RESEARCH PAPER

Towards the FAIRification of Scanning Tunneling Microscopy Images

Tommaso Rodani[1], Elda Osmenaj[2], Alberto Cazzaniga[1], Mirco Panighel[2], Cristina Africh[2] & Stefano Cozzini[1†]

[1]AREA Science Park, Padriciano 99, 34149, Trieste, Italy

[2]CNR-IOM, Consiglio Nazionale delle Ricerche – Istituto Officina dei Materiali, S.S. 14 Km 163.5, Basovizza, 34149, Trieste, Italy

Keywords: Provenance, STM images, data management, FAIR, metadata for nanoscience, W3C PROV

Citation: Rodani T., Osmenaj E., Cazzaniga A., Panighel M., Africh C. & Cozzini S. Towards the FAIRification of Scanning Tunneling Microscopy images. Data Intelligence 5(1), 27-42 (2023). doi: 10.1162/dint_a_00164
Received: Dec. 25, 2021; Revised: Jan. 17, 2022; Accepted: May 10, 2022

ABSTRACT

In this paper, we describe the data management practices and services developed for making FAIR compliant a scientific archive of Scanning Tunneling Microscopy (STM) images. As a first step, we extracted the instrument metadata of each image of the dataset to create a structured database. We then enriched these metadata with information on the structure and composition of the surface by means of a pipeline that leverages human annotation, machine learning techniques, and instrument metadata filtering. To visually explore both images and metadata, as well as to improve the accessibility and usability of the dataset, we developed "STM explorer" as a web service integrated within the Trieste Advanced Data services (TriDAS) website. On top of these data services and tools, we propose an implementation of the W3C PROV standard to describe provenance metadata of STM images.

1. INTRODUCTION

Data management procedures are fundamental for high-quality research, especially in the case of great volumes of scientific data produced. A critical role to ensure good data management is given by metadata and provenance information which add value to the data and allow data to be found, interpreted, re-used and reproduced. For these reasons, annotating data by means of a specific metadata standard improves their ability to meet the FAIR (Findable, Accessible, Interoperable, Reusable) guiding principles [1].

† Corresponding author: Stefano Cozzini (E-mail: stefano.cozzini@areasciencepark.it; ORCID: 0000-0001-6049-5242).

In this paper, we report on the activities carried out on a scientific archive of scanning tunnelling microscopy (STM) images with the objective of organizing it in a more structured and convenient dataset from a FAIR point of view [2]. Since the experimental technique has not substantially changed over the last 20 years, our effort towards the FAIRification of legacy data is relevant both for current research activity in STM and for guiding the FAIR-by-design workflow under active development following current standards [3, 4]. To achieve this goal metadata is a key driver; in the following, we will present our approach in collecting metadata for our STM dataset and our initial effort in defining our own metadata schema.

The images were generated using an Omicron Variable Temperature STM (VT-STM) microscope [5] located at the Istituto Officina dei Materiali (CNR-IOM) in Trieste, Italy. In total, researchers generated about 420,000 images over twenty years of research activity, consisting in 228 GB of raw data. From this sample, an initial batch of about 110,000 STM images recorded in constant current mode was selected and curated in an organized dataset along with 59 instrument metadata for each image. These metadata alone provide valuable information about the conditions in which images were obtained and are useful to make data findable and accessible. Unfortunately, the type of materials that compose the sample, the most relevant information associated with STM images, has been historically registered on a paper logbook. In such a state, it is unfeasible to integrate this information into an automated data management system. To improve the scientific value and FAIRness of the dataset, we annotated images with this specific metadata with a pipeline that leverages human annotation, machine learning (ML) techniques, and instrument metadata filtering. After this labelling procedure, the final dataset consists of 7,287 STM images assigned to three categories of materials, with a total size of 4.7 GB and organized with data files and original instrument metadata files for each individual image, along with provenance metadata for the whole dataset.

Another crucial improvement for the accessibility and usability of the dataset consisted in the creation of a metadata explorer, developed as an integrated service within the TriDAS website [6], which allows users to visually explore images' metadata through interactive and downloadable plots. The core logic of the web service is initially designed around a subset of 11 image metadata, carefully selected together with nanoscience researchers to provide significant information about image characteristics and microscope settings relevant to image quality and context. The functionality of the web application, and its relevance as an interactive tool with the dataset, are then further improved to include images visualization on the browser without the need for any additional software. Besides these activities carried out to increase the dataset usability, we present an application of a provenance standard for the case study of STM images. Intended as a type of structured metadata, provenance tracks the origin and all the intermediate procedures applied to produce a data product, thus becoming fundamental for the reproducibility of the scientific experiment and for the analysis and interpretation of the results. During the FAIRification workflow, the W3C PROV standard [7] is applied to describe the provenance of metadata, from the original STM images to the ones curated and available on the TriDAS website.

We made available the dataset containing 7,287 STM images together with their provenance description [8] and all source code used in this paper [9].

2. METHODS AND RESULTS

2.1 STM Dataset

STM images presented in the dataset were recorded, over twenty years of research activity, by the Surface sTructure and Reactivity at the Atomic Scale (STRAS) research group of the CNR-IOM Institute in Trieste, using an Omicron Variable Temperature STM (VT-STM) microscope. Raw data are composed of forward and backward topography scan arrays stored in binary format in files with extension *.tf0* and *.tb0*, and a *.par* file that contains instrument variables and other information in text format. For some of these topographic images, the related tunneling current images, stored in files with extension *.tf1* and *.tb1*, are also present. By filtering metadata of images, we retrieved a reference dataset of 111,415 constant-current STM images from a vast collection of measurements. The structure and composition of the imaged surface cannot be recorded in an automated way, as such, it has been historically registered on a paper logbook. To obtain this crucial information for STM images, we developed a workflow based on human annotation, machine learning techniques and metadata information. The starting point was to manually label groups of images into different categories according to the sample material. Researchers, within the same day, typically measured samples of the same material, and, considering the typical workflow of the group, it is then reasonable to assume that samples should be of the same category also within a limited time period. Given these assumptions, we created a total of 188 plots composed of at most 100 images sampled from each month of activity. This collection was manually labeled and used to obtain a broad division of the dataset in 18 sample material categories, as shown in Figure 1.

