CriticalMAAS Milestone 3 report – Macrostrat TA4 team

Daven Quinn and the UW–Madison / Macrostrat CriticalMAAS team

February 27, 2024

Contents

Introduction

This **Milestone 3** report covers UW-Madison – Macrostrat contributions to the CriticalMAAS project for the period from December 21, 2023 through February 23, 2024 (Months 4-6). We describe progress during that period towards the research and technical goals defined in our Phase 1 Research Plan and Milestone 2 code/documentation release, gaps in addressing the program goals, and the results of baseline integrations and capability demonstrations from the CriticalMAAS Month 6 Hackathon (*Denver, CO*).

Our approach to CriticalMAAS is based around Macrostrat, a principled, analysis-oriented geologic framework to harmonize maps and other geological knowledge products (e.g., stratigraphic columns) into a descriptive model of crustal rocks. This system is already accessible via stable, performant, and open web services that support wide use, with geologically-oriented user interfaces that allow exploration. We are enhancing this system and tuning it for use in CriticalMAAS by adding new data ingestion capabilities and tools for feedback and expert contribution. Additionally, we seek to incorporate literature-connected machine learning pipelines to augment this system with structured data (e.g., rock unit lithologies) that can further inform querying and analysis tasks specific to CriticalMAAS modeling requirements. We have made significant progress in each of these domains.

1 Research and technical progress: Macrostrat geologic framework

Our core task for CriticalMAAS is to develop pipeline to integrate TA1 mapping data into a geologic data product that can be used by TA3. Our work has largely proceeded as laid out in the Phase 1 Research Plan, and we are midway through the development of a system that will support rapid map ingestion and standardization. At the **Month 6 Hackathon**, we demonstrated an end-to-end pipeline for ingesting geologic mapping from TA1 into Macrostrat, and providing a standardized representation of those datasets to TA3 for critical minerals modeling. This represents the initial demonstration of the system integrations and data throughput required for the success of CriticalMAAS.

By the end of Phase 1, we are committed to delivering a complete, runnable Macrostrat map curation system to USGS; we are on track to do so. The prototype system is housed in the `UW-Macrostrat/macrostrat` repository, which integrates formerly separate components into a unified, open-source representation of Macrostrat's database schema, map ingestion pipelines, and geological data integration tooling.

The successful end-to-end demonstration of this system relied on several areas of development that we have pursued over the last several months; we detail the progress in each of these areas below. In the coming months, we will streamline and extend this pipeline to support quick, accurate capture of geologic mapping datasets (Sec. 3.1) and ingestion of more maps into the system (Sec. 3.2).

1.1 TA1 GeoPackage library

Macrostrat entered the CriticalMAAS program with a well-established understanding of the most effective structure for geologic map data, and we have been working to convey this understanding to TA1 teams. Following our success in leading the design of TA1 data schemas, we have led the effort to establish specific formats for TA1 data products, in order to establish authoritative, consistent standards for candidate map interchange and archival.

The `DARPA-CriticalMAAS/ta1-geopackage` Python library provides a common interface for TA1 to deliver geologic maps to Macrostrat. The library wraps the GeoPackage format, which is a recognized "best-practice" standard for geospatial data delivery. We additionally bake the TA1 schemas into the file format, to ensure that data types and references are consistent between teams. This schema is supplemented by a set of utilities based on `fiona` and `geopandas` Python libraries to write data into the format. The resulting file contains all data that the TA1 teams generate for a given map, and it can be opened natively in QGIS, ArcGIS, and other GIS environments. We have designed Macrostrat ingestion/export scripts to work with the file format as well.

We conceived of this direction just before our **Milestone 2** and began socializing the idea with TA1 teams in December 2023. The library was developed and refined in early January. Integration with TA1 teams started in mid-January, and by the **Month 6 Hackathon**, all TA1 teams had shifted to using this file format to deliver candidate maps to Macrostrat via our S3 buckets; ~90 maps have been provided and ingested into Macrostrat's data pipelines (Sec. 1.2). At the moment, TA1 teams have implemented the schema to varying degrees, and we have been working to support their full compliance with the format. The library is now under collaborative refinement with TA1 workers, who have contributed bug reports, test cases, and pull requests; it will likely underpin internal integration between TA1 teams and storage of TA1 data in the CDR.

