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

Kiri L. Wagstaff, Ellen Riloff, Nina L. Lanza, Chris A. Mattmann, and Paul M. Ramirez

AAAI Knowledge Extraction from Text Workshop

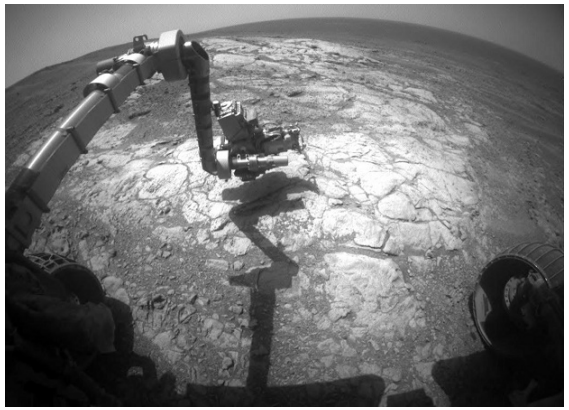
February 12, 2016

Collect data

Downlink data

Uplink commands

Opportunity Data Archive



…

MER: 4M Pancam, 1.7M Navcam, 230K MI

4200+ per-sol summaries of data collected

What have we learned about these targets?

"round-trip" tracking of data products and targets

Vision: A Mars Target Encyclopedia

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

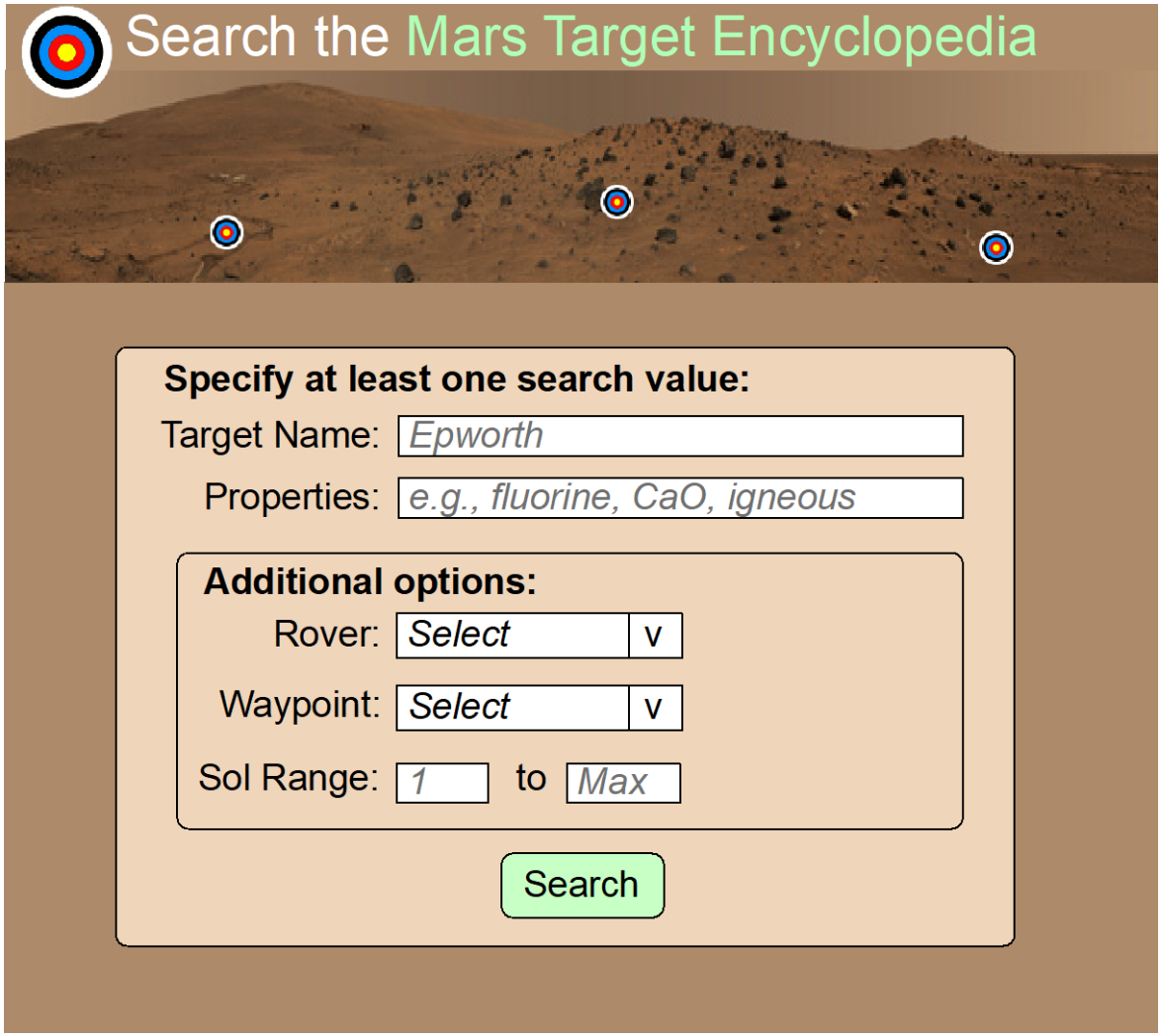
▶ Forni et al. (2014b), "First fluorine and chlorine detection with ChemCam on-board MSL," 8th International Conference on Mars

