

Creating a Mars Target Encyclopedia by Extracting Information from the Planetary Science Literature

Kiri L. Wagstaff

Jet Propulsion Laboratory
California Institute of Technology
4800 Oak Grove Drive, Pasadena, CA 91109
kiri.l.wagstaff@jpl.nasa.gov

Ellen Riloff

School of Computing
University of Utah
Salt Lake City, UT 84112
riloff@cs.utah.edu

Nina L. Lanza

Los Alamos National Laboratory
Los Alamos, NM 87545
nlanza@lanl.gov

Chris A. Mattmann

Jet Propulsion Laboratory
California Institute of Technology
4800 Oak Grove Drive, Pasadena, CA 91109
chris.a.mattmann@jpl.nasa.gov

Paul M. Ramirez

Jet Propulsion Laboratory
California Institute of Technology
4800 Oak Grove Drive, Pasadena, CA 91109
paul.m.ramirez@jpl.nasa.gov

Abstract


Staying up to date with the latest discoveries is a challenge in any scientific field. In planetary science, new observation targets on the surface of Mars are identified and named every day, and new publications announcing new discoveries and conclusions provide frequent updates about these targets. We are constructing a system that uses information extraction and retrieval methods to mine the steadily growing body of planetary science publications about Mars surface targets and automatically construct a concise summary of what is known about each target. The Mars Target Encyclopedia will provide a central, continually updated resource for use by planetary scientists and the interested public. We describe our use of Tika, Sundance, and AutoSlog to extract and summarize information, some of the challenges associated with this domain, and our plans for maturing the system.

1 Introduction and Motivation

The rovers that have been sent to Mars have been extraordinarily active and productive. The Mars Science Laboratory rover has generated > 3500 observation targets in three years, and the Mars Exploration Rover Opportunity has generated even more over its 11+-year mission. There are hundreds of associated scientific publications reporting new discoveries. The downside of this productivity is that as the number of data and publications grow, it becomes nearly impossible for any single person to read, understand, organize, and recollect the amount of information available.


We focus on a specific challenge, which is that of staying up-to-date with everything known (published) about identified surface targets on Mars. Each time an instrument (camera, spectrometer, laser, etc.) is aimed at a target (soil, rock, formation, etc.), that target is given a unique name. Mission planners and planetary science researchers must be aware of an increasing number of names, locations, and facts so

Copyright © 2016, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.



Epworth

First observed: October 19, 2012
MSL Sol 72, Site: 5
Soil target



Observations

- Sol 72 - ChemCam, Remote Micro-Imager

Properties

- high Ca phases were observed [Clegg et al., 2013]
- contains F [Forni et al., 2014a; 2014b; Forni et al., 2015]

Publications

- ▼ Clegg et al. (2013), "High calcium phase observations at Rocknest with ChemCam," LPSC "Specifically, these high calcium phases were observed on Pearson (sol 60), Epworth (a soil on sol 72) and Rocknest_6a (sol 87)."
- ▼ Forni et al. (2014a), "First fluorine detection on Mars with ChemCam on-board MSL," LPSC "Epworth-5 contains a relative large amount of fluorine since the main atomic lines of F I are also observed."

Other targets mentioned: Crest, Goulburn, Link, Measles Point, Pearson

- Forni et al. (2014b), "First fluorine and chlorine detection with ChemCam on-board MSL," 8th International Conference on Mars
- Forni et al. (2015), "First detection of fluorine on Mars: Implications for Gale crater's chemistry," Geophysical Research Letters

More like this

- Epworth_3
- Black Rock (CaF₂)

Figure 1: Example hand-constructed MTE entry for target Epworth. Users can connect to the original observations, properties that were extracted from scientific publications, and the publications themselves. Each publication provides excerpts with relevant context and a list of other targets mentioned.

that new observations can be appropriately interpreted in the context of what is already known. For example, one might ask: Does the observation of high manganese content at a particular location represent a confirmation of an existing trend or an anomalous new discovery?

Current text search tools cannot meet the knowledge needs of planetary scientists and mission planners. Mars surface target names have no naming convention, and names are often borrowed from Earth locations (e.g., "Cumberland," "Ithaca"), people (e.g., "Jake", "Darwin"), or apparent

whimsy (“Frood,” “Worldbeater”). Using Google or journal text search interfaces with these names yields many irrelevant results. This situation presents an opportunity for NLP and information extraction (IE) and retrieval (IR) methods to make a major contribution that can help advance the field of planetary science. The number of targets, amount of data, and number of associated publications are too large for a manual solution to be feasible. This task also presents important challenges that can motivate advances in IE that can benefit other domains.