1.2 Map ingestion pipeline

Macrostrat's map ingestion pipeline allows geologic maps from a variety of sources to be integrated into a harmonized representation, which is critical for leveraging mapd data in regional and continental-scale mineral assessments. In CriticalMAAS, we have committed to extend this pipeline to rapidly integrate high-resolution, high-quality geologic maps from TA1 and other sources (Sec. 3.2), thereby applying standardized representations of age, lithology, and named geological units and bringing them into a unified geological representation to support critical minerals modeling. We have made significant progress on streamlining this map ingestion pipeline to support the rapid ingestion of geologic maps, and have staged almost 200 maps into the system. Notably, at the **Month 6 Hackathon**, we successfully ingested several TA1 maps fully into Macrostrat's harmonized web representation, making them available alongside existing maps in Macrostrat's API and web interfaces.

Since **Milestone 2**, we have greatly updated the infrastructure supporting Macrostrat's map ingestion pipeline. The pre-CriticalMAAS system was designed around expert-led ingestion in a manual, ad-hoc fashion, relying on direct database inserts and custom SQL scripts for each ingested map. This system was effective for ingesting a few dozen maps per year (Macrostrat's recent ingestion rate) with a high degree of control, but it was not scalable to the thousands of maps that we expect to ingest over the course of CriticalMAAS; it also only minimally tracked data provenance. We are midway through retooling this system into a more robust, scalable, and automated pipeline that includes storage buckets for candidate maps (TA1 GeoPackages and other vector datasets), automated ingestion scripts, and a repeatable HITL workflow for legend correction and standardization. The core of this pipeline is being built in the `UW-Macrostrat/macrostrat` repository; it is supplemented by source-specific scripts for acquiring potential vector maps. Going forward, we seek to integrate our map acquisition process with Jataware's web-scraping approach (Sec. 3.2).

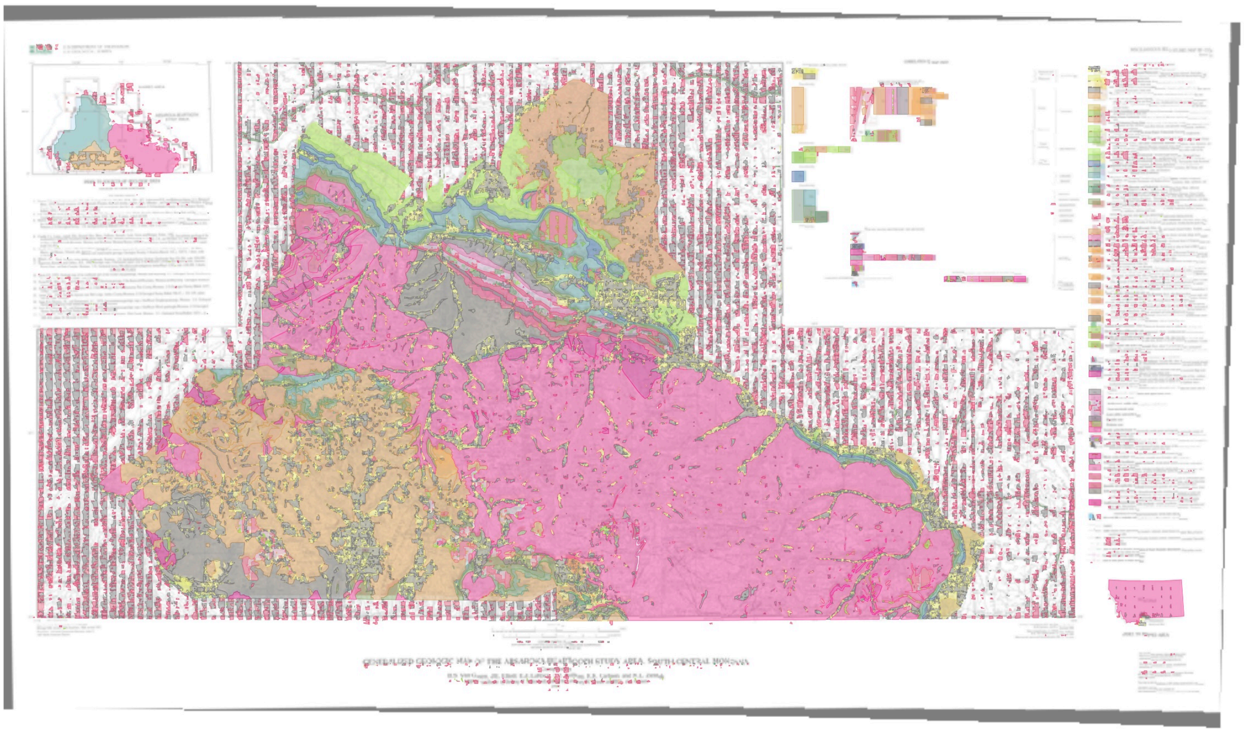
We have validated the Macrostrat map ingestion pipeline both for staging new datasets into the system and for HITL curation and assimilation into our standardized data services. Broadly, map staging is now rapid and automated, and there is a bottleneck in HITL legend curation that we are working to address (Sec. 3.1). The new web scraping/ingestion pipeline has been tested on several collections of maps spanning a range of input data formats and levels of completeness:

