Challenges in Automatic Metadata Extraction for Experimental Techniques in Material Science

Introduction

In the realm of materials science, metadata, which includes crucial information about experimental conditions, instrument parameters, and sample details, plays a significant role in understanding and reproducing experiments. In applications which involve correlative characterization, the metadata is of utmost importance as it is needed to properly configure subsequent measurements. As opposed to simulations in materials science where the metadata are already available in digital form, in many experimental scenarios, the metadata are on lab notebooks, hidden in proprietary formats, or simply not existing outside the mind of the researcher. The main challenges in extraction of metadata in materials science are explained the next section.

Challenges

1. Diverse Characterization Techniques

   The challenge of automatically extracting metadata from diverse datasets is made more difficult by the myriad of characterization techniques used across different domains, underscoring the impracticality of a one-size-fits-all solution. Each dataset, depending on its origin, content type, and structure, may require a unique approach or a combination of techniques for effective metadata extraction. When multiple methods are employed in tandem, it becomes necessary to navigate through various extraction types, ranging from text analysis and image recognition to complex data parsing. Moreover, a significant hurdle often overlooked is the requirement for additional preparation and cleaning of the extracted metadata to render it in a usable format. This step is crucial as raw metadata, especially when amalgamated from diverse sources or extracted through complex procedures, may be loaded with with inconsistencies, inaccuracies, or irrelevant information (sometimes set by default) that must be refined.

2. Heterogeneous Data Formats (also proprietary)

   Each data format and each manufacturer would require a matching extractor and metadata mapper to homogenise the extracted terms for interoperability and further use. An additional layer of complexity is introduced when some of the metadata is hidden within proprietary formats, which are not readily accessible or decipherable without specific tools or permissions. Companies tend to encode their data and

metadata to ensure that only proprietery software can be used to work with them, thereby forcing the users to buy the additional software. This creates an issue for scientist wanting to perform correlative characterisation with the data obtained from different types of instruments from different manufactureres.

3. Lack of Unified Terminology

   The localization or lack of unified terminology impacts the extraction and subsequent utilization of metadata, particularly in fields where specific terms or measurements are critical to understanding and using the data correctly. For instance, in the context of Scanning Electron Microscopy (SEM), manufacturers may refer to the acceleration voltage—a crucial parameter affecting the microscope's resolution and depth of field—using different terms. While some may label it as "EHT" (Extra High Tension), others might use "HV" (High Voltage). This variation in terminology can pose challenges for the automatic extraction of metadata because the extraction algorithms must be capable of recognizing that these different terms refer to the same concept. When metadata extracted from various sources is to be aggregated or used in a centralized system, the lack of unified terminology necessitates additional processing to map disparate terms to a common vocabulary. This inconsistency not only complicates the extraction process but also affects the usability of the metadata. Users or downstream applications expecting one term may not recognize or correctly interpret the metadata if it is labeled with an alternative term.

4. Integration with Existing Workflows

   Integrating software for automatic extraction of metadata into existing workflows poses significant challenges due to compatibility issues, diverse data formats, performance considerations, and the need for minimal disruption. Ensuring the new software works seamlessly with various operating systems, databases, and applications already in use requires extensive customization, which can be both time-consuming and technically complex. Certain workflows are well established monolithic systems but very closed to interactions from software from the outside, which might cause a real bottleneck in integration. If the extraction software is situated in a separate server, the network transfer speed of the data might also create an additional bottleneck.

Considering these challenges, the task of automatic extraction of metadata in materials science can be split into 3 tasks as explained below:

Tasks

1. Extraction of Metadata

   Initially, it's essential to compile a comprehensive list of the necessary metadata, incorporating all potential synonyms. Ideally, this should take the form of a structured metadata schema, essentially a set of parameters derived from an ontology. This ensures that all terms are precisely defined, and their relationships

clearly established. Utilizing the metadata schema enables the thorough extraction of all relevant information. Following the establishment of the metadata schema, it is necessary to extract metadata from various data formats. Metadata may sometimes be found in file headers or provided as a separate file. If the data formats are open and not restricted by proprietary limitations, extraction becomes relatively straightforward once the data structure is fully understood.

2. Data Cleaning and Preparation

   The subsequent phase involves the data's cleaning and preparation. Frequently, default values assigned in the metadata do not accurately reflect the specifics of a given output. Hence, conducting a sanity check to identify and eliminate such inaccurate metadata from the extracted list is crucial. Additionally, to align with the requirements of the metadata schema, it might be necessary to perform certain calculations or string manipulations to accurately derive the needed information. This process is inherently variable and cannot be standardized, as the nature of the metadata often varies significantly depending on its source and structure.

3. Mapping the Extracted Terms to Community Agreed Terminology

   Following the extraction, cleaning, and preparation of the metadata, the next task is to align it with terminology that has consensus within the community. Essentially, this means renaming the extracted metadata in accordance with a metadata schema that includes terms widely accepted by the user community. By undertaking this mapping, the interoperability of the metadata significantly improves, ensuring that individual terms are easily comprehensible to users, all thanks to the standardized framework provided by the metadata schema.

With a solid foundation laid in understanding the crucial steps of metadata extraction, cleaning, preparation, and mapping, we can now delve into practical implementations within the realm of materials science, particularly in electron microscopy. Transitioning from theoretical concepts to practical applications, we explore the diverse array of tools and software available to work with electron microscopy data but can also be used for metadata extraction.