▶ Forni et al. (2015), "First detection of fluorine on Mars: Implications for Gale crater's chemistry," Geophysical Research Letters

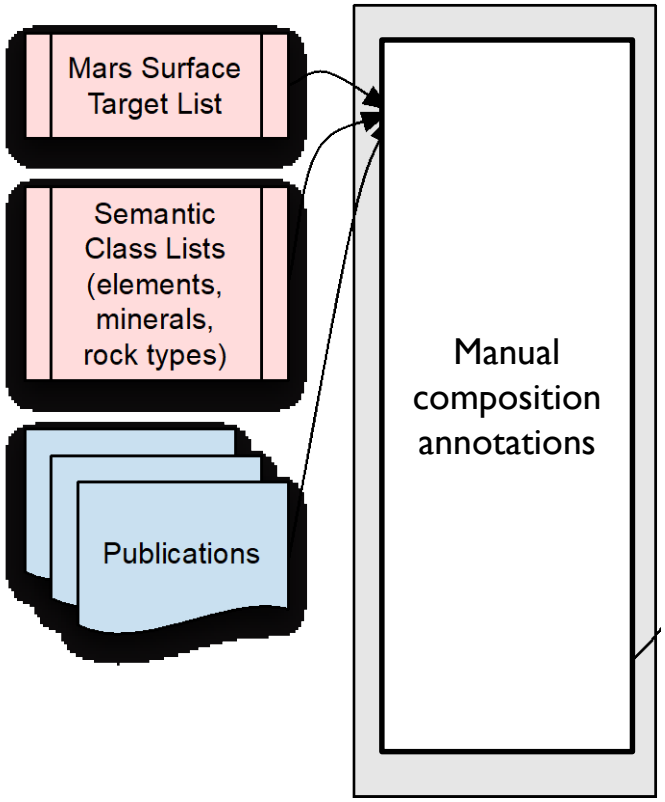
More like this

- Epworth_3
- Black Rock (CaF_2)

