Data Curation Workflow

By: Mathias Heider
Mentor: Mike Deagen

Motivation for Project

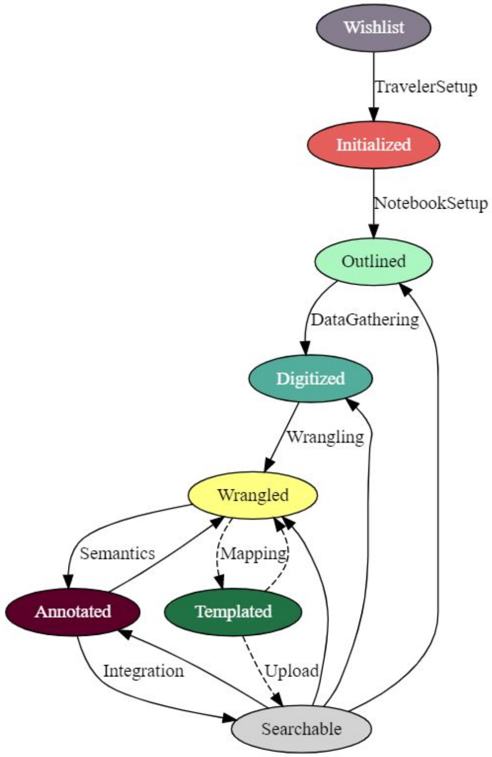
- Materials scientists and engineers generate vast amounts of data through experiments and computation.
- This data was made FAIR (*Findable*, *Accessible*, *Interoperable*, *Reusable*) through the NanoMine database
- the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data
- Data Curation Workflow was created to add version control to the process, add documentation and, improve the ability to trace back errors.

Data Curation Workflow

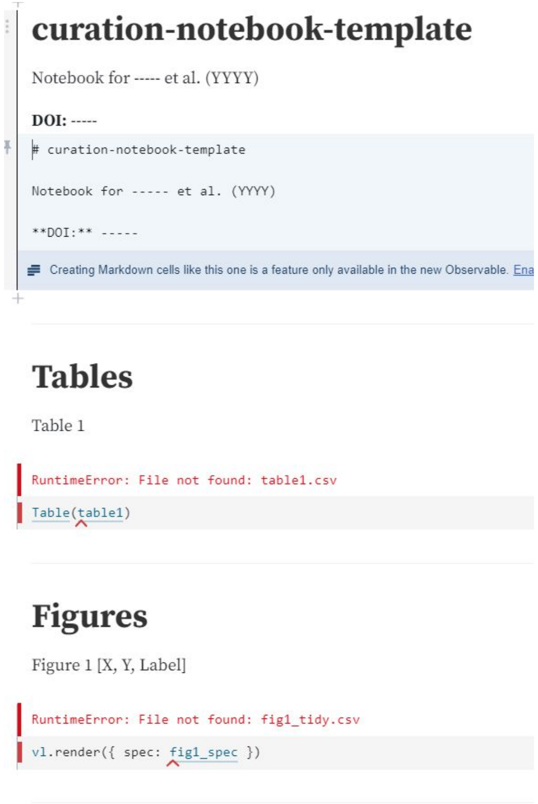
Data curation involves selecting, organizing, and looking after items in a collection.

You can think of the data curation workflow as a large ETL (Extract Transform Load) process which contains many smaller ETL processes.

The goal of this workflow is to make research data on nanomaterials machine interpretable so it can eventually be put and into a database and searched.



Steps of the Data Curation Workflow



curation-notebook-template

Notebook for ----- et al. (YYYY)

DOI: -----

```
# curation-notebook-template

Notebook for ----- et al. (YYYY)

**DOI:** -----
```

Creating Markdown cells like this one is a feature only available in the new Observable. Ena

Tables

Table 1

```
RuntimeError: File not found: table1.csv
Table(table1)
```

Figures

Figure 1 [X, Y, Label]

```
RuntimeError: File not found: fig1_tidy.csv
vl.render({ spec: fig1_spec })
```

Step 1: (Wishlist) Any research article with a known DOI that people want searchable sits in a wishlist to be eventually curated.

Step 2: (Initializing) A job ID and curation traveler is created to keep track of the work done on data curation workflow for the specified research article

Step 3: (Outlined) A observable notebook is created for the research article that will hold the different tables and graphs.

