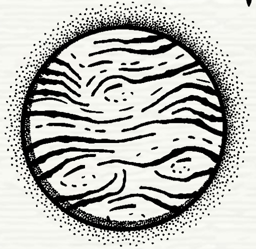
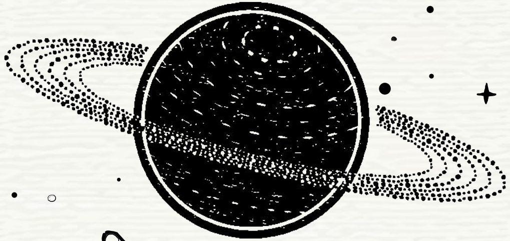
Chaos to Cosmos

The Martian Explorer

Annie Bhalla 3638974





Incoming Signal from "Mars..."

- NASA's robotic explorers like Mars 2020 Perseverance, MAVEN, to name a few, constantly send mission updates and are published as stories/news for public.
- 34 Mars Missions with over 600 stories.
- These reports are rich in content: science, engineering, and operational details.

BUT

- These reports are trapped **unstructured HTML** and buried in websites for public
- No centralized monitoring, no trend detection, and no structured archive for public for comprehensive chronological understanding and analysis.

Incoming Signal from "Mars…"

Why ?

We are the Mars Generation

WORKFLOW OF EXPLORER

01

COLLECT

NASA Website
Status Reports
Web Scraping
JSON

02

PREPARE

XSLT Transform
XML Validation
XML Generation
eXist DB

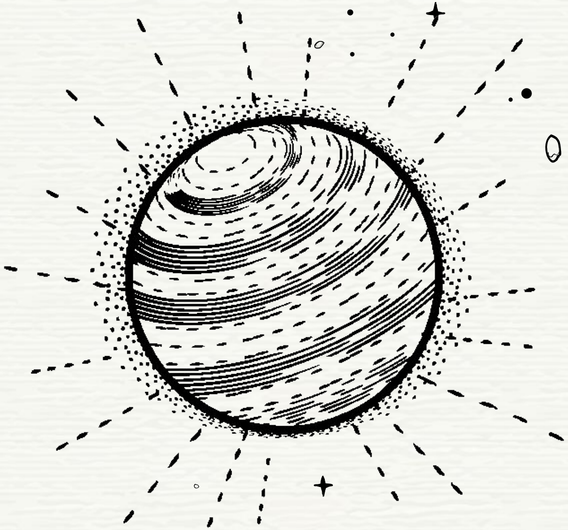
03

ACCESS

XQuery
Next Js UI

01

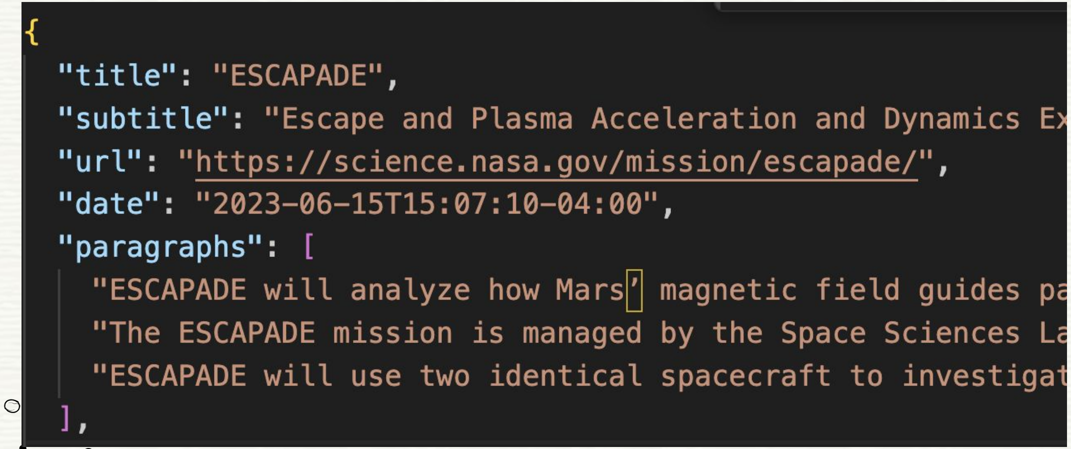
COLLECT MODULE



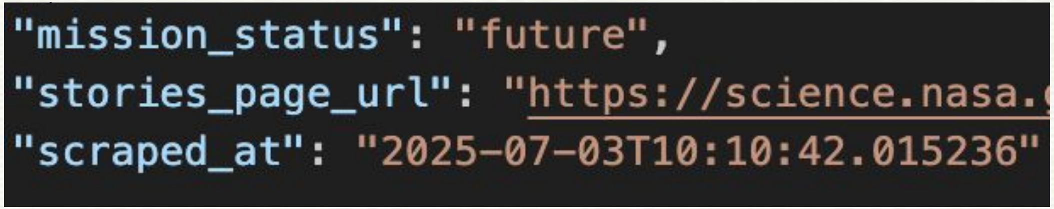
Scraping the “Martian Frontier”

- **Data Sources:** Mars 2020, Perseverance, Maven mission status feeds
- *Extension: to all Mars mission status feed*