We are working to construct a Mars Target Encyclopedia (MTE) that will contain knowledge about Mars surface targets. The MTE will provide access to the data and publications associated with each target (see Figure 1 for a hand-constructed example). Each entry will also include a list of properties that were automatically extracted from the publications, providing a high-level summary of relevant knowledge. The associated excerpts will be highlighted for each publication, serving to (1) provide support for each of the extracted properties and (2) enable users to quickly determine which papers are of the most interest.

The MTE project is in an early stage. In this paper, we report on the motivation, concept, approach, and early results that we have obtained. We also discuss the associated challenges that are of interest to the NLP and IR communities, and we describe the next steps that we will pursue.

2 Constructing a Mars Target Encyclopedia

2.1 Data Set Description

For our initial study, we constructed a corpus that consists of all papers presented at the 2015 Lunar and Planetary Science Conference (LPSC)¹. Each of the 1,991 papers is a maximum two-page extended abstract with a common structure: title, authors with affiliations, a two-column main body that may contain figures and/or tables, and references. The language is academic and makes heavy use of complex noun phrases and parenthetical expressions. The passive voice is often used.

We used the Apache Tika parser (Mattmann and Zitting 2011) to read the PDF files and convert them to plain text. Four documents could not be parsed by Tika, yielding 1,987 documents. The number of extracted words per document ranged from 142 to 2433.

We also obtained a seed list of Mars surface target names identified by the ChemCam science team. ChemCam is an instrument on the Mars Science Laboratory (MSL) rover. It fires a laser at rock or soil targets and then uses a spectrometer to record the emitted energy at 6,144 wavelengths (Maurice and others 2012). The current list of all ChemCam observations can be obtained online². The version we used contained 16,267 observations that spanned 656 distinct targets.

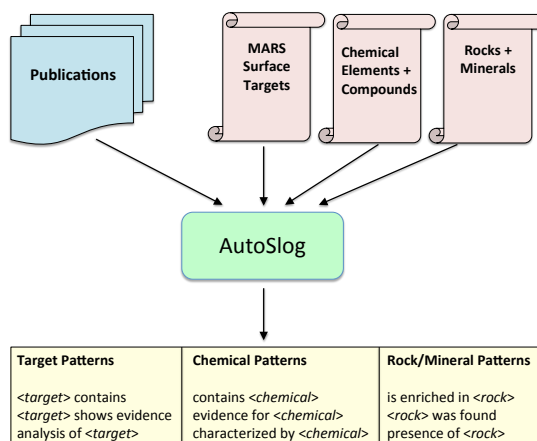


Figure 2: The extraction pattern generation process. AutoSlog learns patterns from publications and lists of target names, chemical elements and compounds, rocks, and minerals.

2.2 Information Extraction Approach

The planetary science literature is large, providing us with an abundance of text for this domain. However, annotated texts are not readily available, and obtaining human annotations from planetary science experts would be time-consuming and expensive. Our goal is to design an information extraction (IE) process that uses weakly supervised learning methods to extract knowledge from planetary science articles.

Figure 2 depicts the first step in our information extraction process. As input, we provide a text corpus of publications along with lists of terms for relevant semantic categories, such as ChemCam Mars surface target names, chemical elements/compounds, and rocks/minerals. The AutoSlog extraction pattern generator (Riloff 1993; 1996) is applied to each list, producing a set of lexico-syntactic patterns associated with the corresponding semantic category. AutoSlog is a weakly supervised pattern learner that uses heuristic rules and coarse statistics, without requiring annotated documents. The learned patterns can then be applied to new texts, to extract information that we will store in the MTE.