- 95 maps from the Nevada Bureau of Mines and Geology (NBMG), including 84 with Shapefiles that we could import
- 9 maps from the National Geologic Map Database (NGMDB) that had GIS data linked directly from their main page (6 had shapefile data that we could import)
- 93 maps (including partial/incomplete datasets) submitted by TA1 performers UIUC, UMN/Inferlink, and Uncharted during the 6-month hackathon, covering maps relevant to the "MANIAC" magmatic nickel challenge

The successful bulk staging of maps is a significant step for Macrostrat's infrastructure. The 210 maps staged since the beginning of February represent about 70% of the total number of maps ingested into Macrostrat from 2016-2023 (294). Although most of these new maps are not fully through the legend curation pipeline, their ingestion represents a substantial increase in the rate of Macrostrat's ability to assimilate new mapping data.

The hackathon provided a valuable opportunity to "stress-test" this ingestion system, and we verified that the basic structure of the pipeline is sound under load. The system can handle both maps provided in a controlled way with a standardized schema (e.g., TA1 datasets uploaded to a S3 bucket) and other map formats that are not yet standardized (e.g., vector geologic maps in NGMDB with varying attribute schemas). However, we also identified several areas that need substantial improvement. The pipeline, as currently implemented, has far too many human touch points to be effective when well-packaged maps are coming in quickly. Some steps need more attention to data pipeline design others require more interaction with TA1 to ensure schema compliance. In particular, the HITL legend-curation step, which is managed in a table-based web interface, is unwieldy and slow due to both the complexity of the task and poor user interface design. We are working to address these issues in the coming months.

Despite the slow pace of HITL curation, we were able to demonstrate the ingestion of maps from TA1 into Macrostrat, successfully staging several maps produced during the hackathon into the system. Once data-format concerns were addressed with TA1 teams, the ingestion process was straightforward and quick, though it required a significant amount of manual intervention. Although there are clear gaps to be filled, successful demonstration of our map ingestion system represents a major proof of concept for our overall approach to integrating mapping data.



Hackathon result: *TA1 map extraction atop Jataware-indexed raster map in Macrostrat's web user interface*

1.3 Macrostrat geologic exploration interface

Once ingested into Macrostrat, maps become part of a user-facing, web-available cyberinfrastructure designed for data distribution and analysis. This system is already widely used by geologists and the public for rapid exploration of geologic information in rich web and mobile user interfaces. As part of CriticalMAAS, we have been working to extend our interfaces with additional capabilities for geologic data exploration relevant to mineral systems (see the Macrostrat development website).

Key capabilities that were discussed in the **Milestone 2** report (e.g., single-map views and paleogeographic reconstructions) continue to be developed. One small but important addition is hooks to Jataware's raster storage system, allowing georeferenced Cloud-Optimized GeoTIFF maps to be displayed alongside their vector equivalents in the Macrostrat web interface (See example map). We have also made substantial progress on new visualizations for our geologic lexicon (Sec. 2.3) and, most notably, **stratigraphic columns**.

Macrostrat's stratigraphic column dataset is a rich index of the subsurface rocks of North America; it is unique in its scale and consistency and provides another "view" of geologic infor-

