ANIMO: Annotation of Biomed Image Modalities

Juan Trelles Trabucco
Dept. Comp. Sci.
University of Illinois at Chicago
Chicago, IL 60607
Email: jtrell2@uic.edu

Pengyuan Li
Dept. Comp. Info. Sci.
University of Delaware
Newark, DE 19716
Email: pengyuan@udel.edu

Cecilia Arighi
Dept. Comp. Info. Sci.
University of Delaware
Newark, DE 19716
Email: arighi@udel.edu

Daniela Raciti
Div. Biology Biological Eng.
California Inst. Technology
Pasadena, California 91125
Email: draciti@caltech.edu

Hagit Shatkay
Dept. Comp. Info. Sci.
University of Delaware
Newark, DE 19716
Email: shatkay@udel.edu

G. Elisabeta Marai
Dept. Comp. Sci.
University of Illinois at Chicago
Chicago, IL 60607
Email: gmarai@uic.edu

2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 978-1-6654-0126-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/BIBM52615.2021.9669898

***Abstract*—Figures within biomedical articles present essential evidence of the relevance of a publication in a curation workflow. In particular, visual cues of the image modality or experimental methods can help expert curators identify relevant papers from an increasing number of publications. Automating the identification of these content-bearing images can thus be helpful in computer-assisted curation. However, the paucity of labeled datasets and the specialized training required to label such images hinder the development of such tools. To address this problem, we present the design of ANIMO, a labeling system that integrates extraction and segmentation tools to ease the annotation burden. We first introduce two taxonomies of image modalities and experimental methods, derived in collaboration with curators. On the back-end of the system, we process batches of documents and create a labeling task per document. At the front-end, expert curators can access these tasks through a web interface and access the article of interest. We describe the evaluation of this system by a group of biocurators, and the human factor lessons learned from this interdisciplinary experience.**

I. Introduction

Biomedical research efforts, ranging from investigating treatment options to uncovering mechanisms underlying diseases, require targeted access to available information, primarily from published literature. At the core of this information-centric process, domain experts curate documents relevant to their domain of expertise. Such curation workflow encompasses the selection, organization, presentation, and annotation of relevant biomedical publications. However, with increasing numbers of scientific publications, there is significant interest in automating parts of the curation workflow, by providing semi-automated techniques to assist human curation in the principled collection and annotation of this type of biomedical data [1], [2].

Most automated methods for biomedical document curation use text-mining techniques to retrieve information from within articles [3], [4]. However, images provide essential evidence for processes and experimental findings in these publications [5]. For instance, image content reveals relevant cues such as the target imaged, the image modality, or the experimental methods, which curators can then associate with the document topic. Therefore, automating the identification of content-bearing images and their image modalities (or image types) can be helpful in computer-assisted curation. At the same time, the paucity of large labeled datasets makes it challenging to train effective machine learning models, creating a chicken-and-egg problem: too few labeled images cannot support effective models, and ineffective models cannot assist in the labeling of images. Furthermore, existing modality classes may not match the granularity of the curator's domain requirements, and creating these taxonomies and then labeling images accordingly requires a high level of training and expertise on the side of the human user. Leveraging the domain expertise of curators through user-friendly, interactive labeling tools can help alleviate this problem.

Although most labeling tools provide capabilities to label portions of an image (e.g., Label-Studio [6]), they are not tailored to the biomedical curation task. First, figures within articles often comprise several subfigures requiring the repetitive segmentation of the original image (Fig. 1), a task that adds significant overhead on time spent by the few available domain experts. Second, curators often need context to label a sample correctly. This context includes figure captions, neighboring panes (Fig. 1), and even access to the original article. Yet most general-purpose labeling tools [6] only show isolated images and corresponding labels. For biomedical curation, there are no tools for labeling image modalities that incorporate this required context within one system. Furthermore, a better characterization of the user domain and documentation of the human factors behind the labeling process can increase the success rate of such a tool, while leveraging domain knowledge for downstream tasks.

In this work, we present the joint efforts of a team of biomedical curators and computer scientists to build a system for labeling images within biomedical publications, at the level of individual subfigures. The main contributions of this work are: 1) A description of a design process centered on the curator activities, leading to the requirements of a tool for labeling images from biomedical papers. 2) The design

process of two hierarchical taxonomies of image modalities and experimental setups, one specialized for biomed modalities and one specialized for Covid-19. 3) The implementation of the resulting design in a novel system named ANIMO (ANnotation of Image MOdalities), which integrates a back-end pipeline of image extraction and segmentation tools, a centralized database, and a front-end interface (Fig. 3). 4) A qualitative and quantitative evaluation of the system with domain experts and a discussion of the human factors lessons learned from this experience.

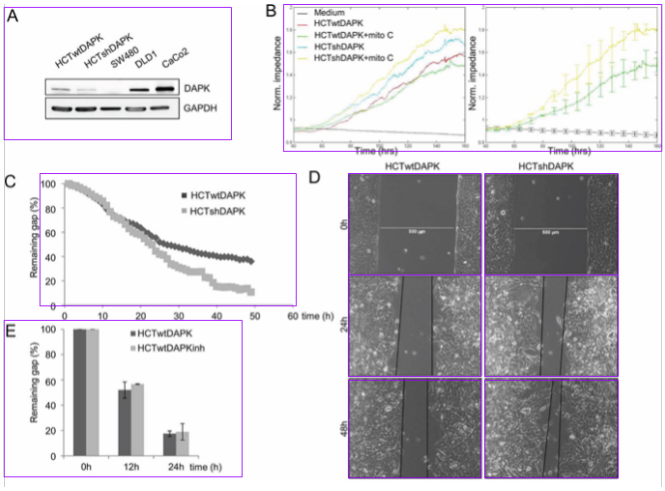

Fig. 1. Example figure (edited Figure 2 from Ivonavska *et al.* [7], shared under a Creative Commons 3.0 license) showing subfigure extraction instances with bounding boxes. Subfigures A and E are correctly extracted, and retain all of their content. Subfigure B has two panes with the same type of graph. Subfigure C suffers from over-cropping. Subfigure D suffers from over-fragmentation; each pane was extracted independently. Over-cropping and over-fragmentation cases depend on the space between components.