- **Tools:** Python, BeautifulSoup, requests
- **Extract:** Mission name, date, title, body, keywords…
- **Output:** Stored as JSON objects for processing

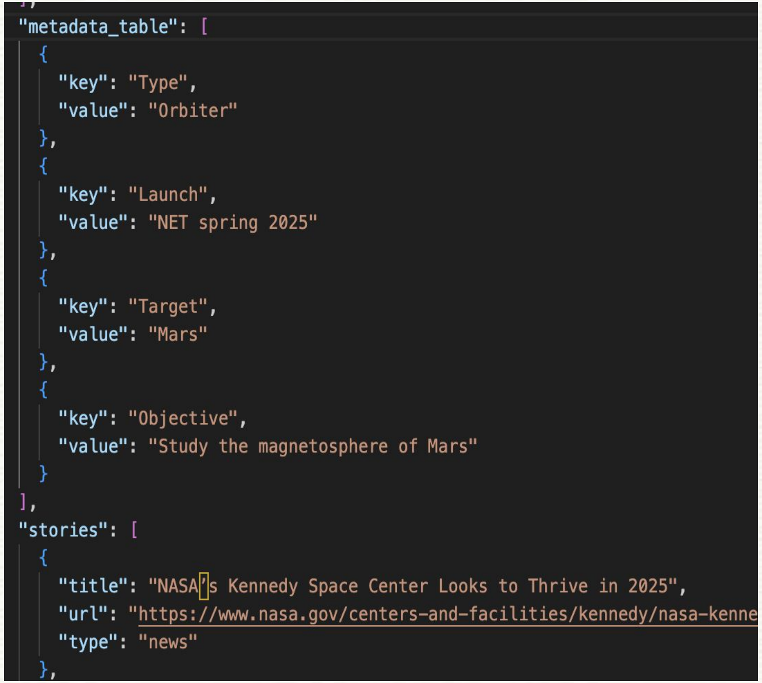
Scraping the “Martian Frontier”



```
{
  "title": "ESCAPADE",
  "subtitle": "Escape and Plasma Acceleration and Dynamics Ex
  "url": "https://science.nasa.gov/mission/escapade/",
  "date": "2023-06-15T15:07:10-04:00",
  "paragraphs": [
    "ESCAPADE will analyze how Mars’ magnetic field guides pa
    "The ESCAPADE mission is managed by the Space Sciences La
    "ESCAPADE will use two identical spacecraft to investigat
  ],
```



```
"mission_status": "future",
"stories_page_url": "https://science.nasa.
"scraped_at": "2025-07-03T10:10:42.015236"
```



```
"metadata_table": [
  {
    "key": "Type",
    "value": "Orbiter"
  },
  {
    "key": "Launch",
    "value": "NET spring 2025"
  },
  {
    "key": "Target",
    "value": "Mars"
  },
  {
    "key": "Objective",
    "value": "Study the magnetosphere of Mars"
  }
],
"stories": [
  {
    "title": "NASA’s Kennedy Space Center Looks to Thrive in 2025",
    "url": "https://www.nasa.gov/centers-and-facilities/kennedy/nasa-kenne
    "type": "news"
  },
```