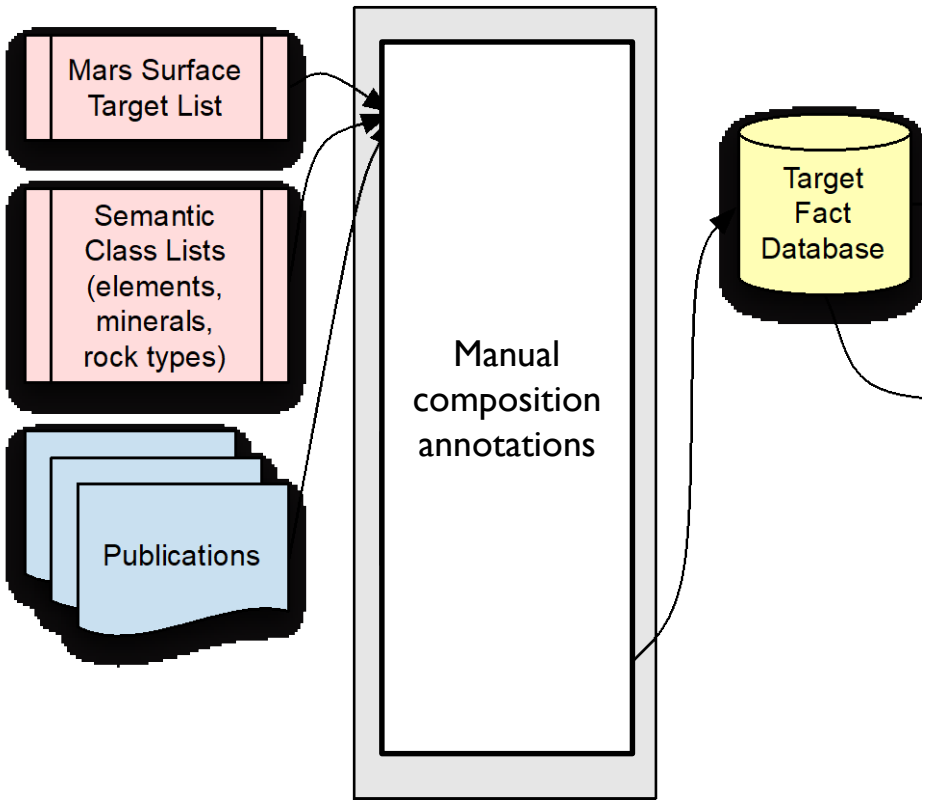
Searchable interface



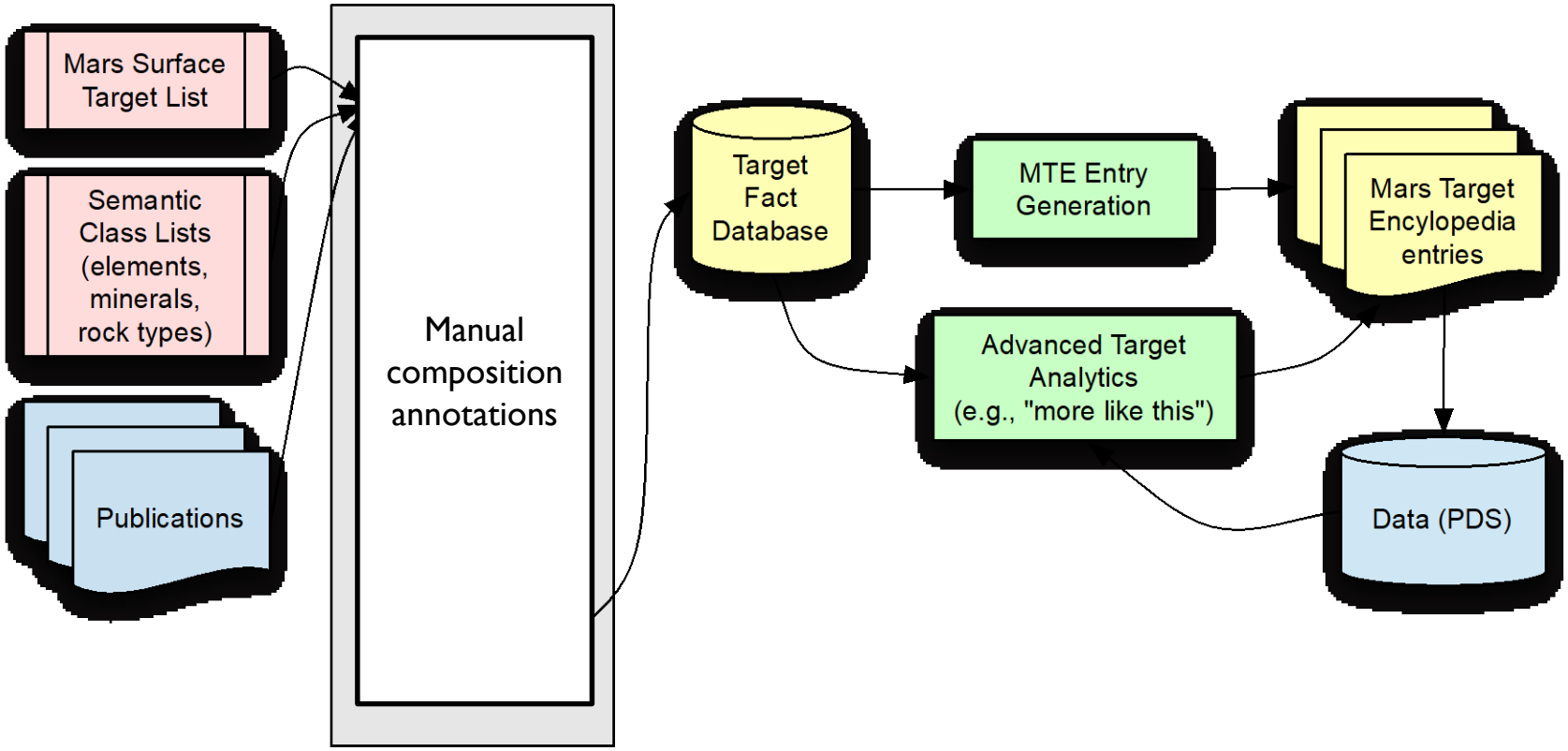
From text to knowledge



From text to knowledge



From text to knowledge



The corpus

- LPSC 2015 abstracts (n=1987)
  - Parsed PDFs using Apache Tika
  - # words parsed per document: 142 to 2433