II. Background and Related Work

Document Triage. During document triage, curators evaluate the relevance of a document for a particular domain. To support this process, researchers have proposed binary classifiers based on features from titles, abstracts and post-publication annotations such as MeSH terms [8]. Other approaches added features from image captions to the vector representations [9], [10]. However, none of these two approaches used features from the figure content besides captions. In contrast, Shatkay *et al.* [5] identified the presence of experimental and graphical images to build feature vectors. Similarly, Li *et al.* [11] used a more complex taxonomy and convolutional neural networks to tag the images and create a vectorized representation. Our work supports this line of research where image features complement textual features.

Taxonomies for Image Modalities. In our work, we follow ImageCLEF's [12] definition for the term modality to denote the biomedical modality that originated the figure (e.g. light microscopy), or a general image type (e.g., line chart). ImageCLEF's taxonomy for the subfigure classification task consisted of 30 categories for diagnostic and generic biomedical illustrations; and several modality classifiers have flourished since then to solve this task [13], [14], [15], [16], [17], [18], [19], [12]. For document triage, Shatkay *et al.* [5] used a smaller taxonomy to classify graphical images, gels and microscopy images. We derived our biomedical curation and CORD-19 taxonomies mostly from these pieces of work. Other ad-hoc taxonomies have been used for classifying images by modality [20], and for organizing content in search engines like Open-i [21].

Tools such as SourceData [22] provide image curation tools focused on experimental details (e.g. biological entities). In our work, experimental methods are embedded in the taxonomy, specifically for microscopy and gel categories (e.g. InSitu Hybridization and Western Blot). We furthermore describe the process through which we developed the taxonomies.

Image Labeling. A variety of open-source tools support data labeling tasks for image classification, object detection, or segmentation. In many cases, a single toolkit [6] supports these and more general labeling options. However, for image classification, such general domain tools consider an image in isolation as a labeling task. Conversely, our work considers **a publication** through the lens of a labeling task, and we extract and arrange the content within the publication to provide context to support this manual operation.

In more specific domains, labeling techniques attempt to make such manual processes more efficient. For example, active learning strategies reduced the labeling workload by allowing a classifier to tackle unlabeled dataset. Gur *et al.* [23] used such a strategy with ultrasound images. In our work, we attempt to reduce the workload by integrating a segmentation back-end pipeline to extract the image panes from article figures. Another technique guides the domain expert through stepwise labeling to reduce annotation time [24]. In contrast, we organize labels in a matrix, provide context, and support multi-instance labeling to tackle the task faster. Also, although several techniques improve the quality of crowdsourced annotations [25], we do not explore crowdsourcing due to the required expertise to identify modalities. To the best of our knowledge, ANIMO is the first tool that supports interactive labeling of figures from biomedical publications based on the publication PDF content.

Visualizing Image and Text Data. Document figures provide cues related to the relevance of a publication. Yet, our experience suggests that biocurators may require access to textual data such as captions to identify the modality of an image. In the data visualization field, some approaches combine image and text data to provide convenient access. For instance, Document cards [26] provide a compact representation of a mixture of images and extracted key terms in different layouts. Another approach creates captions for every image in a collection [27], where image features and text features fed a co-embedding projection to 2D space; the resulting representation resembles a galaxy metaphor. In our interface design, we built on these approaches and improved data presentation by including extracted captions, in order to avoid dropping relevant infrequent words or presenting to the biocurators an unfamiliar, complex visual encoding [28], [29].

III. Methods

A. Collaboration Setting and Design Process

Over a period of four years, we participated in a multi-site collaboration with researchers from two different organizations (Protein Information Resource at the University of Delaware, and WormBase at the California Institute of Technology). One senior biocurator collaborator specializes in the curation of proteins for Uniprot, and another senior biocurator specializes in gene expression curation for the model organism *Caenorhabditis elegans*. Our team further included two data mining researchers and two visual computing researchers. Members of this team are all co-authors of this publication. During this collaboration, we also gathered information from a site visit to the Jackson Laboratory (Bar Harbor, ME, USA) where biologists study genetic mutations in mice, rats, fruitflies, and zebrafish. Although the first step of our project focuses on generating labeled data, we also collected information on the researchers' current curation workflows for further downstream tasks.

In our system design, we followed an activity-centered design (ACD) paradigm, an extension of human-centered design (HCD), because of ACD's proven success rate in interdisciplinary collaboration projects (63% for ACD compared to 25% for HCD) [30]. Furthermore, the ACD paradigm places emphasis on user activities and workflows, and as such, the final product gets its value from the importance of the activities it supports, rather than from the number of users using a system. The emphasis on activities fit our scientific collaboration where the curator teams were relatively small and featured high levels of expertise hard to replicate in naive users. We implemented the ACD paradigm through a series of tight iterations, where we met with curators to define functional specifications; define, adapt and revise taxonomies; evaluate prototype designs; and validate changes in the specifications. We used an online shared journal to allow curators to keep track of their questions, suggestions and concerns about the system, the user interface and taxonomies. Our team met monthly to evaluate progress, and we held frequent smaller group meetings with the curators to continuously update the system design or solve any lingering issues.

