CriticalMAAS Milestone 6 report (final) – Macrostrat TA4 team

Daven Quinn and the UW–Madison / Macrostrat CriticalMAAS team

November 4, 2024

Contents

1 Report period

This **Milestone 6** report covers progress by the UW–Madison/Macrostrat team during the period from 2024-05-07 to 2024-11-03, including progress at the 9-month hackathon and the final part of Macrostrat's Phase 1 delivery (months 8-12). It also includes progress during a 3-month no-cost extension to the project, which pushed final delivery from August to November 2024.

2 Research and technical progress

During this reporting period, the Macrostrat–UW Madison team has focused on supporting mineral prospectivity modeling for TA3 teams, enhancing HITL pipelines for maps and

literature-derived information, contributing design insights to the CDR based on our capabilities and geological expertise, and completing the final code and infrastructure elements of our Phase 1 delivery. Our research has led to advancements in several areas targeted by CriticalMAAS, which we discuss here.

2.1 Infrastructure development and CDR integration

During the project, Macrostrat's core systems were generalized, modularized, and containerized to increase their flexibility to incorporate new and high-volume data sources, especially CriticalMAAS maps. Additionally, elements of the system were planned to be integrated into the CDR, either through API-based access or as overlays to the database. The services that make up the core of Macrostrat were moved a Kubernetes-based infrastructure and released as open-source software to facilitate use by other teams. This new version of Macrostrat is now running at https://v2.macrostrat.org and can be stood up as a local instance using a "seed" database dump and `docker-compose` using the `UW-Macrostrat/macrostrat` repository.

Macrostrat has substantial software capabilities that overlapped with CriticalMAAS needs, including management of geologic map data at scale, integration of multiple data sources into a unified modeling framework, and the ability to serve maps to users with high performance to thousands of users daily. Additionally, the xDD side of our team has substantial expertise with processing documents. Our team made several key contributions to CDR design, including:

- Leading the development of project's data integration schemas `DARPA-CRITICALMAAS/schemas` (months 1-4)
- The `UW-xDD/document-store` system for storing and querying documents (months 3-6)
- Tile server schema and API design for serving and visualizing geologic maps (months 9-12)

These elements were mostly superseded and not directly integrated into the CDR, but in many cases the basic design principles and philosophy were adopted by the program. (A key example of this is the core data fields used to describe geologic units in the CDR, which were based on Macrostrat's existing data model.) More direct and substantial integration of our capabilities would have been beneficial to the program had it been a priority beyond our team.

2.2 Geologic map integration

One of the major elements of Macrostrat's delivery is an integration pipeline to align geological maps into Macrostrat's database systems and data expectations. Though the primary target is CriticalMAAS maps, the pipeline is designed generally to allow geologic maps to be brought in from a variety of sources, including from GIS-ready file formats. (Most maps of high interest for prospectivity modeling are already digitized but do not include structured geological data.)

At the 9-month hackathon, we demonstrated a refined version of this pipeline targeting CriticalMAAS maps that was able to make them "analysis-ready" and web-accessible quickly after publication to the CDR. The usefulness of our pipeline was limited by the fact that, at the time,

