MouseScholar: Evaluating an Image+Text Search System for Biocuration

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

Carla Floricel
Dept. Comp. Sci.
University of Illinois Chicago
Chicago, IL 60607
Email: cflori3@uic.edu

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

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

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

Martin Ringwald
The Jackson Laboratory
Bar Harbor, Maine 04609
Email: Martin.Ringwald@jax.org

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

***Abstract*—Biocuration is the process of analyzing biological or biomedical articles to organize biological data into data repositories using taxonomies and ontologies. Due to the expanding number of articles and the relatively small number of biocurators, automation is desired to improve the workflow of assessing articles worth curating. As figures convey essential information, automatically integrating images may improve curation. In this work, we instantiate and evaluate a first-in-kind, hybrid image+text document search system for biocuration. The system, MouseScholar, leverages an image modality taxonomy derived in collaboration with biocurators, in addition to figure segmentation, and classifiers components as a back-end and a streamlined front-end interface to search and present document results. We formally evaluated the system with ten biocurators on a mouse genome informatics biocuration dataset and collected feedback. The results demonstrate the benefits of blending text and image information when presenting scientific articles for biocuration.**

***Index Terms*—document search, biocuration**

I. Introduction

In biology and biomedical research, analyzing scientific articles to integrate biological data into well-structured data repositories is a process known as biocuration [1]. Although biocuration databases support many researchers who leverage these data to catalyze advancements in related domains [2], only biocurators seek to deal with a body of literature by reviewing and organizing it in taxonomies and ontologies. Due to the high level of expertise and training required to become a biocurator, there are significantly fewer biocurators than specialists in other biomedical fields, which leads to enormous biocurator workloads when processing information. Unsurprisingly, automation is fervently desired [3], and several tools exist to aid biocuration, for example, by triaging scientific articles [4]–[7] or annotating entities within the full text [8]. Still, while document search engines help biocurators access documents, existing options do not leverage image data in these documents, whereas biocurators routinely access essential image and figure information, such as the imaging modality, to make curation decisions.

Most document search engines specialized in scientific content, such as PubMed [9], Google Scholar, or PubTator [8], allow queries only over text features like titles and abstracts. PubTator extends the support for biocuration by tagging bioentities and enhancing the presentation of the search results with color-coded information. Few alternatives have explored searching over figures [10], yet they highlight the importance of mixing images and text when presenting search results. In particular, Trelles et al. [11] discuss the importance of integrating a domain-specific taxonomy to support more complex biomedical queries, and when perusing biomedical document query results [11], but do not specifically address the biocuration process.

In this work, we generalize a hybrid image+text retrieval approach, and apply it specifically to support biocuration. In this process, we address several significant challenges that have so far impeded the development of biocuration systems that would integrate figure data. First, figures are typically located inside PDF documents and seldom available as standalone images, impeding automated image analysis. Biocurators are also less interested in the figure content (e.g., "two mice") compared to the figure imaging modality (e.g., "light microscopy"). Second, figures in biomedical documents are usually compound images (i.e., they contain multiple sub-figures) where each subfigure could be of a different modality or type. Third, biocuration often relies on hierarchical ontologies and taxonomies, requiring hierarchical analysis of figure modalities. At the same time, labeled images to support machine learning of hierarchical modalities are uncommon, resulting in a need for either expert-level manual labeling (e.g., a "Northern blot" image is also a "gel" image) or leveraging incomplete labels, where labels for only a node in the taxonomy but not for higher levels are provided. In

U.S. National Institutes of Health, U.S. National Science Foundation

addition, data provenance and trustworthiness are relevant in biocuration. In contrast, for example, searching the web for "light microscopy" images returns microscope images, instead of images generated using the desired microscopy modality. Fourth, the image modality is often included in the figure caption but also frequently omitted, for example, a caption describing the progress of a disease may not mention that the figure includes microscopy images. Last, displaying figures without the context information located on the same document page has limited value from a biocuration perspective.

Based on a six-year collaboration with biocurators at four major sites (University of Delaware [UniProt], Caltech [WormBase], Princeton [BioGRID], and Jackson Laboratory [Mouse Genome Informatics, Gene Expression Database]), we introduce a biocuration search system, MouseScholar, which leverages image biomodality information in a subset of publications from the Mouse Genome Informatics Database (MGI, Jackson Laboratory). Image modalities denote the method of creation or acquisition of an image [12], ranging from laboratory methods using microscopes or radiology imaging to computer-generated plots to summarize experimental results. The modality information serves as a summary for the experiments in a study, and can serve as a proxy for the actual image content.

In this system we leverage a biocuration-specific taxonomy of image modalities, which serves as the system's connecting tissue, and we integrate document and image-derived data extracted through a processing back-end, including captions and previews for pages containing figures. A front-end interface allows for complex biocuration searches, and presents the relevant document results, enhanced with text highlighting in the abstract and captions, along with the figure-related data, including previews for those pages containing figures, subfigures in compound figures, and image biomodality information. The result is a first-in-class biocuration system to support document searching over figure modalities, hierarchical taxonomies and compound figures, and to present this evidence in conjunction with captions, relevant text, subfigure and context information. We demonstrate MouseScholar on a subset of documents in the Mouse Genome Informatics Database at the Jackson Laboratory, Maine, U.S.A.. We evaluate the system usability with 10 biocurators and we report the evaluation results. Our source code is publicly available and can be found at: https://github.com/uic-evl/bio-search.