02

PREPARE MODULE

From HTML Chaos to Structured Insight

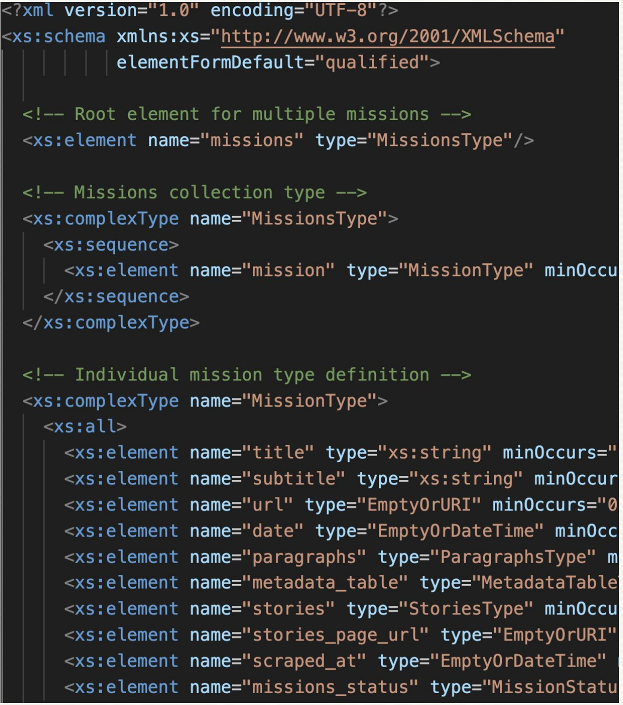
- **Data Preparation:**
  - Text cleaning + formatting
- **Data Conversion:**
  - Converted to structured XML format
  - Validated using a custom XSD schema
- **Data Storage:**
  - XML data stored in eXist-db

From HTML Chaos to Structured Insight

```
<mission>
    <title></title>
    <subtitle></subtitle>
    <url></url>
    <date></date>
<stories_page_url></stories_page_u
rl>
    <scraped_at></scraped_at>
    <paragraphs>
      <paragraph></paragraph>
</paragraphs>
```

```
    <metadata_table>
      <metadata>
        <key></key>
        <value></value>
      </metadata>
     </metadata_table>
    <stories/>
    <missions_status></missions_status>
</mission>
```