B. System Requirements

According to the activity-centered paradigm, we compiled our analysis using the following dimensions: activities and tasks, humans, data, workflow, and non-functional requirements. Data consists of biomedical article figures and captions from more than 100 journals: each figure can comprise several subfigures, and a subfigure can comprise one or multiple panes (Figure 1, D). Our workflow actors, the curators, have typically earned a Ph.D. and have several years of curatorial experience.

We identified two major activities and their respective tasks. The first activity *(A1)* initiates a labeling task for each document. Its first task *(T1.1)* extracts each figure and caption from a PDF file. Then, for each figure, a task *(T1.2)* splits the figure into its constituent subfigures. The next task *(T1.3)* places the content and metadata into a database. The final task, *(T1.4)*, starts the labeling task for a user. The second activity *(A2)* focuses on labeling subfigures in a document. Activity tasks include: *(T2.1)* access each figure and subfigure extracted from a document; *(T2.2)* annotate each subfigure based on the assigned taxonomy; *(T2.3)* annotate each subfigure based on the quality of the extracted content; *(T2.4)* whenever possible, annotate a group of subfigures at once; and *(T2.5)* display the PDF content when necessary.

Based on the non-functional requirements we gathered, we could assume that a third-party is responsible for placing the PDF documents to process and for starting the process. Activity A1 could be supported by the back end of the system, where it could run offline and without interactions from the end user. For activity A2, the requirements specified access through a web-browser in a desktop environment.

C. Taxonomies

Previous experiences of our text-mining experts with biomedical taxonomies facilitated the starting design of the biomedical taxonomy. In particular, our team members used similar taxonomies when they built document triage classifiers for curators working with the Mouse Genome Database [5], and when they developed award-winning image segmentation tools for the ImageCLEF Medical Task competition [31].

After multiple interview rounds and iterations through the labeling interface, we have adapted and finalized the taxonomy for biomedical image curation to comprise five main categories: Experimental, Organs & Organisms, Molecular Structure, Graphics, and Others. Only the Experimental category includes experimental methods (grey boxes in Fig. 2). This category is further divided into microscopy images (Light, Fluorescent and Electron Microscopy), plate images (i.e., techniques that monitor yeast/bacteria growth in plates), and gel-based images (i.e., gel electrophoresis techniques for proteins and DNA/RNA). Although fluorescence microscopy is also a light-based modality, the curators preferred to keep it in a distinct class from light microscopy [32]. Therefore, the reporter genes and immunochemistry, and in situ hybridization and whole mount methods appear under both modalities; however, only fluorescence microscopy includes the Episcopic fluorescence image capturing (EFIC) subcategory. In contrast to the ImageCLEF microscopy taxonomy, the biomedical curators we interviewed grouped scanning and transmission microscopy under electron microscopy. The remaining experimental methods appear under the gel-based modality: northern blot, western blot, reverse transcriptase (RT-PCR) and others (e.g., polyacrilamide and agarose).

The remaining categories did not include experimental setups. Organs and Organisms groups MRI & CT-scans, X-Rays, and visible light photographs. Molecular Structure includes 3D representation of molecules, chemical structures, and macromolecule sequences (protein and DNA). Finally, in the Graphics category we simplified, after several iterations, a vast taxonomy of graphs into scatterplots, line charts, histograms (and bar-based graphs), flowcharts (including pathways) and

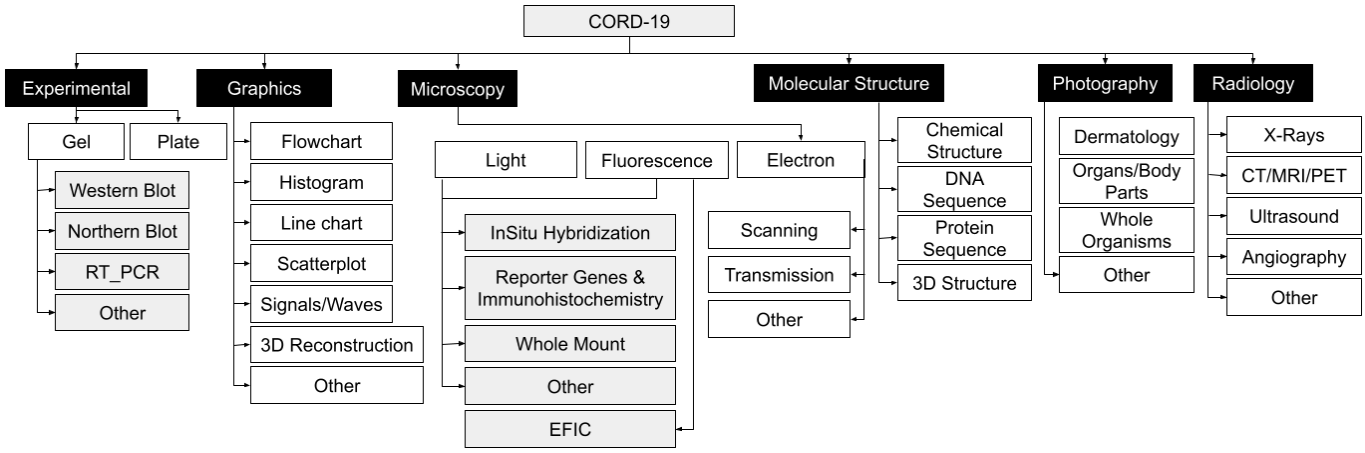

Fig. 2. Taxonomy of image modalities for CORD-19. Experimental setups are indicated by a grey background.

others. This taxonomy is not complete, but includes the most common modalities and experimental methods used in the study of most model organisms whose biological processes and functions are subjects of the curated bio databases. A preliminary version of this resulting taxonomy was used by Li *et al.* [11] for document retrieval.