II. Methods

A. Biocuration Requirements Engineering

Our solution design started with multiple interviews and observation studies we performed with biocurators at our collaborating sites. Through these interviews and observation in the workplace, we gained an in-depth understanding of a typical biocuration workflow, as well as of the variations among different groups and different sites. We observed the use of different ontologies and taxonomies, depending on the problem studied. We noted the use of pre-filtered corpora of publications, and the overall unbalanced relevant/not-relevant distribution of publications in repositories like PubMed [6]. We further noted the high level of expertise demonstrated by biocurators, and we paid close attention to their use of information, and, in particular, of figures. We confirmed that figures played a dominant role in the preliminary analysis of publications, and we observed biocurators painstakingly opening each publication to examine figures and figure-adjacent information [13].

Based on these observations, we inferred and validated the following requirements:

- Support batch-process segmentation of figures from PDFs
- Extract subfigures from compound figures
- Support taxonomies, and use these taxonomies when curating subfigures
- Support automated hierarchy-based modality-labeling of each subfigure
- Extract captions and mine the captions for modality information
- Extract additional relevant information (e.g., in the document title, abstract, etc.)
- Provide context in the form of hybrid text-image information and page previews
- Present the evidence supporting the relevance of a particular document

Additional non-functional requirements included online availability of the system, concurrent multi-user access, protecting proprietary document collections from public use, and a low learning curve.

B. Dataset

We instantiate our approach on a subset of PDF documents curated by the Mouse Genome Informatics (MGI) project between 2012 and 2016. The MGI database provides the most comprehensive resources using laboratory mice as model organisms [14]. MGI comprises several databases, including the Gene Expression Database (GXD) [15], [16], an extensive resource of mouse development expression information. Specifically, GXD focuses on endogenous gene expression data during the development of wild-type and mutant mice, including studies considering the endogenous expression pattern of a gene of interest (e.g., knock-in reporter studies) and excluding studies researching treatment effects or exogenous factors. The majority of these documents are curated manually from the literature [17].

Because all publications are part of MGI, each paper deals with mice, but they may not be relevant to, for example, GXD. Papers that are deemed to be GXD-irrelevant could either have no expression data, or only types of expression data that are not captured as part of GXD, such as studies reporting on ectopic gene expression via the use of transgenes [17].

The subset we use, denoted as GXD_{2000} [4], comprises 2,000 MGI collection documents, whose relevance to the GXD is to be determined during biocuration. The dataset is thus a good target for illustrating the manual component of the biocuration process [4], [6]. In addition, the dataset includes a metadata file with document attributes including titles, PubMed's unique identifiers, publication year, and abstracts.

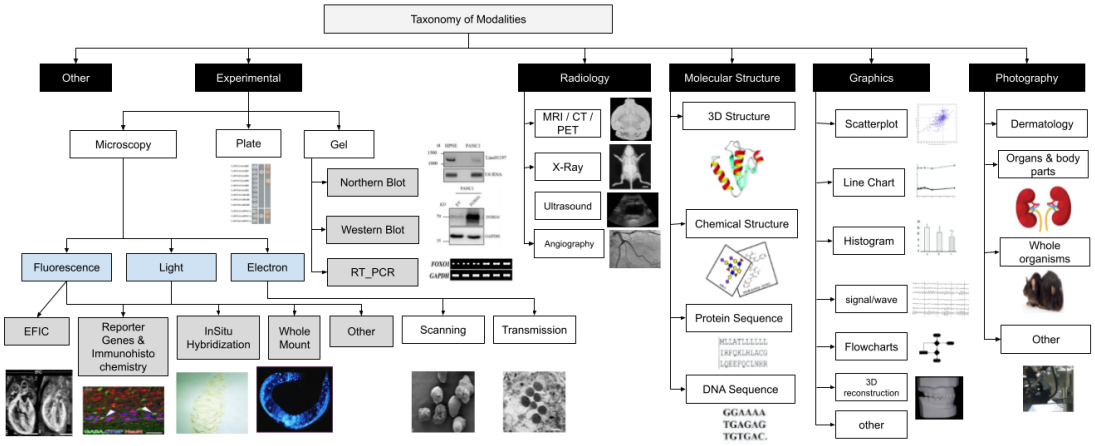

Fig. 1. Biocuration taxonomy of figures in scientific publications, with examples.

C. Instantiating for biocuration

Together with our biocurator collaborators, we gradually designed a taxonomy of image modalities [11], which represents the image acquisition methods and experimental approaches. The hierarchical taxonomy identifies seven coarse categories (experimental, graphics, microscopy, molecular structure, photography, radiology, and others), where each category can be further specialized down to two levels. For instance, experimental images include gels created using different laboratory techniques like Western or Northern blotting.

On a machine with 16GB RAM, 16 cores, and an RTX 2070 GPU, we first cloned the Trelles et al. [11] repository containing the project's front-end and back-end components. In addition, we installed a PostgreSQL database, an Apache Lucene index server, Matlab, PyTorch, and an Nginx web server. Then, we installed the onboarding Python packages from the repository, which includes standalone components: the PDFigCapX tool, developed by Li et al. [18], for identifying and extracting captions and figures, and the FigSplit tool, developed by Li et al. [19], for identifying and extracting subfigures from PDF documents. Once setup was complete, we then performed the extraction of figures and captions from the 2,000 GXD_{2000} PDF documents.