Then we selected a subset of 10 images for each of three specific material categories, namely Gr_Ni100, Gr_Ni111 and N_Gr_Ni111 for a total of 30 images, as shown in Figure 2. In particular, Gr_Ni100 includes images taken on monolayer graphene grown by chemical vapour deposition on Ni (100). Due to the square symmetry of the substrate, the resulting layer is composed of patches of aligned graphene (aligned with the substrate crystallographic structure and showing a typical wavy 1D moiré pattern) and rotated graphene (identifiable in the images by a 2D moiré pattern) [10, 11, 12, 13, 14]. The Gr_Ni111 category contains STM images taken on monolayer graphene grown by chemical vapour deposition on Ni (111) single crystals. The layer is composed of patches of epitaxial graphene (in register with the substrate lattice), appearing as a triangular arrangement of spots, and rotated graphene identifiable in the images by the presence of a 2D moiré pattern [15, 16, 17, 18, 19, 20, 21]. Finally, the N_Gr_Ni111 category represents images taken on monolayer graphene grown by chemical vapour deposition on a Ni (111) single crystal previously doped with atomic nitrogen. During the growth, some nitrogen atoms present in the Ni bulk are trapped in the graphene mesh, doping the layer and originating characteristic defects, visible as dark triangles and bright clover-like features [22, 23].

With the aim of associating the type of material composing the sample to a larger set of images, we developed an approach based on recent developments in *representation learning* [24] for image recognition. Representation learning techniques leverage only the availability of large datasets to train a model that automatically detects features of the images which are relevant for a detection or classification task. The pioneering work of Le Cun [25], as well as more recent progress in the field [26], led to convolutional

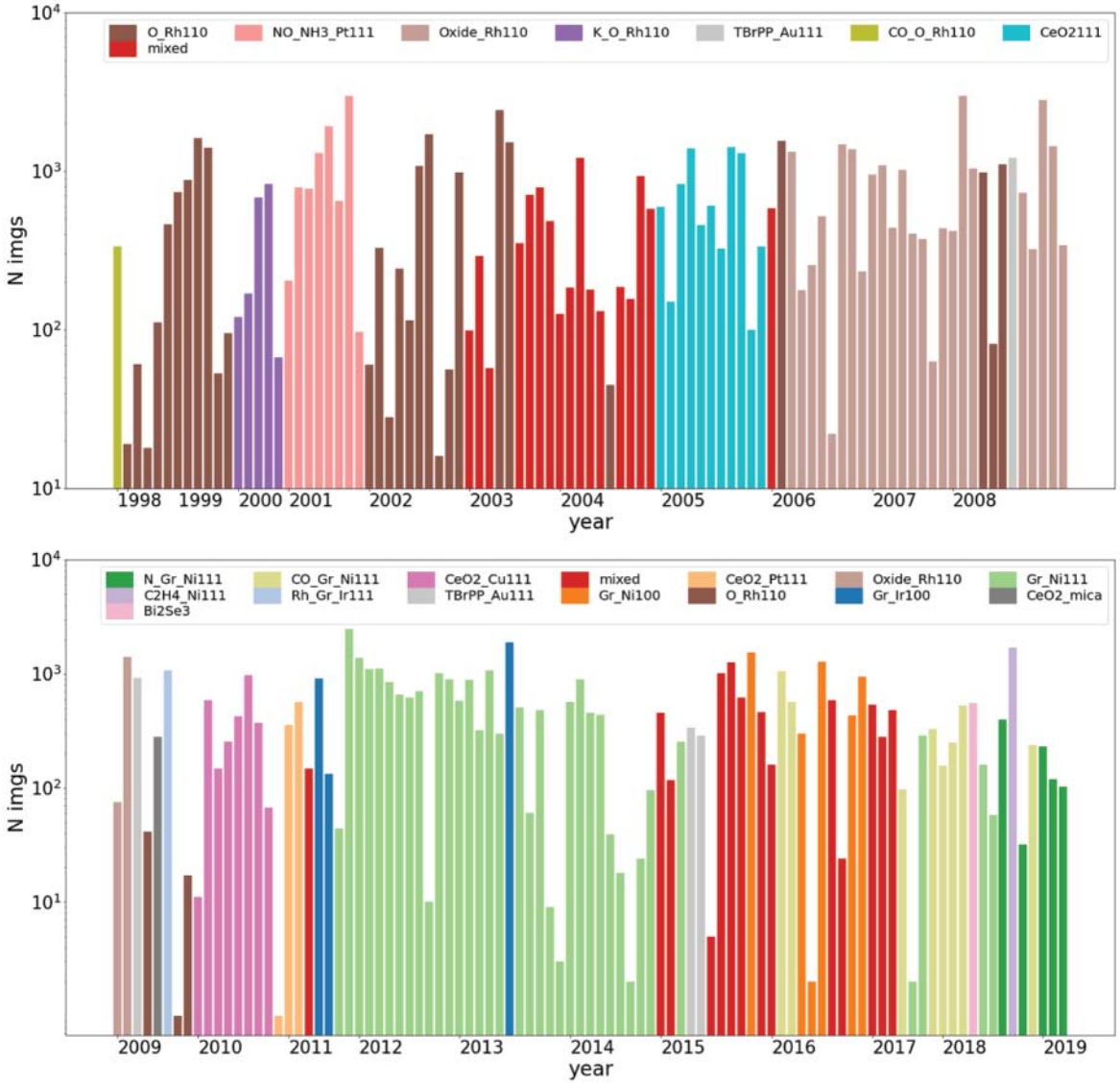

Figure 1. Monthly activity of TASC laboratory color coded by sample material category. Months where more than one sample material was recorded are labelled as "mixed".

neural networks, a family of deep-learning models particularly suited for image feature analysis thanks to their translation equivariance and locality properties.