The advent of the COVID-19 pandemic and its research questions motivated a second taxonomy, adapted to the research publications in the CORD-19 dataset [33]. To better reflect this new dataset, and based on biocurator input, it became necessary to separate the microscopy branch from the experimental category. Because the CORD-19 literature explicitly studied the effect of SARS-CoV-2 on various human body organs, we further divided the Organs & Organisms class into a Radiology-based class and a visible light Photography class. Finally, we included Signals and Waves as a subclass within the Graphics branch. These changes are in agreement with the original ImageCLEF taxonomy; however, the resulting new taxonomy (Fig. 2) does not include other Image CLEF subcategories for signals, nor several of the generic biomedical illustrations categories. Figure 4 shows this CORD-19 taxonomy as used in the ANIMO user interface.

D. System Back-End Design

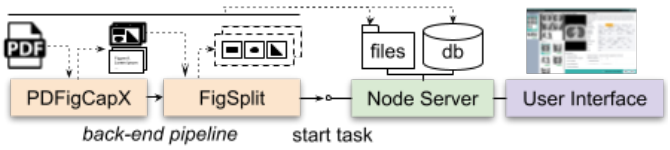

Fig. 3. ANIMO architecture. The pipeline processes a PDF document to extract images and captions (PDFigCapX [34]), and then splits the figures into constituent subfigures (FigSplit [35]). Finally, the pipeline starts a task and stores the metadata and content in a database and file storage, which are then accessed by the front-end.

The back end of the system consists of three components that leverage state-of-the-art solutions for extracting content from PDF documents. Given an input PDF document or a batch of them, the first step uses PDFigCapX [35] to extract the images and captions. Next, as the image-modality is a property of an individual image, we process each output figure to obtain its constituent panels. To do so, we use FigSplit [34], a tool that uses a connected component analysis to identify each panel. In the last step, we store the metadata in a MongoDB database, and assign a task to one curator based on a round-robin strategy. This strategy relies on user membership information (organization and group), which is specified at the start of the pipeline, along with the desired taxonomy. Our complete back-end consists of a MongoDB database, a NodeJS server, and a content extraction pipeline written in Python.

E. Front-End Design

Following discussions with the biocurators, we designed the ANIMO front end to tackle one document at a time. An inbox page provides access to every labeling task, and each task is pre-processed by the back-end pipeline workflow that implements the pre-processing tasks (A1). After selecting a labeling task, the user interface in Figure 4 is shown.

The left side of this interface view allows the user to select a specific figure or subfigure from the document (T1.1). Figures in that document are shown as a list of thumbnails in the leftmost column (Fig. 4 a). Next to it, the selected figure is shown in bigger size, and below, we show the extracted subfigures (Fig. 4 b,c). A colored background indicates the figure elements, whereas a color-coded badge shows whether the current element was reviewed (green), skipped (yellow) or pending (red). A figure is marked as "reviewed" once the curator reviews (i.e., labels) all its subfigures.

On the right side, curators can interact with each subfigure to provide labels and observations, based on the curation taxonomy derived and implemented. The image viewer shows the subfigure assigned for labeling; clicking on it displays the image at full size. Based on subsequent suggestions from the biocurators, the remaining components shown include the figure caption, the labeling matrix, the observations panel, the actions panel, and the PDF viewer (Figure 4 d,e,f,h). These

additional components provide contextual information for the labeling task that was deemed beneficial by the biocurators. We describe each component in detail below.

1) Labeling Matrix: The labeling matrix represents the hierarchical taxonomy without representing each inner level in the tree, which were not deemed useful by the curators. Instead, the matrix representation prioritizes the leaves of the tree, such that each row may be a node from different levels. For instance, each microscopy class gets a row with the corresponding experimental methods, thus reducing three levels of depth to a single row. Curators preferred a matrix representation instead of a list of badges or a search component, due to its quick access characteristics, and as a useful visual reminder of the whole taxonomy. In this manner, we place the taxonomy knowledge "in the world" as opposed to "in the user's head" [36]. To annotate a subfigure, curators then check the corresponding cells in the labeling matrix (T2.1).

Furthermore, our collaborators favored multi-label selection over a single selection, due to ambiguities that arise from the extraction process. For example, for some images, the curators may not be able to decide between two classes; or the image shown may include superimposed elements (e.g., a chemical structure placed in an empty area of a line chart). In addition, the segmentation tool may not correctly separate subfigures. For these irregular cases, the curator is required to input further comments in a text box.