Next, we executed the figure segmentation scripts, which provide a wrapper for MATLAB, to obtain the sub-figures and related metadata, which includes the subfigures coordinates within a figure (i.e., bounding boxes) and page location. After that, we used import scripts to insert the document metadata, figure, and sub-figure attributes into the PostgreSQL database. Document attributes included title, authors, DOI, journal, and publication date. Our code can be extended to consider cases where the documents metadata needs to be fetched from a different source. We kept the raw image content on the local file system while the database stored the asset locations and prediction data that can support related image labeling efforts.

Then, we predicted the image modality per image using convolutional neural network classifiers and the taxonomy scheme we had derived. To this end, we leveraged the pre-trained image classifiers provided by Trelles et al. [11], a set of hierarchical classifiers developed with PyTorch and PyTorch Lightning. After downloading the model weights from the repository release, we used the repository core-backend scripts to infer the modality predictions. To do so, we first matched each desired classifier name with the path to the corresponding classifier weights. Next, we executed the inference scripts on the GXD_{2000} subfigure image collection. The scripts use either an EfficientNet [20] or a ResNet [21] model, depending on previous experimental results.

The result of the classification step generated 84,388 labels distributed across the 26 nodes in the taxonomy. In situations where the target taxonomy did not match the repository provided taxonomy, the configuration supports deleting nodes, although not deleting classes within them. Leveraging a significantly different taxonomy, such as adding new parent or child nodes, would have required acquiring labeled data and training by transfer learning, or creating accompanying classification models from scratch. In our training procedure, we experimented with 80/10/10 and partitions for medium sized-datasets like the microscopy subset (thousands of images) and 90/10/5 for large sized-datasets like the graphics subset (hundreds of thousands). Smaller datasets may require different partitions.

Next, we leveraged the Apache Lucene server to index the document attributes, figure captions, and the predicted image modality labels. These indexes support the queries documents and image-based data. We ran the indexing scripts to create the indexes and deployed the Flask application server, which interacts with the Apache Lucene server.

We finished the installation by installing the npm packages for the front-end React application and a Node.js server to support a basic single-user authentication mechanism. We built the projects and deployed the front-end app in the Nginx web server and the login service on a Node.js instance.

D. Front-end features

The front-end of MouseScholar supports queries using keywords, date ranges, and image modalities (Fig. 2A). Com-

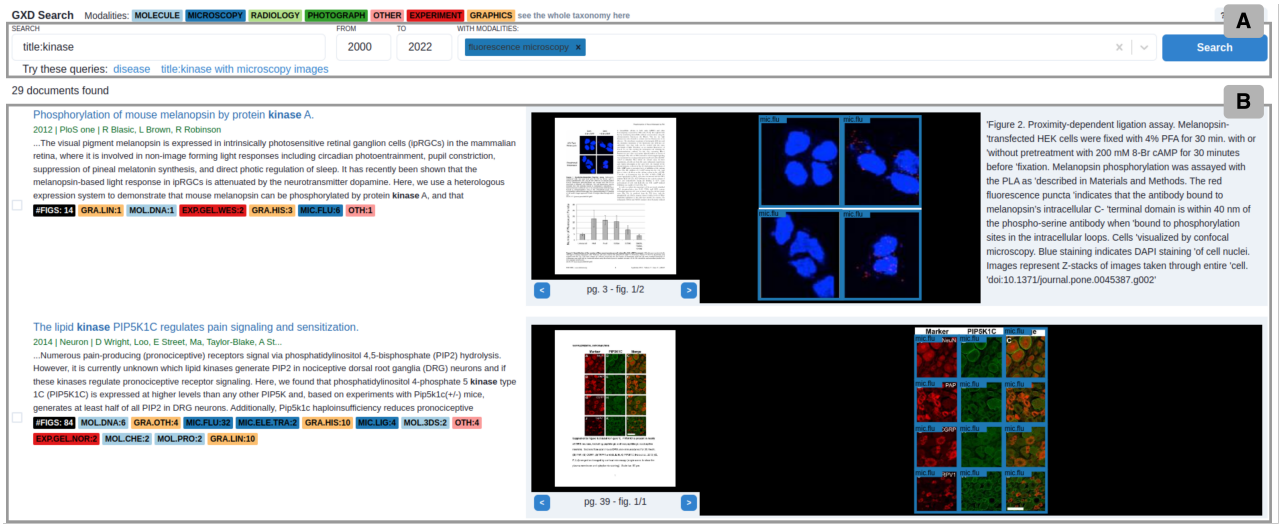

Fig. 2. MouseScholar interface. A) Query options for the search interface, including a field for keyword queries, date range filters, and image modality filters. B) Search results showing text surrogates (left side) which include the found image modalities within each result, and image surrogates (right side) composed of the thumbnails and individual figures with captions.

bining these entries allows a biocurator to target documents containing a keyword, and to include specific sub-figure modalities. In addition, users can form boolean queries with the title, abstract, captions, and even full text (not available for the GXD_{2000} subset) when simple keywords are not enough. We designed this interface through several discussion sessions and feedback sessions with collaborators and domain experts over two years, where we took Google Scholar as a source of inspiration given the simplicity of its design and the familiarity of researchers with it.