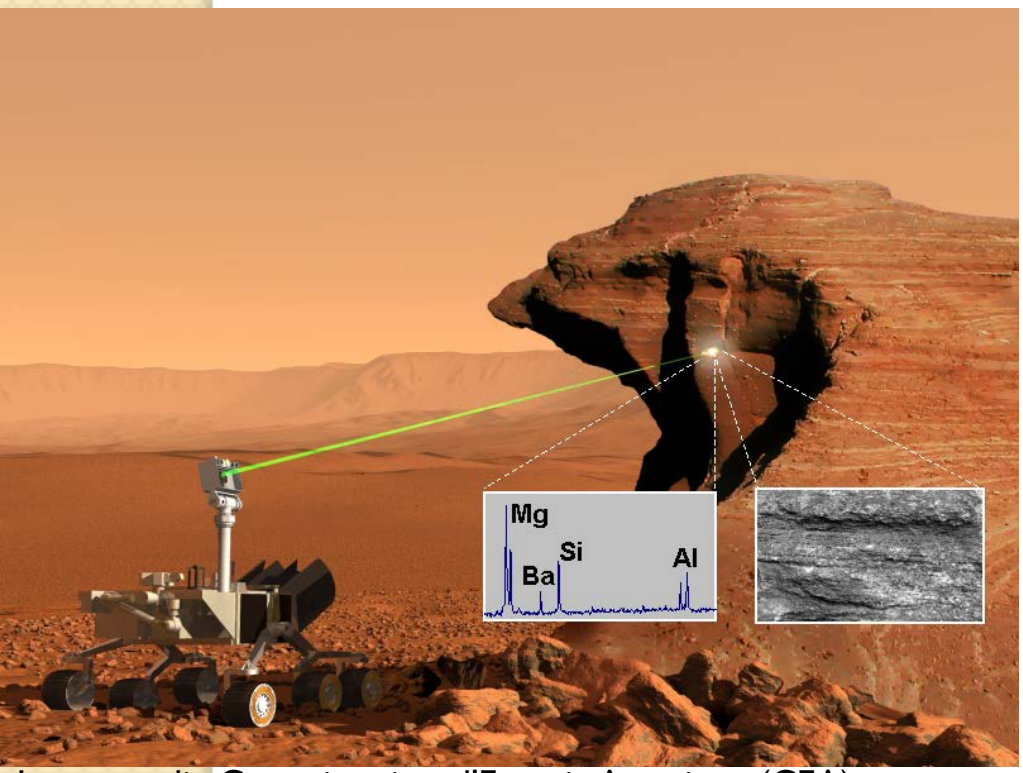
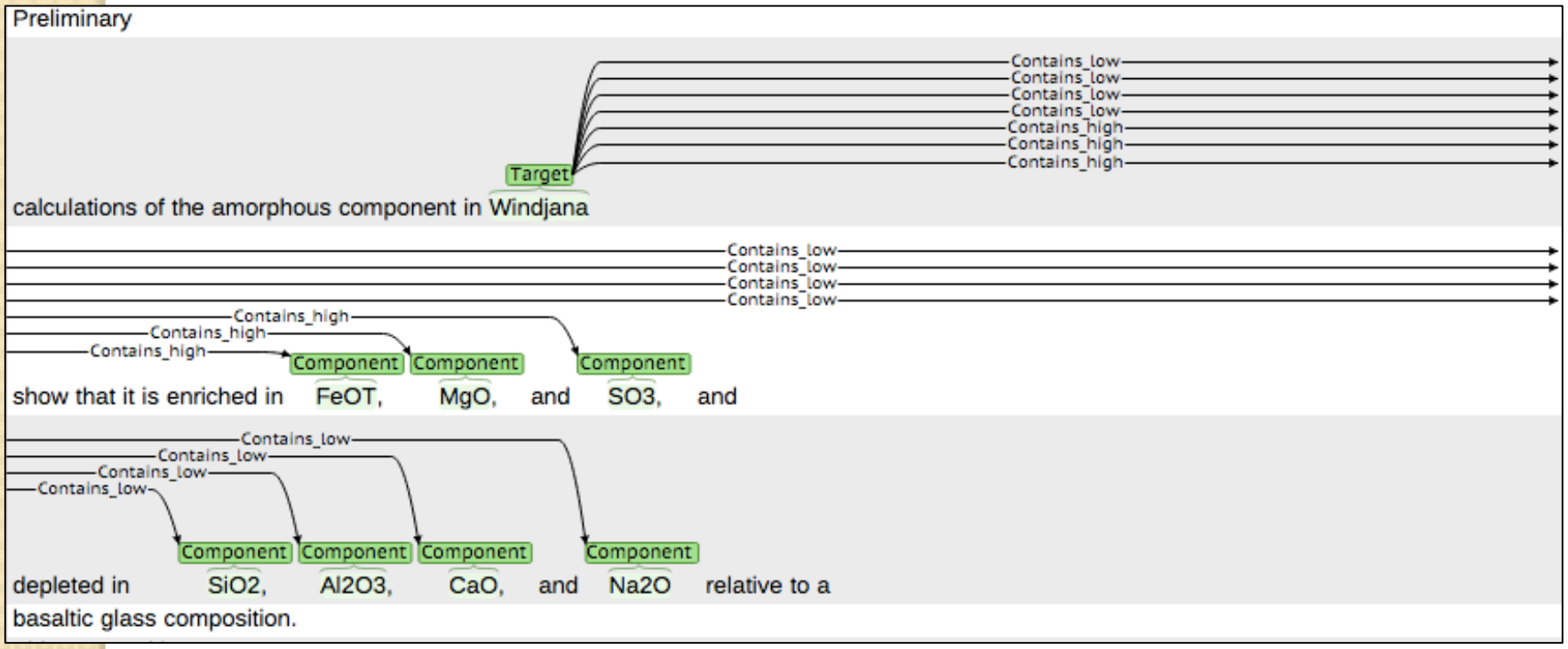


