Figures in Scientific Open Access Publications

Lucia Sohmen[2 [0000−0002−2593−8754]], Jean Charbonnier[1 [0000−0001−6489−7687]], Ina Blümel[1,2 [0000−0002−3075−7640]], Christian Wartena[1 [0000−0001−5483−1529]], and Lambert Heller[2 [0000−0003−0232−7085]]

[1] Hochschule Hannover, Expo Plaza 12, 30539 Hannover
[2] Technische Informationsbibliothek, Welfengarten 1B, 30167 Hannover

Abstract. This paper summarizes the results of a comprehensive statistical analysis on a corpus of open access articles and contained figures. It gives an insight into quantitative relationships between illustrations or types of illustrations, caption lengths, subjects, publishers, author affiliations, article citations and others.

Keywords: Open Access, Scientific Figures, Statistical Analysis

1 Motivation and target

Researchers often reuse figures from other publications for their own work, for example presentations or articles. In order to find those images, it is useful to have a search engine that finds figures from scientific articles.

The goal of the NOA (*Nachnutzung von Open Access Bildern*, Reuse of Open Access Images) project is to build a freely accessible corpus of figures from open access articles, providing links to the original article as well[3]. A first version of a search engine allowing for filtering and searching is available at `http://noa.wp.hs-hannover.de/`. In order to secure access to the images after project completion, they will be uploaded to Wikimedia Commons (commons.wikimedia.org). As a side effect of the mentioned extraction of figures from papers, we use the built-up corpus of images linked to corresponding articles for various analyses and relations to other quantitative data/article such as citations. This paper summarizes the results of a comprehensive statistical analysis on our corpus and gives an insight into quantitative relationships between illustrations or types of illustrations, subjects, publishers, journals, article citations and others.

2 Related Work

Over the years, there have already been attempts at creating search engines for scientific images. So far, all of these have used some subset of articles from the life sciences. FigSearch[7], developed in 2004, claims to be the first of these applications. The Yale Image Finder[9] was developed in 2008 Another search engine is Figuresearch[1] from 2009.Viziometrics[6] from 2016 is the newest application that allows users to directly search for images. Their dataset contains 650 000 articles

Table 1. Publishers (including aggregators), number of papers, figures, percentage of papers with figures and years included in the dataset.

Publisher	# Articles	# Figures	% With Figures	years included
Copernicus	9 592	85 720	71,7	2014-2017
Springer	78 418	310 214	98,0	2003-2018
Hindawi	147 848	1 172 657	80,3	2008-2017
Frontiers	57621	217 897	83,3	2009-2017
PMC	747 839	2 796 271	81,3	1848-2017
all	1 041 318	4 582 759	80,7	

and 4,8 million images from the PubMedCentral (PMC) corpus. Their search engine is the only one that is still available to search in at `viziometrics.org`.

Several statistical analyses of article corpora containing images have been done. [6] analyzes the Viziometrics corpus. [4] extracted 6.4 million figures from 1 million papers in computer science and biomedicine. They found that, over time, figure counts and their captions lengths have increased. There was a small positive correlation between the figure count and the number of citations to a paper. [5]looked at 1133 psychology papers to find out what factors influence the number of citations to a paper. The authors found that the number of graphs had a negative correlation while the number of tables and models had a positive correlation with the citations. [2] analyzed 5180 articles from six journals in different domains to analyze the figure use of multiple authors versus single authors and found that multiple authors use more figures per article.

3 Corpus and analysis method

Our corpus includes figures from open access articles from different sources. Criteria for inclusion were accessibility (difficulty of downloading a large set of articles), format (easy to parse, like XML) and license (suitable for reuse and upload to Wikimedia Commons). A big part of the corpus is a subset from PubMedCentral (PMC) which stores millions of articles from the life sciences. Other articles were downloaded from the publishers as a dump or via API.

All the articles that we downloaded have the XML format with most of them using the JATS-XML specification that is required by PMC. After download, the articles were parsed with a Java program that was developed within our project. It extracts all the relevant data from the documents (for example article metadata, figure URLs and captions) and writes it to the project database. Furthermore, this data has been enhanced with additional information, including journal discipline, corresponding Wikipedia categories and citation data from Crossref. This makes up the dataset on which we base our statistics.

We found 3 million figures in 1 million articles, including articles with zero figures. We counted everything that was embedded in a "figure" tag in the XML form of an article. These do not usually include tables and equations. See Table 1 for an overview of the different publishers and their image count in our dataset.

