doi:10.1017/S1431927620023806
Microsc. Microanal. 26 (Suppl 2), 2020
© Microscopy Society of America 2020

Constructing Self-Labeled Materials Imaging Datasets from Open Access Scientific Journals with EXSCLAIM!

Eric Schwenker[1,2], Weixin Jiang[1,3], Trevor Spreadbury[1], Sarah O'Brien[1], Nicola Ferrier[4], Oliver Cossair[3] and Maria KY Chan[1]

[1]Center for Nanoscale Materials, Argonne National Laboratory, Lemont, Illinois, United States, [2]Department of Materials Science and Engineering, Northwestern University, Evanston, Illinois, United States, [3]Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, Illinois, United States, [4]Mathematics and Computer Science Division, Argonne National Laboratory, Lemont, Illinois, United States

Due to recent improvements in image resolution and acquisition speeds, materials microscopy is experiencing an explosion in imaging data. Yet, despite the volume of images generated, the overall accessibility landscape is highly fragmented, as researchers who do release images to the public, often only do so as snapshots of their larger private dataset in context of scientific journal publications. The effort to automatically consolidate images and descriptive information from web-based platforms has garnered broad attention from the computer vision, language technologies, and chemistry/materials informatics communities [1]–[4]. However, these methods are problematic for scientific figures because over 30% of figures are compound in nature [5], and it is the individual images themselves, paired with relevant context, that are necessary for construction a proper labeled dataset. To this end, we outline the design of a software pipeline for the automatic **EX**traction, **S**eparation, and **C**aption-based natural **L**anguage **A**nnotation of **IM**ages from scientific figures (EXSCLAIM!). Successful consolidation of materials imaging across literature sources will enhance navigation and searchability of materials microscopy images for both novice and experienced researchers, as well as establish the framework necessary for users to search by images, text, or some combination of both.

The EXSCLAIM! pipeline is composed of three main extraction classes: a (1) *JournalScraper*, (2) *CaptionDistributor*, and (3) *FigureSeparator* [6]. These extraction classes are executed sequentially, reading a user query and manipulating a JSON data structure to document all information pertinent to the creation of a large-scale imaging dataset offline. The *JournalScraper* is designed to collect images from open access articles in Nature, American Chemical Society (ACS), and Royal Society of Chemistry (RSC) journal families. The *CaptionDistributor* uses custom adaptations of the tokenization and tagging tools in the spaCy [7] natural language processing (NLP) toolkit to distribute complete grammatical segments of text to the associated image in the figure, alongside any keywords found in the text that are relevant to the initial query. Finally, the figure itself is segmented (separated) at a semantic level into master images, which appropriately represent the entirety of the image that each assigned caption refers to. This could be a single image, or a collection of images highlighting temporal, structural evolution, etc., that are part of the same subfigure description.

As a test case, open access articles in the Nature family journals were collected from a search of electron microscopy images combined with a wildcard-style general search of nanostructures (i.e. nano* = nanoparticle, nanosheet, nanoflake, nanorod, etc.). This particular query returned a total of 13,450 open-source articles with 83,504 figure-caption pairs and over 4300 microscopy images explicitly related to one of the nanostructure types in the search query. To quantify the caption distribution steps, we analyzed images returned from within the top 10% of articles (based on relevance to the query) and found the precision rate at approximately 87% for images related to the query keyword that were classified as microscopy with high-confidence. A schematic summarizing some of the content saved in the image record of the final JSON is presented in Figure 1. Furthermore, the caption text is distributed in a grammatically appropriate way, even with multiple descriptive adjectives (the main keywords describing the individual images) referring to a single noun, “images”.

https://doi.org/10.1017/S1431927620023806 Published online by Cambridge University Press