Image credit: Commissariat a l'Energie Atomique (CEA)

- Target list from MSL ChemCam (n=656)

Manual composition annotations

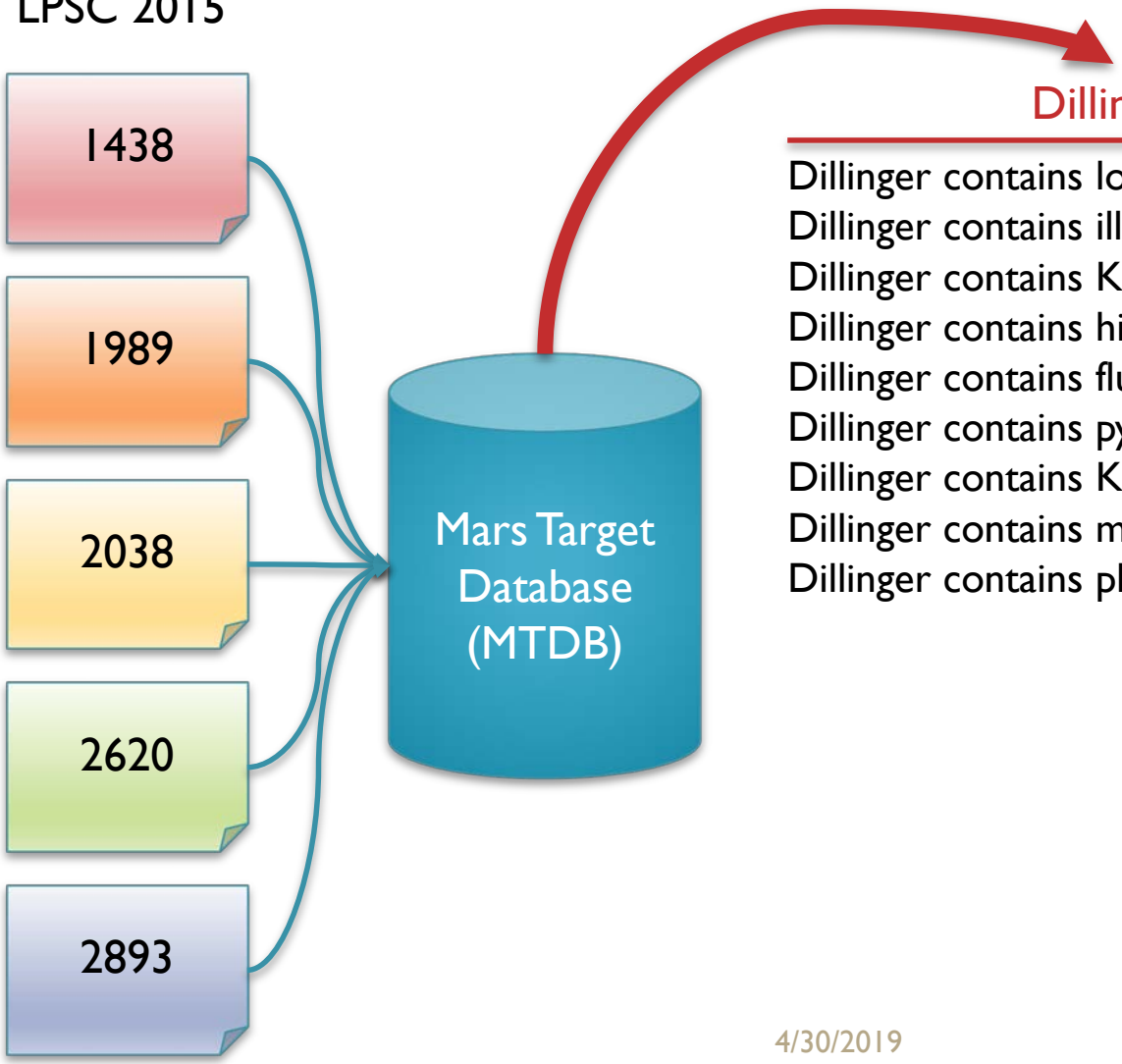
- brat: interactive annotation and visualization