To support the hybrid presentation of text and image data in the query results (i.e., surrogates), our design juxtaposes page thumbnails, figures, and captions (Fig. 2B) to more traditional surrogates for text information used in systems like Google Scholar. However, our text surrogates also include information on the modalities found in the document and count. Overall, the design aims for compact representations while displaying sufficient image data to make an informed decision about the document's relevance to particular research interests.

This compact representation contrasts designs in other systems that show every sub-figure found in the document within the surrogate. Our design displays one figure at a time while overlaying the modalities' information as color-coded boxes, because adjacent panels communicate different story aspects. For instance, a biocurator looking for microscopy images may also be interested in line charts representing experimental results. To rank figures, the system performs a count based on the number of sub-figures matching elements from the modality filter and sorts those figures in descending order. When no modality filter is present, we display figures based on the figure number.

III. Evaluation and Results

We evaluate the system's performance and usefulness through a quantitative and qualitative approach. First, we perform a quantitative evaluation of the image classifiers, to estimate the performance of the system's modality predictions during retrieval, when the retrieved images would not be pre-labeled. Second, we perform a quantitative and qualitative evaluation of the system capabilities with 10 biocurators. Last, we examine the benefits of using MouseScholar in conjunction with a document triage classifier trained on GXD_{2000}.

A. Image classifier results

Due to the influence of the prediction of the image modalities in the retrieval process, we report first the performance of the image modality classifiers on our test labeled data. Table I shows the accuracy, F1-weighted and F1-macro scores for the parent nodes in the taxonomy. We chose to report F1 macro score due to the unbalanced distribution of samples in the training set and contrast it with F1-weighted, which weighs the scores based on the class size. The higher-modality classifier represents the top level classification while the rest of the classifiers output more detailed modality classes. We also show the number of images in the test set, which are proportional to the number of training images, to illustrate the impact of these numbers on the classifier performance. The rightmost column shows the architecture used by each classifier. Accuracy and F1 scores are calculated as:

$Acc = (TP + TN)/(TP + TN + FP + FN)$; $F1 = 2 * TP/(2TP + FP + FN)$; $F1_{macro} = \sum_{i=1}^{n}(F1_i)/(n)$; and $F1_{weighted} = \sum_{i=1}^{n}(F1_i * w_i)/(n)$ where TP, FP, TN, and FN are true/false positive/negative, w_i represents the weight of class i, and n is the number of classes in the classifier.

The reported scores suggest positive expectations for the predictions in the GXD_{2000} images. The high scores of most classifiers indicate that MouseScholar would not suffer from notable mispredictions when filtering images by modalities. However, the results also indicate potential issues with two of the classifiers, where the scores are less strong. The gel

TABLE I
ACCURACY, F1-WEIGHTED AND F1-MACRO SCORES FOR THE PARENT NODE FIGURE CLASSIFIERS AND NUMBER OF IMAGES IN THE TEST SET

Classifier	#Test	Acc.	$F1_m$	$F1_w$	Arch.
higher-modality	11,060	98.61	87.99	98.45	EfficientNet-B1
experimental	1,637	99.27	94.97	99.25	EfficientNet-B0
gel	25	92.00	76.96	92.62	EfficientNet-B1
graphics	5,413	99.33	92.54	99.33	EfficientNet-B1
microscopy	395	93.92	93.14	93.88	EfficientNet-B0
electron	88	84.09	83.68	84.09	EfficientNet-B1
molecular	109	96.33	96.33	96.33	EfficientNet-B1
photography	49	93.88	93.15	93.74	ResNet-34
radiology	3,136	99.71	91.21	99.69	EfficientNet-B0

classifier might produce misclassification between Northern blots, Western blots, and RT-PCR experimental gel images. In addition, misclassifications may also occur between the microscopy electron classes: transmission, scanning, and others. These two classes feature lower sample counts in both the original training set and in our test set. The classifier performance may improve with higher numbers of samples.

B. Expert evaluation

We evaluated MouseScholar in the context of biocuration, where we aimed to understand the system's usability in terms of helping curators find papers of interest. Specifically, we assessed the system's ability to support more efficient extraction of data from the literature, and the curation of experimental results into scientific databases.

In addition, we aimed to understand how our system helps curators to identify relevant search results. In theoretical frameworks for information retrieval [22], users evaluate the relevance of a surrogate based on the information shown to the user and the level of analysis: looking at surrogates, skimming through the document, and close reading. Most search engines support this task by allowing researchers to identify an interesting search result, open the details, and later download the document. By displaying figure information, we expected researchers to avoid skimming through the document and instead successfully assess the document's relevance from the presented surrogate.

Therefore, we asked 10 biocurator researchers to evaluate MouseScholar, including three biocurators from GXD, four biocurators from the BioGRID Database, one biocurator from Xenbase, one from WormBase, and one from the Rat Genome Database. No personal data from the researchers was collected. To leverage the GXD_{2000} dataset, we framed the baseline evaluation task as follows: "Imagine you are looking for scientific documents about mice gene expression worth curating. Use the system to perform queries that would lead to those papers, and for each query, inspect the first page of results and assess the relevance of the results towards your goal". Evaluators were also asked to attempt queries related to their interests, as described further below.

