NOA: A Search Engine for Reusable Scientific Images beyond the Life Sciences

Jean Charbonnier[1], Lucia Sohmen[2], John Rothman[1], Birte Rohden[1], and Christian Wartena[1]

[1] Hochschule Hannover, Expo Plaza 12, 30539 Hannover
Jean.Charbonnier@hs-hannover.de
[2] Technische Informationsbibliothek, Welfengarten 1B, 30167 Hannover

Abstract. NOA is a search engine for scientific images from open access publications based on full text indexing of all text referring to the images and filtering for disciplines and image type. Images will be annotated with Wikipedia categories for better discoverability and for uploading to WikiCommons. Currently we have indexed approximately 2,7 Million images from over 710 000 scientific papers from all fields of science.

Keywords: Open Access, Image Retrieval

1 Reusing Scientific Images

Images play an important role in scientific publications. In some cases images are specific for the paper, but in many cases images are general illustrations that could be reused in several papers or for illustrating a presentation. In order to effectively find reusable images one would need a search engine that allows for filtering scientific disciplines and image types, and that searches in scientific images only, or allows to filter images from scientific journals. NOA (*Nachnutzung von Open Access Bildern*, Reuse of Open Access Images) is such a search engine. A first version of this search engine is available at `http://noa.wp.hs-hannover.de/`.

2 Sustainability: Uploading Images to Wikimedia

The goal of the NOA project is to build a freely accessible corpus of images from open access articles and upload them to Wikimedia Commons which is a collection of freely reusable images that has existed for many years. Thus, access to the images will be secured even after the project is over and an already established user base will be able to make use of the images. In addition we will make them available through a dedicated search engine.

3 Data

Since Wikimedia Commons only accepts images with a Creative Commons-Attribution or a Creative Commons-Attribution-Share Alike license, our article

aggregation has focused on publishers using these licenses, including Hindawi, Frontiers, Copernicus, Springer Open and parts of PubMed Central.

Currently we have collected 2,7 Million images from over 710 000 papers in over 5 000 journals. 2 Million more images will be added in the upcoming months.

4 System Description

We retrieved all articles from the mentioned publishers through their public APIs or as a complete dump. We store the articles and their metadata in a MongoDB instance. This information is enriched and fed to an Apache Solr instance which delivers all data to a web frontend. Currently the system is hosted on a virtual machine with 32 GB RAM and 5 TB SSD Space. 32 GB are used to store the data in MongoDB, 237 GB are used to store all images on the file system.

4.1 Text Based Image Retrieval

Images in scientific journals often show specific and abstract objects, graphs, and drawings. Image recognition will not be effective in indexing these images. Instead, we use the text from the article and the metadata for retrieval. We add the image caption, the title of the paper, the journal title and the author names to the index. Words from the caption get the highest weight using Solr's eDisMax relevance score. Thus, we will get results based on matches from different fields, but usually ranked below images that have the query terms in the caption.

Hong Yu [10] argues that information about an image in a scientific paper is found all over the paper. However, they try to generate a complete explanation of the picture. For indexing we need words that directly refer to concepts shown in the image. Moreover it is important that users understand why an image was found. This is easy in case the query terms are found in the meta data that are displayed in the result list, but not if these terms are found somewhere in the paper. Nevertheless, we will use the information from text regions with references to the image indirectly, as explained below.

Often image captions use specific abbreviations. In our collection we find on average almost one acronym per caption. We try to expand acronyms and add the definitions to the search index as well. Thus a query for *Fast Fourier Transformation* will also return images annotated solely with *FFT*.

Definitions for abbreviations are searched in the corresponding paper. If no definition was found, we take the definition found in another paper from the same journal, but only when the abbreviation is unambiguous within all papers from that journal. We found 2 838 713 occurrences of words written in all capitals that thus are likely to be abbreviations or acronyms. For 25 336 abbreviations with a total of 643 231 occurrences we found a definition in the paper. For 379 509 more we could find an unambiguous definition in another paper from the same journal. Thus a total of about 36% of all potential abbreviations could be expanded.

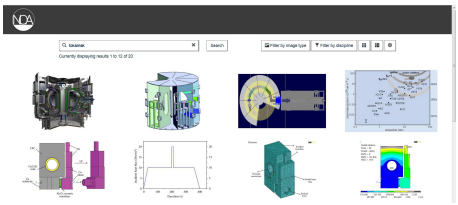

Fig. 1. Grid view

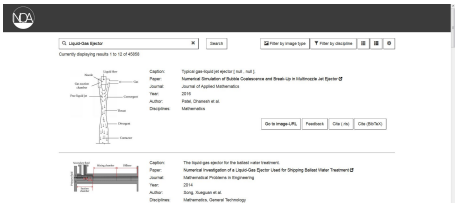

Fig. 2. List view

4.2 User Interface

The user interface has a typical design with a search field on top and search results below. Results are either displayed in a grid or as a list. The user can easily switch between the views and store their preference. In the first view (Fig. 1) only the images are displayed and the metadata is shown when an image is clicked. In the second view (Fig. 2) the metadata is shown next to the images.