- Python script to import to PostgreSQL DB

Initial results

LPSC 2015

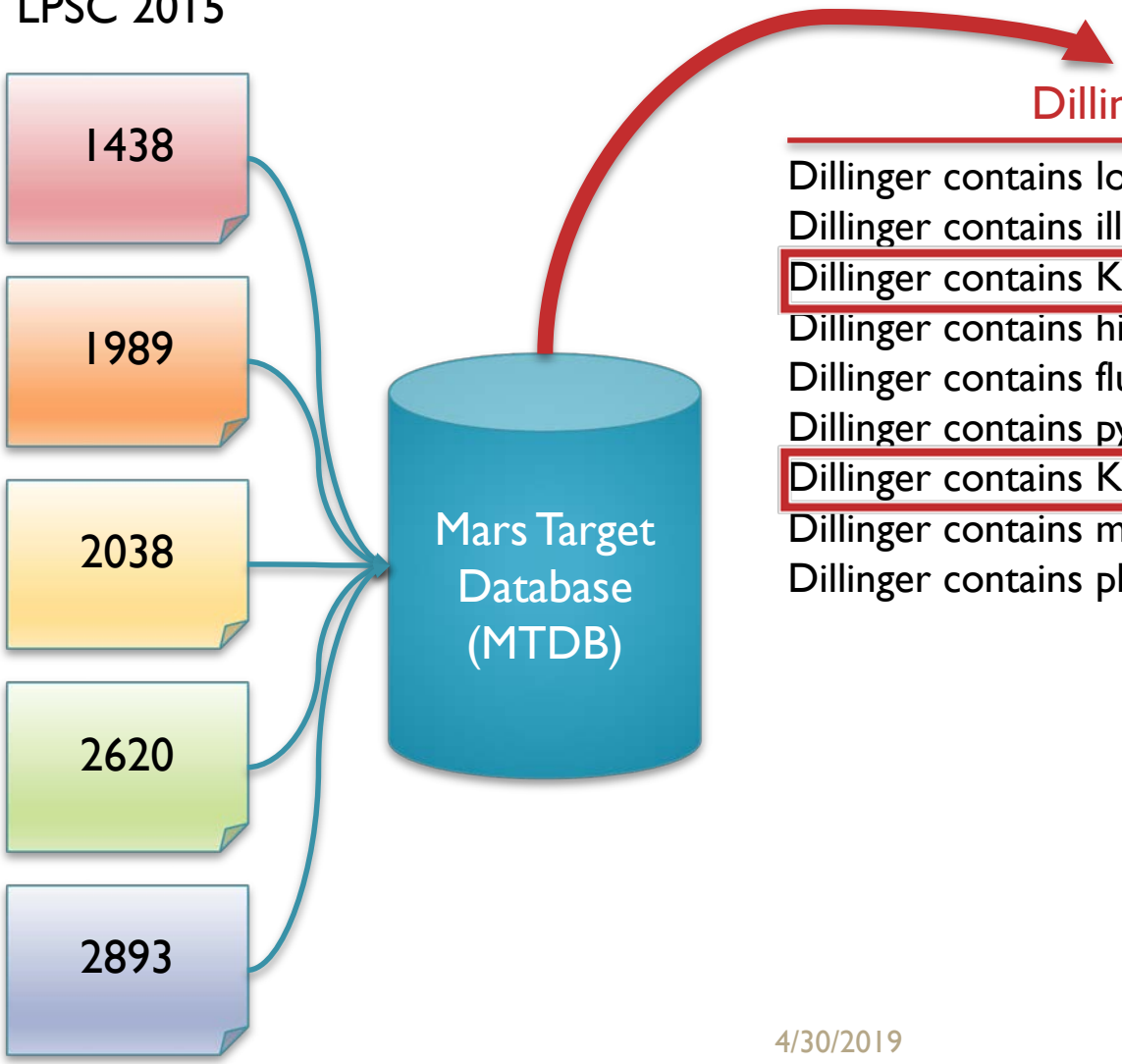


Dillinger

Dillinger contains low Na/K ratio (doc 1438)
Dillinger contains illite clays (doc 1438)
Dillinger contains K-feldspar (doc 1438)
Dillinger contains high K (doc 1438)
Dillinger contains fluorine (doc 1989)
Dillinger contains pyroxene (doc 2038)
Dillinger contains K-feldspar (doc 2038)
Dillinger contains magnetite (doc 2038)
Dillinger contains phyllosilicate (doc 2038)