2) Observations Panel: This panel shows the list of potential issues arising from the subfigure extraction process that could hinder the use of the subfigure as a training sample. The panel also includes a comments box for any feedback that the curator may give and a 'close-up' checkbox to indicate whether the subfigure represents a zoomed-in view of an entity. Common issues with the extraction pipeline include:

- Compound-image: the extracted image contains two or more subfigures from the original image (e.g., subfigures 'A' and 'B' are shown together). The curator needs to indicate the number of subfigures and whether they all belong to the same modality (homogeneous vs. heterogeneous).
- The image needs further cropping: Segmentation targeted the subfigure correctly but failed to remove content from its neighbours. This problem is common when there is not enough separation between subfigures.
- Over-cropped: Segmentation misses boundary annotations from the subfigure. Though overcropping commonly affects the text on graph's axis or subfigure organizers (e.g., A, B, (a), (b), and so on), these mistakes do not hurt the identification of the modality.
- Multipane: Segmentation identifies the subfigure, but the subfigure has several panes. The element may still be a suitable training sample only when all panes belong to the same modality.
- Over-fragmented: Segmentation identifies a pane within a subfigure due to the spacing between its elements. Over-fragmented samples can be suitable training samples when they target a pane. When over-fragmentation results in a subsection of the subfigure (e.g., a node in a graph), the element loses the related semantics and we recommend to skip it.

3) Actions Panel: Curators can save the updates to the selected labels and annotations by interacting with the action panel on the right bottom corner. The 'apply to all subfigures' checkbox provided multi-instance labeling (T2.4), but only for all the subfigures of the selected figure. For example, a majority of subfigures may belong to the same modality. In this case, curators may use the same label for all subfigures by checking their thumbnails, and then proceed to make any corrections. For images not worth labeling due to incorrect extraction, labelers can use the Skip button. Examples of skipped subfigures include pieces of text wrongly extracted or over-fragmented elements that break the semantics of the subfigure.

4) PDF Viewer: While designing the user interface, we expected that the figures, subfigures and captions could provide enough context to determine the modality label. However, curators explained that sometimes they need to refer to the PDF documents for paragraphs mentioning the figure (T2.5). Therefore, we added a PDF Viewer as a collapsible component displaying the page where the figure appears.

Finally, we complemented our web interface with a search page. Users can search labeled images by modality, state, or by the presence of observations. Potentially incorrectly labeled images can be flagged for later inspection. In addition, a bar chart shows the number of samples labeled, by modality.

The front-end of ANIMO was implemented using React, with code available at github.com/uic-evl/ANIMO.

IV. Evaluation

We evaluated ANIMO through a qualitative and quantitative approach, as appropriate in this type of collaborative design [37], [38], [39]. Because no design approach is failproof, we first report results provided by the two domain experts who participated in ANIMO's design. Because we recognize these experts as co-authors, we minimize reports of subjective feedback. Instead, we include a factual report of their analysis, and report the difference in capabilities relative to our collaborators' previous process [40]. We furthermore report qualitative and quantitative feedback from an evaluation with a set of researchers who were not involved in our design and development process.

A. Expert Usage Evaluation and Case Study

Quantitatively, our two curator collaborators tested ANIMO intermittently for eight months and labeled approximately 6,000 images, with approximately 5600 reviewed images and 400 skipped images. During this time, the domain experts recorded labeling concerns in the shared online journal document, including observations to use during labeling, explanations of exemplar cases for certain labels, and the granularity of the taxonomy.

We furthermore performed a detailed case study with these experts, where we asked each curator to work through a couple

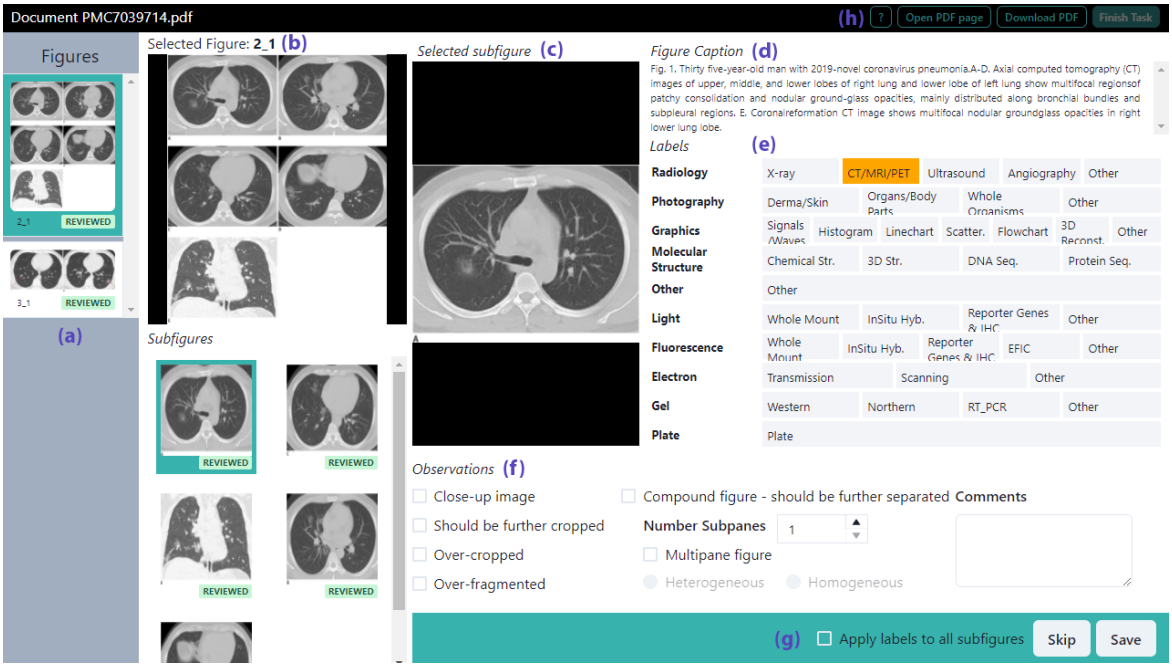

Fig. 4. Front-end interface for labeling a document showing a taxonomy adapted for the CORD-19 dataset. Thumbnails show document figures and subfigures (a-c), captions and access to the PDF document (d, h). Annotation features (e-f) capture modality labels and extraction errors.

of articles from their field of expertise. We processed those four documents through the back-end pipeline such that each curator received four tasks. We then asked the curators to label each document during a Zoom meeting with screen sharing while using a think-aloud protocol. The objective was to define a common ground for observations during labeling, evaluate the interface's ease of use, and reflect on the taxonomy. For this case study we used ANIMO with the first taxonomy.