The evaluation protocol consisted of four steps, which followed a similar evaluation structure as proposed at the BioCreative VII demonstration [23]. Each participant received an activity document detailing the instructions for each step. First, we required the participants to get familiar with the interface by reading an interface tutorial document, which introduced the taxonomy of modalities and the system features. The second step introduced a guided activity with instructions to follow. Biocurators executed text and text+image queries in this step, including boolean operations. Next, biocurators followed an exploratory activity where they were asked to perform a number of open queries to find documents worth curating for mouse gene expression. We requested that biocurators not search online for the full text of the document, and only rely on the information provided in the surrogates. In addition, we asked biocurators to write down the queries performed and annotate any element on the interface that they found helpful, or that presented a hindrance.

In the last step, we presented a questionnaire divided into four sections. The first section presented open-ended questions to collect feedback and identify improvement points. The second section presented eleven questions based on the system usability score guidelines (SUS) [24], where the first ten questions used a Likert scale from one to five, and the last question a Likert scale from one to ten. The third section presented two questions to inquire about the system's capabilities for identifying documents for curation without requiring biocurators to skim through the document. Lastly, additional questions on a Likert scale (1 to 5) inquired about the usefulness of the surrogates' components.

C. Questionnaire results

MouseScholar obtained a usability score of 80.75, which indicates high satisfaction of the biocurators with our system (SUS scores between 80-90 are equivalent to "excellent" [25]). This result is consistent with the biocurators' comments about the intuitiveness of the interface. All biocurators indicated that the results and data format were useful. While the respondents were not aware of a similar system, they suggested PubMed and Textpresso as the most similar systems to ours. In particular, biocurators agreed that they would use the system frequently (M=4.4$\pm$.84), that the system was easy to use (M=4.2$\pm$.79), that the functions were well integrated (M=4.2$\pm$.63), and that most people would learn to use it very quickly (M=4.5$\pm$.71). Biocurators were more neutral in their confidence in using the system (M=3.9$\pm$.74). Biocurators disagreed on the following items: the system was unnecessarily complex (M=1.7$\pm$.67), they would need support from a developer to use the system (M=2$\pm$1.25), the system had too much inconsistency (M=1.8$\pm$.71), the system was cumbersome to use (M=1.4$\pm$.73), and they need to learn many things before starting to use it (M=2$\pm$1.33).

Most biocurators agreed that inspecting only the surrogates provided enough information (as opposed to opening the PDF document) to assess whether a result is worth curating (M=3.9$\pm$1.1). Two biocurators were neutral, and one disagreed. Biocurators also estimated that 60 to 100% of the results retrieved were GXD-relevant.

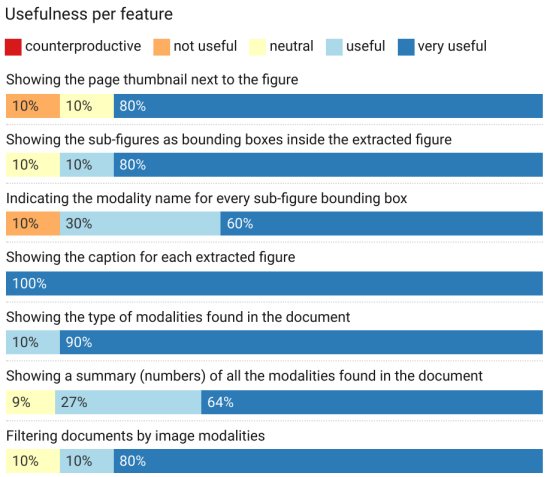

Fig. 3. MouseScholar's usefulness per feature.

In addition, biocurators agreed that they would recommend our system to their colleagues (8.3/10), that the system met their expectations (M=4.1±.57), and that they got a very positive impression of the system (M=4.4±.69). Results further show that biocurators agreed on the importance of the elements within the surrogates (Fig. 3). Showing captions next to the figures was the most useful feature (M=5), followed by the summary of the type of modalities in the document (M=4.9±.32), showing bounding boxes to indicate the subfigures (M=4.7±.68), the modality count (M=4.6±.69), displaying page thumbnails (M=4.5±1.08), and the modality name (M=4.4±.96). Finally, the capability of the system to filter by modalities was perceived as very useful (M=4.7±.67).

The biocurators also identified directions of future research. One biocurator suggested that the images in the surrogate need to be ranked based on query matches on modalities and captions, as opposed to only ranking by matches on modalities. Four biocurators suggested further support for query formulation, including feedback to validate syntax, handling stemming and synonyms, and support for wildcards. Another biocurator suggested annotating the documents by species found in the text or the document type (e.g., reviews vs. experimental studies). One suggestion included refining the taxonomy for graphics to match experiments with CRISP-based phenotypes and gene expression data, for instance, by identifying heatmaps and volcano plots, which are a type of scatterplots. Other biocurators wanted support for showing only images with an input modality and hiding bounding boxes to ease reading. Finally, one biocurator suggested that our system allow users to input PMIDs; the system should then process the documents and allow users to export them when ready.

When inquired about ways they see themselves using the system, biocurators suggested triaging documents for particular collections. "The system will not only help scientists find papers of interest but will also allow for more efficient extraction of data from the literature and curation of experimental results into scientific databases which facilitates research", one biocurator commented. Another biocurator commented: "MouseScholar could help find evidence for genetic markers of cell types in legacy uncurated documents". Biocurators further noted they could use the MouseScholar pre-processing pipeline to expand the document annotation capabilities of existing systems and to organize their documents into relevant taxonomies of modalities.