In absence of a sufficiently large set of STM images in the dataset carrying information on the sample material, we focused on the technique of transfer learning [27]. Transfer learning consists in employing the weights learned on a network trained on a generic enough dataset, to target a compatible task on a

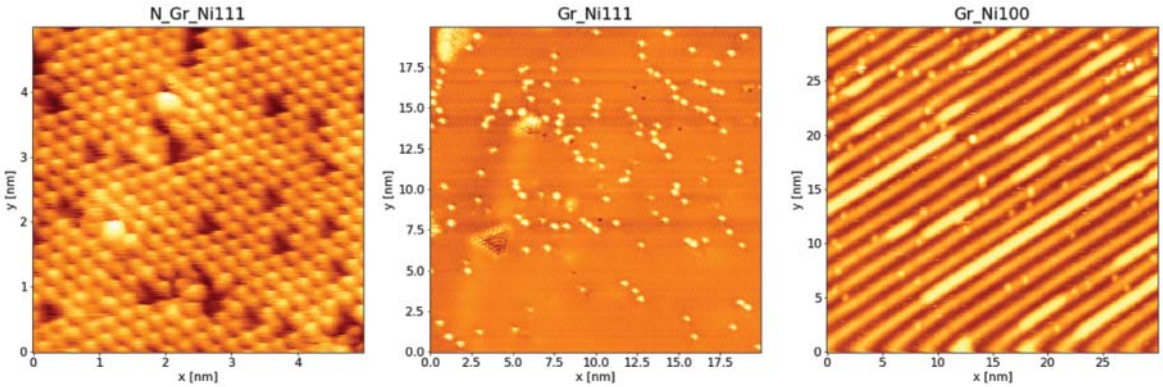

Figure 2. From left to right: example images of (a) N_Gr_Ni111, (b) Gr_Ni111 and (c) Gr_Ni100 categories from the labeled set of 30 images.

different set of images. A plethora of theoretical results [28], as well as applications to datasets of microscopy images [29, 30, 31], show that models trained on ImageNet [32] capture features that are relevant in an extremely heterogeneous set of image classification tasks.

From preliminary analysis, it emerges that a Resnet50 model trained on ImageNet [33] has sufficient expressive power for extracting relevant features in the specific case of STM images. More specifically, the representation extracted from the input of the last-but-one linear layer of the network, consisting of a vector of length 4,096, encodes attributes of the images that are sensitive to their thematic content. Formally, this very construction yields a non-linear map

$$f_\theta : \mathbb{R}^{224\times224\times3} \to \mathbb{R}^{4096}, \quad f_\theta : x \mapsto f_\theta(x), \tag{1}$$

sending each image of 224×224 pixels and 3-color channels to the corresponding representation. Since the visual characteristics of an image and the nature of the material composing the sample are strongly correlated, images with similar representations are likely to correspond to the same material category.

Following this line of thought, we started from a set of 30 elements of the STM dataset, composed of 10 manually labeled images for each of the three material categories described above.

Given two images x_1 and x_2, their similarity in content is well described by the cosine similarity between the corresponding representations defined in Equation 2.1:

$$S_{cos}(x_1, x_2) = \frac{f_\theta(x_1) \cdot f_\theta(x_2)}{\|f_\theta(x_1)\| \|f_\theta(x_2)\|}. \tag{2}$$

For each image x, the elements of the dataset on which the function $S_{cox}(x,\_)$ assumes a higher value corresponding to putative images in the same class of x. For each of the 30 labeled images, 24 images were selected with this automatic method and manually verified. On the 720 images obtained following this procedure a further manual verification has been applied to avoid the following behaviours: choice of

images which are almost identical to the retrieval seed, choice of images in different material classes from the seed but visually similar as taken at a different scale. This procedure leaves us with a final collection of 290 images labeled with the corresponding material category recorded in 64 days. Using this collection, we selected images recorded in the same days and labeled them correspondingly, and after a final manual verification, we obtained the final dataset of 7,287 images.

Despite our strategy being tailored to the specific case of the STM Dataset, each aspect of the selection process, from the manual annotation to the ML procedure, can be generalized to similar contexts upon slight modification of [9], in particular when dealing with annotations of microscopy images required for a FAIRification workflow. A more detailed description of the methodology, the technical specification and the validation criteria of the entire pipeline is available in the master thesis of the first author of this article [34].

2.2 STM metadata explorer

The STM dataset is enriched with useful metadata that increase the findability of relevant images. However, it is fundamental to provide scientists with a web service to facilitate and simplify the search process. Here, we present STM Metadata Explorer, an easy-to-use and interactive web service developed as an integrated service within Trieste Advanced Data Services (TriDAS) to visually explore images' metadata through interactive and downloadable plots. The core logic of the web service is designed around the metadata that users can select through the platform to find the relevant images. We selected a small subset of metadata that provide significant information about image characteristics and microscope settings, listed and described in Table 1.

Table 1. STM Explorer metadata available.