The AutoSlog software package includes the Sundance shallow parser and information extraction engine (Riloff and Phillips 2004), which applies the lexico-syntactic patterns generated by AutoSlog. The Sundance parser performs tokenization, sentence segmentation, morphological analysis, part-of-speech disambiguation, syntactic chunking, syntactic role assignment, and clause segmentation. Although it was originally designed to process news articles, Sundance has also been successfully applied to scientific publications

¹Downloaded from <http://www.hou.usra.edu/meetings/lpsc2015/>

²http://pds-geosciences.wustl.edu/msl/msl-m-chemcam-libs-4_5-rdr-v1/mslccm_1xxx/document/msl_ccam_obs.csv

from the biomedical literature (Ramakrishnan et al. 2010; Pokkunuri et al. 2011). A key attribute of Sundance is that its dictionaries and data files are easily customizable for specialized domains, which has allowed us to tailor it for planetary science texts. Our domain-specific customizations include (1) the specification of two words that, despite ending in a period, do not indicate the end of a sentence (“Mt.” and “wt.”) and (2) domain-specific vocabulary including element names (e.g., “chlorine”, “Cl”), mineral names (e.g., “akaganeite”, “feldspar”, “MnO”), target names (from the list described above), and unusual terms (e.g., “sol” (Martian day), “MSL”).

2.3 Illustrative Results

In this section, we provide example results for extracting information for a Mars surface target named Windjana. Windjana is a sandstone that was named after the Windjana Gorge in Western Australia (Anderson, Beegle, and Abbey 2015).

Consider the following four sentences, taken from different documents, and the extraction patterns generated by AutoSlog using the target, chemical, and mineral lists:

1: “Windjana is remarkable in containing an abundance of potassium feldspar (and thus K in its bulk chemistry) combined with a low abundance of plagioclase (and low Na/K in its chemistry).”

Target:	<subj>_CONTAINING_ABUNDANCE
Mineral:	ABUNDANCE_OF_<np>

2: “The high abundances of K-feldspar and iron oxides in Windjana, also reflected in the APXS chemical analysis as high K and Fe (Table 2) [7], are unusual.”

Mineral:	ABUNDANCE_OF_<np>
Target:	OXIDES_IN_<subj>

3: “The Windjana sandstone contains high magnetite along with 2:1 phyllosilicates [9].”

Target/Mineral:	<subj>_CONTAINS_MAGNETITE
-----------------	---------------------------

4: “The abundance of K-spar and the potential presence of illite in Windjana must be considered when interpreting the formation of the Dillinger sandstone because these phases can form in diagenetic K-rich environments on Earth.”

Mineral:	ABUNDANCE_OF_<np>
Target/Mineral:	ILLITE_IN_<np>

If we synthesize all the information found in these sentences by the patterns, we can produce a summary of known properties about Windjana:

Target Windjana
Properties: - contains abundance of potassium feldspar - high abundances of K-feldspar and iron oxides - contains high magnetite - contains illite

The synthesis was done manually in this example, and we plan to develop methods for merging the information extracted from different patterns automatically.

This summary compiles knowledge extracted from multiple papers. While each author team might choose to focus on different aspects in their individual papers, information extraction from the corpus as a whole enables a collective picture of the composition of this target to emerge. This sum-

mary could inspire further discoveries or hypotheses, as the reader considers, for example, what it means for feldspar, magnetite, and illite to be jointly present and whether this is unusual.

3 Challenges

In addition to the typical challenges of conducting information extraction in a new domain (e.g., domain-specific vocabulary), there are several NLP/IE challenges of special interest involved in this project.

First, we must maintain high precision in the extracted information. For this application, precision is more important than recall; users would rather have an incomplete summary than one that contains incorrect information. Previous work on high-precision IE has emphasized the importance of human review (Caruana, Hodor, and Rosenberg 2000). The AutoSlog-TS system ranks candidate extraction patterns based on their prevalence in a set of (unannotated) relevant versus irrelevant documents, enabling human review to be focused on the most likely patterns first (Riloff 1996). This process also provides an assessment of the system’s current precision in extracting patterns. We plan to likewise incorporate weak supervision via human review of the extracted patterns and properties.

Second, we must identify and handle cases where the extracted information within a summary is not consistent. If the system extracts a pair of properties such as “high in magnetite” and “low in magnetite”, for the same target, the consistency of the collective result is reduced. In some cases, it is non-trivial to determine whether two facts are consistent or contradictory. For example, consider two hypothetical sentences: “Target Oliphaunt is feldspar-rich” and “Target Oliphaunt has low Si.” By definition, a feldspar contains a lot of silicon, but domain knowledge is required to detect that these statements are in conflict.