D. Integration with triage using a dataset-specific classifier

To further investigate the potential benefits of MouseScholar, we performed an additional quantitative evaluation using the GXD_{2000} dataset, this time in conjunction with a triage classifier. The GXD_{2000} dataset has been manually curated over several years by biocurators at the Jackson Laboratory, resulting in 1000 GXD-relevant documents and 1000 GXD-irrelevant documents. A state of the art document triage classifier trained on this dataset that leverages image modalities had previously yielded good classification [4].

We analyze the queries executed by biocurators during the exploratory activity section of the evaluation. Figure 4 (left) shows query samples, ranked by the recall@10 metric, from biocurators associated with the GXD project, including queries using keywords, boolean operators, and filters on modalities. For each query, we show the results on the GXD_{2000} dataset, capped at 50 results per query. Black boxes represent GXD-relevant results, and gray boxes represent GXD-irrelevant results, as determined during manual biocuration. A vertical red line marks results displayed on the first page of results (the activity required analysis of the first page of results). On the right side, we show the results of pre-filtering the dataset with the state of the art classifier, followed by the same MouseScholar queries. Both approaches are effective in filtering results. The recall@50 per query increases slightly with the use of the classifier, and we observe two changes in the recall@10: one increment (green arrow) and one decrement (purple arrow). Decrements in recall@10 are due to false positive results in the document classifier predictions.

For completeness, Figure 4 (right) shows query samples from biocurators not affiliated with GXD (black arrows indicate change in ranking but no change in recall@10). As expected, the query results are not necessarily aligned with the GXD manual curation results, since the biocurator interests in this case were not GXD-centered. The recall statistics are, accordingly, less meaningful here. Queries that yielded zero results are not included in the figure. As before, the figure illustrates the ability of MouseScholar to effectively filter the document collection, and to retrieve and present a relatively small number of potentially relevant results. These results confirm the biocurators' estimate that a large proportion of the results shown were relevant to their queries.

IV. Discussion and Conclusion

Interacting with search interfaces to triage documents is a core biocuration activity. Many systems operate only on full text and document metadata. In addition, search results

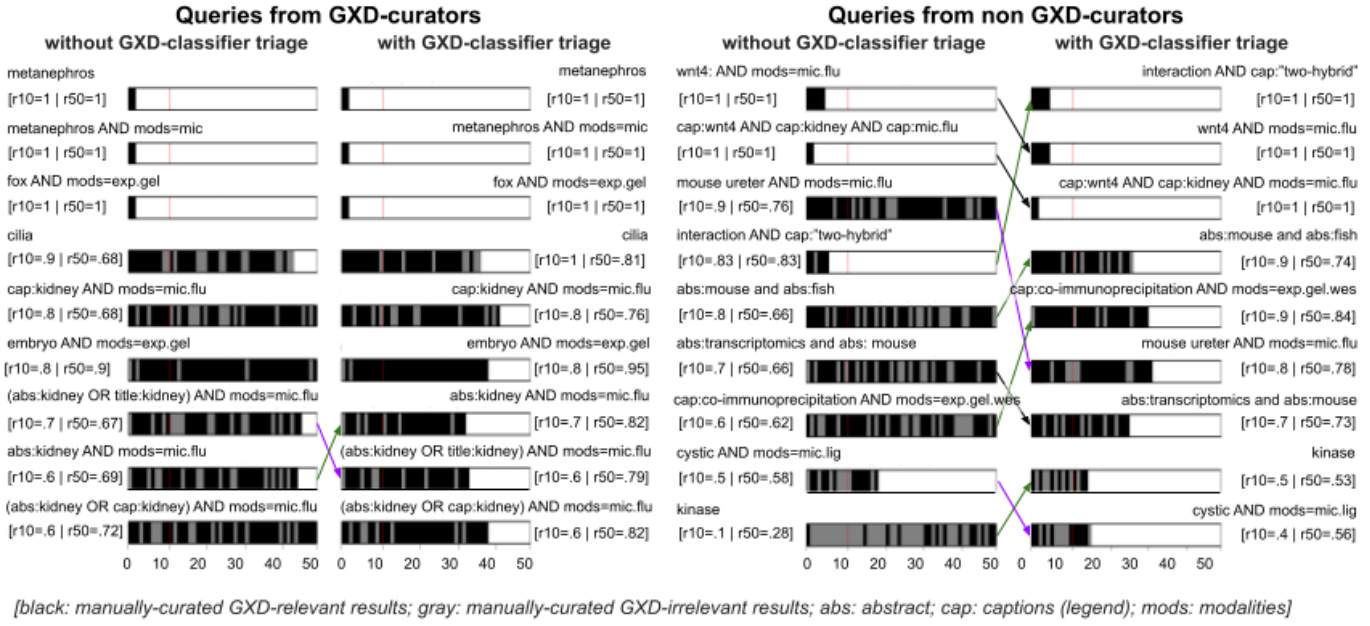

Fig. 4. MouseScholar retrieval results for queries executed by GXD curators (left) and non GXD-curators (right), versus GXD-classifier triage followed by MouseScholar retrieval for the same queries. Results are capped at 50. A red line marks the top 10 results shown on the first page. Black indicates dataset entries manually labeled as GXD-relevant; gray indicates dataset entries manually labeled as GXD-irrelevant; white indicates no other results. Green arrows indicate improvement in the recall@10; purple arrows indicate a decline in the recall@10; black arrows represent the lack of changes in recall@10 but change in order.

typically appear in vertical lists of cards that display metadata and highlight sentences matching the input queries. Existing biocuration systems do not display nor provide alternatives to search figures and subfigures within documents, and do not identify modalities, as we do.