Metadata	Description
Date	image acquisition date
FieldXSizeinnm	*X* dimension in nanometers of the scan size
FieldYSizeinnm	*Y* dimension in nanometers of the scan size
XOffset	*X* coordinate of the tip offset in nanometers from the center of the scan axes
YOffset	*Y* coordinate of the tip offset in nanometers from the center of the scan axes
ScanSpeed	speed measured in nanometers per second of the scan
ScanAngle	rotation angle of the fast scan direction in the *XY* plane measured in degrees
GapVoltage	bias voltage applied between tip and sample in the constant current scan mode, measured in Volts
LoopGain	integral term of the PID feedback loop controller of the tunneling current
FeedbackSet	setpoint of the tunneling current, measured in nanoamperes
Label	sample material composition

The web service workflow is summarized in Figure 3 and allows users to visually explore images' metadata through a quantile plot for a single metadata field and a scatter plot that shows the distribution of images between two chosen metadata fields. In both plots, hovering on top of plot objects creates a pop up that shows information about that object: the number of images and value intervals for quantiles plots and the number of images and metadata values for each field in the scatter plot. The right toolbar lets users interact with the plots by moving, zooming in and out, and saving plots as images.

These features are useful for a first exploration of the dataset which then should be downloaded for further analysis and image visualization. To address this issue, the scatter plot features the selection of a specific metadata combination to retrieve a new page containing a table with metadata fields for each image in that subset. On this page, researchers can select, order, filter, and search images based on their metadata values. Moreover, the *ID* column consists of each image unique identifier in the database and, by clicking on it, the corresponding STM image is rendered and shown in a new page, where a download feature is included to obtain data, metadata, plot and provenance metadata for each image.

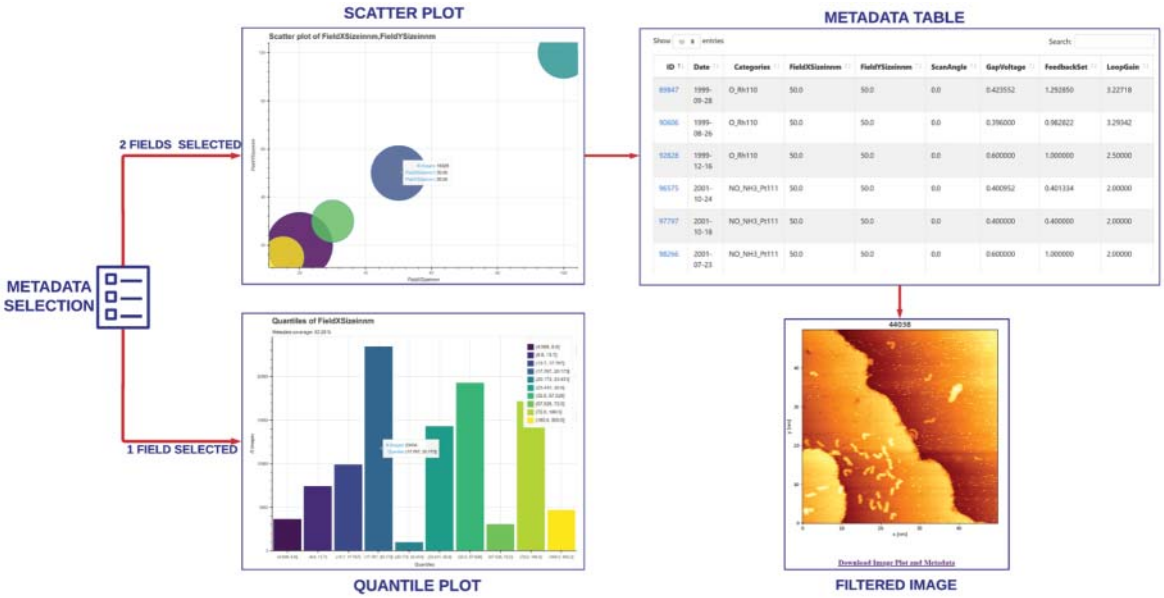

Figure 3. The web service workflow on the TriDAS website. Users select metadata and based on the fields selected, can explore the images metadata through a quantile plot for a single metadata field and a scatter plot that shows the distribution of images between two chosen metadata fields.

TriDAS is implemented in Python, use *Bokeh* [35] for data visualization, *spym* [36] to process and plot images, and *Flask* [37] framework as backend. The source code, as well as a list of all used software packages are publicly available [9] and reported in the provenance metadata.

2.3 Application of W3C PROV to STM case study

Provenance is a kind of metadata that describes the history of data from the original data sources to data products. Provenance information, that tracks the processes applied to data, from the origin to the final results, is critical to enable reproducibility [38] and reusability in scientific research experiments. In relation to these needs, we present an approach to describe the provenance of our use case on STM images by applying the PROV-DM [39], a generic data model of the W3C PROV standard [40].

As a first step, we designed the workflow of the principal events performed during the FAIRification process of STM images, from the raw data folder generated by VT-STM measurements to the final image that can be visualised on the TriDAS website.

For each of the above activities, we first identified the actors responsible together with the generated outputs, and secondly, mapped them with the W3C PROV core concepts described in Table 2.

Table 2. Mapping between the elements of STM case study with W3C PROV concept types and relations.