Existing Tools

- Digital Micrograph / Gatan Microscopy Suite[1]

  Digital Micrograph also known as Gatan Microscopy Suite stands as one of the most important software for researchers and professionals working with Gatan devices in the field of electron microscopy. Its comprehensive suite of features is specifically

[1] https://www.thermofisher.com/de/en/home/electron-microscopy/products/software-em-3d-vis/avizo-software.html

designed to complement Gatan's hardware, offering users a seamless experience in data acquisition, analysis, and visualization. However, its utility is somewhat constrained by its compatibility; images or data obtained from devices manufactured by other companies cannot be always processed using Digital Micrograph, limiting its versatility across the broader spectrum of electron microscopy equipment. Moreover, while an offline version of the software exists, it offers a reduced set of functionalities compared to its full version, potentially hindering the depth of analysis possible. Though it is perfect for metadata extraction from Gatan devices, the same cannot be said for other manufacturers.

- Thermo Scientific Avizo Software[2]

Thermo Scientific Avizo boasts impressive capabilities for 3D visualization and analysis across diverse scientific fields like that of SEM-FIB Tomography, but its true strength lies in handling data generated by Thermo Fisher instruments. For tasks specifically involving images from other manufacturers, its functionality might prove less comprehensive. Although Avizo can open various file formats, complexities arise when dealing with specialized formats or data structures unique to non-Thermo Fisher equipment. This can require additional configuration, scripting, or even external tools to achieve optimal analysis. For example, reading metadata from generic TIFF tags of images, in addition to Thermofisher or FEI specific tags is possible, but the same is not applicable for other vendor specific tags.

- Autoscript TEM[3] and Autoscript 4[4]

Autoscript TEM and Autoscript 4 Software are Python-based APIs designed to facilitate communication with Thermo Fisher scientific instruments. These software packages come with a comprehensive set of predefined metadata accessible through Autoscript, which can be used by metadata extraction scripts. Autoscript TEM is tailored specifically for Thermo Fisher's transmission electron microscopes, whereas Autoscript 4 is aimed at their scanning electron microscopes, including those equipped with ion beams. As commercial products, these software solutions require a separate purchase to unlock the extensive flexibility they offer. It's important to note that due to their customization for Thermo Fisher equipment, Autoscript software is not compatible with instruments or data from other manufacturers.

- ImageJ suite IMBalENce[5]

The IMBalENce is a suite of plugins for ImageJ mainly catering to scale adjustment and metadata extraction of scanning electron microscopy TIFF images. Currently it

[2] https://www.thermofisher.com/de/en/home/electron-microscopy/products/software-em-3d-vis/avizo-software.html
[3] https://www.thermofisher.com/order/catalog/product/de/en/AUTOSCRIPT-TEM
[4] https://www.thermofisher.com/de/en/home/electron-microscopy/products/software-em-3d-vis/autoscript-4-software.html
[5] https://github.com/IMBalENce/EM-tool

covers metadata extraction from TIFF images generated by ZEISS, FEI (currently Thermo Fisher) and Tescan scanning electron microscopes. It is a quite practical tool as ImageJ is widely used by researchers for further processing of their SEM data and these plugins will enable extraction of metadata and thereby saving the metadata along with the images. The extracted metadata are saved with the parameter names as specified by the manufacturer, and not according to a structured metadata schema which is agreed upon by a community. This creates a problem for interoperability, as the same parameter would be called by different names in metadata files depending on their origins.

- pyem[6]

UCSF pyem is a suite of Python programs for data analysis in electron microscopy of biological samples. It supports metadata queries or in other words metadata extraction from the images. But it is highly specific for cryogenic transmission electron microscopy, and hence cannot be hence used for the more general scenarios.

- Hyperspy[7]

Hyperspy is an open-source python library which is very versatile and can be used for data analysis using multidimensional datasets. It has its own metadata schema which is very limited in the number of parameters. The functions which extract the metadata according to hyperspy's metadata schema do a decent job of mapping metadata which are differently named by different manufacturers. This selection of parameters can unfortunately not comprehensively cover the need of having descriptive metadata as required by FAIR principles. But hyperspy can also be used to extract the original metadata from most electron microscopy image formats. Therefore, hyperspy can be used as a good metadata extractor. But data cleaning and mapping still need to be done for the sake of interoperability.

To achieve this aim, it is recommended to use a stand-alone mapping service[8], which can be extended with plugins, which perform, extraction, data cleaning and preparation and mapping of metadata to a metadata schema. The mapping service can either be locally installed while dealing with sensitive and huge volumes of data or can be used as an online service which is remotely hosted. The versatility plugins offer infinite possibilities for the application of the mapping service, as it can cover multiple materials science techniques. For example, for scanning electron microscopy, the plugins make use of hyperspy for metadata extraction, and further process the extracted metadata and map them according to a published metadata schema.

[6] https://github.com/asarnow/pyem
[7] https://zenodo.org/records/10412190
[8] https://matwerk.datamanager.kit.edu/mapping-service-ui.html

Another alternative which makes use of hyperspy is the scythe module[9] for electron microscopy, which also maps the metadata to a different metadata schema[10] for electron microscopy. Note that this metadata schema covers most electron microscopy techniques in a single metadata schema, as opposed to the mapping service in which each technique has a different schema and corresponding plugin. The mapping service even has an extractor and mapper for SEM-FIB Tomography which is not covered under the scythe metadata schema.

[9] https://materialsio.readthedocs.io/en/latest/_modules/scythe/electron_microscopy.html
[10] https://github.com/materials-data-facility/scythe/blob/master/scythe/schemas/electron_microscopy.json