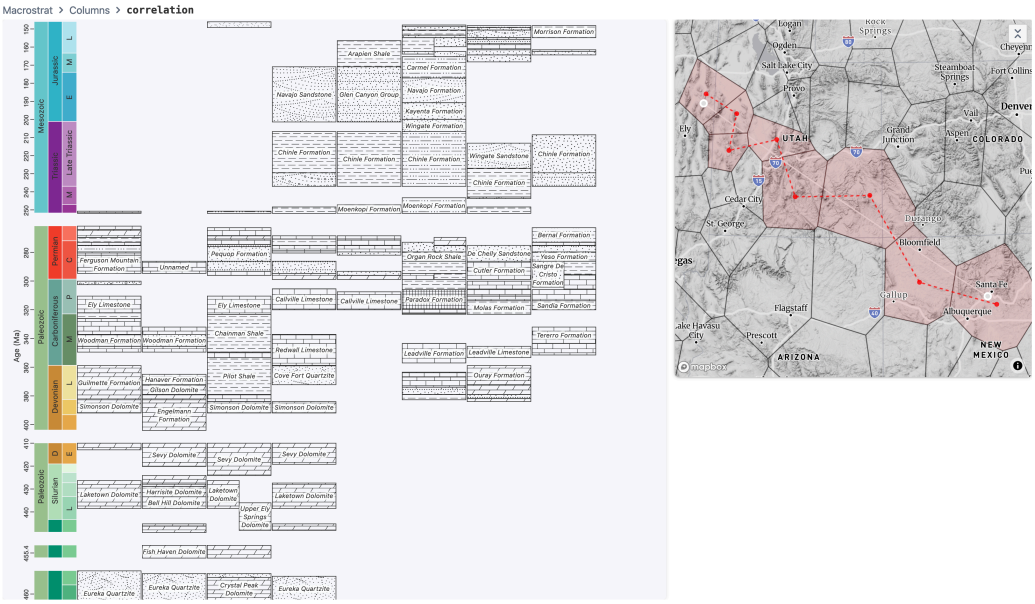

Figure 1: Column correlation diagram showing a depth-oriented view of geologic domains in Macrostrat

the CDR had only recently gained the ability to store projected geologic maps and held only a handful. Still, we were able to successfully ingest maps from the CDR and serve them to TA3 using Macrostrat's tile-based API, providing substantial capabilities for both analysis and visualization.

2.3 Data products in support of mineral prospectivity modeling

During Phase 1, the UW–Madison team created a number of data products and tools that are useful to mineral prospectivity modeling. The first of these is our geological mapping API, which has been successfully used through the course of the program to generate harmonized, regionally-consistent geologic maps for visualization and analysis. During the 9-month hackathon, we were able to provide geologic maps to TA3 teams using Macrostrat's APIs, via Python-based tools to query and stitch the results produced by MTRI (DARPA-CRITICALMAAS/macrostratpy). We also inaugurated capabilities for server-side filtering that support basic subsetting by rock unit age or lithology. These capabilities proved impactful at the 9-month hackathon, where the pipeline from TA1 maps to TA3 analysis was fully connected, using Macrostrat's API endpoints as a bridge. While these tools met the analytical needs of the TA3 teams, they did not fit into the program-level CDR plan, so were subsequently de-emphasized. However, some tileserver elements created by Macrostrat were incorporated into the CDR after the 9-month event.

In addition to harmonized geologic map datasets, we provided unique stratigraphic summary products that cannot be easily constructed by other software. We demonstrated the production of thickness isopach maps of specific rock types and stratigraphic intervals, based on Macrostrat's unique stratigraphic dataset and API (DARPA-CriticalMAAS/macrostrat-isopachs).

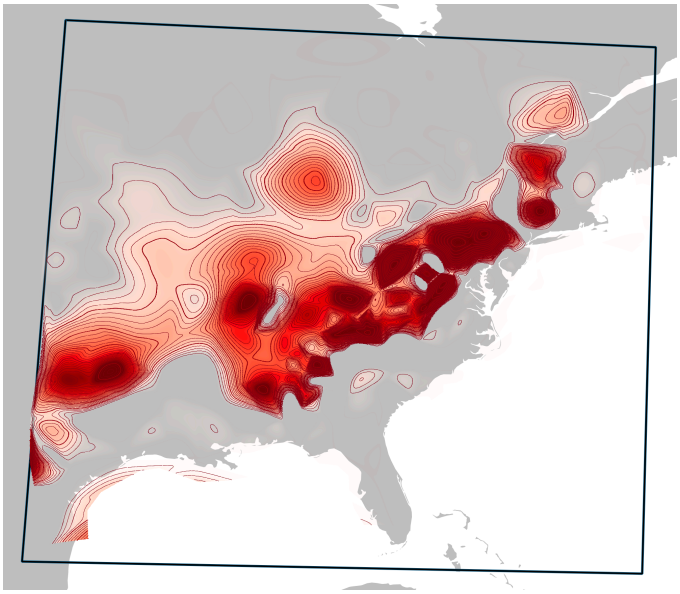

Figure 2: Contoured thickness (isopach) map of Cambrian and Ordovician carbonate produced using the Macrostrat API. Integrating structural surfaces into Macrostrat would greatly improve the fidelity of these data products.