Our evaluation shows that MouseScholar effectively leverages the figure content to improve the experience when searching for relevant papers in biocuration. Our classifier experiments demonstrate that categorizing the figure data into meaningful imaging modality categories yields good results. The quantitative evaluation of the classifiers yielded excellent accuracy and F1 scores, indicating the strengths of the taxonomy approach. The quantitative and qualitative feedback from domain experts highlights the benefits of including figures in the search. The high usability scores related to potential adoption and frequency of use, and to recommending the system to peers are strong indicators of the potential of this hybrid text+images approach. The questionnaire results suggest that integrating image modalities when querying documents and presenting enhanced surrogates have the potential to accelerate biocuration workflows, which can be time consuming, and to expand the scope of documents worth curating.

MouseScholar successfully meets several of the challenges impeding biocuration from leveraging image-based data. On the back end, we defined and leveraged a taxonomy of figure modalities. We integrated a preprocessing pipeline, which is able to extract image and subfigure data from PDFs, and to classify subfigures based on this taxonomy, leveraging related efforts for interactively labeling datasets [13], [26]. On the front-end, search options allow biocurators to examine relevant information, while providing context to support relevance assessment. Evaluator feedback indicates that MouseScholar retrieval is effective, and can help identifying content relevant to biocuration. The system furthermore requires low effort to expand familiar query formulations with modalities.

Although most biocurators thought the surrogate information shown was sufficient, outlier feedback indicates the system may benefit from providing access to the PDF document or means to further explore such content. Showing image evidence eased relevance-determination: the biocurators' feedback strongly shows that MouseScholar is effective when biocurating documents. There was also significant interest in using MouseScholar in conjunction with text-based classifiers, which provide a prediction of the document relevance to a curation database. As shown in our integration of MouseScholar with a document classifier, MouseScholar has the potential of complementing triage or to be used as a standalone application, such as PubTator.

Compared to related experiments with image+text retrieval in general biomedical research [11], in our evaluation biocurators tended to need advanced support for inputting complex queries and supporting synonyms. In addition, some biocurators preferred seeing more explicit filtering boxes (similar to GXD's Expression Literature Search tool) instead of our filters, which leverage a quick search layout. Further instantiation with specific taxonomies and filtering options may be desirable for other biocuration groups.

In terms of design lessons learned from this experience,

we conclude that images should be shown within the search results wherever possible. Furthermore, our design leverages the clean and sparse design of typical search interfaces, which was well received by biocurators. Last, our results confirm that images should play a part in automated triaging, which was first suggested by Shatkay et al. [27] and by Li et al. [4].

In terms of generalizability, our work shows how to leverage a hybrid text+image search infrastructure for the biocuration domain. Our taxonomy aims to be general enough to cover most of the modalities in the biocuration space. Feedback from biocurators suggests that further refinement and specification may help accommodate particular needs. The cost to adapt the taxonomy depends on the data availability. For instance, projects like Pathway Figure OCR [28] provide a vast collection of pathway figures that can enhance our graphics category. In contrast, identifying volcano plots depends on exploratory tools to locate those samples within the most similar category (scatterplot).

Regarding limitations, our study considered usability aspects of the system and did not collect biocurator-dependent metrics such as task scores or task completion time. Further evaluation may include ablation experiments, which would however require a golden set of documents to evaluate queries with image data and without image data. This is laborious work that would require onerous effort from biocurators. Future quantitative evaluations could also include keylogging and the use of eye tracking to further understand where biocurators look when evaluating the relevance of a surrogate. In terms of scalability, our system supports handling expanding collections and multiple users. In terms of long-term support, our laboratory and university have the resources to maintain and expand the system [29].

In conclusion, we described the design, implementation, and evaluation of MouseScholar, a first-in-class instantiation for biocuration of a hybrid text+image search system. We presented the challenges to integrate image data into the biocuration workflow, we extracted and documented the requirements behind leveraging figure information as part of the biocuration workflow, we instantiated and formally evaluated a biocuration search system, and we described the process to extending this approach to other biocuration domains. The result is a biocuration system with unique capabilities, able to support document searching over figure modalities, hierarchical taxonomies and compound figures, and to present this evidence in conjunction with captions, relevant text, subfigure and context information. Our evaluation with biocurators at four sites demonstrates the benefits of using enhanced surrogates when presenting query results, and documents the preferences of biocurators for these representations.

Acknowledgment

We thank the biocurators that participated in this study. This work was supported by the awards from the U.S. National Institutes of Health (NLM R01LM012527, NCI R01CA258827) and the U.S. National Science Foundation (CNS-1828265, CDSE-1854815).

References

[1] S. Burge, T. K. Attwood, A. Bateman, T. Z. Berardini, M. Cherry, C. O'Donovan *et al.*, "Biocurators and Biocuration: surveying the 21st century challenges," *Database*, 2012.