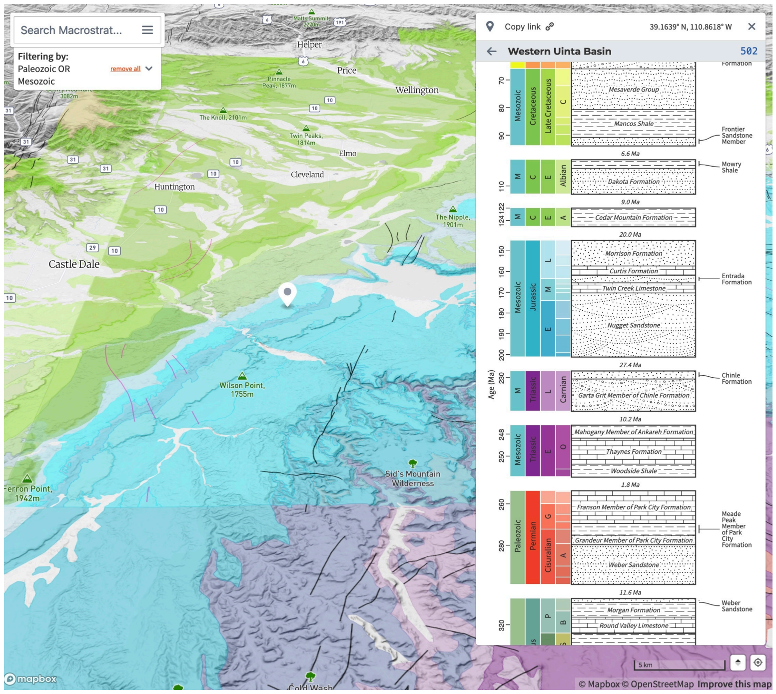


Figure 1: Stratigraphic column visualization in Macrostrat's web interface

mation about the crust alongside maps. This dataset is queryable via Macrostrat's API, but a lack of straightforward visualization has limited its usefulness for exploration and modeling. Macrostrat's column visualization, now publicly available at `dev2.macrostrat.org/columns`, allows more intuitive and geologically sophisticated interaction with Macrostrat's column dataset. Columns can be explored alongside the geologic map interface, and navigation between datasets will be made more explicit in future iterations of the tool. In the future, this new capability can support the integration of geologic maps as well. Since geologic maps also have a time-stratigraphic element alongside their spatial footprint, stratigraphic visualizations can provide a useful capability to evaluate the quality and consistency of TA1 maps and legend extractions.

1.4 Providing TA1 geologic maps to TA3

Macrostrat's geologic map APIs support standardized querying of our geologic maps and present a stable target for integration with TA3. Importantly, these APIs are designed to work similarly for both single maps and the Macrostrat "harmonized" map, which is a composite from maps at multiple scales. Thus, the TA1 maps that we have ingested into Macrostrat's system can be accessed in a straightforward way. MTRI has built a library for querying Macrostrat that allows our tile-based API output to be straightforwardly brought into a QGIS environment and "re-merged" into a coherent product, which reduces some of the complexities of working with streaming tiles. This capability was tested early in the program, but we validated it during the hackathon by pulling a collection of rocks matching "mafic" lithologies into QGIS using MTRI's tools. This mostly went well, validating the basic

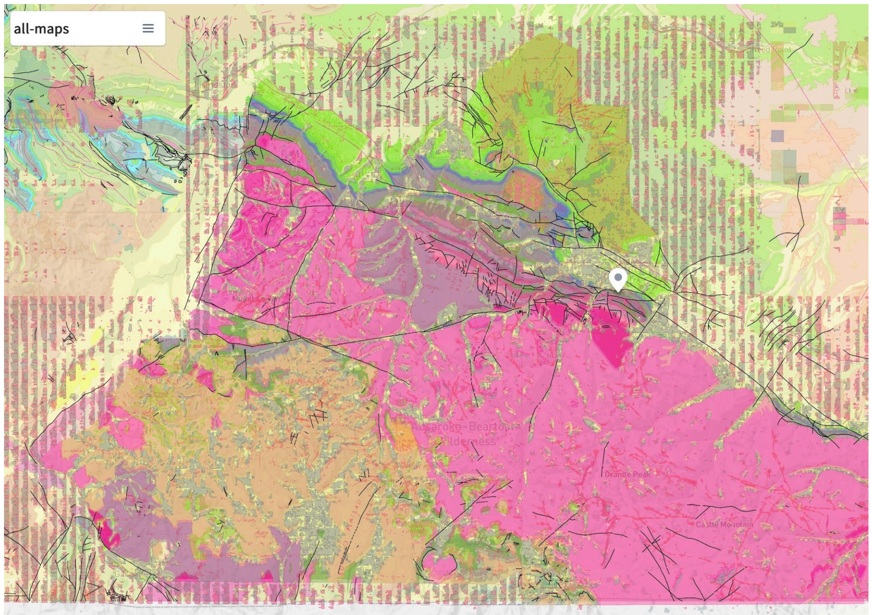


Figure 2: TA1 output in the context of other maps in Macrostrat's web interface — including an already-ingested vector dataset for the same map.

capability of the system to support TA3 querying. Although improvements must be made in efficiency and specificity of filtering, the basic capability in place is the final piece of an end-to-end system for TA1-3 integration, a significant milestone for Macrostrat's role in CriticalMAAS.