Steps of the Data Curation Workflow

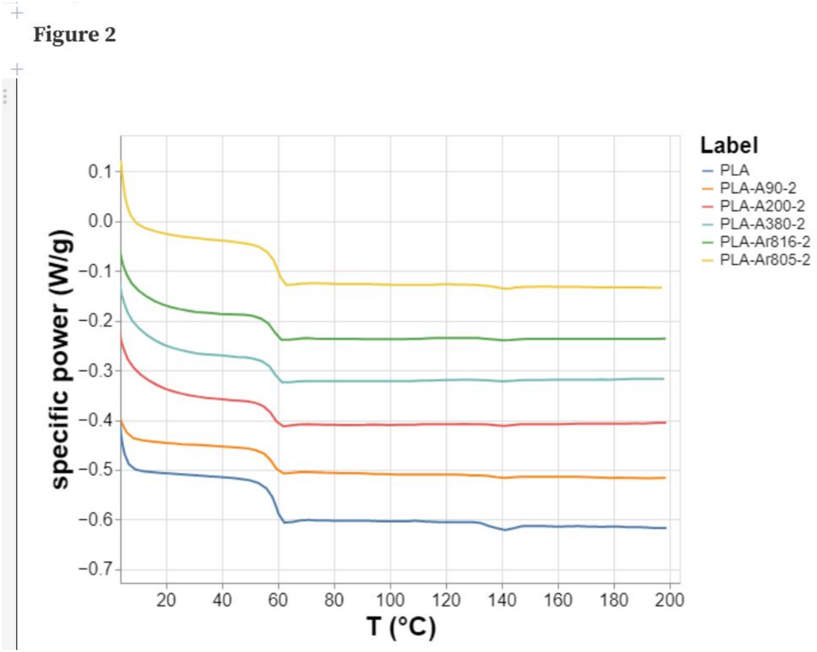


Figure 2

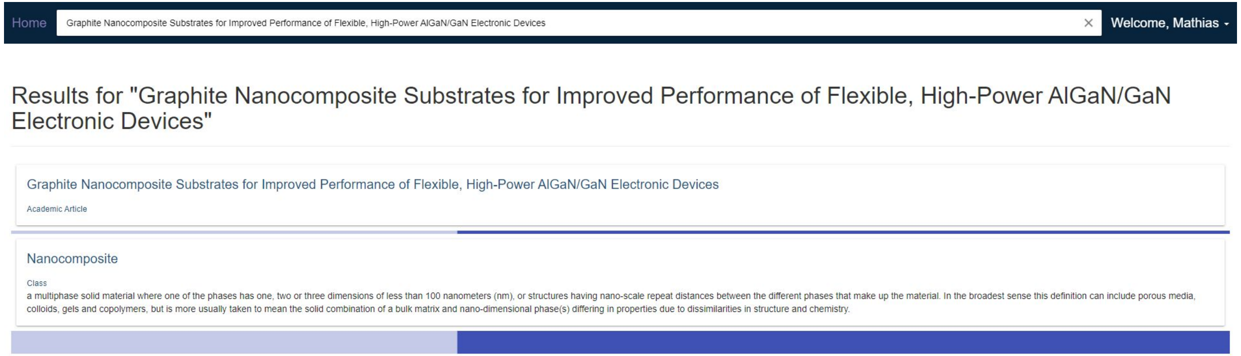
Step 4: (Digitized) Data is extracted from the specified research article using tools like WebPlotDigitizer and Tabula. This information is then uploaded to the observable notebook for visual inspection.

Step 5: (Wrangled) Although most data can extracted from charts or tables. Some data has to be wrangled in order to be able to be further processed. Once wrangled these datasets can are ready for semantic annotation.

Steps of the Data Curation Workflow

Step 6: (Templated) Data points from the different tables and graphs are mapped into different excel sheets using a Python mapping script in order to simplify the process. After this process the data is ready to be uploaded to the NanoMine database.

Step 7:(Searchable) Data at this stage have been uploaded to the NanoMine database allowing the data points from the research article to being searched.


Home
Graphite Nanocomposite Substrates for Improved Performance of Flexible, High-Power AlGaN/GaN Electronic Devices
Welcome, Mathias

Results for "Graphite Nanocomposite Substrates for Improved Performance of Flexible, High-Power AlGaN/GaN Electronic Devices"

Graphite Nanocomposite Substrates for Improved Performance of Flexible, High-Power AlGaN/GaN Electronic Devices

Academic Article

Nanocomposite

Class

a multiphase solid material where one of the phases has one, two or three dimensions of less than 100 nanometers (nm), or structures having nano-scale repeat distances between the different phases that make up the material. In the broadest sense this definition can include porous media, colloids, gels and copolymers, but is more usually taken to mean the solid combination of a bulk matrix and nano-dimensional phase(s) differing in properties due to dissimilarities in structure and chemistry.

Next Steps for the Project

- Data Curation Workflow is a continuous process to expand the NanoMine database with more data.
- Curating new types of data that might not yet exist in NanoMine (Schema/Ontology).
- Make more programmatic approaches to the Workflow to different parts more automated.
- Provide examples for researchers to upload their own data immediately without having to be trained through the entire process. Then incrementally adding to the database/knowledge graph to make the data searchable.

Progress Report

- Since the Data Curation Workflow is a continuous process I have been adding more and more research articles to the NanoMine database.
- Began creating a metadata observable notebook that will show data from the wish list in a more visual way.
- Also adding in-depth documentation about the Data Curation Workflow to the metadata Notebook.
- Provide a step by step process with pictures to the notebook so people previous not acclimated with the project could understand how to upload data points.