[2] D. Howe, M. Costanzo, P. Fey, T. Gojobori, L. Hannick, W. Hide *et al.*, "The future of biocuration," *Nature*, vol. 455, no. 7209, 2008.

[3] A. Tang, K. Pichler, A. Füllgrabe, J. Lomax, J. Malone *et al.*, "Ten quick tips for biocuration," *PLoS Comp. Bio.*, vol. 15, no. 5, 2019.

[4] P. Li, X. Jiang, G. Zhang, J. Trelles, D. Raciti, C. Smith *et al.*, "Utilizing image and caption information for biomedical document classification," *Bioinf.*, vol. 37, 2021.

[5] X. Jiang, P. Li, J. Kadin, J. A. Blake, M. Ringwald, and H. Shatkay, "Integrating image caption information into biomedical document classification in support of biocuration," *Database*, 2020.

[6] X. Jiang, M. Ringwald, J. A. Blake, C. Arighi, G. Zhang, and H. Shatkay, "An effective biomedical document classification scheme in support of biocuration: addressing class imbalance," *Database*, 2019.

[7] K. van Auken, P. Fey, T. Z. Berardini *et al.*, "Text mining in the biocuration workflow: applications for literature curation at WormBase, dictyBase and TAIR." *Database*, vol. 2012, 2012.

[8] C.-H. Wei, A. Allot, R. Leaman, and Z. Lu, "Pubtator central: automated concept annotation for biomedical full text articles," *Nucleic acids res.*, vol. 47, no. W1, 2019.

[9] J. White, "PubMed 2.0," *Med. Ref. Serv. Quarterly*, vol. 39, no. 4, pp. 382–387, 2020.

[10] P. Lee, J. West *et al.*, "Viziometrics: Analyzing visual patterns in the scientific literature," *Trans. Big Data*, 2017.

[11] J. Trelles, C. Arighi, H. Shatkay, and G. E. Marai, "Enhancing biomedical search interfaces with images," *Bioinf. Adv.*, 2023.

[12] A. Garcia Seco de Herrera, R. Schaer, S. Bromuri, and H. Muller, "Overview of the ImageCLEF 2016 medical task," in *CLEF*, 2016.

[13] J. Trelles, P. Li, C. Arighi, D. Raciti, H. Shatkay, and G. E. Marai, "ANIMO: Annotation of Biomed Image Modalities," in *BIBM*, 2021.

[14] M. Ringwald, J. E. Richardson, R. M. Baldarelli, J. A. Blake *et al.*, "Mouse Genome Informatics (MGI): latest news from MGD and GXD," *Mammalian Genome*, vol. 33, no. 1, pp. 4–18, 2022.

[15] J. H. Finger, C. M. Smith, T. F. Hayamizu, I. J. McCright, J. Xu, J. T. Eppig *et al.*, "The mouse gene expression database: new features and how to use them effectively," *Genesis*, vol. 53, no. 8, pp. 510–522, 2015.

[16] C. M. Smith, T. F. Hayamizu, J. H. Finger, S. M. Bello, I. J. McCright, J. Xu *et al.*, "The mouse Gene Expression Database (GXD): 2019 update," *Nucleic Acids Res.*, vol. 47, no. D1, 2018.

[17] X. Jiang, M. Ringwald, J. Blake *et al.*, "Effective biomedical document classification for identifying publications relevant to the mouse gene expression database (GXD)," *Database*, vol. 2017, 2017.

[18] P. Li, X. Jiang, and H. Shatkay, "Figure and caption extraction from biomedical documents," *Bioinf.*, 2019.

[19] P. Li, X. Jiang, C. Kambhamettu *et al.*, "Compound image segmentation of published biomedical figures," *Bioinf.*, 2018.

[20] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proc. ICML*, 2019, pp. 6105–6114.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, 2016, pp. 770–778.

[22] F. Loizides and G. Buchanan, "Towards a framework for human (manual) information retrieval," in *Proc. IRFC*, 2013, pp. 87–98.

[23] A. Chatr-Aryamontri, L. Hirschman, K. Ross, R. Oughtred *et al.*, "Overview of the COVID-19 text mining tool interactive demonstration track in BioCreative VII," *Database*, 2022.

[24] J. Brooke, "SUS-A quick and dirty usability scale," *Usab. Eval. Indus.*, vol. 189, no. 194, 1996.

[25] A. Bangor, P. Kortum, and J. Miller, "Determining what individual sus scores mean: Adding an adjective rating scale," *J. Usability studies*, vol. 4, no. 3, pp. 114–123, 2009.

[26] J. Trelles, A. Wentzel, W. Berrios, and G. E. Marai, "BI-LAVA: Biocuration with Hierarchical Image Labeling through Active Learning and Visual Analysis," *arXiv preprint*, 2023.

[27] H. Shatkay *et al.*, "Integrating image data into biomedical text categorization," *Bioinf.*, vol. 22, no. 14, pp. e446–e453, 2006.

[28] WikiPathways, "Pathway figure ocr," https://pfocr.wikipathways.org/.

[29] G. E. Marai, J. Leigh, and A. Johnson, "Immersive analytics lessons from the electronic visualization laboratory: A 25-year perspective," *IEEE Comp. Graph. Apps.*, vol. 39, no. 3, pp. 54–66, 2019.