During the 9-month hackathon, this capability was deployed to produce a thickness map of "Cambrian and Ordovician carbonates", a rock type and time interval of particular interest for Mississippi Valley-type lead-zinc deposits. This capability is applicable to arbitrary rock types and depositional settings; it can be improved by integrating cross sections and structural surfaces (e.g., "depth to formation top" maps) into Macrostrat's geologic framework. This future work will allow isopachs generated in a similar manner to the above to be more accurate and applicable to structurally complex domains, providing highly targeted input data layers for TA3.

Macrostrat's design is well-suited to producing the harmonized mapping and stratigraphic data products that are needed for TA3, and it does so in a scalable, performant, and conceptually sophisticated way. Both data products benefit from substantial work to standardize lithologic and age descriptors for rock units described in maps and columns. The capabilities developed during CriticalMAAS Phase 1 to support mineral prospectivity modeling will be integrated with Macrostrat's other offerings and will remain available for similar modeling tasks in the future.

2.4 Knowledge graph development

Via xDD, the UW–Madison team originated CDR records for ~50,000 mining reports and scientific papers relevant to Critical Minerals problems. More interestingly, these were paired with extractions (e.g., the coordinates of figures, tables, and maps within PDFs) correlated with in-document mentions of specific minerals and deposit types. We made some progress to pushing these data to the CDR at the 9-month hackathon, but the APIs were not quite ready

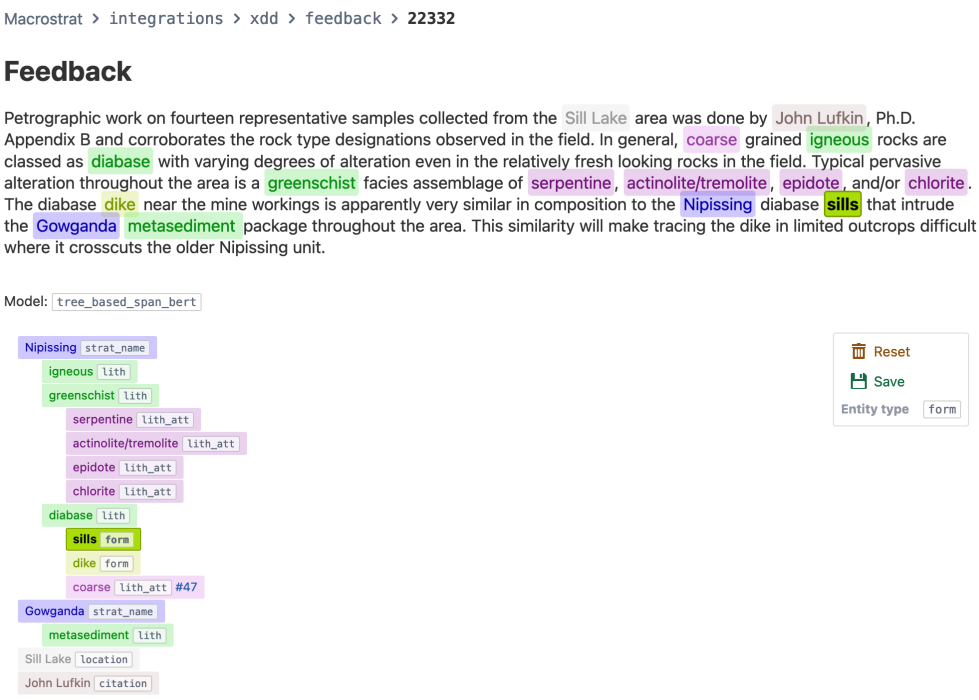

Figure 3: User feedback interface for knowledge graph construction

to accept this data at that point. Still xDD remains well-placed to supply document information to the CDR based on the patterns established during the program.