4 Results

4.1 Licences and figures with source reference

The license type of the figures is of interest for re-usability. CC-BY clearly dominates the corpus: CC-BY-4.0 came to a number of 351694, -3.0 to 75729, -2.5 to 30036 and -2.0 to 216472. CC0 was only assigned 1986 times. Although we did not filter out CC-BY-SA type licenses, none of the articles in the corpus are under that license type. 7878 times no license was found.

To identify figures that were reused from an external source and are therefore not under the same license as the article, we spotted keywords in the captions to find out whether an external source is cited. This algorithm identified about 5% of all images. Manual inspection revealed that roughly 8/9 of those results were false positives, so the actual rate of reused images is about 0,55 percent. Recall was valued over precision to avoid violation of copyright.

4.2 Figure types

Table 2 shows the average number of charts (including charts and graphics) and images (including photos, microscopy and other imaging methods) per paper for disciplines with 2000 or more papers. The often much higher proportion of charts is noticeable in almost all disciplines, especially in the subjects belonging to the field of Engineering and Technology[3]. In total, Engineering and Technology subjects contain the highest number of figures, followed by Natural Sciences and Medical and Health Sciences. All disciplines with less than 2000 papers can be derived from the underlying raw data[8].

4.3 Figure caption length

Since the captions are usually the most important source for information about an image, we determined the caption length for all images. In Table 3 we can see that there are large differences in the average caption length per discipline. While life sciences usually have long captions, mathematics and technical sciences tend to use shorter captions. In Fig. 1 we see the distribution of caption lengths.

4.4 Citations

We investigated whether the number of figures correlates with the citations to an articles as suggested by [5] and [6]. This information was added using the Crossref API. Those numbers were compared with other services. Although they were a bit lower overall, they correlated strongly. We assumed that more figures lead to more readers. Interestingly, the number of figures in an article does not correlate with the number of citations it has received (correlation: $6.19 \cdot 10^{-3}$, Fig. 4.3). This does not change considerably even after excluding all outliers with over 20 figures and over 100 citations (Table 4). However, articles with a figure count of 6-10 have the highest median citation count of 4. See [8] for details.

[3] We refer to the Revised Field of Science and Technology (FOS) classification at http://www.oecd.org/sti/inno/38235147.pdf.

Table 2. Average number of charts and images for disciplines with 2000 or more papers.

Discipline	#Papers	Charts/Paper	Images/Paper
all	932542	3.6	0.7
Medicine	432424	2.4	0.8
Biology	136655	3.9	0.6
Chemistry and Pharmacy	78525	3.7	0.3
Mathematics	34668	4.8	0.4
Physics	29900	5.7	0.8
Geosciences	25845	2.2	0.1
Process Engineering, Biotechnology	24019	4.6	1.4
Science in General	21779	6.4	1.1
Computer Science	19563	5.9	0.4
Electrical Engineering	14648	7.0	0.8
Energy, Environmental Protection	13321	4.7	0.8
General Technology	11587	9.7	1.0
Measurement and Control Engineering	14648	7.0	0.8
Mechanical Engineering	11052	8.6	3.2
Materials Science	11052	8.6	3.2
Agriculture and Forestry	12444	2.6	0.5
Nuclear Engineering	13297	4.7	0.8
Earth Sciences	7388	6.5	0.7
Psychology	5755	2.0	0.3
General Engineering	3375	6.7	1.3
Sports	3144	1.5	0.1
Architecture, Civil Engineering and Surveying	2774	12.7	1.5
Education	2736	1.4	0.1
Economics	2337	3.3	0.1

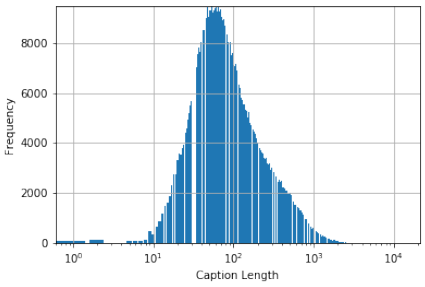

Fig. 1. Distribution of caption length on a logarithmic scale.

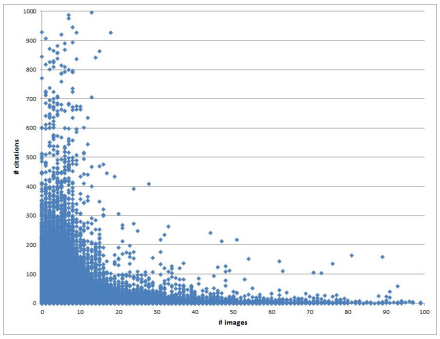

Fig. 2. Count of References.

