Change Log

| Revision | Submission Date | Affected Sections or Pages | Change Summary |
| --- | --- | --- | --- |
| Initial | 18-10-2017 | All | Initial issue of document. |

Table of Contents

1 Document Overview 3

1.1 Identification 3

1.2 Purpose 3

1.3 Terminology and Notation 3

1.4 References 3

2 Overview of the System 3

3 Instructions for Operations 4

3.1 Initial Setup 4

3.2 Ingesting Documents 5

3.3 Searching the MTE 5

3.4 MTE Web Server 5

4 Input Descriptions 6

5 Output Descriptions 6

A. Error message summary 6

# 

# Document Overview

## Identification

| **Property** | **Value** |
| --- | --- |
| Element | Instrument Data Systems (IDS) |
| Program Set | MTE |
| Version | 1.0 |

## Purpose

The purpose of this document is to provide a guide for the installation and operation of the Mars Target Encyclopedia.

## Terminology and Notation

* “Component” refers to a geochemical component of a target. It could be an element or a mineral.
* “MSL” refers to the Mars Science Laboratory (Curiosity) rover.
* “MTE” is an abbreviation for Mars Target Encyclopedia.
* “Target” refers to named observation targets (e.g., rocks, soils) on the surface of Mars.

## References

Table 1: Applicable JPL Rules documents

| Title | DocID |
| --- | --- |
| Software Development | 57653 |

Table 2: Applicable MGSS documents

| Title | Document Number |
| --- | --- |
| MGSS Implementation and Maintenance Task Requirements | DOC-001455 |

# Overview of the System

The Mars Target Encyclopedia (MTE) is an information extraction system that takes in a set of scientific publications (in PDF format) and extracts geochemical information about Mars targets (e.g., rocks, soils, drill holes). This information is represented as a set of compositional relationships between and elements (e.g., iron, fluorine) or minerals (e.g., hematite, plagioclase). The extracted relationships, along with source publication meta-data (e.g., authors, title, venue) are stored in a Solr database. Figure 1 illustrates the system architecture. The information extraction process is separated into two phases: named entity recognition (to find elements, minerals, and targets in the documents) and relation extraction (to find relationships between targets and components).



Figure 1. Mars Target Encyclopedia information extraction pipeline.

Users can interact with the MTE in the following ways:

* **Search the MTE.** The MTE includes a simple web server interface that supports searching for a specific target (e.g., “Dillinger”), an element (e.g., “hydrogen”), or a mineral (e.g., “hematite”). Search results allow users to access further information about each returned target, including all known geochemical relationships for that target. Each relationship is supported with a citation to the source document, link to the original PDF, and a single-sentence excerpt from which the relationship was extracted.
* **Update the MTE with content from new documents.** The MTE is intended to be able to grow and evolve as new publications become available. Users can point the MTE ingestion pipeline at a new collection of PDF documents. Their contents will be analyzed and any new geochemical relationships that are found will be added to the MTE database.

# Instructions for Operations

## Initial Setup

The MTE requires that the following software packages be installed on the computer system:

* Python version 2.7
* Apache Tika 1.13: <http://tika.apache.org/>
* Grobid: <https://github.com/kermitt2/grobid>
* CoreNLP: <https://stanfordnlp.github.io/CoreNLP/>
* jSRE 1.1: <https://hlt-nlp.fbk.eu/technologies/jsre>
* Apache Solr 6.x: <http://lucene.apache.org/solr/>
* parser-indexer: <https://github.com/USCDataScience/parser-indexer-py>
* MTE: <https://github-fn.jpl.nasa.gov/wkiri/mte>

Once the repository has been cloned, the parser-indexer server can be started. It is recommended that this be run by a team or daemon account so that it can be cleanly started/restarted as needed. Full installation instructions are provided at <https://github-fn.jpl.nasa.gov/wkiri/mte/wiki/Parser-Indexer-Pipeline>.

## Ingesting Documents

The MTE database is populated using an information extraction pipeline that processes a set of input PDF documents to extract meta-data, identify relevant entities, and extract geochemical relationships between the entities.