The main thrust of our work with xDD, however, was research into ways to more effectively integrate literature data about geologic units and mineral occurrences into Macrostrat's data models. Such a pipeline would allow rapid extraction of information from documents while preserving the ability to query and search over a structured representation of the knowledge. The main reason that we sought a no-cost extension was to complete this work, which depended on the academic schedule of several students that led the modeling elements of this work.

Our goal for CriticalMAAS Phase 1 was to build an approach to extract basic information (e.g., lithology, grain sizes, mineral contents) about named geologic units from papers and mining reports that mention them. Our final delivery of Phase 1 is an end-to-end pipeline for extraction, evaluation, and retraining of knowledge-graph construction models to support this type of data extraction.

We've tested both BERT-based knowledge-graph construction approaches and LLM prompting approaches, which can both output properly formatted data, but often with a low degree of accuracy in their classification of tokens. To improve this, we've built a feedback user interface that will allow geoscientists to classify relationships to build a tree-like set of relationships (e.g. `Bonneterre Formation > dolomite > laminated`) from paragraphs that can be used as training data for the models. The current pipeline supports continuous retraining for the knowledge graph construction models, including linking structured model extractions to

Macrostrat's data dictionaries. All elements of this pipeline have been tested, including retraining steps. However, user feedback has not yet been constructed in enough volume to produce model improvements.

Software elements

Macrostrat's knowledge-graph construction pipeline is now a functional distributed software system, with elements summarized below:

Modeling frameworks Two modeling approaches for knowledge-graph extraction are supported:

1. A LLM-based modeling framework (UW-Macrostrat/llm-kg-generator)
2. A BERT-based modeling framework (UW-Macrostrat/unsupervised-kg)

These software pipelines are responsible for running models xDD document extractions (based on Critical Minerals-related corpora) on CHTC GPU resources, and for retraining based on structured data.

Data extraction and evaluation The model runners above post their outputs into a result curation system, consisting of:

1. A database of extractions from xDD documents , knowledge graph links and nodes, and references to structured data dictionaries, housed in Macrostrat's PostgreSQL database (UW-Macrostrat/macrostrat).
2. An API server for accepting new knowledge graph entities and relationships from models and user feedback, validating them and linking to structured data dictionaries, storing in the database, and subsetting for retraining (UW-Macrostrat/macrostrat-xdd).

User feedback interface

1. A web-based interface for geoscientists to classify relationships extracted from documents, which can be used as training data for the models (UW-Macrostrat/web and UW-Macrostrat/web-components)
2. An ORCID-based authentication system that allows feedback by authorized collborators from multiple institutions atop Macrostrat's web infrastructure. Currently deployed at dev2.macrostrat.org/integrations/xdd (UW-Macrostrat/api-v3)

Next steps

The pipeline is now ready to deploy for feedback and model retraining on our initial Critical-MAAS corpus as well as a wider corpus constructed for more general geological units. While we have not yet successfully produced high-quality, structured model outputs at scale, we are confident that this pipeline represents an approach that can be applied towards multiple scientific goals. Ongoing work will focus on training the models by collecting more feedback data,