From HTML Chaos to Structured Insight

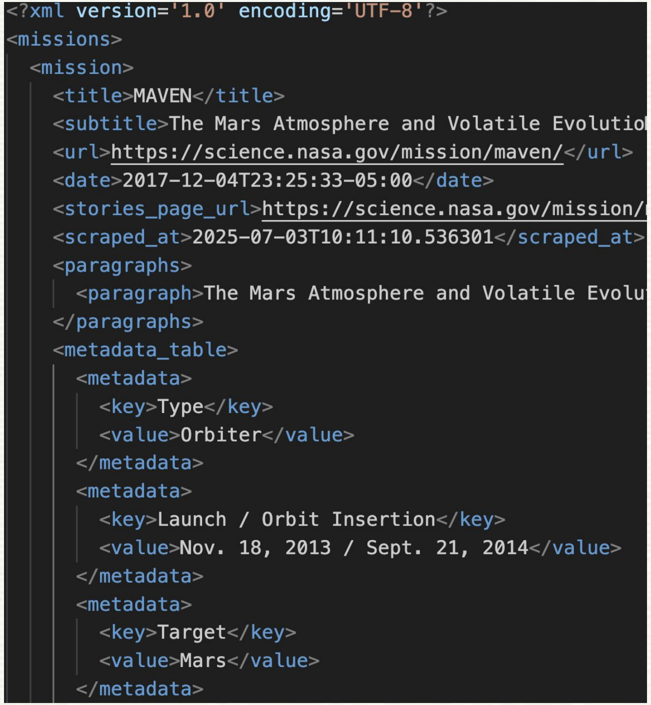


```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
           elementFormDefault="qualified">

  <!-- Root element for multiple missions -->
  <xs:element name="missions" type="MissionsType"/>

  <!-- Missions collection type -->
  <xs:complexType name="MissionsType">
    <xs:sequence>
      <xs:element name="mission" type="MissionType" minOccu
    </xs:sequence>
  </xs:complexType>

  <!-- Individual mission type definition -->
  <xs:complexType name="MissionType">
    <xs:all>
      <xs:element name="title" type="xs:string" minOccurs="
      <xs:element name="subtitle" type="xs:string" minOccur
      <xs:element name="url" type="EmptyOrURI" minOccurs="0
      <xs:element name="date" type="EmptyOrDateTime" minOcc
      <xs:element name="paragraphs" type="ParagraphsType" m
      <xs:element name="metadata_table" type="MetadataTable
      <xs:element name="stories" type="StoriesType" minOccu
      <xs:element name="stories_page_url" type="EmptyOrURI"
      <xs:element name="scraped_at" type="EmptyOrDateTime"
      <xs:element name="missions_status" type="MissionStatu
```



```
<?xml version='1.0' encoding='UTF-8'?>
<missions>
  <mission>
    <title>MAVEN</title>
    <subtitle>The Mars Atmosphere and Volatile Evolutio
    <url>https://science.nasa.gov/mission/maven/</url>
    <date>2017-12-04T23:25:33-05:00</date>
    <stories_page_url>https://science.nasa.gov/mission/
    <scraped_at>2025-07-03T10:11:10.536301</scraped_at>
    <paragraphs>
      <paragraph>The Mars Atmosphere and Volatile Evolu
    </paragraphs>
    <metadata_table>
      <metadata>
        <key>Type</key>
        <value>Orbiter</value>
      </metadata>
      <metadata>
        <key>Launch / Orbit Insertion</key>
        <value>Nov. 18, 2013 / Sept. 21, 2014</value>
      </metadata>
      <metadata>
        <key>Target</key>
        <value>Mars</value>
      </metadata>
```

03

ACCESS MODULE

Querying for Clarity with XQuery

- XML data stored in eXist-db
- Need to query questions like
  - What are the unique mission types

```
'''xquery version "3.1";
    distinct-values(
      doc("/db/martian-explorer/missions.xml")
        /missions/mission
        /metadata_table/metadata
          [key = 'Type']
          /value/text()
    )
'''
```