Throughout a two-hour session, the domain experts discussed their perspectives on using annotations and how an image classification algorithm would benefit from these labels. For example, in a group of line chart figures, one subfigure was severely overcropped. From other similar samples, the curator could guess that the image was also a line chart; however, the facilitator and curators agreed to disregard severely affected samples from the training set. In this case, the context helped the curator but the resulting label would most likely affect the classifier. In a separate case, participants agreed that a subfigure missing any text indicating its subfigure label (e.g. Figure 'A') represented an overcropped sample where the observation may not have had a bad effect on the classifier. Furthermore, curators identified the need to annotate over-fragmentation cases, and the need of indicating whether multi-pane cases contained homogeneous or heterogeneous panes. The subjective expert feedback was enthusiastic, and both curators commented on how intuitive the interface was, how easy it was to label images, and how helpful the error reporting and multi-instance labeling features were. In addition, we observed that curators often provided further information about the image in the comments box; such comments resembled keywords from finer taxonomy subcategories, or objects identified within the image. As a consequence of this case study, one of the curators released a set of labeling guidelines for their students and collaborators to use as future reference.

We asked both labs to briefly explain how ANIMO changes their workflow.

Dr. Raciti responded: *"While WormBase does not classify images based on type (e.g. electron microscopy, fluorescent microscopy, etc..), [instantiating ANIMO with our ontologies and entities (e.g., genes)] will be extremely useful for curation. [It] will be very easy to use the modality taxonomy to further classify the images. It will also be instrumental to append annotations to the images themselves. In addition, in our current image curation workflow, curators need to manually crop relevant panes, save the image as a separate file, and annotate it. ANIMO provides an intuitive interface and does all the cropping and file naming work behind the scenes, saving substantial curator time."*

Dr. Arighi responded: *"Biocurators greatly need tools to assist in the annotation/tagging process of publications. ANIMO interface greatly facilitates the image visualization and tagging tasks in my project. It is interoperable with tools that process the publications in PDF to extract the figures and segment these into their corresponding panels. It has great features. I have been using ANIMO for tagging images to train systems for image classification in biomedical domain. However, this is just one application of ANIMO. Once a dataset of manually and/or automatically tagged images is available, this system would be ideal for retrieval of publications with images based on specific image types, which would be very useful for biocuration of specific biomedical data. This system will become very powerful when combined with a text mining*

component, as it will provide a very efficient method for biocurators to find experimental information about their topic of interest and the underlying methods providing evidence. "

B. Quantitative Evaluation

We asked seven biocurators, including five not affiliated with the project, to interact with ANIMO and provide qualitative and quantitative feedback. From this group, three were researchers working with different model organism databases (WormBase, ZFIN), two were senior graduate students working on the curation of proteins, and two were our collaborators. To provide a better motivation and alignment with the researchers' activities and workflows, we uploaded to the system publications from each researcher's domain of expertise. No personal data or further information about the evaluators was collected.

Each evaluator was provided with the following set of instructions: First, they were asked to select a labeling task from the interface inbox, based on their research interests. Second, they read a brief manual (one page) that described the main features of the labeling interface. Third, they proceeded to label the subfigures, and finally, they filled out a questionnaire (5-point Likert scale) about ANIMO.

Results from the questionnaire were overwhelmingly positive. We found that the majority of biocurators believed it was easy to label all the figures in the document ($M = 3.86 \pm .83$), although one curator found the task hard to complete. Similarly, they found the taxonomy easy to understand ($M = 3.71 \pm .88$), and biocurators found its representation as a matrix very helpful ($M = 4.29 \pm .69$), as opposed to other common representations (e.g., lists or dropdowns). Other features like reporting segmentation issues ($M = 4.29 \pm .45$) and labeling multiple subfigures at the same time ($M = 4.14 \pm 1.12$) had also a positive impact. Notably, biocurators considered contextual features like reading captions ($M = 4.57 \pm .5$), seeing thumbnails ($M = 4.57 \pm .5$) and accessing the PDF document ($M = 4.0 \pm .76$) to be very useful; however, they reported that they were often able to assign a modality label to the subfigure without looking at any of these features ($M = 3.85 \pm .35$).

We also received several suggestions, such as having a glossary of definitions and examples, or issues with figure shared axes/legends when those elements are not always repeated for all subfigures. Therefore, it is relevant to also show the original figure. Other suggestions included reminders to save any changes before selecting another figure, allowing multi-instance labeling for a selected group of images as well as skipping that selected group if needed, and adding open text when selecting the 'other' category on any taxonomy branch.

V. Discussion and Conclusion

One of the main lessons learned from this project was that fine-tuning a taxonomy of biomedical images is intensive, and highly dependent on the biocurators' focus and interests. During several months, we iterated through the categories and subcategories to be captured in the labeling interface taxonomy. The curators themselves changed their views about the relevant aspects of the taxonomy as they used the interface. Different biocuration groups also expressed different preferences concerning the taxonomy organization and even regarding the depth of the taxonomy. For example, our collaborator stated: *"In the past, we were interested in identifying pathways images from the biomedical literature. If the images were already labeled as such, it would have been very easy to pull them out. So, adding a graphics pathway tag would help"*. Although we started with one fixed taxonomy, we later made the taxonomy assigned to a task configurable. Figure 4 shows the taxonomy used for the CORD-19 dataset.