W3C PROV concepts		STM elements
PROV types	Entities	• Raw data • Reference dataset • Structured & FAIR dataset • Filtered image
	Activities	• VT-STM measurements • Image selection & retrieval • Image labelling process • Metadata selection
	Agents	• STRAS research group • Data scientist • Research user • VT-STM microscope • Analysis software
	Usage	• Image selection & retrieval **used** Raw data • Image labelling process **used** the Reference dataset • Metadata selection **used** the Structured & FAIR dataset
PROV relations	Derivation	• Reference dataset **derived from** Raw data • Structured and FAIR dataset **derived from** the Reference dataset • Filtered image **derived from** Structured & FAIR dataset
	Generation	• Raw data **was generated by** VT-STM measurements • Reference dataset **was generated by** Image selection & retrieval • Structured & FAIR dataset **was generated by** Image labelling process • Filtered image **was generated by** Metadata selection
	Attribution	• Raw data **was attributed to** STRAS research group • Reference dataset **was attributed to** STRAS research group and Data scientist • Structured & FAIR dataset **was attributed to** STRAS research group and Data scientist • Filtered image **was attributed to** Research user
	Association	• VT-STM measurements **were associated with** STRAS research group and VT-STM microscope • Image selection & retrieval **was associated with** Data scientist • Image labelling process **was associated with** STRAS research group and Data scientist • Metadata selection **was associated with** Research user
	Delegation	• VT-STM microscope **acted on behalf of** STRAS research group • Analysis software **acted on behalf of** Data scientist

Part of the terms we used in the provenance workflow has been already agreed upon among the NFFA-Europe community as they have been defined in the NFFA-Europe Glossary [41] developed in collaboration with the Joint Lab "Integrated Model and Data Driven Materials Characterization" (MDMC) of the Helmholtz Association of German Research Centers [42]. For the mapping, we considered three components of PROV-DM: entities and activities, derivations, and agents with their responsibilities.

Entities: In PROV, an Entity is defined as "*a physical, digital, conceptual, or other kind of thing with some fixed aspects*" [39]. From PROV-DM core descriptions, we identified the following entities: Raw data, Reference dataset, Structured & FAIR dataset and Filtered image. In our case study, Raw data refers to the unorganized collection of 420,000 STM images acquired using the VT-STM microscope. Reference dataset groups together 110,000 images acquired in constant-current mode, while Structured & FAIR dataset includes 7,287 images manually labeled in three sample material categories. Finally, Filtered image corresponds to single images downloadable from the STM Metadata Explorer on the TriDAS website.

Activities: An Activity is "*something that occurs over a period of time and acts upon or with entities*" [43]. In our case, we mapped as Activities four events represented by: VT-STM measurements, Image selection & retrieval, Image labeling process, and Metadata selection. The first activity, VT-STM measurements, corresponds to image acquisition at CNR-IOM. It is followed by Image selection & retrieval, which describes the actions taken to obtain the Reference dataset from Raw data. The image labeling process is the pipeline used to enrich a subset of the Reference dataset with material composition metadata and finally, Metadata selection represents the workflow of the web APP to find a particular image of interest from the Structured & FAIR dataset.

Agents: In PROV, an Agent [39] can be a person, an organization, software or other entity that has some responsibility for a given activity or entity. We identified STRAS research group, VT-STM microscope, Data scientist and Research user as prov:Agents and Analysis software as prov:softwareAgent. STRAS research group indicates the researchers of the laboratory where the Raw data were generated. Data scientist is the person responsible for the FAIRification of the dataset while the Research user is the person interested in the data collected from the Structured & FAIR STM dataset.

The roadmap of the FAIRification activities and the subsequent mapping with W3C components leads to the provenance workflow presented in the graphical illustration (Figure 5).

PROV Activities are represented as lilac rectangles, PROV Agents as light orange pentagons and PROV Entities in light yellow ovals. The responsibility properties are depicted in pink. The workflow starts with VT-STM measurements attributed to STRAS research group and is associated with both STRAS research group and VT-STM microscope that acts on behalf of STRAS research group. VT-STM measurements generated Raw data that were used during Image selection & retrieval to generate the Reference dataset. The Reference dataset, which was derived from Raw data, was attributed both to STRAS research group and Data scientist. Analysis software acts on behalf of Data scientist and Research user. The image labelling process, associated with Data scientist and STRAS research group, used the Reference dataset to generate the Structured & FAIR dataset. Therefore, Structured & FAIR dataset derived from Reference dataset. At last,

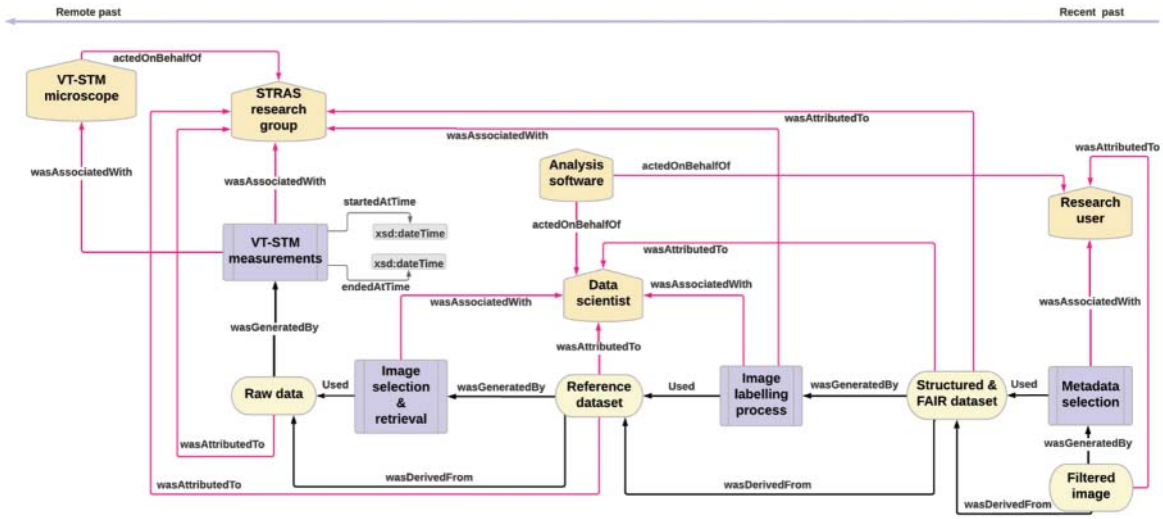

Figure 4. Graphical representation of the provenance workflow of STM images based on PROV-DM structures.

