Extracting data from vector figures in scholarly articles

CHJ Hartgerink (Tilburg University, c.h.j.hartgerink@uvt.nl)

06 July, 2017

Introduction

Information from figures in scholarly articles can be harvested for the underlying data (see also the "In Brief" report; Hartgerink, C. H. J. 2017). Figures are typically presented in order to communicate something about the underlying data, but in a static way. As such, reshaping this communication is not readily possible, because the data are not available. Examples of reuse if the data are available could be as simple as joining data across figures, standardizing axes across figures for easy comparison, or using the data to compute relative numbers instead of absolute numbers. Moreover, considering the current low rates of data sharing (Wicherts et al. 2006; Vanpaemel et al. 2015; Krawczyk and Reuben 2012) and rapid decrease of the odds of successfully requesting those data (Vines et al. 2014), reusing those data in the long run becomes effectively impossible. Hence, we find it important to be able to have alternative ways of collecting data from results presented in a scholarly report.

Some figures are stored in bitmap format whereas others are stored in vector format. In a bitmap format the image is stored by saving the color code for each pixel. This means that information about overlapping datapoints is lost, because a pixel in a bitmap does not differentiate between different layers. However, in a vector format, information is stored on the shape and its position on the canvas, which is unrestricted to a specific pixel size. As such, these images can be enlarged without loss of image quality. Moreover, the position of those shapes can be retraced in order to reconstruct data points in a figure. This can even be done when data points overlap, because unlike in the pixel format, overlapping shapes are stored separately in a vector image.

In the current report, we share the results of software (alpha stage) to extract raw data from vector based images. More specifically, we report the method of data extraction and the effectiveness and provide documentation for increasing ease-of-use for the software. Finally, we review the potential of using vector based images to extract data from scholarly reports.

Method

Extraction procedure

At the highest level, typical figure components are the body, header, footer, and axes. Figure 1 provides a conceptual depiction of the figure components. In order to extract data, recognition of some these components is mandatory, whereas recognition of others is optional. For example, the header and footer are irrelevant to data extraction, but are relevant to data comprehension; hence these are optional. Left- and bottom axes are mandatory, because these typically depict the scale of the presented plots. Right- and top axes are optional because they, as far as we know, rarely are used as the main axes and mostly are just to delimit the plotbox. Logically, the body of the plot, containing the depicted data, is mandatory for extraction.

Based on the plot body, absolute locations of the individual data points are extracted. Not all vector images are created in a similar way, but in the best case scenario, the vector gives three parameters: the x coordinate of the centre, the y coordinate of the centre, and the radius r. As such, for a simple circle the underlying vector code might look as follows:

```
<circle cx="103.71" cy="121.22" r="25.234" fill-opacity="0" stroke="#cf1d35" stroke-width=".26458"/>
```

The current alpha software is primarily developed to operate on circles of similar size within one plot.

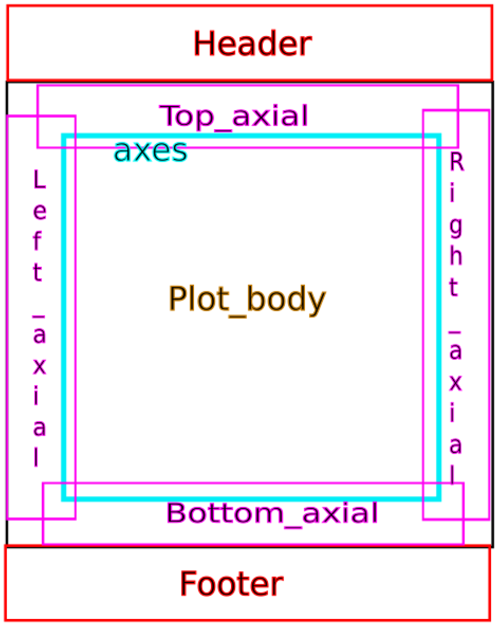


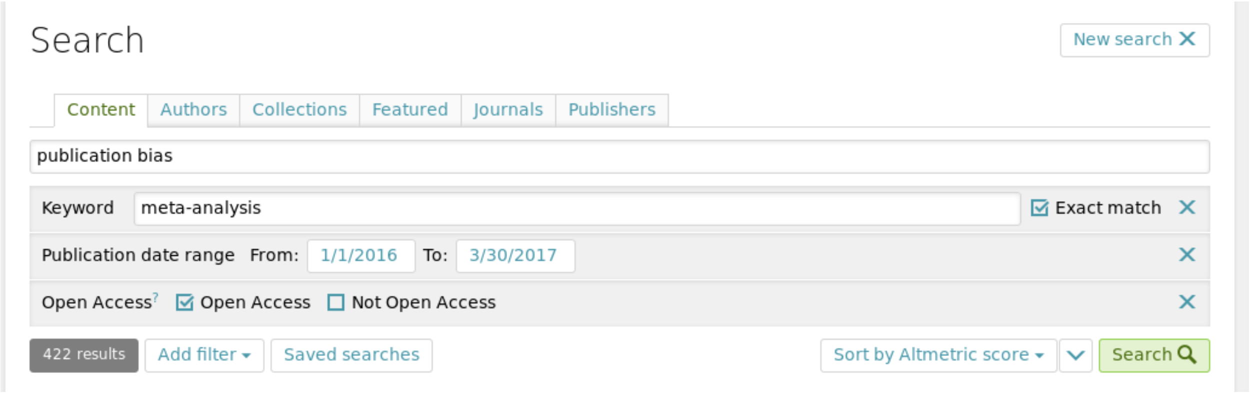
Figure 1: Conceptual representation of the typical components to a data based plot. This serves as the basis of the software to extract data from the plot body.