Another lesson learned was that different laboratories might interpret the same terms differently. For example, one group argued strongly, not incorrectly, for Charts being a type of Experimental data since some "charts" capture time-series of organ measurements, like EKGs. In another situation, a curator questioned whether 'histogram' provides the best description for bar-based charts. In several cases, clarifications and edits to the terminology used in the interface were necessary. Tooltips and sample images can ease the understanding of elements in the taxonomies.

Interaction-wise, curators agree on the usefulness of multi-instance labeling, given the composition of figures in publications. Curators preferred to label every subfigure in an image based on the majority class and then perform corrections. As a limitation, we only support multi-instance labeling per figure but, as seen in Figure 4, multi-instance labeling for all subfigures may speed up the process. Curators mentioned that in some cases, it would be helpful to provide multi-instance labeling for skipping figures with extraction errors.

An additional factor emerging from this experience was that the distribution of labeled images among the different classes is not uniform. As such, having the two-layer pipeline and assigning additional labeling tasks to the curators to generate more samples for a particular class is extremely helpful. In this respect, it was essential to manage annotation tasks for multiple batches of documents and multiple users. This observation also reinforces the importance of user expertise in biomedical image curation and the fact that crowdsourcing is a less viable option.

Given the variability in figure composition within biomedical articles, understanding which observations are helpful for labeling purposes remains a challenge. For instance, we noticed that a considerable number of over-cropped images are suitable for training purposes. Similarly, over-fragmented samples (panes in subfigures) can also be helpful for training purposes. However, our figure splitting component targets subfigures primarily. For this reason, our modeler had to explore the resulting dataset manually to flag undesired training samples. Adding a further step in the pipeline can solve this problem by including tools to edit the image bounding boxes and match them to the user-provided labels. Alternatively, we are training modality classifiers to pre-label subfigures and cut down curation time [32].

In conclusion, in this work, we introduced and evaluated

a novel labeling tool, ANIMO, to annotate image subfigures from the biomedical literature. The tool was designed with a focus on the user workflows and the human factors behind the biocuration process. ANIMO integrates extraction and segmentation tools to ease the annotation burden, and introduces taxonomies of image modalities and experimental methods revised in collaboration with curators. It supports multiple taxonomies, and includes the relevant controlled vocabulary for image types, which could be adapted for other projects. It offers the ability to add the same figure type tag to all panels at once when needed. It provides a quick way to report on image segmentation problems (e.g., when figures are over cropped, or need further separation into panels) that can be used to provide feedback to developing teams. Finally, ANIMO allows fast access to the publication content and full figure view to provide context when needed. A quantitative and qualitative evaluation with domain experts demonstrates that ANIMO effectively supports the annotation of biomedical literature figures.

ACKNOWLEDGMENT

This work was supported by awards from the National Institutes of Health (NLM R01LM012527, NLM U01GM120953) and the National Science Foundation (CNS-1828265, CNS-1625941). We thank our collaborators at MGI and GDX at Jackson Labs, in particular Judith Blake and Martin Ringwald, and members of the biocuration research groups at Caltech (WormBase), at the University of Oregon (ZFIN), and at the University of Delaware. We would also like to acknowledge Paul W. Sternberg, WormBase (NIH U24 HG002223).