In the meta data pane the title of the paper contains a hyperlink to the official publication. Similarly, the original image can be downloaded. Finally, we offer the possibility to download the metadata of the paper in RIS and BibTeX format to enable easy import to common reference management programs.

The search results can be filtered by discipline and by image type.

5 Related Work

Other search engines for scientific images have been developed in the past, although they usually only index biomedical images extracted from articles in the PMC corpus. An early example of this is Figsearch [6], a prototypical search engine from 2004 that lets users search figure legends. BioText [2], which was developed in 2007, is a very basic search interface for figures and articles. The developers of the Yale image finder [9] from 2008 did research on text extraction from images [8][9] and made this text searchable. FigureSearch [1] from 2009 has a focus on image classification [3], [7], [4] and automatically generated text summaries for the images [10] [11] [1]. The newest search engine is Viziometrics [5] which implements automatic classification and crowd-source tagging of images from the PMC corpus. Only FigureSearch (`http://figuresearch.askhermes.org/`) and Viziometrics (`http://viziometrics.org/`) still have working instances.

6 Future Work

For most components basic algorithms have been used. Much of future work will deal with improving image classification, keyword extraction, abbreviation expansion, etc. We also plan to use OCR to index text in the images. In order to integrate the figures into Wikimedia Commons and enhance retrieval, we are currently working on annotating images with categories from the English Wikipedia using image captions and relevant sections from the papers.

Currently the database is static. In the upcoming months we will implement continuous updating with recently published papers. Another major topic will be evaluation of the service including what potential functions are useful for users.

Acknowledgements

We would like to thank Frieda Josi, Lambert Heller, Ina Blümel for many helpful comments. This research was funded by the DFG under grant no. WA 1506/4-1.

References

1. Agarwal, S., Yu, H.: FigSum: automatically generating structured text summaries for figures in biomedical literature. AMIA ... Annual Symposium proceedings. AMIA Symposium 2009, 6–10 (Nov 2009)
2. Hearst, M.A., Divoli, A., Guturu, H., Ksikes, A., Nakov, P., Wooldridge, M.A., Ye, J.: BioText Search Engine: beyond abstract search. Bioinformatics 23(16), 2196–2197 (Aug 2007), https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btm301
3. Kim, D., Ramesh, B.P., Yu, H.: Automatic figure classification in bioscience literature. Journal of Biomedical Informatics 44(5), 848–858 (Oct 2011), http://www.sciencedirect.com/science/article/pii/S1532046411000943
4. Kim, D., Yu, H.: Hierarchical Image Classification in the Bioscience Literature. AMIA Annual Symposium Proceedings 2009, 327–331 (2009), http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2815366/
5. Lee, P.s., West, J.D., Howe, B.: Viziometrics: Analyzing visual information in the scientific literature. IEEE Transactions on Big Data (2017)
6. Liu, F., Jenssen, T.K., Nygaard, V., Sack, J., Hovig, E.: FigSearch: a figure legend indexing and classification system. Bioinformatics 20(16), 2880–2882 (Nov 2004), https://academic.oup.com/bioinformatics/article/20/16/2880/236814/FigSearch-a-figure-legend-indexing-and
7. Rafkind, B., Lee, M., Chang, S.F., Yu, H.: Exploring Text and Image Features to Classify Images in Bioscience Literature. In: Proceedings of the Workshop on Linking Natural Language Processing and Biology: Towards Deeper Biological Literature Analysis. pp. 73–80. BioNLP '06, Association for Computational Linguistics, Stroudsburg, PA, USA (2006), http://dl.acm.org/citation.cfm?id=1567619.1567632
8. Xu, S., Krauthammer, M.: A New Pivoting and Iterative Text Detection Algorithm for Biomedical Images. Journal of Biomedical Informatics 43(6), 924–931 (Dec 2010), http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3265968/
9. Xu, S., McCusker, J., Krauthammer, M.: Yale Image Finder (YIF): a new search engine for retrieving biomedical images. Bioinformatics 24(17), 1968–1970 (Sep 2008), http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2732221/
10. Yu, H.: Towards answering biological questions with experimental evidence: automatically identifying text that summarize image content in full-text articles. AMIA ... Annual Symposium proceedings. AMIA Symposium pp. 834–838 (2006)
11. Yu, H., Lee, M.: BioEx - A Novel User-Interface that Accesses Images from Abstract Sentences (2006)