To process a new set of documents:

1. Create a text document containing the full paths to the PDF files, one per line.

find <pdf-dir> -name \*.pdf > input-pdfs.list

1. Parse the PDF files with Apache Tika:

NER\_MODEL=$MTE\_HOME/trained\_models/ner\_model\_train\_62r15v3\_emt\_gazette.ser.gz

JSRE\_BASE=[where jSRE is installed locally]

JSRE\_MODEL=$MTE\_HOME/trained\_models/jSRE-lpsc15-merged-binary.model

PARSER=$MTE\_HOME/parser-indexer-py/src/parserindexer/parse\_all.py

python $PARSER -n $NER\_MODEL -j $JSRE\_BASE -m $JSRE\_MODEL -li input-pdfs.list -o corenlp-out.jl

1. Add the extracted information into the Solr database:

SOLR\_URL=http://localhost:8983/solr/docsdev

python $MTE\_HOME/parser-indexer-py/src/parserindexer/indexer.py -i corenlp-out.jl -s $SOLR\_URL

## Searching the MTE

Users can search the contents of the MTE using Solr query URLs, assuming that the Solr server is running on the local host at port 8983.

Examples:

1. Generate a histogram of all components (elements, minerals) found:

<http://localhost:8983/solr/docs/query?q=source:corenlp&fq=type:contains&rows=0&facet=on&facet.field=cont_names_ss&facet.mincount=1&facet.limit=-1&facet.sort=count>

1. Generate a histogram of all targets found in LPSC 2014 documents:

[http://localhost:8983/solr/docs/query?q=id:lpsc14\*&fq=type:target&fq=source:corenlp&rows=0&facet=on&facet.field=can\_name&facet.mincount=1&facet.limit=-1&facet.sort=count](http://localhost:8983/solr/docs/query?q=id:lpsc14*&fq=type:target&fq=source:corenlp&rows=0&facet=on&facet.field=can_name&facet.mincount=1&facet.limit=-1&facet.sort=count)

1. Show all relationships extracted for documents from LPSC 2014, including sentence exerpts:

[http://localhost:8983/solr/docs/query?q=id:lpsc14\*&fq=source:corenlp&fq=type:contains&fl=excerpt\_t&fl=cont\_names\_ss&fl=target\_names\_ss](http://localhost:8983/solr/docs/query?q=id:lpsc14*&fq=source:corenlp&fq=type:contains&fl=excerpt_t&fl=cont_names_ss&fl=target_names_ss)

## MTE Web Server

The MTE also includes a Flask-based web server that provides users with an easy-to-use search interface to the contents of the Solr database. It provides an interface to query the Solr DB and returns results in JSON format, which are processed and displayed using Javascript. It also connects to MMGIS to provide a spatial map of the MSL rover’s traverse with search results marked with red dots on the map.

The web server is currently deployed on buffalo.jpl.nasa.gov and accessible at <https://mte.jpl.nasa.gov/>. Access is restricted to a specified list of JPL users which is maintained by system administrators available at [mliahelp@jpl.nasa.gov](mailto:mliahelp@jpl.nasa.gov). The base URL accesses a small database that consists only of manually created relationships. The full set of automatically extracted relationships (currently covers ~6000 LPSC abstracts from 2014—2016) is available at <https://mte.jpl.nasa.gov/corenlp/>. Users can search on targets, elements, minerals, or author names. Typing in the search box also generates a list of possible completions to aid finding results.

Full details about how to configure a system to deploy an MTE web server are available at <https://github-fn.jpl.nasa.gov/wkiri/MTE/wiki/MTE-Web-Interface>.

# Input Descriptions

New input documents are required to be in PDF format. They are processed by the Apache Tika parser to extract the text content using UTF-8 to preserve special characters and math symbols. Currently, the MTE assumes that the input PDFs were published by the Lunar and Planetary Science Conference. PDFs generated by other sources may not be successfully analyzed.

The MTE also requires access to some reference files. They include:

* MSL ChemCam target list
* International Mineralogical Association mineral list
* Element list

These files are included in the refs/ directory of the git repository.

# Output Descriptions

The result of the MTE information extraction processing is stored in a JSON file, as specified when parse\_all.py is run. This is an intermediate product that can be inspected using any JSON viewer. The contents are then stored in Solr via the indexer script as described above. At that point they can be queried using Solr searches as indicated in the examples above.

1. Error message summary
2. Input file to parse cannot be found (-i option).

Error: Could not find input PDF file t.pdf.

1. File containing list of input files cannot be found (-li option).

Error: Could not find file containing input paths pdfpaths.list.

1. A file in the list of input files could not be found.

Error: Could not find PDF file d.pdf.

1. A file in the list of input files could not be parsed.

Error: Could not parse PDF file d.pdf.

1. Named entity (NER) model cannot be found.

Error: Could not find NER model /proj/mte/trained\_models/ner\_model\_train\_62r15v3\_emt\_gazette.ser.g.

1. Relation extraction (jSRE) model cannot be found.

Error: Could not find jSRE model /proj/mte/trained\_models/jSRE-lpsc15-merged-binary.mode.