WG6 Report

SciELO 20 Years

WG6 presentation title:	Workshop on the use of data from the SciELO database
Lecturer/speaker/rapporteur:	Danilo J. S. Bellini
Group coordinator:	Gustavo Fonseca
Executive secretary:	Carolina Tanigushi
WG6 date:	2018-09-24
Report date:	2018-09-25
Venue:	Tivoli Mofarrej São Paulo Hotel

Summary

During the workshop, these had been done:

- Brief introduction to the data analysis processes, emphasizing data access, data cleaning and exploratory data analysis
- *Hands-on* examples using Python and R, with emphasis in the *Pandas* library resources, showing how data munging, normalization, data visualization and other data analysis processes can be performed
- Explanation of data analyses previously performed on data coming from SciELO and external sources

Most of the time was spent on exploratory data analysis, interpretation of descriptive statistics and visualization. Some highlights of what had been studied in more depth include:

- Hirsch index
  - Google Scholar's `h5-index` and `h5-median`
  - Calculation from raw Dimensions' data
  - SCImagoJR's H index
- Field Citation Ratio (FCR) from Dimensions

Tools

Two programming languages had been used, besides several:

- Python
  - IPython / Jupyter Notebook
  - Python built-in modules (`csv`, `statistics`, `urllib`, `json`, `glob`, `os`, `re`, `collections`, `itertools`, `pprint`)
  - `numpy`
  - `pandas`
  - `matplotlib`
  - `seaborn`
  - `openpyxl`, to open XLSX files
  - `scipy.stats`, to calculate the Pearson's correlation coefficient
  - NetworkX, a graph manipulation library including an API to draw graphs with matplotlib
- R
  - R built-in modules (`base`, `utils`, `stats`, `graphics`)
  - R Studio, an IDE for R, for creating R Markdown notebooks
  - `dplyr`, to perform grouping operations similar to SQL's `GROUP BY` and Pandas' `DataFrame.groupby`

Data sources

Several analyses were performed before the workshop, whose processes and results were part of it. The data that had been studied in every analysis performed during and before the workshop came from these sources:

- SciELO's JSON APIs (RESTful) from:
  - *ArticleMeta*, to get journal metadata
  - *Ratchet*, to get access data
- SciELO's `articlemeta` Python library, an alternative way to access the ArticleMeta API
- Reports from the *SciELO Analytics*
- *SciELO Citation Index* entries from the *Web of Science*
- *Dimensions* data regarding two journals: *Nauplius* and *Brazilian Journal of Plant Physiology*
- *SCImagoJR*'s CSV with all SJR and H indices for 2017
- *Scopus*' XLSX with all the data they make available

Introduction to the analysis

Besides:

- Identifying which collections have data in SciELO analytics (all certified and development collections, besides the active independent collections)
- Downloading all SciELO analytics reports
- Evaluating if the network reports have everything from the remaining reports
- Simplifying the column names
- Normalizing/cleaning the ISSN when dealing with multiple collections
- Normalizing the thematic area (dealing with unfilled data)

It had been seen how to plot data, with a strong emphasis on data interpretation and multiple types of plots (bar plots, line plots, box-and-whisker plots, heat maps, scatterplots, etc.), as well as subplot splitting/grouping.

Previously prepared analyses

- Number of indexed journals in the SciELO network
- Deindexing reason in the SciELO Brazil collection
- Evaluating the daily access in the SciELO Brazil collection
- Three indices in Scopus 2017: CiteScore, SNIP and SJR
- SCImago Journal Rank in 2017, including SJR and H index
- FCR and H index in Dimensions
- Google Scholar indices
- Languages of research articles in SciELO Brazil, by thematic area, document publication year and journal indexing year
- Citations in the SciELO CI
- Proportion of Brazil in affiliation institutions in research articles from journals in the SciELO Brazil collection

Results

- The proportion of Brazilian affiliations of research articles in the SciELO Brazil collection is decreasing
- Most citations of research articles in the SciELO network come from documents/journals that aren't in the SciELO network. For research articles written in English, 76% of the received citations comes from documents external to the SciELO network
- The normalization step when calculating the FCR and its non-standard average calculation can easily *push down* the result (e.g. a journal with 10 documents receiving 15 citations and 3 documents with zero citations would have an average of citations of less than 7.5, before this number gets normalized by the year and field of research), making it an index best fit to evaluate older journals that are no longer publishing
- We should always look for the mathematics that defines an index, as that evaluation can already give us some insights regarding its bias towards some documents/journals

Results

- Scopus indices should be taken with care: mixing the data from all countries makes it hard to compare data from SciELO and from other journals
- In SciELO Brazil, 95% of the journals marked as *deceased* were actually just *renamed* to a new journal title/ISSN
- Matching data with external sources is difficult without a common standardized index such as the ISSN and DOI
- 0.8% of SciELO journals are marked in Scopus as not open, which seem to be an issue regarding Scopus data