Metadata selection associated with Research user used the Structured & FAIR dataset to generate a Filtered image that was attributed to the Research user.

As a final step, we conducted a practical implementation of the above workflow, by using a PROV Python Library for W3C Provenance Data Model [44]. The provenance document created in Python was then exported in a JSON representation for PROV, PROV-JSON, thus providing a compact and accurate representation of PROV that is particularly suitable for interchanging PROV documents, allowing reproducibility.

3. CONCLUSIONS AND OUTLOOK

In this paper, we describe tools and services designed to improve the overall value of a scientific dataset of STM images by implementing different aspects of FAIR principles.

To address findability and accessibility, we used extracted metadata of each image as a filter to create a structured dataset from a raw data folder. We focused then on the annotation of images with sample material composition. As a result, we obtained a final dataset of 7,287 images of the surface of three materials, Gr_Ni100, Gr_Ni111 and N_Gr_Ni111.

We then created a web service to visually explore this information through intuitive graphical representations. The crucial component of this web service is metadata enrichment with information on sample composition, obtained with machine learning techniques. Moreover, we improved the usability of the dataset by including visualization and download functionalities directly in the web browser.

To address reproducibility, as well as interoperability and reusability we then focused on provenance metadata. The use of provenance standards is fundamental to achieve interoperability and encouraging the reuse of datasets. For these reasons, we applied to our case study the W3C PROV standard, which is a general, high-level standard for provenance. We used an open-source tool called F-UJI [45] to verify and assess the level of FAIRness achieved, which supports a programmatic FAIR assessment of research data based on a set of core metrics. The FAIR level result was 'advanced'. Even if we have a total score on findability and accessibility, the level of interoperability and reusability is moderate, showing some aspects we should improve in future work.

We foresee several directions for the future development of this case study: generalization of our provenance implementation, development of a domain-specific metadata schema for scanning probe microscopy, implementation of a FAIR-by-design workflow for the newly acquired data, continuous development of the STM Metadata explorer service and, more specifically concerning our case study, label propagation with semi-supervised learning [46, 47].

The implementation details of this work, in particular the PROV implementation, are somewhat specific to the present STM case study, but, in principle, they can be easily generalized and applied to a large number of scanning microscopy experiments (SEM, AFM, etc.), with the possibility to include active provenance capture [48].

The FAIRification process described in this work is applied to the legacy data acquired in the past twenty years in a STM laboratory. For newly acquired data, we started to actively implement a FAIR-by-design workflow starting from data acquisition. This process includes the use of an Electronic Laboratory Notebook (ELN) for reusability and provenance and the development of an open-source Python package for data reading to improve accessibility and interoperability [36]. A key activity in data management, especially in light of compliance with FAIR principles, is the development and adoption of metadata schemas. Currently, a metadata schema for STM is missing, and no standards are adopted for data and metadata acquired with this technique. Motivated by this lack, we started a coordinated effort to develop a standard STM metadata schema [49] with the final aim, after sharing and approval by the involved scientific community, to make it a de-facto standard in the field, openly available for reuse and a further extension to other scanning (probe) microscopy techniques. With this respect, we are planning to continue the work presented in this paper by converting the obtained structured dataset to make it compliant with the new STM metadata schema, as soon as it will be defined, thus further carrying on its process of FAIRification. We finally mention that an open-source software [50] is already available for loading and performing extensive data processing and analysis on several STM data formats, including those reported in this manuscript.

The STM Metadata explorer presented in this work was developed to improve the usability of the dataset. We plan to add analytics to assess user experience to further develop the service towards user needs.

Finally, we plan to extend the labelling to the whole dataset by label propagation, a powerful semi-supervised learning technique. Currently, the labelled samples are a small fraction of the total and its collection required extensive human annotation. Extending the labelling to the whole dataset will enable the development of more advanced services (such as advanced queries) and large-scale experimentation.

In conclusion, we believe that this work will inspire and engage a large scientific community in addressing the problems of data provenance, metadata schema development and, more in general, the FAIRification of scientific data. We are sure that this is an essential endeavour for the development of future research.

ACKNOWLEDGEMENTS

This work has been supported by funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 857650, EOSC-Pillar project and European Union's Horizon 2020 research and innovation programme under grant agreement No. 101007417 within the framework of the NFFA-Europe Pilot Joint Activities.

AUTHOR CONTRIBUTIONS

All authors contributed to the writing and review comments of the manuscript. Stefano Cozzini (0000-0001-6049-5242, stefano.cozzini@areasciencepark.it), conceived the idea of writing the manuscript. Tommaso Rodani (0000-0003-0570-3509, tommaso.rodani@areasciencepark.it)and Elda Osmenaj (0000-0002-4300-3012, osmenaj@iom.cnr.it) led the writing process. Tommaso Rodani (0000-0003-0570-3509, tommaso.rodani@areasciencepark.it) and Alberto Cazzaniga (0000-0001-6271-3303, alberto.cazzaniga@areasciencepark.it) contributed to the web service development and together with Mirco Panighel (0000-0001-8413-5196, panighel@iom.cnr.it) and Cristina Africh (0000-0002-1922-2557, africh@iom.cnr.it) obtained the final and FAIR dataset. Elda Osmenaj (0000-0002-4300-3012, osmenaj@iom.cnr.it) and Tommaso Rodani (0000-0003-0570-3509, tommaso.rodani@areasciencepark.it) contributed to the implementation of W3C PROV for provenance description.

