MARS TARGET ENCYCLOPEDIA: INFORMATION EXTRACTION FOR PLANETARY SCIENCE. K. L. Wagstaff[1], R. Francis[1], T. Gowda[1], Y. Lu[1], E. Riloff[2], and K. Singh[1], [1]Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA 91109, kiri.wagstaff@jpl.nasa.gov, [2]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 surface science targets (such as rocks or soils) on Mars. Users can search for all targets that contain a given element (e.g., "calcium") or mineral (e.g., "hematite") and see a map 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 publications about its composition (see Fig. 2).

The information in the MTE was mined from the planetary science literature using information extraction technology. Rather than analyzing instrument data, we analyzed publications about findings. All MTE entries link to source publications, so users can easily browse the full context in the original document (Fig. 2).

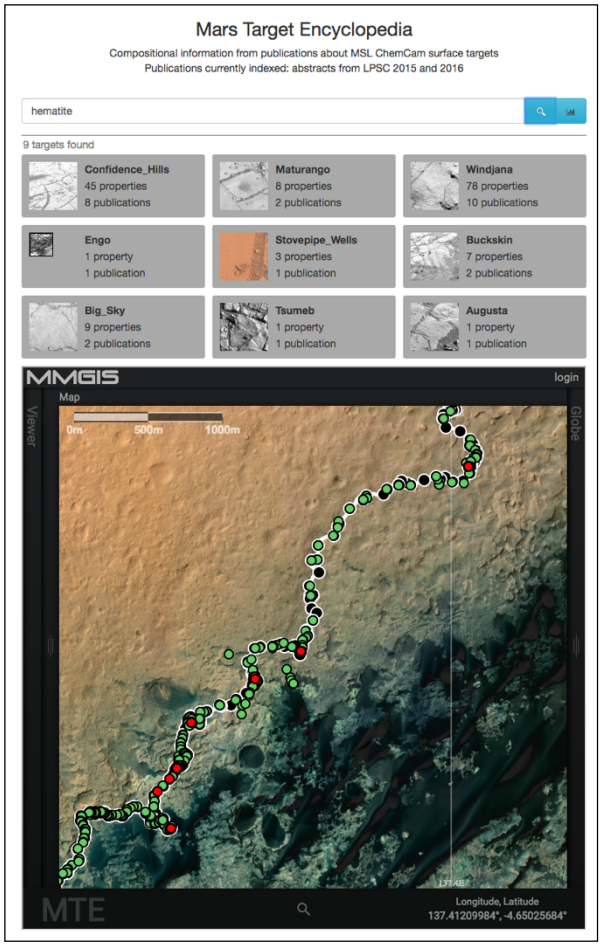

Figure 1. MTE search results for "hematite."

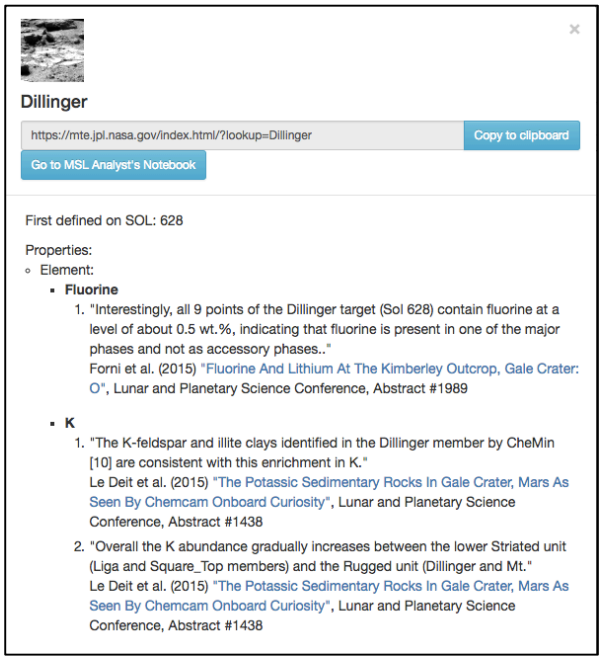

Figure 2. Page 1 of the MTE entry for the Dillinger target. Each component (e.g., fluorine) has a source citation.

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, and staying up-to-date is difficult.

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).

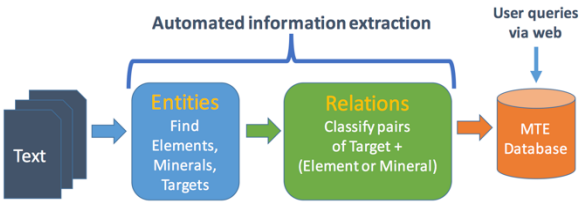

Figure 3. Information extraction process for the MTE.

We trained and evaluated the system 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 all documents that mentioned the MSL ChemCam instrument from LPSC 2015 (n=63) and 2016 (n=55). We trained the system on the LPSC 2015 hand-labeled documents 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/corrected. We evaluated the system on 35 hand-labeled documents from LPSC 2016 (the remaining 20 documents were used for development only).

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 uppercase/lowercase letters and digits) to identify the class of each word (entity). Performance (F-measure) was high overall (nearly 0.90; see Figure 4). Both methods performed about the same for the Element and Mineral classes, but the list-based method performed better than the CoreNLP NER for Targets. However, the CoreNLP model learned several new terms that were not in the training documents nor on the lists (e.g., "aluminium" and targets such as "Buckskin" and "Hoanib").

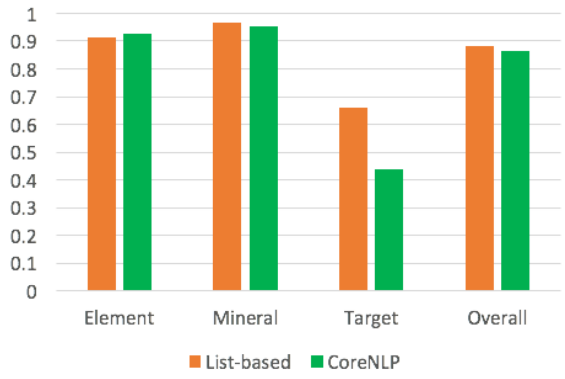

Figure 4. NER performance on LPSC 2016 documents.

Relation extraction. We used the jSRE package [4] to train a classifier to decide whether a compositional relation exists for a given (Target, Component) pair in the text (e.g., ("Epworth", "calcium") from "Target Epworth contains calcium" -> yes). A Component is any Element or Mineral. The classifier uses a "bag-of-words" representation (i.e., ignores the order of words) of the full sentence 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 numbers of candidates for each Component type are shown in Table 1.

Table 1. Number of candidates for relation extraction.

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

Relation extraction performance (F-measure) is shown in Figure 5. We compared the trained classifier to a simple baseline method that classifies all Target-Component pairs as having a compositional relation ("All-yes"). This baseline performs quite well: 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*.