of Utah, 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). They can also look up information about a specific target of interest (e.g., “Dillinger”) to browse previous findings about its composition.

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.

**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., “Frood”). 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 (see Fig. 2).

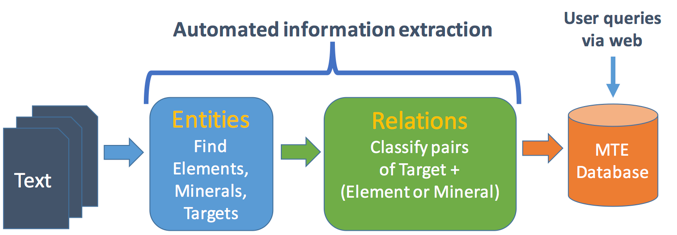


Figure . Information extraction for the MTE.

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 [] 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 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).

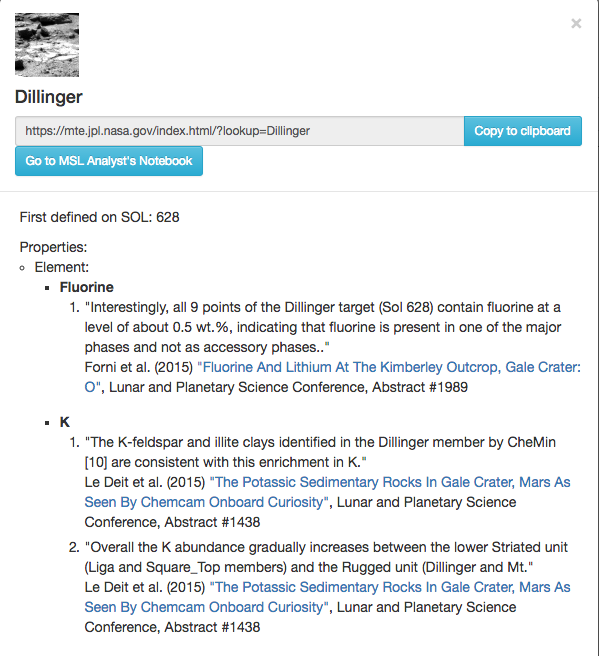
*Named entity recognition.* We used the Stanford CoreNLP system [] to train a classifier to recognize elements, minerals, and targets. The classifier uses local context, entity type frequency, spelling, and “word shape” (patterns of vowels and consonants) to identify the class of each word (entity). Performance was excellent for the Element and Mineral classes but lower for the Target class.

Figure . MTE search results for "hematite."

*[insert performance plot here]*

*Relation extraction.* We used the jSRE package [] to train a classifier to decide whether a compositional relation existed for each candidate (target, component) pair in the text. 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.

*[insert performance plot here]*

**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] CoreNLP. [2] Giuliano et al. (2006). [3] Wagstaff, K. L. et al. (2015) *AAAI Workshop on Knowledge Extraction from Text*. [TBD]