When contradictions are detected, we must determine how to reconcile them. The MTE inherently requires robust *information fusion* to generate each encyclopedia entry. Conflicts are especially likely since the information is extracted from different authors and publications rather than a single source. Existing approaches to cross-document information fusion include assigning a confidence proportional to the number of sources that agree on a fact and estimating the reliability of the sources themselves (Ji 2010).

Since there is a temporal component to the papers, it can also be the case that an interpretation could be overturned or negated by later findings or a more careful examination of the available data. The same challenge appears when performing information extraction for news articles (McKeown and Radev 1995; Ji 2010). A simple strategy would be to let the most recently reported information, as determined by publication date, supersede older information. However, since apparent conflicts could be created by an incorrect extraction (see above), it may be best to flag the facts as conflicting but let the user review them.

Third, coreference is especially challenging because of the diversity in how authors refer to the same targets or properties. Windjana is variously referred to as “Windjana”, “Windjana drill fines”, “Windjana drill tailings”,

“Windjana sample”, “Windjana sandstone”, “it”, etc. Elements and minerals have multiple manifestations; “Cl” vs. “chlorine”, “potassium feldspar” vs. “K-feldspar” vs. “K-spar”, etc. Coreference resolution remains an unsolved problem, although new advances in active learning are promising (Sachan, Hovy, and Xing 2015).

Fourth, when interpreting observations, scientific conclusions range from solid facts to speculation about causes. In some cases, evidence even for basic properties (e.g., “contains Mn”) may fall into a gray area. Scientific language may therefore employ epistemic modifiers such as “likely” or “probably” or “possibly.” Examples from the LPSC 2015 corpus related to the Windjana target include “*the Windjana drill tailings likely contain a spectrally opaque material (e.g., magnetite, ilmenite)*” (Johnson et al. 2015) and “*the potential presence of illite in Windjana*” (Rampe and others 2015). This type of language is known as *hedging*, and some methods have been developed for automatically detecting hedges (Medlock and Briscoe 2007; Agarwal and Yu 2010). Distinguishing this information from more confidently stated conclusions is vital to preserving the nuance reported in the documents, as previously studied in the context of medical discussion forums (Sokolova et al. 2013).

4 Conclusions and Next Steps

There is a growing need for a comprehensive, up-to-date compilation of Mars surface targets and knowledge of what has been discovered about them. We are working to automatically construct a Mars Target Encyclopedia by applying information extraction methods to planetary science publications. This is a cross-document information extraction task that will benefit ongoing science investigations by providing the full context of previously published knowledge. Browsing the synthesized encyclopedia entries, with their summaries of target knowledge, could also reveal previously undetected connections or similarities between targets.

We have employed Tika, Sundance, and AutoSlog to extract basic information about Mars surface targets. We plan to employ a small amount of human review to provide light supervision/feedback and evaluation of the system’s precision. Important open questions remain about how to detect and accommodate inconsistency in information extracted from different documents at different times. The system also must correctly capture epistemic modifiers so that the level of confidence in the extracted knowledge is preserved.

The initial term lists that we give to AutoSlog were manually compiled, but we plan to expand them automatically using bootstrapping methods for semantic lexicon induction. The goal is to learn additional terms that can refer to Mars targets, chemicals, and rocks/minerals, such as missing terms (because manually compiled lists are inevitably incomplete) as well as name variants (e.g., “World-beater” vs. “Worldbeater”), general terms (e.g., “target”), abbreviations (e.g., “F” for fluorine), and shorthand terms (e.g., “Fe-oxide” vs. “iron oxide”). We will use the Basilisk semantic lexicon bootstrapping algorithm (Thelen and Riloff 2002), using the manually compiled lists as seed terms and a large collection of planetary science articles as the text corpus. Basilisk automatically learns new semantic class members

by identifying terms that consistently co-occur in related contexts with the seeds, in an iterative bootstrapping process. Basilisk has been successfully used to generate semantic dictionaries for a variety of domains, including disease outbreaks (Phillips and Riloff 2007; Qadir and Riloff 2012), terrorist events (Thelen and Riloff 2002), and subjectivity analysis (Riloff and Wiebe 2003). We will manually review Basilisk’s proposed additions to maintain high integrity of the lists.

In addition to these plans, we welcome suggestions that can help guide the evolution and maturation of the MTE to maximize its utility and impact.