Corpus

Using ScienceOpen, we searched for meta-analytic reports that mention “publication bias”. Throughout this report, we focus on funnel plot figures from meta-analyses, hence this search approach. ScienceOpen aggregates several databases and preprint services, resulting in over 30 million records. Moreover, the ScienceOpen database allowed us to restrict search results to Open Access publications, in order to be able to freely redistribute those publications in our Github repository for this project. This facilitates reproducibility of the data based on the newly developed software. We searched the ScienceOpen database on March 30 2017, resulting 422 reports based on the search criteria (Figure 1), but the webpage presented only 368 reports.

2. Collect corpus and data on vector images

- Determine vector figure based on selectable axes
-



Search

Content | Authors | Collections | Featured | Journals | Publishers

publication bias

Keyword: meta-analysis ☑ Exact match

Publication date range From: 1/1/2016 To: 3/30/2017

Open Access: ☑ Open Access □ Not Open Access

422 results

Sort by Altmetric score

Figure 2: Screenshot of the search criteria used to search ScienceOpen.

Documentation

1. Have pdfs
2. Make cproject
3. Convert pdfs to svg
4. Manually clip pages with plots of interest to only retain plot of interest
5. Convert figures to data

Results

Through searching on ScienceOpen, we identified 16 meta-analytic reports containing vector based funnel plots. Upon manual inspection of those 368 meta-analytic reports, only 136 contained funnel plots. Of those 136, we identified 16 reports with vector based images. As mentioned in the methods section, we determined whether a funnel plot was vector based by trying to select the ticks on the axes; if these were selectable, we deemed the funnel plot a vector.

2. Extracting data

- Identified vectors == vector: 13
- Number of vectors in those papers: 25
- DOIs of papers that were extracted: 10.1186/s12885-016-2685-3, 10.1186/s12889-016-3083-0, 10.1186/s13027-016-0058-9, 10.1186/s40064-016-3064-x, 10.1515/med-2016-0052, 10.1515/med-2016-0099, 10.1590/S1518-8787.2016050006236, 10.21053/ceo.2016.9.1.1, 10.2147/BCTT.S94617, 10.3349/ymj.2016.57.5.1260, 10.3390/ijerph13050458, 10.5114/aoms.2016.61916, 10.5812/ircmj.40061
- Number of vectors with data extracted: NA
- data mapped correctly: 0

Table 1: Papers

doi
10.1186/s12885-016-2685-3
10.1186/s12889-016-3083-0
10.1186/s13027-016-0058-9
10.1186/s40064-016-3064-x
10.1515/med-2016-0052
10.1515/med-2016-0099
10.1590/S1518-8787.2016050006236
10.21053/ceo.2016.9.1.1
10.2147/BCTT.S94617
10.3349/ymj.2016.57.5.1260
10.3390/ijerph13050458
10.5114/aoms.2016.61916
10.5812/ircmj.40061

Discussion

As the results indicate, vector based images are a .

However, the usage of vec

Author notes

All presented materials in this report are version-controlled and freely available on Github (https://github.com/chartgerink/2015ori-3), and were preserved upon submission at Zenodo [@]. This research was funded by the Office of Research Integrity under grant NUMBER_INSERT.

References

Hartgerink, C. H. J. 2017. "Developing the Potential of Data Extraction in Scholarly Articles for Research Verification." *D-Lib Magazine*. http://www.dlib.org/dlib/may17/05inbrief.html#HARTGERINK.

Krawczyk, Michal, and Ernesto Reuben. 2012. "(Un)Available Upon Request: Field Experiment on Researchers' Willingness to Share Supplementary Materials." *Accountability in Research* 19 (3): 175–86. doi:10.1080/08989621.2012.678688.

Vanpaemel, Wolf, Maarten Vermorgen, Leen Deriemaecker, and Gert Storms. 2015. "Are We Wasting a Good Crisis? The Availability of Psychological Research Data After the Storm." *Collabra* 1 (1). University of California Press. doi:10.1525/collabra.13.

Vines, Timothy H., Arianne Y.K. Albert, Rose L. Andrew, Florence Débarre, Dan G. Bock, Michelle T. Franklin, Kimberly J. Gilbert, Jean-Sébastien Moore, Sébastien Renaut, and Diana J. Rennison. 2014. "The Availability of Research Data Declines Rapidly with Article Age." *Current Biology* 24 (1). Elsevier BV: 94–97. doi:10.1016/j.cub.2013.11.014.

Wicherts, Jelte M., Denny Borsboom, Judith Kats, and Dylan Molenaar. 2006. "The Poor Availability of Psychological Research Data for Reanalysis." *American Psychologist* 61 (7). American Psychological Association (APA): 726–28. doi:10.1037/0003-066x.61.7.726.