REFERENCES

[1] L. Hirschman, G. A. Burns, M. Krallinger *et al.*, "Text mining for the biocuration workflow," *Database*, 2012.
[2] Z. Lu and L. Hirschman, "Biocuration workflows and text mining: overview of the BioCreative 2012 Workshop Track II," *Database*, 2012.
[3] H. Shatkay and M. Craven, *Mining the biomedical literature*. MIT Press, 2012.
[4] S. Ananiadou and J. Mcnaught, *Text mining for biology and biomedicine*, 2006.
[5] H. Shatkay, N. Chen, and D. Blostein, "Integrating image data into biomedical text categorization," *Bioinformatics*, pp. e446–e453, 2006.
[6] M. Tkachenko, M. Malyuk, N. Shevchenko *et al.*, "Label Studio: Data labeling software," 2020. [Online]. Available: github.com/heartexlabs/label-studio
[7] J. Ivanovska, I. Zlobec, S. Forster *et al.*, "Dapk loss in colon cancer tumor buds: implications for migration capacity of disseminating tumor cells," *Oncotarget*, 2015.
[8] G. Schneider, S. Clematide, and F. Rinaldi, "Detection of interaction articles and experimental methods in biomedical literature," *BMC bioinformatics*, pp. 1–11, 2011.
[9] X. Jiang, P. Li, J. Kadin *et al.*, "Integrating image caption information into biomedical document classification in support of biocuration," *Database*, 2020.
[10] G. Burns, X. Li, and N. Peng, "Building deep learning models for evidence classification from the open access biomedical literature," *Database*, 2019.
[11] P. Li, X. Jiang, G. Zhang *et al.*, "Utilizing image and caption information for biomedical document classification," *Bioinformatics*, 2021.
[12] A. García Seco de Herrera, R. Schaer, S. Bromuri *et al.*, "Overview of the ImageCLEF 2016 medical task," in *CLEF*, 2016.
[13] A. Kumar, J. Kim, D. Lyndon *et al.*, "An ensemble of fine-tuned convolutional neural networks for medical image classification," *IEEE Biomed Health Informatics*, pp. 31–40, 2016.
[14] S. Koitka and C. M. Friedrich, "Optimized convolutional neural network ensembles for medical subfigure classification," in *CLEF*, 2017, pp. 57–68.
[15] Y. Yu, H. Lin, J. Meng *et al.*, "Deep transfer learning for modality classification of medical images," *Information*, p. 91, 2017.
[16] S. Koitka and C. M. Friedrich, "Traditional feature engineering and deep learning approaches at medical classification task of ImageCLEF 2016." in *CLEF*, 2016, pp. 304–317.
[17] J. Zhang, Y. Xie, Q. Wu *et al.*, "Medical image classification using synergic deep learning," *Medical Image Analysis*, pp. 10–19, 2019.
[18] V. Andrearczyk and H. Müller, "Deep multimodal classification of image types in biomedical journal figures," in *CLEF*, 2018, pp. 3–14.
[19] O. Pelka and C. M. Friedrich, "FHDO Biomedical Computer Science Group at medical classification task of ImageCLEF 2015," in *CLEF*, 2015.
[20] S. Singh, K. Ho-Shon, S. Karimi, and L. Hamey, "Modality classification and concept detection in medical images using deep transfer learning," in *IEEE IVCNZ*, 2018, pp. 1–9.
[21] D. Demner-Fushman, S. Antani, M. Simpson *et al.*, "Design and development of a multimodal biomedical information retrieval system," *Journal Comp Science and Eng*, pp. 168–177, 2012.
[22] R. Liechti, L. George *et al.*, "Sourcedata: a semantic platform for curating and searching figures," *Nature methods*, pp. 1021–1022, 2017.
[23] Y. Gur, M. Moradi, H. Bulu *et al.*, "Towards an efficient way of building annotated medical image collections for big data studies," in *CVII-STENT/LABELS@MICCAI*, 2017, pp. 87–95.
[24] J. Son, S. Kim, S. J. Park, and K.-H. Jung, "An efficient and comprehensive labeling tool for large-scale annotation of fundus images," in *CVII-STENT/LABELS@MICCAI*, 2018, pp. 95–104.
[25] J. Yuan, C. Chen, W. Yang *et al.*, "A survey of visual analytics techniques for machine learning," *Computational Visual Media*, pp. 1–34, 2020.
[26] H. Strobelt, D. Oelke, C. Rohrdantz *et al.*, "Document cards: A top trumps visualization for documents," *IEEE Trans. Vis. Comp. Graph.*, pp. 1145–1152, 2009.
[27] X. Xie, X. Cai, J. Zhou *et al.*, "A semantic-based method for visualizing large image collections," *IEEE Trans. Vis. Comp. Graph.*, pp. 2362–2377, 2018.
[28] G. E. Marai, "Visual Scaffolding in Integrated Spatial and Nonspatial Analysis," in *EuroVis Workshop Visual Analytics (EuroVA)*. The Eurographics Association, 2015.
[29] G. E. Marai, C. Ma, A. T. Burks, F. Pellolio *et al.*, "Precision Risk Analysis of Cancer Therapy with Interactive Nomograms and Survival Plots," *IEEE Trans. Vis. Comp. Graph.*, p. 1732–1745, 2019.
[30] G. E. Marai, "Activity-centered domain characterization for problem-driven scientific visualization," *IEEE Trans. Vis. Comp. Graph.*, pp. 913–922, 2017.
[31] P. Li, S. Sorensen, A. Kolagunda *et al.*, "UDEL CIS at ImageCLEF medical task 2016," in *CLEF*, 2016, pp. 5–8.
[32] J. Trelles, P. Li, C. Arighi, H. Shatkay, and G. E. Marai, "Modality-classification of microscopy images using shallow variants of deep networks," in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2020, pp. 2379–2385.
[33] L. L. Wang, K. Lo, Y. Chandrasekhar *et al.*, "Cord-19: The covid-19 open research dataset," *ArXiv*, 2020.
[34] P. Li, X. Jiang, and H. Shatkay, "Figure and caption extraction from biomedical documents," *Bioinformatics*, pp. 4381–4388, 2019.
[35] P. Li, X. Jiang, C. Kambhamettu *et al.*, "Compound image segmentation of published biomedical figures," *Bioinformatics*, pp. 1192–1199, 2018.
[36] T. Luciani, A. Burks *et al.*, "Details-first, show context, overview last: supporting exploration of viscous fingers in large-scale ensemble simulations," *IEEE Trans. Vis. Comp. Graph.*, p. 1225–1235, 2018.
[37] P. Hanula, K. Piekutowski *et al.*, "DarkSky Halos: use-based exploration of dark matter formation data in a hybrid immersive virtual environment," *Front. in Robo. and AI*, p. 11, 2019.
[38] G. E. Marai, A. G. Forbes, and A. Johnson, "Interdisciplinary immersive analytics at the electronic visualization laboratory: Lessons learned and upcoming challenges," in *2016 VR Workshop on Immersive Analytics (IA)*. IEEE, 2016, pp. 54–59.
[39] G. E. Marai, J. Leigh, and A. Johnson, "Immersive analytics lessons from the electronic visualization laboratory: a 25-year perspective," *IEEE Comp. Graph. Appl.*, pp. 54–66, 2019.
[40] C. Floricel, N. Nipu *et al.*, "THALIS: Human-Machine Analysis of Longitudinal Symptoms in Cancer Therapy," *IEEE Trans. on Vis. and Comp. Graph.*, 2021.