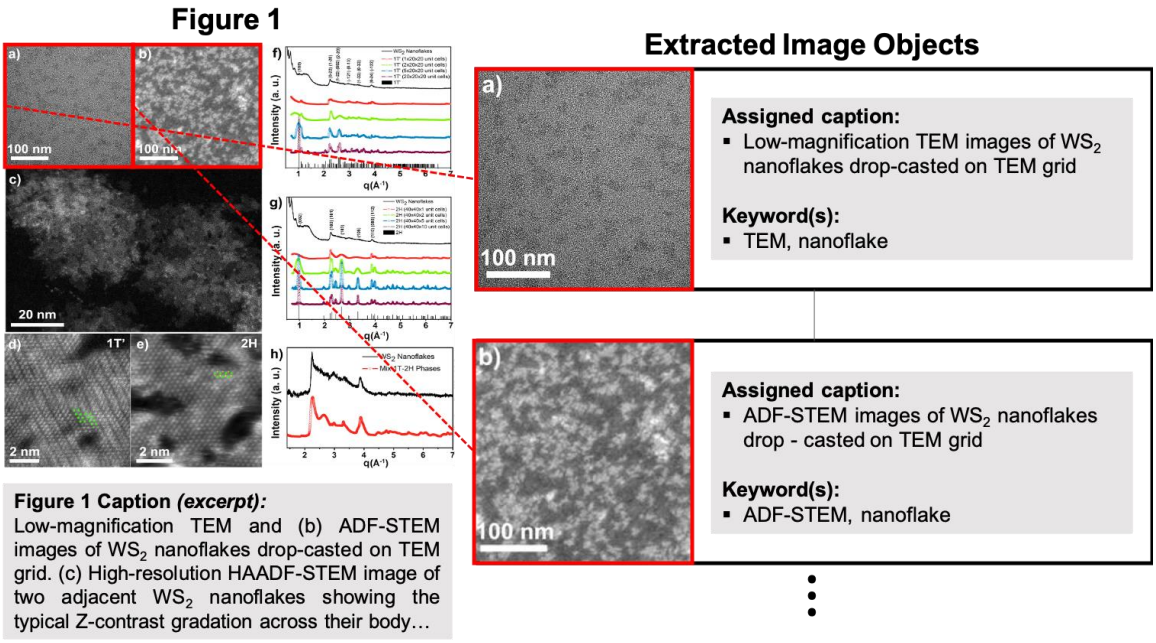

Figure 1. Schematic showing the relation between the initial figure/caption pair and subsequent image objects extracted in the EXSCLAIM! software pipeline. The highlighted microscopy image objects demonstrate a successful pairing between the images and their respective caption descriptions. Demo 'Figure 1' obtained from [8].

References

[1] F. Schroff, A. Criminisi, and A. Zisserman, "Harvesting image databases from the web," IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, no. 4, pp. 754–766, 2010

[2] L.-J. Li and L. Fei-Fei, "Optimol: automatic online picture collection via incremental model learning," Int. J. Comput. Vis., vol. 88, no. 2, pp. 147–168, 2010.

[3] X.-S. Hua and J. Li, "Prajna: Towards recognizing whatever you want from images without image labeling," in Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015.

[4] Y. Yao et al., "Towards Automatic Construction of Diverse, High-quality Image Datasets," IEEE Trans. Knowl. Data Eng., 2019.

[5] P.-S. Lee, J. D. West, and B. Howe, "Viziometrics: Analyzing Visual Information in the Scientific Literature," *IEEE Trans. Big Data*, 2017.

[6] Jiang *et al.*, "Semantic Segmentation for Compound Figures", *arXiv preprint arXiv:1912.07142*, 2019.

[7] https://spacy.io/

[8] R. Mastria *et al.* "In-plane Aligned Colloidal 2D WS 2 Nanoflakes for Solution-Processable Thin Films with High Planar Conductivity." *Sci. Rep.* 2019.

[9] This material is based upon work supported by Laboratory Directed Research and Development (LDRD) funding from Argonne National Laboratory, provided by the Director, Office of Science, of the U.S. Department of Energy under Contract No. DE-AC02-06CH11357. Use of the Center for Nanoscale Materials, an Office of Science user facility, was supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, under Contract No. DE-AC02-06CH11357. We gratefully acknowledge the computing resources provided and operated by the Joint Laboratory for System Evaluation (JLSE) at Argonne National Laboratory.

https://doi.org/10.1017/S1431927620023806 Published online by Cambridge University Press