DATA AVAILABILITY STATEMENT

The data that support the findings of this work are openly available in Zenodo at https://doi.org/10.5281/zenodo.5799773, under Creative Commons Attributions 4.0 International license. The source code used is openly available on GitHub at https://doi.org/10.5281/zenodo.4019640, under Apache License 2.0.

REFERENCES

[1] Wilkinson, M., et al.: Comment: The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3, 1–9 (2016)
[2] GO FAIR: FAIRification Process. Available at: https://www.go-fair.org/fair-principles/fairification-process/
[3] Jacobsen, A., et al.: A generic workflow for the data FAIRification process. Data Intelligence 2(1–2), 56–65 (2020)
[4] Sinaci, A.A., et al.: From raw data to FAIR data: the FAIRification workflow for health research. Methods of Information in Medicine 59.S 01, e21–e32 (2020).
[5] Scienta Omicron: VT SPM Lab. Available at: https://scientaomicron.com/en/products-solutions/SPM/VT-SPM-Lab

[6] TriDAS: Trieste Advanced Data Services. Available at: https://tridas.nffa.eu/
[7] Missier, P., Belhajjame, K., Cheney, J.: The W3C PROV family of specifications for modelling provenance metadata. In: Proceedings of the 16th International Conference on Extending Database Technology, pp. 773–776 (2013)
[8] Tommaso, R., et al.: Dataset of Scanning Tunneling Microscopy (STM) images of graphene on nickel. Version 1.0. Zenodo (2021)
[9] Tommaso, R.: t0m-R/STM images 0.1.0. Version 0.1.0. (2020)
[10] Zou, Z.Y., et al.: Strain release at the graphene-Ni(100) interface investigated by in-situ and operando scanning tunnelling microscopy. Carbon 172, 296–301 (2021)
[11] Sala, A., Zou, Z., Carnevali, V., et al.: Quantum Confinement in Aligned Zigzag "Pseudo-Ribbons" Embedded in Graphene on Ni(100). Advanced Functional Materials p. 2105844 (2021)
[12] Zou, Z., et al.: Honeycomb on Square Lattices: Geometric Studies and Strain Analysis of Moir'e Structures at a Symmetry-Mismatched Interface. The Journal of Physical Chemistry C 124(46), 25308–25315 (2020)
[13] Zou, Z., et al.: Operando atomic-scale study of graphene CVD growth at steps of polycrystalline nickel. Carbon 161, 528–534 (May 2020)
[14] Zou, Z., et al.: Graphene on nickel (100) micrograins: Modulating the interface interaction by extended moir'e superstructures. Carbon 130, 441–447 (2018)
[15] Carnevali, V., et al.: Doping of epitaxial graphene by direct incorporation of nickel adatoms. Nanoscale 11(21), 10358–10364 (2019)
[16] Patera, L.L., et al.: Real-time imaging of adatom-promoted graphene growth on nickel. Science 359(6381), 1243–1246 (2018)
[17] Africh, C., et al.: Switchable graphene-substrate coupling through formation/dissolution of an intercalated Ni-carbide layer. Scientific Reports 6(1), 19734 (2016)
[18] Patera, L.L., et al.: Temperature-Driven Changes of the Graphene Edge Structure on Ni(111): Substrate vs Hydrogen Passivation. Nano Letters 15(1), 56–62 (2014)
[19] Bianchini, F., et al.: Atomic Scale Identification of Coexisting Graphene Structures on Ni(111). The Journal of Physical Chemistry Letters 5(3), 467–473 (2014)
[20] Patera, L.L., et al.: In Situ Observations of the Atomistic Mechanisms of Ni Catalyzed Low Temperature Graphene Growth. ACS Nano 7(9), 7901–7912 (2013)
[21] Puppo, S.D., et al.: Tuning graphene doping by carbon monoxide intercalation at the Ni(111) interface. Carbon 176, 253–261 (2021)
[22] Fiori, S., et al.: Inside out growth method for high-quality nitrogendoped graphene. Carbon 171, 704–710 (2021)
[23] Perilli, D., et al.: Mechanism of CO Intercalation through the Graphene/Ni(111) Interface and Effect of Doping. The Journal of Physical Chemistry Letters 11(20), 8887–8892 (2020)
[24] Bengio, Y., Courville, A., Vincent, P.: Representation learning: a review and new perspectives. IEEE transactions on pattern analysis and machine intelligence 35(8), 1798–1828 (2013)
[25] Le Cun, Y., et al.: Handwritten digit recognition: applications of neural network chips and automatic learning. IEEE Communications Magazine 27(11), 41–46 (1989)
[26] Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press, 2016. Available at: http://www.deeplearningbook.org
[27] Pan, S.J., Yang, Q.: A Survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering 22(10), 1345–1359 (2010)
[28] Huh, M., Agrawal, P., Efros, A.: What makes ImageNet good for transfer learning? NeurIPS LSCVS 2016 Workshop (2016)