Initial results

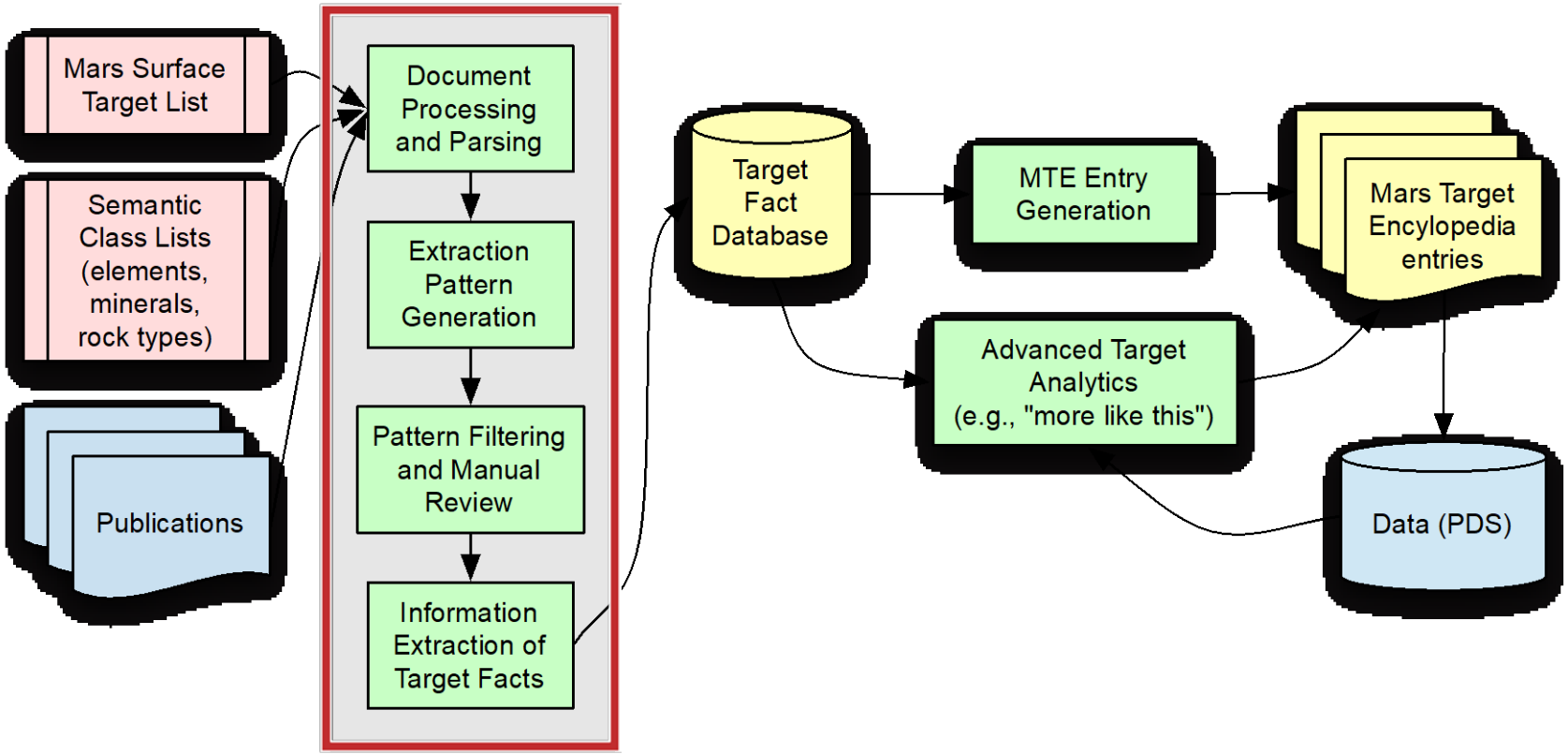
LPSC 2015



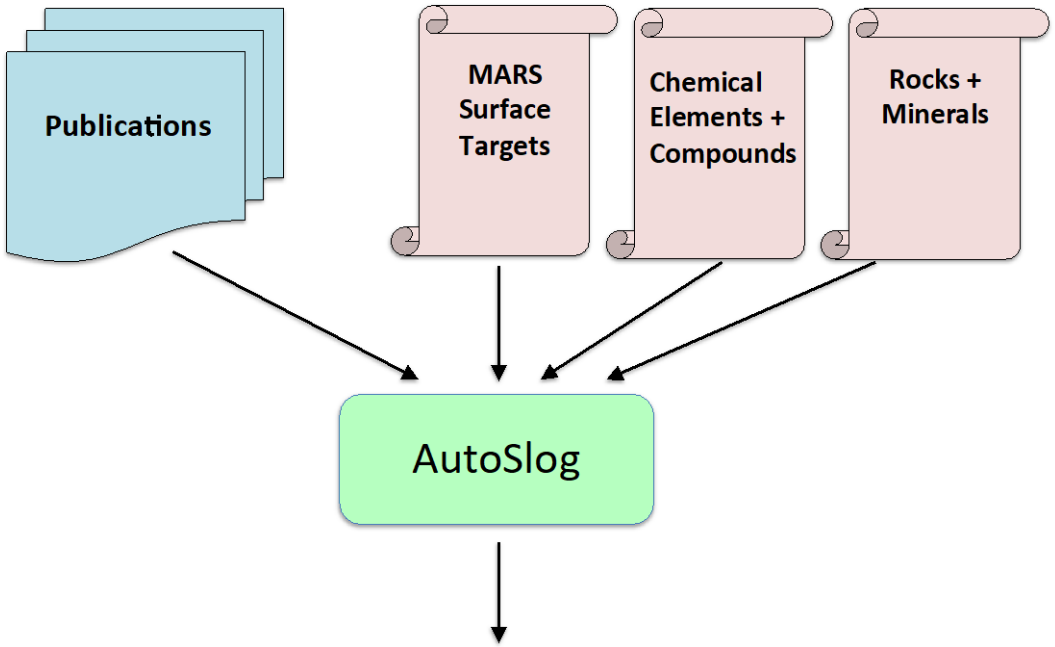
Dillinger

Dillinger contains low Na/K ratio (doc 1438)
Dillinger contains illite clays (doc 1438)
Dillinger contains K-feldspar (doc 1438)
Dillinger contains high K (doc 1438)
Dillinger contains fluorine (doc 1989)
Dillinger contains pyroxene (doc 2038)
Dillinger contains K-feldspar (doc 2038)
Dillinger contains magnetite (doc 2038)
Dillinger contains phyllosilicate (doc 2038)

Automating text to knowledge



Automated annotation



Target Patterns	Chemical Patterns	Rock/Mineral Patterns
<*target*> contains <*target*> shows evidence analysis of <*target*>	contains <*chemical*> evidence for <*chemical*> characterized by <*chemical*>	is enriched in <*rock*> <*rock*> was found presence of <*rock*>

Initial automated results (Windjana)

- "The Windjana sandstone contains high magnetite along with 2:1 phyllosilicates [9]."