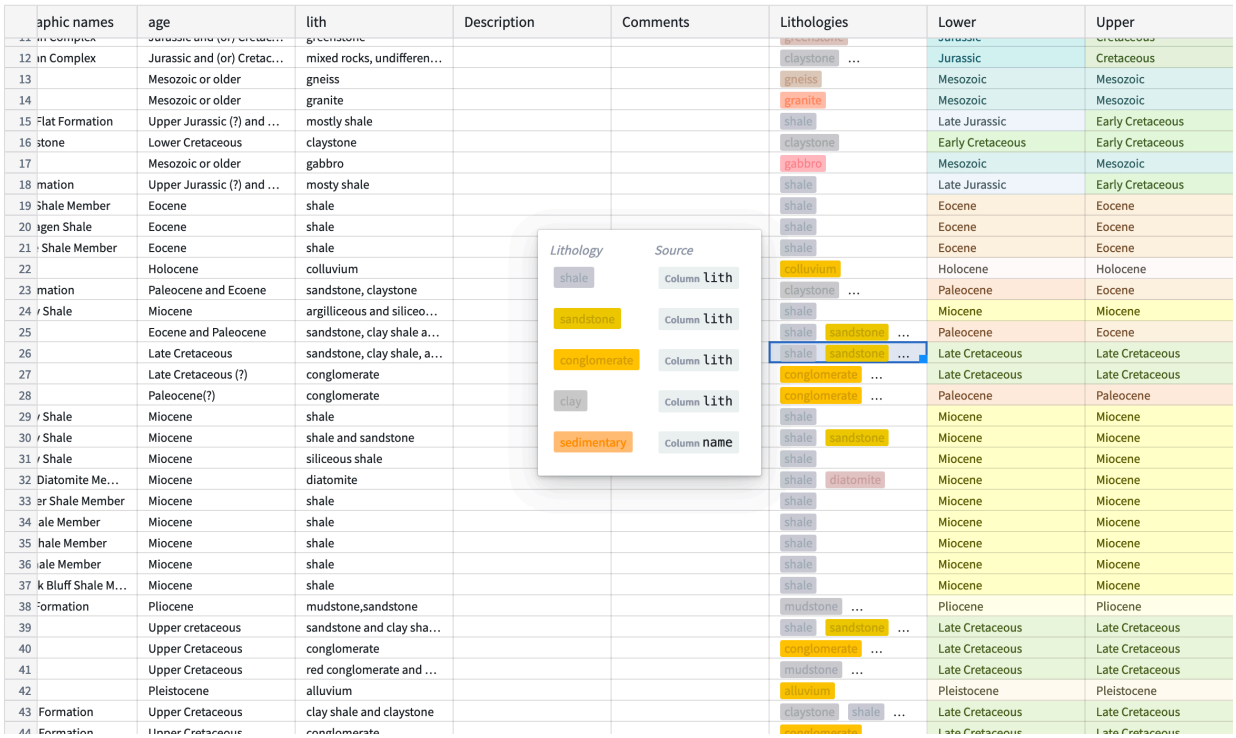

Figure 4: Table view of lithology data

and on exploring ways to include LLM-based models in the training steps, such as improving prompt design and context selection. We anticipate that this research will ultimately lead to a principled approach to assembling structured datasets from descriptive text, with substantial supporting tools.

2.5 Human-in-the-loop interfaces

In addition to building data pipelines to support prospectivity modeling and literature data extraction, we have built exploratory human-in-the-loop (HITL) interfaces that allow interaction with different elements of geologic data. Some notable successes are summarized below:

- User interfaces that support evaluation of geological information and constraints, such as:
  - Web-based, inspectable visualizations of TA1 map datasets
  - Column correlation diagrams.
- Embedding-based search tools (in collaboration with Meng Ye, SRI) for ranking geological units based on their similarity to a user-provided queries, with embeddings trained over critical minerals documents in xDD (`UW-Macrostrat/embedding-tiler`).

