BUILDING A SPATIAL INDEX OF ORBITAL MARS SCIENCE USING MACHINE-READING APPROACHES.

D. P. Quinn[1], S. E. Peters[1], and I. A. Ross[2], [1]Department of Geoscience, University of Wisconsin–Madison (daven.quinn@wisc.edu), [2]Department of Computer Science, University of Wisconsin–Madison.

Introduction: As the volume and topical breadth of Earth and planetary science research output increases, new tools are needed to assist researchers in finding, curating, and synthesizing information from the published literature. A key need within the geosciences is to link publications to their geographical focus area, which would allow searching for previous work relevant to a particular region of interest and to result in large-scale spatial summaries of data extracted from the literature.

"Machine reading" of the geoscientific literature using cyberinformatics approaches represents one way to automatically build spatial indices. The *xDD* (formerly, *GeoDeepDive*) project is a digital library and supporting cyberinfrastructure that seeks to build new machine-assisted pathways for discovery and utilization of published scientific information [1, 2]. *xDD* holds the full text of 14.5M published articles from all domains of science and maintains agreements with key publishers (notably Elsevier, Wiley, and AGU among many others) for continual assimilation of new material. The text, figures and tables of each article are extracted and parsed using natural language and machine-learning approaches [3, 4]. *xDD* exposes an application programming interface (API) with textual search capability; this public interface is the basis for the work presented here.

Even with strong machine-reading tools, automatically extracting spatial context from papers is a multifaceted problem. Terrestrial papers refer to their locations in many ways, and efforts to extract spatial information from papers has thus far been limited. The most common approach is to correlate named entities mentioned in text with curated geospatial datasets. This approach was used to correlate stratigraphic units mentioned in scientific publications to their spatial extent [5, 6]; a similar approach was used to estimate the frequency of stromatolite occurrences in the geologic record [7]. Another approach, searching for direct references to spatial coordinates using pattern matching, has shown positive results but requires a brittle rule-based approach [8, 9]. In practice, most direct in-text mentions of locations in terrestrial papers are named locations. Even when location information is extracted, lack of knowledge of the context of a mention often inhibits the usefulness of a mention.

References to Mars orbital imagery: The planetary science literature presents an appealing target for validating approaches for establishing spatial context for the literature. Orbital planetary science research is typically based on imagery datasets that are given unique identifiers at the time of creation; these are tied to the spatial footprint of the dataset. These datasets are referenced by ID in studies of a particular area. With automated access to full-text search of the literature, joining publications with the regions they considered becomes a relatively straightforward text-matching exercise.

We used the PDS Mars ODE dataset coverage shapefiles [10] to assemble a list of unique dataset IDs against which to match information extracted from publications. These image IDs (typically strings of a form similar to "ESP_012601_1400"), were fed to the xDD snippets API (https://xdd.wisc.edu/api/snippets), which returned a list of mentions in papers indexed by the system, typically in body text and data tables. The full 14.5M-document corpus indexed by xDD is queried for each image. Over 200,000 image IDs for the HiRISE, CTX, and CRISM instruments on the Mars Reconnaissance Orbiter (to November 2020) were checked, and results were stored in a PostgreSQL database for further analysis. We attempted to also include HRSC images, but they are typically referred to by an integer orbit number that is not unique enough to be effectively retrieved using simple string matching. Other older data products (e.g., OMEGA, THEMIS, and Viking MDIM) are referred to inconsistently in the literature and are not included at this stage of analysis.

Results: We found literature mentions of 6,845 images (~3.5% of the total archive), with a total of 9,257 appearances in the literature across 980 papers. 4,321 HiRISE, 1,536 CTX, and 1,078 CRISM images (~6.5% of the HiRISE and ~5% of CRISM data archives) are mentioned in a publication. The average image has 1.35 mentions, and papers mention 9.5 images on average. The most-mentioned single dataset is HiRISE image PSP_001513_1655, which was mentioned in 18 separate publications.

The publication–image links assembled by these literature text searches were correlated with the extent of each image footprint to build a set of spatial coordinates linked to each paper, which form a rough proxy for the area on which each scientific study is based. Evaluation of the spatial patterns of images mentioned in individual studies is ongoing, but these range from tight geographic clusters for regional geological studies to widely scattered references for global geomorphic and mineralogic surveys and engineering calibration studies.