  - `<subj>_CONTAINS_MAGNETITE`
- "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."
  - `ABUNDANCE_OF_<np>`
  - `OXIDES_IN_<subj>`
- "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)."
  - `<subj>_CONTAINING_ABUNDANCE`
  - `ABUNDANCE_OF_<np>`

Initial automated results (Windjana)

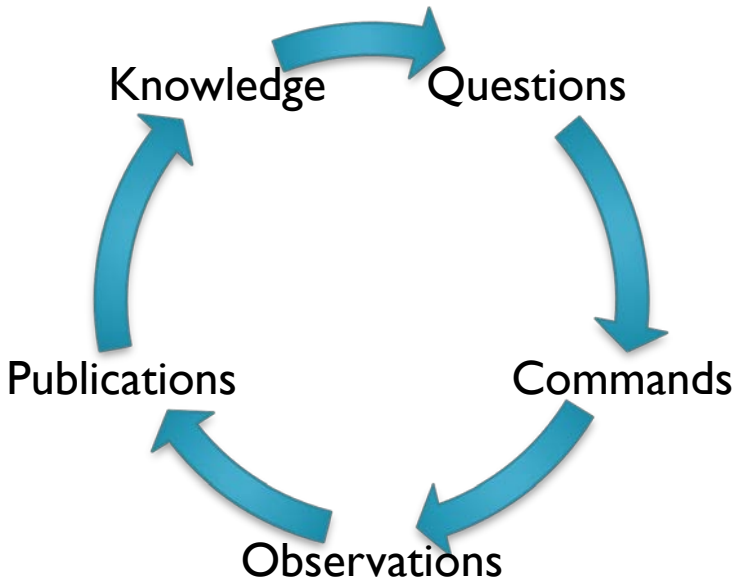
- Synthesized:

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

Open questions

- How to combine extractions into coherent entry? (information fusion)
  - Multi-source IE: papers can disagree, be superseded
  - Detect and resolve contradictions
  - Handle coreference
- How to handle "epistemic modifiers"
  - E.g. "probably", "likely contain", "potential presence of"

MTE benefits



- Answer questions such as:
  - What is already known about target X?
  - Is there consensus about target X's properties?
  - What other targets are like X?
  - What targets are most unusual?