[29] Modarres, M., et al.: Neural Network for Nanoscience Scanning Electron Microscope Image Recognition. Scientific Reports 7 (2017)
[30] Aversa, R., et al.: Deep Learning, Feature Learning, and Clustering Analysis for SEM Image Classification. Data Intelligence 2(4), 513–528 (2020)
[31] Cazzaniga, A.: Representation Learning and Hierarchical Clustering for microscopy images. MHPC Thesis, Scuola Internazionale Superiore di Studi Avanzati (2020). Available at: http://hdl.handle.net/20.500.11767/119181
[32] Deng, J., et al.: Imagenet: A large-scale hierarchical image database. 2009 IEEE conference on computer vision and pattern recognition. IEEE, pp. 248–255 (2009)
[33] He, K., et al.: Deep Residual Learning for Image Recognition. (2016). https://doi.org/10.1109/CVPR.2016.90
[34] Rodani, T.: Machine Learning techniques and visualization tools for STM images at CNR-IOM labs (2020). Available at: https://doi.org/10.5281/zenodo.5801169
[35] Bokeh Development Team: Bokeh: Python library for interactive visualization (2020). Available at: https://bokeh.org/
[36] Panighel, M.: rescipy-project/spym: v0.7.0. Version v0.7.0. (2021). https://doi.org/10.5281/zenodo.5792910
[37] Grinberg, M.: Flask web development: developing web applications with python. O'Reilly Media, Inc., (2018)
[38] Gil, Y., et al.: Examining the Challenges of Scientific Workflows. Computer 40(12), 24–32 (2007)
[39] Belhajjame, K., et al.: Prov-dm: The prov data model. W3C Recommendation 14, 15–16, (2013)
[40] PROV-Overview: An Overview of the PROV Family of Documents. Project Report (2013). Available at: https://eprints.soton.ac.uk/356854/
[41] NFFA: NFFA Glossary. Available at: https://www.nffa.eu/apply/data-policy/glossary/
[42] MDMC: Integrated Model and Data Driven Materials Characterization. Available at: https://jl-mdmc-helmholtz.de
[43] Lebo, T., et al.: PROV-O: The PROV Ontology. English. W3C Recommendation. United States: World Wide Web Consortium, (2013)
[44] Trung Dong, H.: Prov Python (2014). Available at: https://github.com/trungdong/prov
[45] Devaraju, A., Huber, R.: F-UJI – An Automated FAIR Data Assessment Tool. Version v1.0.0, (2020). https://doi.org/10.5281/zenodo.4063720
[46] Zhu, X., Ghahramani, Z.: Learning from labeled and unlabeled data with label propagation. Tech. Rep., Technical Report CMU-CALD-02–107, Carnegie Mellon University (2002)
[47] Murphy, K.P.: Machine learning: a probabilistic perspective. MIT press (2012)
[48] Spinuso, A., Atkinson, M., Magnoni, F.: Active provenance for Data-Intensive workflows: engaging users and developers. 2019 15th International Conference on eScience (eScience), IEEE, pp. 560–569 (2019)
[49] Panighel, M.. rescipy-project/nxstm: v0.3.0. Version v0.3.0. (2021), https://doi.org/10.5281/zenodo.5792931
[50] Němcas, D., Klapetek, P.: Gwyddion: an open-source software for SPM data analysis. Open Physics 10(1), 181–188 (2012)

AUTHOR BIOGRAPHY

Tommaso Rodani is a Research Fellow at AREA Science Park, working on the development of machine learning algorithms for genomics data in HPC infrastructure. He received a Master's Degree in Data Science and Scientific Computing at the University of Trieste, Italy.
ORCID: 0000-0003-0570-3509

Elda Osmenaj is a Research Fellow in Scientific Data Management at National Research Council-IOM Materials Foundry (CNR-IOM) in Trieste. She is coordinating Task 6.1 activities for the Use Case 1 of the EOSC-Pillar project and is involved in the Joint Activity 6 (JA6) on Implementing FAIR data approach within the NFFA-Europe PILOT project. She is a member of the ICDI-CC (Italian Computing and Data Infrastructure-Competence Centre) for Open Science, FAIR and EOSC.
ORCID: 0000-0002-4300-3012

Alberto Cazzaniga is a permanent Researcher at the RIT Institute at AREA Science Park in Trieste, where he works and jointly coordinate the activities of the LADE research group focused on applications of machine learning and deep learning techniques in life sciences. After completing a DPhil in Pure Mathematics at the University of Oxford and post-doctoral studies at AIMS-SA funded by the Claude Leon Foundation, he moved to Trieste, where he completed a Master in High Performance Computing at SISSA and ICTP, and transitioned to research in advanced statistical modelling.
ORCID: 0000-0001-6271-3303

Mirco Panighel is a Post doctoral fellow at CNR-IOM in Trieste and his current research activity involves variable temperature scanning tunneling microscopy (STM) of graphene nano-structures in ultra-high vacuum (UHV) and the development of Python packages for scientific data analysis. He is part of the Data Management team in the NFFA-Europe Pilot project and is responsible for the Open and FAIR data implementation in the STRAS laboratory.
ORCID: 0000-0001-8413-5196

Cristina Africh is senior research scientist at CNR-IOM and head of the Structure and reactivity at the atomic scale (STRAS) group at CNR-IOM. Her research focuses on the investigation of surfaces at the atomic scale, mainly by scanning tunneling microscopy. Cristina Africh is also coordinator of the NFFA-Europe interoperable distributed research infrastructure (IDRIN), whose interoperability relies on FAIR data management.
ORCID: 0000-0002-1922-2557

Stefano Cozzini is presently director of the Institute of Research and Technologies at Area Science Park where he coordinates several scientific infrastructures and projects at national and international level. He has more than 20 years' experience in the area of scientific computing and HPC/Data e-infrastructures. His main scientific interests are scientific computing and machine learning techniques applied to scientific data management. He is presently actively involved in the master's degree on Data Science and Scientific Computing master at University of Trieste, Italy.
ORCID: 0000-0001-6049-5242