2 Research and technical progress: xDD literature integration

2.1 The CriticalMAAS Document Store

Published documents of various types, ranging from USGS Professional Papers to mining company reports, contain information pertinent to critical minerals assessments and geological rock unit characterization. CriticalMAAS workflows, therefore, require a solution for finding, retrieving, and processing relevant documents for any given mineral system. We realized prior to **Milestone 2** that a clearer delineation was needed between efficient and effective search/discovery of documents, which xDD provides, and the deep access to document content that allows **TA2** to test ML tools and HITL workflows. This distinction is particularly salient in CriticalMAAS, which relies largely on documents that are freely and openly licensed unlike most of xDD's corpus.

The needs of CriticalMAAS presented an opportunity to build a new system that can complement xDD's capabilities for openly licensed documents. As such, starting in December 2023, we started building the **CriticalMAAS Document Store** (`UW-xDD/document-store`) to provides full-page/document services operating over open-licensed documents managed by CriticalMAAS. It operates alongside existing xDD infrastructure for indexing and retrieval, and will be geared to supporting annotation and feedback interfaces, including Jataware's *Silk* document annotator, which already can read the Document Store APIs.

The Document Store is designed to be a standalone CDR component that can be used independently. Integration with xDD confers several advantages: The Document Store is "seeded" with a set of documents that xDD has been working with the USGS to acquire, store, and index; xDD's search APIs can be used to navigate this set. However, the only information from xDD that we depend on in the document store is title and DOI; in principle, other knowledge curation systems (e.g., the TA2 knowledge graph) can be integrated with the Document Store along similar lines.

Currently, the Document Store relies heavily on xDD for metadata search and document discovery. Title-based search and retrieval is currently possible, but more complicated search tasks fall back to xDD. If a need is demonstrated for a truly independent and fully searchable collection of user-contributed program documents not indexed by xDD, certain xDD services (e.g, ElasticSearch) can be brought over into the Document Store context.

2.2 Document entity extraction

In addition to providing a baseline for program storage of documents, xDD provides services over documents that can be valuable for CriticalMAAS. In particular, the COSMOS pipeline, which identifies and contextualizes document entities (e.g., figures, tables, equations, abstracts), has been successfully used for tasks such as recalling all tables from documents that mention a specific element, or for finding geologic maps relating to surveys for specific minerals (*see figure*). This capability was used during the hackathon to rapidly forward maps related to Nickel to TA1 for ingestion. We are working to run this pipeline over all CriticalMAAS-relevant documents to provide an searchable index of document extractions. When combined with the Document Store, these extractions can "pre-seed" TA2 data curation UIs and pipelines with relevant snippets of documents, allowing efficient workflows to be built. At the hackathon, we planned such an integration with Jataware (user interfaces) and SRI (machine-learning pipelines); we will stand up APIs to provide extractions in the Document Store framework in the coming months.

In addition to being useful within TA2 pipelines, xDD and COSMOS were demonstrated as user-facing tools for document search and discovery. We received valuable feedback from USGS scientists on the system, including the need for more search capabilities for closely grouped terms, more API documentation and usage examples, and improved labeling of maps and other figure types. We will work to assimilate these suggestions over time to make xDD a more useful service for discovering literature related to critical minerals

2.3 Geologic unit characterization

As in most geologic map datasets, structured data about Macrostrat rock units is mostly limited to basic lithologic classifications that only record major components. This gap is important for mineral systems modeling, since mineral occurrences are correlated with infrequent

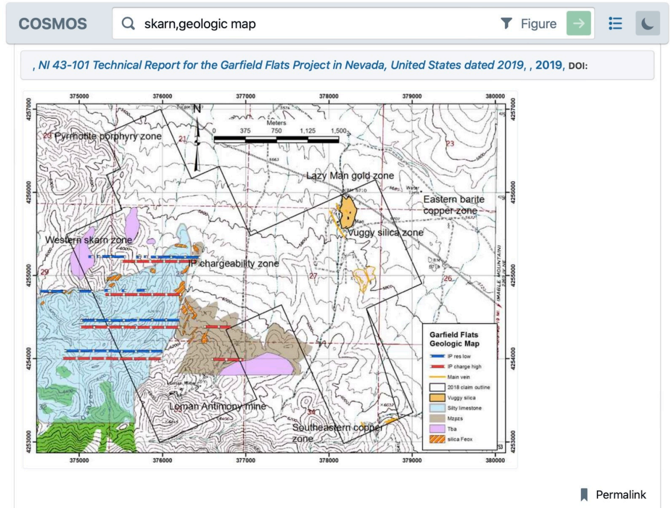


Figure 3: COSMOS UI for surfacing geologic entities from the scientific literature

features that do not show up in "bulk" descriptions of map units (e.g., localized veins or small pegmatite bodies). Geologists typically solve such problems by hand-labeling datasets for specific problems; this was done for the State Geologic Map Compilation (SGMC) project; Lawley et al, 2023 reports that assembling the necessary lithologic classification took months of overhead even using only regional-scale maps. To extend such an effort to local maps at the scale desired by CriticalMAAS would be infeasible.