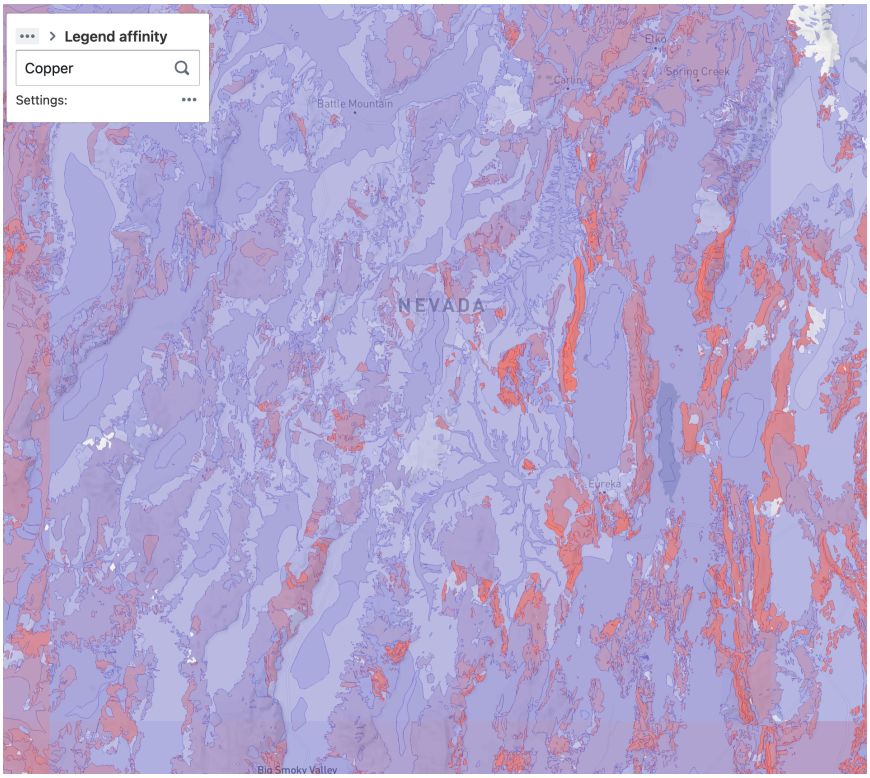

Figure 5: Embedding-based web search for geologic units

- Table views of lithology data that make it easier to inspect and filter data holdings.
- Tools for iteratively constructing map topology (`Mapboard/topology-manager`) and rapidly editing geologic maps (`Mapboard/mapboard-platform`). Although there is some overlap in functionality with the Polymer product from Jataware, they are performant and flexible and will likely be useful for constructing "publication-ready" maps from TA1 data.

All of our HITL interfaces are constructed using modern web mapping technologies and are designed for broad accessibility and use by geoscientists. Development of these HITL tools will continue alongside other Macrostrat activities, and we will continue to refine them as modular web components (`UW-Macrostrat/web-components`) that can be used in other geologically rich applications going forward.

3 Summary

In this last portion of the CriticalMAAS Phase 1 project, the UW–Madison team has made substantial progress in developing Macrostrat's core infrastructure and data products to support critical minerals modeling going forward. The no-cost extension allowed us to bring the knowledge-graph construction pipeline to a well-integrated finish, as forecast during the endpoint conversation in July. More generally, during Phase 1, we believe that we were able to contribute substantially to CriticalMAAS, by leading TA4 and data integration efforts early in

the program and influencing the design of key elements based on our expertise.

We've been able to demonstrate the utility of our software approaches and data holdings to TA3 at various points. However, we have struggled to integrate our capabilities with the core program-level software deliverables, and we accomplished less than planned in several domains. Some areas of notable weakness in our delivery were heightened by the handling of program integration. For instance, we struggled to produce polished user interfaces with quick turnaround times, while Jataware excelled at this. However, the geological basis of our designs was unmatched among the performers. The program would have benefited substantially from a more explicit synthesis of the respective strengths of the TA4 teams. While we worked in good faith towards this outcome, changing and poorly communicated priorities made it difficult to exercise leadership toward this goal consistently.

During CriticalMAAS, the UW–Madison/Macrostrat team has developed sophisticated, geologically advanced, and in many cases production-ready software for managing geologic data at scale. These systems lay the groundwork for future advances in critical minerals prospecting and geologic data integration more generally, and remain available for integration by the USGS or other organizations. Macrostrat will continue to develop these approaches. Ultimately, CriticalMAAS has provided Macrostrat with an opportunity to rapidly expand its focus and capacity, and we thank DARPA for the opportunity to participate.