Table 3. Caption length in characters for disciplines with over 10.000 figures. Disciplines are counted according to assignment of journals. Figures from journals assigned to more than one discipline are counted for each of these disciplines.

discipline	n	mode	median	mean
all	2963059	54	265	411.9
General Technology	124131	52	81	119.8
Mathematics	179023	52	84	126.3
Architecture Civil Engineering and Surveying	36931	70	89	117.6
Electrical Eng., Measurement and Control Eng.	115415	50	101	141.7
Energy, Environmental Protection, Nuclear Eng.	74368	68	116	175.5
Mechanical Eng., Materials Science	137878	69	125	174.0
Computer Science	126198	43	133	243.5
Geosciences	58875	111	140	159.5
General Engineering	27018	83	198	269.4
Agriculture and Forestry	29942	59	201	291.7
Earth Sciences	54480	86	220	294.2
Chemistry and Pharmacy	319335	111	228	416.5
Physics	199369	75	274	468.0
Psychology	13142	117	338	443.9
Process Eng., Biotechnology	144268	123	355	440.3
Medicine	1374680	69	357	471.8
Science in General	162051	47	513	697.1
Biology	615226	330	524	652.8

Table 4. Number of images and related citation counts

Articles in set (f=figures, c=citations)	number of papers	Median citation count	Mean citation count	Correlation between citation count and figure count
all	1048575	3	8,3	0,006192715
0-20 f., 0-100.c	984284	3	7	0,037702209
0 f.	211441	1	5,3	not possible
1-5 f.	519924	3	9	0,022292513
6-10 f.	238525	4	9,8	-0,008327417
11-678 f.	78688	2	6,1	-0,008684956

5 Discussion

The study gives an insight into a large data set based exclusively on open access articles.The dataset consists of articles with CC-BY-licenses that were available for mass download in an XML-format. The majority of figures within our corpus are charts. This figure type often visualizes research results and can range from the very standardized form of a graph with an x- and y-axis to drawings that can show abstract concepts in different formats. These figures could be used for research in the field of automatic information extraction. Images, on the other hand, are the more likely candidates for reuse since they usually do not show numbers that are only relevant for one paper. Researchers that work in analyzing

images should consider the average caption length in each discipline. Our paper shows a clear trend towards shorter captions in technology and longer captions in the life sciences. This could mean that captions in the life sciences generally contain more information and are therefore a better source for analysis than captions in other disciplines. However, it could also mean that this field needs more words to explain a single concept. Our results on the citation numbers do not match what [6] found. These differences could be explained by our inclusion of different disciplines or the slightly different way of ordering the numbers. This invites more study into the question whether figure use is a predictor for scientific impact, possibly with a focus on different disciplines. The result of our study is that the number of figures in a paper is not a good predictor for scientific impact. However, it seems like papers with between 1 and 10 figures, which are the most common, receive the most citations. Further research should include a more faceted classification of figure types and how they relate to different disciplines and citations.

Acknowledgment

This research was funded by the DFG under grant no. 315976924.

References

1. Agarwal, S., Yu, H.: FigSum: automatically generating structured text summaries for figures in biomedical literature **2009**, 6–10
2. Cabanac, G., Hubert, G., Hartley, J.: Solo versus collaborative writing: Discrepancies in the use of tables and graphs in academic articles **65**(4), 812–820
3. Charbonnier, J., Sohmen, L., Rothman, J., Rohden, B., Wartena, C.: NOA: A search engine for reusable scientific images beyond the life sciences. In: Advances in Information Retrieval. pp. 797–800. Lecture Notes in Computer Science, Springer, Cham. https://doi.org/10.1007/978-3-319-76941-7_78
4. Clark, C., Divvala, S.: PDFFigures 2.0: Mining figures from research papers. In: Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries. pp. 143–152. JCDL '16, ACM. https://doi.org/10.1145/2910896.2910904
5. Hegarty, P., Walton, Z.: The consequences of predicting scientific impact in psychology using journal impact factors **7**(1), 72–78. https://doi.org/10.1177/1745691611429356
6. Lee, P., West, J., Howe, B.: Viziometrics: Analyzing visual patterns in the scientific literature
7. Liu, F., Jenssen, T.K., Nygaard, V., Sack, J., Hovig, E.: FigSearch: a figure legend indexing and classification system **20**(16), 2880–2882. https://doi.org/10.1093/bioinformatics/bth316
8. Sohmen, L., Charbonnier, J., Blümel, I., Wartena, C., Heller, L.: Figures in scientific open access publications - underlying data (2018). https://doi.org/10.5281/zenodo.1295579
9. Xu, S., McCusker, J., Krauthammer, M.: Yale image finder (YIF): a new search engine for retrieving biomedical images **24**(17), 1968–1970. https://doi.org/10.1093/bioinformatics/btn340