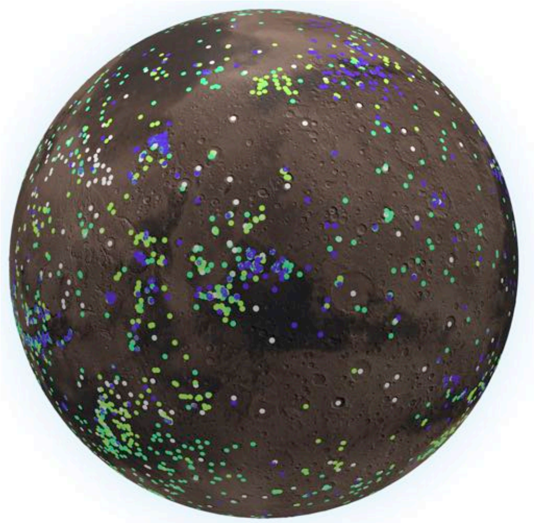

Figure 1: Links between publications and named spatially-resolved features on Mars. Purple: CRISM imagery; green: HiRISE and CTX imagery; and gray: crater names.

Crater names: Many studies of Mars also establish location context based on named landform features. These names are less unique than PDS image IDs, resulting in a "named entity recognition" problem that has parallels to location recognition against the terrestrial literature. On Mars, the maintenance of a standardized nomenclature over the entire planet by the International Astronomical Union greatly simplifies the extraction of unique spatial matches. We conducted a text matching between publications and crater names using the USGS / IAU planetary names database [11] to search the xDD snippets route for strings such as "Lyot Crater." Although we focused only on craters, matching required more manual intervention than for PDS image IDs: to avoid cluttering the result set with papers outside of Mars science, we omitted several craters that have hosted landed science missions (Gale, Gusev, Jezero, and Victoria), as these locations are often referred to in astrobiology and analog studies. A few crater names are non-unique across planetary bodies, with a few lunar craters (e.g., "Copernicus") and some volcanic edifices on Earth sharing the same name as Martian craters.

Results: We searched for publications matching 1122 crater names from the USGS database. Of these, 483 (~43%) were mentioned in the literature. 5,645 mentions of craters were matched in the literature, with the most common matches being Endeavour crater with 334 mentions and Eagle crater with 252. Both of these craters were visited by landed missions. The next few craters by reference count were participants in landing-site selection process. In general, the most-referenced craters in the literature are mentioned by far more papers than PDS image IDs. This suggests that some studies using common datasets in these regions do not mention specific source images, and that some craters are only mentioned in passing for comparative purposes. Further work will be required to distinguish between these scenarios. 3,071 unique publications were linked to locations based on crater names, and 3,434 publications across both crater names and PDS image IDs.

Discussion: The recovery of links to thousands of publications suggests that we were able to effectively find a sizable portion of the Mars orbital literature based on text matching from public datasets (PDS image IDs and crater names). The larger number of publication matches returned by crater name matching reflects the more natural usage of these terms in descriptive text; matching by PDS image ID requires the listing of these resources in the main body of the publication.

These spatial indices are immediately useful for searching for literature relevant to a specific spatial location on Mars. Going forward, the Mars datasets assembled here can also be used to train machine-learning analysis of terrestrial location names. They can also underpin further automated spatial analysis. For instance, searching for mentions of "sulfates" across the papers that mention PDS images returns 395 unique publications that may tie these deposits to individual spatial locations. With additional inference and machine learning, such mappings can be used to build global indices of geologic phenomena.

Recommendations: Current data citation practices for planetary orbital imagery are relatively well-adapted to machine-reading approaches compared to Earth observation. However, we advise several guidelines for how image IDs are constructed and mentioned in text. Imagery datasets should have highly unique IDs and be mentioned in body text, tables, or appendices. Supplementary material is not indexed by xDD and images referenced from there will be missed. As more research is conducted atop composite imagery overlays [e.g., 12], new mechanisms (such as quadrangle unique identifiers) should be created to ensure that scientific studies remain mappable to spatial locations.

References: [1] Peters SE et al. Eos. 2017;98; [2] xDD. https://xdd.wisc.edu; [3] Manning C, et al., Association for Computational Linguistics 2014., doi:10.3115/v1/P14-5010; [4] Goswami A, et al. arXiv:1910.12462. http://arxiv.org/abs/1910.12462; [5] Peters SE, et al. G^3. 2018;19(4):1393–1409; [6] Macrostrat. https://macrostrat.org; [7] Peters SE, et al., Geology. 2017;45(6):G38931.1. doi:10.1130/G38931.1; [8] Marsicek J, et al., AGU Pages. 2018;26(2):70–70. doi:10.22498/pages.26.2.70; [9] Goring S, et al., EarthArXiV. 2021, doi:10.31223/X54312; [10] PDS Mars ODE coverage shapefiles, https://ode.rsl.wustl.edu/mars/coverage; [11] IAU planetary names database, https://planetarynames.wr.usgs.gov; [12] Dickson et. al, LPSC 2020 # 2309.