Rover

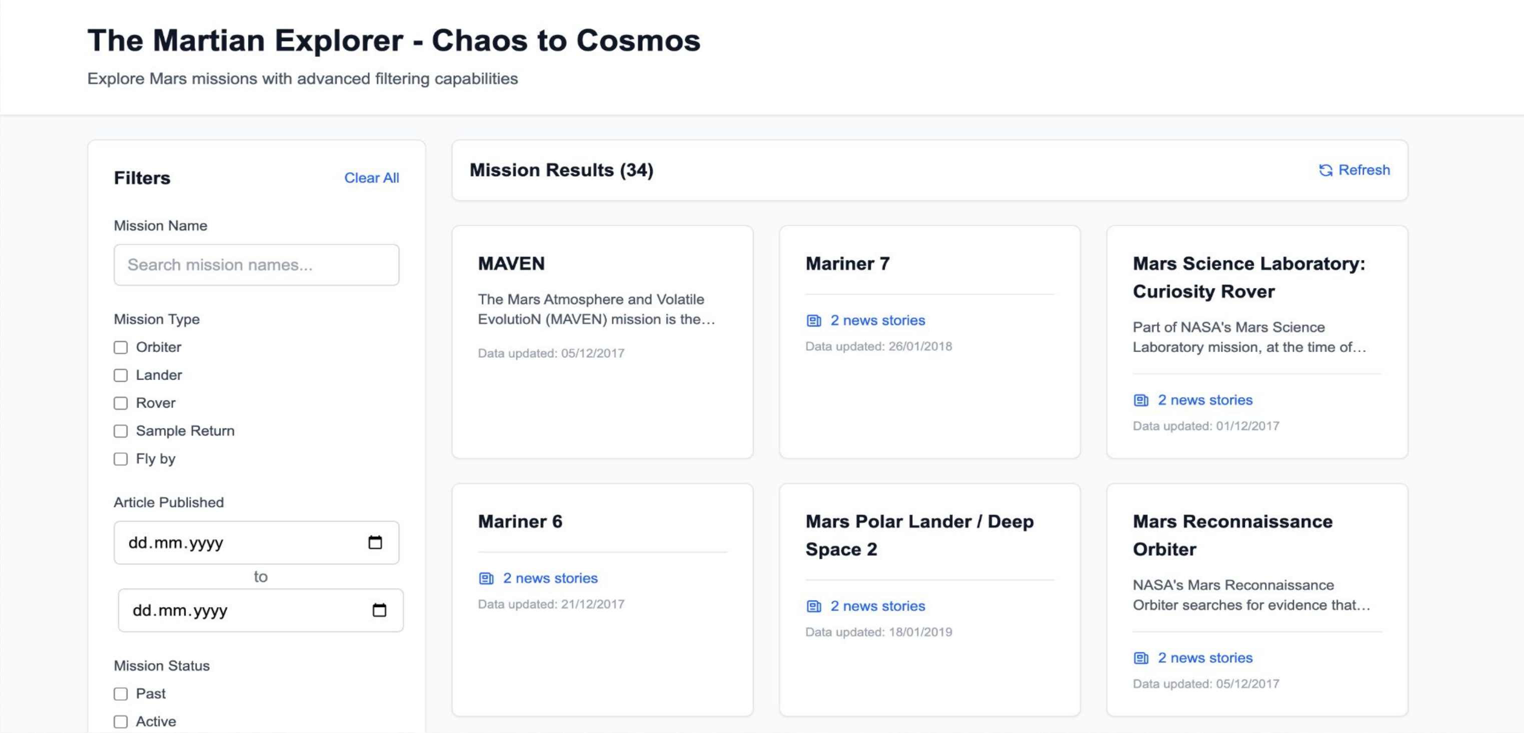
Lander

Orbiter

Sample Collector

Fly By

Querying for Clarity with XQuery



The Martian Explorer - Chaos to Cosmos

Explore Mars missions with advanced filtering capabilities

Filters Clear All

Mission Name

Search mission names...

Mission Type

□ Orbiter
□ Lander
□ Rover
□ Sample Return
□ Fly by

Article Published

dd.mm.yyyy
to
dd.mm.yyyy

Mission Status

□ Past
□ Active

Mission Results (34) Refresh

MAVEN
The Mars Atmosphere and Volatile EvolutioN (MAVEN) mission is the…
Data updated: 05/12/2017

Mariner 7
2 news stories
Data updated: 26/01/2018

Mars Science Laboratory: Curiosity Rover
Part of NASA's Mars Science Laboratory mission, at the time of…
2 news stories
Data updated: 01/12/2017