Acknowledgments

This work 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

- Agarwal, S., and Yu, H. 2010. Detecting hedge cues and their scope in biomedical literature with conditional random fields. *Journal of Biomedical Informatics* 43(6):953–961.
- Anderson, R. C.; Beegle, L.; and Abbey, W. 2015. Drilling on Mars: What we have learned from the Mars Science Laboratory Powder Acquisition Drill System (PADS). In *Proceedings of the 46th Lunar and Planetary Science Conference*, Abstract 2417.
- Caruana, R.; Hodor, P. G.; and Rosenberg, J. 2000. High precision information extraction. In *Proceedings of the KDD-2000 Workshop on Text Mining*.
- Ji, H. 2010. Challenges from information extraction to information fusion. In *Proceedings of the 23rd International Conference on Computational Linguistics*, 507–515.
- Johnson, J.; Wiens, R.; Maurice, S.; Blaney, D.; Gasnault, O.; Cloutis, E.; Mouélic, S. L.; and Bender, S. 2015. Chemcam passive reflectance spectroscopy of ferric sulfates and ferric oxides near the base of Mt. Sharp. In *Proceedings of the 46th Lunar and Planetary Science Conference*, Abstract 1433.
- Mattmann, C., and Zitting, J. 2011. *Tika in Action*. New York: Manning Publications.
- Maurice, S., et al. 2012. The ChemCam instrument suite on the Mars Science Laboratory (MSL) rover: Science objectives and mast unit description. *Space Science Reviews* 170(1):95–166. doi:10.1007/s11214-012-9912-2.
- McKeown, K., and Radev, D. R. 1995. Generating summaries of multiple news articles. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, 74–82.
- Medlock, B., and Briscoe, T. 2007. Weakly supervised learning for hedge classification in scientific literature. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 992–999.
- Phillips, W., and Riloff, E. 2007. Exploiting role-identifying nouns and expressions for information extraction. In *Pro-*

ceedings of the Conference on Recent Advances in Natural Language Processing.

Pokkunuri, S.; Ramakrishnan, C.; Riloff, E.; Hovy, E.; and Burns, G. 2011. The role of information extraction in the design of a document triage application for biocuration. In *Proceedings of the ACL/HLT-2011 Workshop on Biomedical Natural Language Processing*, 46–55.

Qadir, A., and Riloff, E. 2012. Ensemble-based semantic lexicon induction for semantic tagging. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, 199–208.

Ramakrishnan, C.; Baumgartner Jr., W.; Blake, J.; Burns, G.; Cohen, K. B.; Drabkin, H.; Eppig, J.; Hovy, E.; Hsu, C.; Hunter, L.; Ingulfsen, T.; Onda, H.; Pokkunuri, S.; Riloff, E.; Roeder, C.; and Verspoor, K. 2010. Building the scientific knowledge mine (SciKnowMine): A community-driven framework for text mining tools in direct service to biocuration. In *Proceedings of the LREC-10 Workshop on New Challenges for NLP Frameworks*, 9–14.

Rampe, E. B., et al. 2015. Potential cement phases in sedimentary rocks drilled by Curiosity and Gale Crater, Mars. In *Proceedings of the 46th Lunar and Planetary Science Conference*, Abstract 2038.

Riloff, E., and Phillips, W. 2004. An introduction to the Sundance and AutoSlog systems. Technical Report UUCS-04-015, University of Utah School of Computing.

Riloff, E., and Wiebe, J. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of the Conference on Empirical methods in natural language processing*, 105–112.

Riloff, E. 1993. Automatically constructing a dictionary for information extraction tasks. In *Proceedings of the 11th National Conference on Artificial Intelligence*, 811–816.

Riloff, E. 1996. Automatically generating extraction patterns from untagged texts. In *Proceedings of the 13th National Conference on Artificial Intelligence*, volume 2, 1044–1049.

Sachan, M.; Hovy, E.; and Xing, E. P. 2015. An active learning approach to coreference resolution. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, 1312–1318.

Sokolova, M.; Ioshikhes, I.; Poursepanj, H.; and MacKenzie, A. 2013. Helping parents to understand rare diseases. In *Proceedings of the Workshop on NLP for Medicine and Biology associated with RANLP 2013*, 24–33.

Thelen, M., and Riloff, E. 2002. A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 214–211.