We seek to address this data limitation in a more automated way with a pipeline to acquire lithologic information about specific rock units from the geologic literature using mentions in xDD's corpus. Two approaches are being developed in parallel: an LLM-based approach (`UW-Macrostrat/factsheet-generator`; *Bill Xia*) and a knowledge-graph curation approach (`UW-Macrostrat/unsupervised-kg`; *Devesh Sarda*). Since **Milestone 2**, we have built infrastructure supporting "end-to-end" extraction of lithologic descriptors for rock units and presentation as candidate relationships in Macrostrat's "lexicon explorer" user interface. Forthcoming feedback tools (Sec. 3.4) will allow this dataset to be augmented by expert geologists; if successful, this line of work will allow descriptions of the geologic units to be automatically assembled from geologic reports and papers. Having successfully demonstrated this pipeline, we will continue to improve the quality of extractions and their integration with Macrostrat's data services and user interfaces.

3 Gaps

Although we have made substantial progress on building infrastructure and systems to support program needs, several major gaps in functionality have become apparent as our systems underwent stress-testing in the **Month 6 Hackathon**.

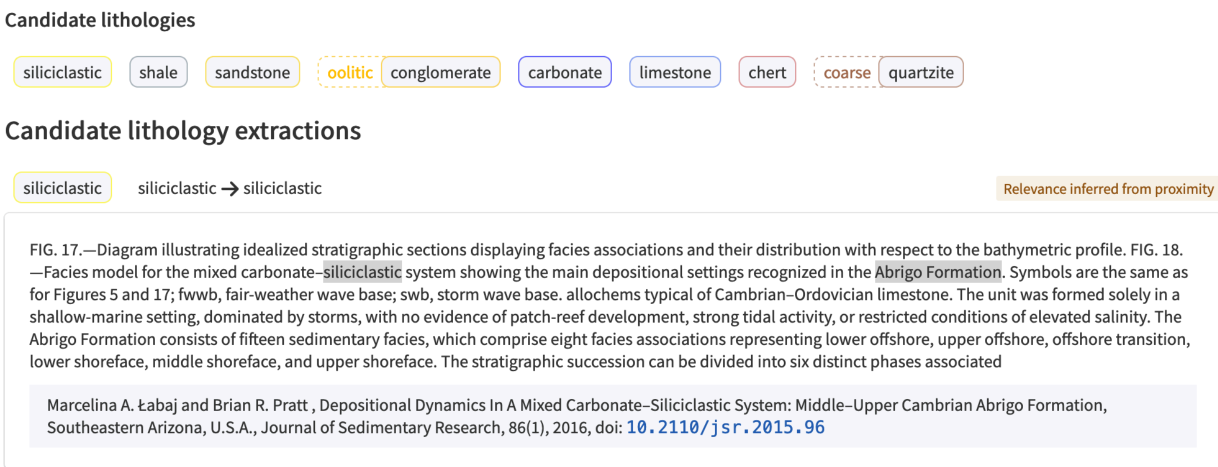


Figure 4: Candidate lithology extractions from scientific literature, represented in a prototype feedback interface

3.1 Map ingestion throughput

At the **Month 6 Hackathon**, the standardized data format and TA1 focus on producing results allowed us to stage a large number of maps into Macrostrat's ingestion pipeline. However, the throughput of this pipeline was not sufficient to quickly process TA1 maps into forms that could be evaluated visually, integrated with other mapping, and used by TA3. Some bottlenecks, such as inefficient scripts, have already been addressed, but the major problems remain an too many human touch points for each map dataset, the lack of reporting and metrics within our pipeline, and an inefficient HITL legend-curation interface. Our key goal for the immediate future is to retool the pipeline to expose a well-defined set of mostly-automated tools to ingest, process, and update map datasets. Ideally, much of the ingestion of TA1-provided maps should occur without human intervention, and the system should be able to handle a large number of "candidate" maps at various stages of assimilation. Although the core task has been demonstrated, achieving a reliable throughput and simple process is critical to ensuring that map curation is straightforward and can be taken over by USGS staff.