Mariner 6
2 news stories
Data updated: 21/12/2017

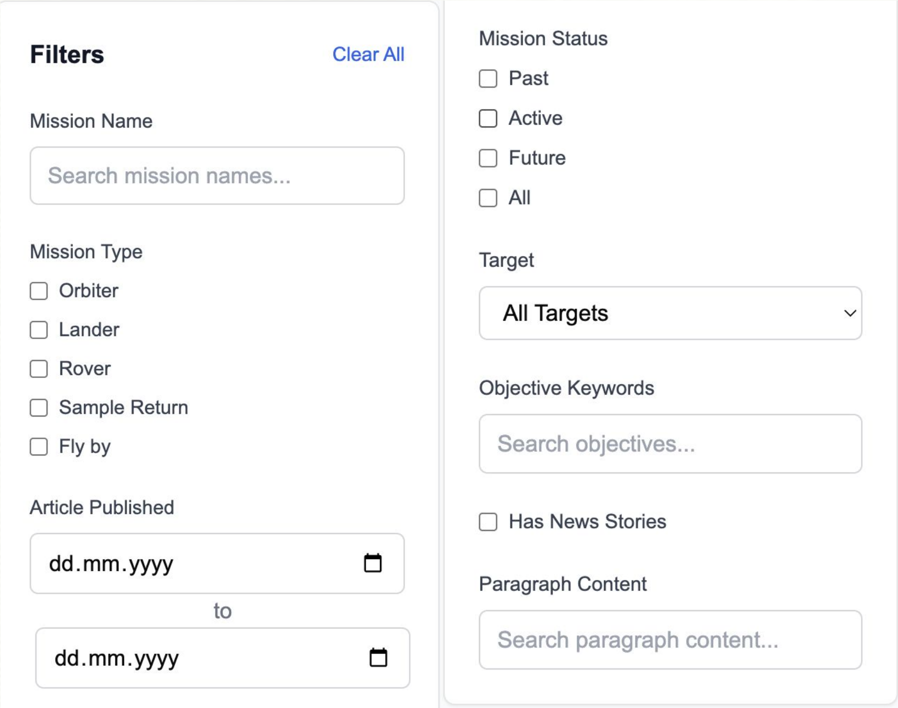
Mars Polar Lander / Deep Space 2
2 news stories
Data updated: 18/01/2019

Mars Reconnaissance Orbiter
NASA's Mars Reconnaissance Orbiter searches for evidence that…
2 news stories
Data updated: 05/12/2017

Querying for Clarity with XQuery

Features:

1. Multi boolean filtering
2. Date range pickers
3. Free text search
4. Toggle and Checkbox Controls



Filters Clear All

Mission Name

Search mission names...

Mission Type

□ Orbiter
□ Lander
□ Rover
□ Sample Return
□ Fly by

Article Published

dd.mm.yyyy

to

dd.mm.yyyy

Mission Status

□ Past
□ Active
□ Future
□ All

Target

All Targets

Objective Keywords

Search objectives...

□ Has News Stories

Paragraph Content

Search paragraph content...

Extensions

1. **XQuery:**
    - SQL-like query access to XML documents for extracting text and aggregations
    - Built on XPath
        - Tree-like document structure with simplified access

Extensions

2. eXist-DB

- designed to store, validate, and query XML documents directly.
- No need to flatten or convert into rows/tables (like in SQL databases)
- Supports XQuery, language for querying hierarchical XML data.
- Easy to:
  - Search deeply nested structures
  - Filter by tags, attributes, and values
  - Aggregate and transform XML directly

Challenges

1. **Inconsistent HTML Structures**
   a. NASA's mission status pages don't follow a strict or unified HTML format across mission feeds.
   b. Elements like <div>, <p>, and <span> vary — requiring mission-specific scraping logic.
   c. Occasional missing fields (e.g., no date or malformed titles).
   d. Data is printed in a wide variety possibilities. For instance: launch / landing date - Not before 2026, Jun. 16 2025

Solution: Scrapers with fallback parsing.

General Challenges

2. **Schema Design Complexity**
   a. Designing the XSD to allow flexibility while still enforcing structure.
   b. Early versions of XML failed validation due to missing elements or typos.

Solution: Iterated on schema, added default values, made elements optional or allow empty values

RESOURCES

NASA Mission Data Sources:

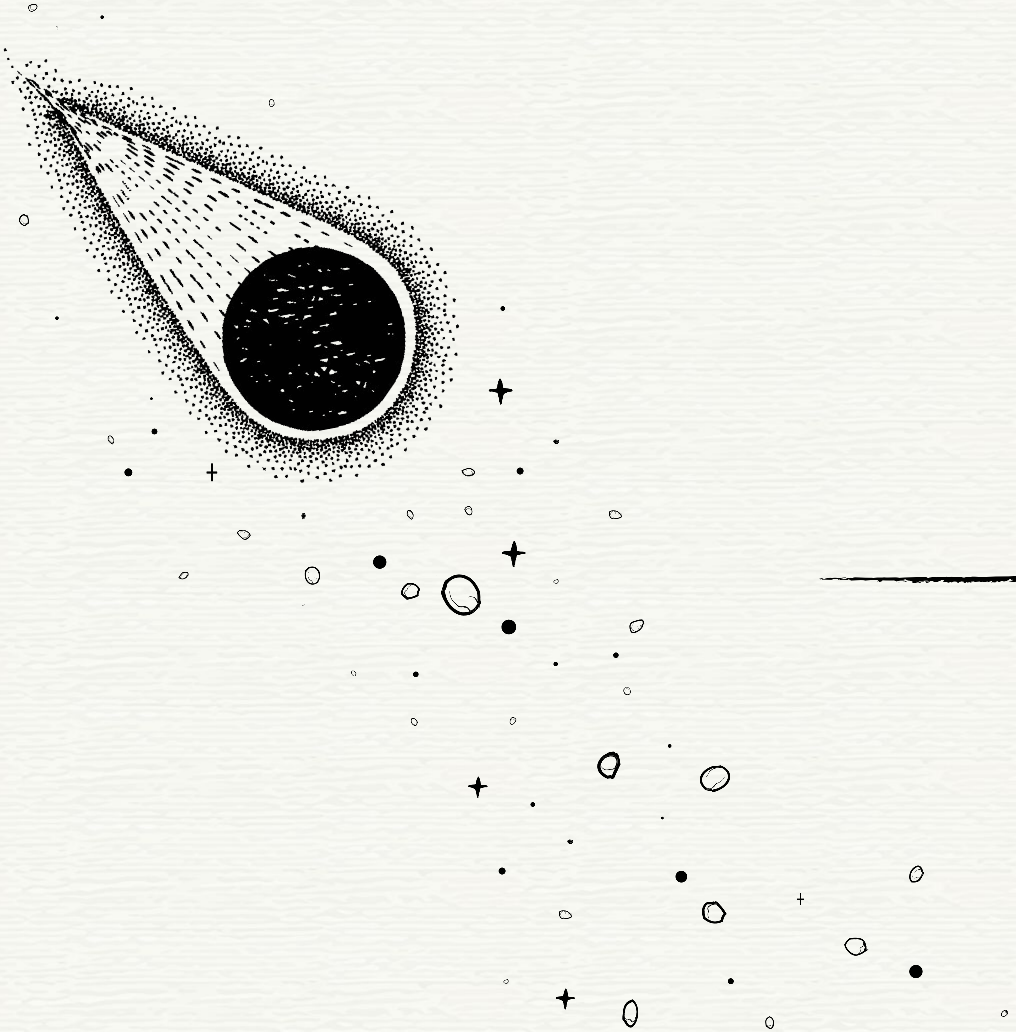
- Mars 2020 Mission Status (link)
- Maven Mission Updates (link)
- Mars Science Missions (link)

Python Libraries & Tools:

- BeautifulSoup: HTML parsing and scraping
- lxml: XML building and validation
- Next Js & React Components, Tailwind, Typescript for user interface

Find the explorer

https://github.com/Anniebhalla16/TheMartianExplorer



Q&A!