3.2 Accessing vector maps

Map ingestion is not only relevant for TA1 datasets: the USGS and other organizations (e.g., state and international geologic surveys) maintain large archives of vector mapping, both for modern "born-digital" maps and for older maps that have been manually digitized (often at great expense and a high level of quality). Importantly, since they were either produced recently or targeted for expensive re-processing, these maps are often among the most important to access for mineral modeling — a "best-in-class" analytical capability should incorporate the best modern mapping to the degree possible. This potential gap has been on our radar since the start of CriticalMAAS, but gained salience during the **Month 6 Hackathon**, when one of the maps ingested by TA1 as a "high-value target" for Nickel was actually already in Macrostrat, as part of a 2005 USGS compilation.

Hurdles to accessing vector maps appear to be largely organizational, with NGMDB unable or unwilling to provide a centralized, direct access point to their datasets. Given the state and fragmentation of USGS systems, these maps cannot be accessed straightforwardly. We have started working on approaches to access these maps via public web pages, but progress has been episodic. One possible path forward is to integrate with Jataware's map scraping approach, particularly taking advantage of their augmentation of the USGS's map viewer. Working with Jataware to develop an approach to vector map access will ensure that two teams are not duplicating tedious web-scraping work. We may also elect to expose a map upload tool for USGS staff and others to directly contribute.

3.3 TA3 query workflow

Macrostrat's standardized API allows querying both a harmonized, multiscale geologic map and single maps, which was an acceptable initial target for providing geologic map layers to TA3. As originally conceived, based on the methods of Lawley et al, 2023, TA3 would query Macrostrat for a categorized representation of the geologic map, and all categories would be used as layers in the mineral modeling process. However, TA3 has elected to go a somewhat different route of pre-filtering the geologic map to high-potential rock types, and using only those polygons for modeling. Both of these approaches are principled and fundamentally compatible with Macrostrat's system design; however, because the current filtering approach did not come into focus until recently, we have not invested much time in developing the user experience of querying Macrostrat for map subsets. Several imperfections need to be addressed: the Macrostrat tile APIs do not support provider-side filtering and API calls cannot be efficiently batched to "build up" a complex query. The latter problem will have to be solved in conjunction with MTRI's Python library and QGIS tools. By Month 9, we will demonstrate and document this query process to allow arbitrary slicing of Macrostrat's dataset, easily and with provenance tracking.

In addition to these gaps in API design and documentation, we need to adjust and simplify our lithologic vocabularies to better match implicit relationships. In the current version of our APIs, rocks can be missed if they are tracked with more specific terms than the user is querying for. We will construct a querying approach that navigates hierarchical lithologic vocabularies to ensure that matching is intuitive. These improvements will be further augmented by automated entity extraction (Sec. 2.3).

3.4 Feedback interfaces

One major gap in our current system is the lack of tools to provide feedback over geologic data being accumulated in the system, in particular those datasets being ingested from TA1 (Sec. 1.2) and accumulated by geologic entity characterization (Sec. 2.3). We need to start building and exposing feedback tools to allow data augmentation and correction. We have many of the infrastructure needed to build such tools (e.g., login systems, data visualization

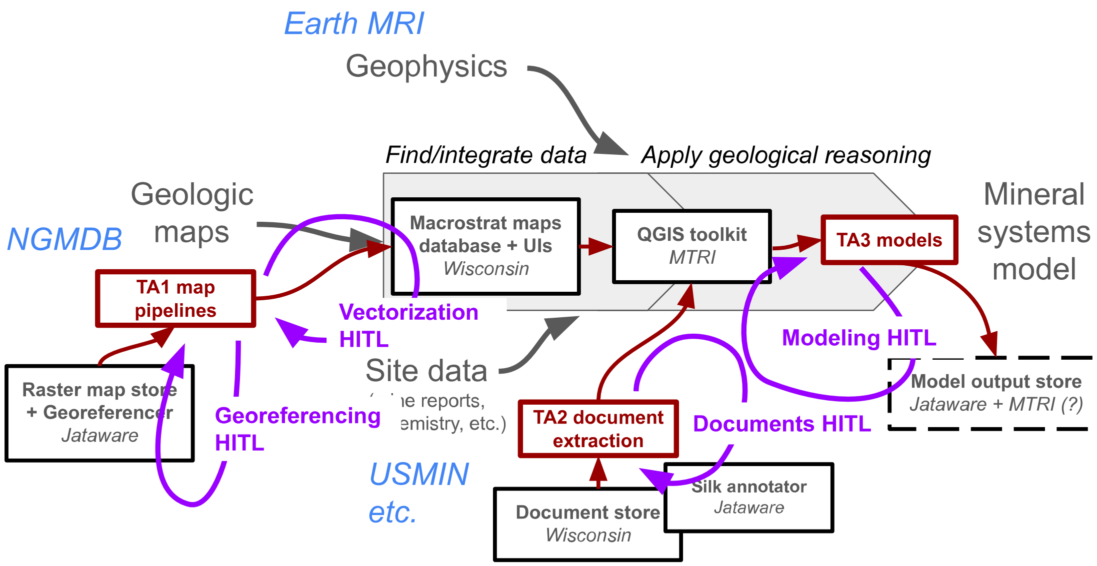


Figure 5: CriticalMAAS program flow chart showing software (black boxes), data management workflows (red), and HITL workflows (purple) being built by TA4 teams

interfaces) but we have not yet built specific UI "widgets" for feedback. The one exception is the table interface for map curation, but this is not yet ready for external testing. Map feature correction interfaces are possible with the Mapboard GIS app, but we need to understand the structures and failure modes of TA1 datasets before committing to specific approaches for feedback; after getting TA1 maps into the system at the **Month 6 Hackathon**, making these plans is now possible. Similarly, now that TA4 has access to TA2 mineral sites we will be able to build feedback interfaces in the coming months. For HITL tools supporting both TA1 and TA2, we will coordinate closely with Jataware to ensure that we are not building redundant systems.

4 Issues and concerns

4.1 Design requirements for program integration

One of the main goals of TA4 is to facilitate data interchange across the CriticalMAAS program. Indeed, building HITL tooling that links such a disparate set of performers requires careful attention to integration design, although such "data-pipelining" activities are mostly implicit in the BAA. The "CriticalMAAS Data Repository" (CDR) plan attempts to make data and storage requirements more explicit, but the specific capabilities and design expected for CDR systems has been made clear. Some aspects of the CDR plan seem geared towards satisfying reporting requirements and evaluation (driven by DARPA/MITRE), while others are more focused on technical needs, in particular for data searchability/web accessibility, that seem more geared towards USGS end users.

Since the start of the CriticalMAAS program, the UW-Macrostrat team has led development several infrastructure components oriented towards establishing program-level shared capabilities. In our view, these explicitly respond to *integration* goals (i.e., addressing needs outside of our own tools). We have sought alignment with Jataware and other performers to ensure that we were not producing duplicate functionality. At **Milestone 2**, we established the broad need to work in these directions, received buy-in from other teams, and commenced work.

- `DARPA-CriticalMAAS/ta1-geopackage`: a GeoPackage-based data format for validating and storing TA1 output (Sec. 1.1)
- `UW-xDD/document-store`: A supplemental store for public/user provided PDFs that provides full-text access, integrates with xDD APIs (Sec. 2.1)
- The Macrostrat system itself has "CDR-like" capabilities (e.g., persistent storage, high availability, and web-based data search and access); parts of our goals for the system are explicitly focused around being the center point of TA1–TA3 integration (Sec. 1.2)

Broadly, we have been successful in building these components and scaffolding data-integration workflows around them. Taken together, these systems represent significant time spent on building shared backend services by the UW–Madison team. However, given the lack of consistent communication from DARPA about the CDR, there seems to be a risk that we will fail to meet specific goals of program leadership, forcing a disruptive late-stage change in work plan.

A shared TA4 goal for the next phase of work is to solidify integration plans and scope final CDR functionality; we hope to develop these plans in coordination with Jataware and MTRI with effort to ensure continuity with capabilities and work already in progress (based on our mostly-productive working relationships within TA4, we expect such discussions to go well). However, without substantial input into CDR design and open, ongoing communication about specifics, we are exposed to risk of needing to reorient around a system design "handed down" from DARPA based on requirements that were not communicated. Pivoting into an unanticipated integration work plan could seriously impact our capacity and timelines to deliver key functionality.

4.2 Inefficient reporting

The DARPA CriticalMAAS program has a substantial reporting burden, which has been a significant time sink for the UW-Macrostrat team, and I expect for other teams as well. Beyond the Milestone Reports described in the BAA, we are being asked to produce weekly activity reports, posters, descriptions of tools, lists of capabilities, etc. These are useful for communicating program goals and progress, certainly, but they often seem to be promulgated on Slack with unclear deliverables or overlapping/contradictory requirements outlined by different people. Often there seems to be minimal coordination between DARPA and MITRE about the purpose and structure of reports. The time spent on responding to these requests and building the necessary artifacts is substantial, and it is not clear that each of them has independent value. Posters, for example, take several hours to produce, especially when printing time is considered. I would like to see a more streamlined, unified reporting process that is clearly set out in a single document with due dates, and effort made to economize on the number of reports required. This would allow us to focus on coordinating between teams and building the infrastructure and tools that are the core of our work.