plosOpenR
Exploring FP7 funded PLOS publications

Najko JAHN [a,1], Martin FENNER [b,c] and Jochen SCHIRRWAGEN [a]
[a] *Bielefeld University Library, Bielefeld University, Germany*
[b] *Hannover Medical School, Germany*
[c] *Technical Lead PLOS Article Level Metrics, Public Library of Science, USA*

Abstract. This case-study presents possible fields of applications for alternative science measures on grant-supported research publications. The study is based on plosOpenR, a set of functions for the statistical computing environment R. plosOpenR facilitates access to the application programming interfaces (API) provided by Open Acces publisher Public Library of Science (PLOS) and OpenAIRE, a Open Access infrastructure for European funded research. These functions can be furthermore used to visually explore PLOS-Article Level metrics, a collection of alternative science impact measure. 1,166 publications acknowledging funding within the Seventh Framework Programme of the EU are reported. Distributed over 624 projects, alternative metrics based on this sample and corresponding collaboration networks are presented. Findings reveal the potential of reusing data, that are made openly and automatically available by publishers, funders and repository community.

Keywords. Data mining, Article-Level Metrics, Statistical Computing, R, PLOS, OpenAIRE

1. Introduction and Motivation

With the growing number of openly available research services, new opportunities arise for measuring performance and the impact of research output. The quantitative study of scholarly communication no longer solely depends on traditional citation analysis, but is complemented with usage and social media data. Publishers, funders and the repository community are likewise seeking for ways to provide relevant data for the ongoing work on alternative metrics.

Since its launch, the Open Access publisher Public Library of Science (PLOS)[2] has been a strong advocate of alternative ways to publish and measure research. PLOS provides a suite of tools and metrics on every PLOS research contribution that display the citations, usage statistics from PLOS and PubMed Central, and social media activity on the article web page.[3] In November 2012 PLOS launched the Altmetrics collection,

[1] Corresponding Author: Bielefeld University Library, Bielefeld University, Universitaetsstrasse 25, 33602 Bielefeld, Germany; E-mail: najko.jahn@uni-bielefeld.de.
[2] http://www.plos.org/
[3] http://article-level-metrics.plos.org/

a growing collection of papers that study alternative science measures [1] published in PLOS journals.

The FP7 funded project OpenAIRE[4], which aims to set up an Open Access Infrastructure for Research in Europe, features a dedicated work package for exploiting usage metrics as a supplement to conventional citation analysis[2]. To this end, OpenAIRE has started to aggregate data from federated usage data providers.[2]

In order to stress potential fields for data exchange and use cases in the field of alternative metrics, PLOS and OpenAIRE jointly develop a set of tools for the statistical computing environment R[3]. Demonstrated on FP7-funded research output that has been identified in PLOS journal publications, these sources allow the exploration of PLOS Article-Level metrics. The sources are collected in the Open Source GitHub software repository plosOpenR.[5] It is part of PLOS ALM Development collection on GitHub, that provides the PLOS ALM applications in use as well.

Derived from a diverse and fast developing set of research services, a growing community of scientists also collaborate on statistical tools to reuse and analyse research output [4,5]. plosOpenR contributes to the collaborative effort to develop R-based tools for facilitating Open Science (rOpenSci)[6], in particular, it informs the further development of the R package rplos[7].

This case study discusses first results of plosOpenR. After providing an overview on the APIs and existing software solutions used in the next section, preliminary results for 1,166 PLOS research article acknowledging 624 FP7 funded projects are presented. Finally, the potential benefits and limits from the perspectives of research infrastructure and quantitative science studies are discussed.

2. Background and Data

PLOS provides two public available APIs. The Solr-based PLOS Search API[8] gives access to the fulltext-corpus of all PLOS articles published. Search fields correspond to article sections. For our study of FP7 funding acknowledgement visible in PLOS articles, we mainly rely on the fields listed in Table 1.

Field	Description
`id`	DOI (Digital Object Identifier)
`financial_disclosure`	Funding acknolwedgement (free text)
`affiliate`	Affiliation of the authors (free text)

Table 1. PLOS Search API fields

PLOS Article-Level Metrics API can be used to retrieve measures on PLOS articles.[9] The API is written in Ruby. For our work we analysed the following sources listed in Table 2 on the next page.

[4] https://www.openaire.eu/
[5] https://github.com/articlemetrics/plosOpenR
[6] http://ropensci.org/
[7] https://github.com/ropensci/rplos
[8] http://api.plos.org/solr/search-fields/
[9] http://api.plos.org/alm/faq/

Provider	Source	Description
Usage		
PLOS	`counter`	HTML article view, pdf and XML downloads (COUNTER 3)
PubMed Central	`pmc`	HTML article view, pdf and XML downloads PubMed Central
Citations		
PubMed Central	`pubmed`	Times cited for an article from PubMed Central
CrossRef	`crossref`	Times cited for an article from CrossRef
Scopus	`scopus`	Times citedr an article from Scopus
Social media events		
Twitter	`twitter`	Tweets for an article
Facebook	`facebook`	The number of Facebook Likes for an article
Mendeley	`mendeley`	The number of times a user has bookmarked an article in Mendeley
CiteULike	`citeulike`	The number of times a user has bookmarked an article in CiteULike
PLOS Comments	`ploscomments`	The number of times a user has comment an article on PLOS

Table 2. PLOS ALM API fields examined

OpenAIRE exposes FP7 funding information via its OAI-PMH interface.[10] Detailed project information are listed in the set `project`. Fields used to identify and contextualize projects in correspondence with acknowledgement in PLOS articles are

`GrantID, acronym, title, start date, end date, call id, fundedby, fundedhow, sc39`

Analysing the article level metrics for a set of PLOS articles involves three steps:

- Retrieve a set of articles through the Search API
- Collect the metrics for these articles
- Visualize the metrics

The R package rplos[11] provides many useful tools to query and analyse PLOS research output. Many of its functions are used in our case-study. plosOpenR extends rplos' functionality for aggregating financial disclosure and affiliations. It furthermore interfaces with the PLOS APIs transforming the JSON and XML outputs into R `data.frame` structures in order to allow easier statistical analysis.

Article level metrics are much easier to understand through visualizations. R provides powerful graphic devices often used in statistics. Within plosOpenR, different visualisation techniques are demonstrated and documented on the PLOS API webpage[12]. For our case-study, we focus on alternative scatterplots to explore ALM distributions and network visualisations to examine collaboration pattern and maps for spatial analysis. For the latter, map data from http://thematicmapping.org/ is used.

[10] Base URL http://api.openaire.research-infrastructures.eu:8280/is/mvc/openaireOAI/oai.do

[11] https://github.com/ropensci/rplos

[12] http://api.plos.org/2012/07/20/example-visualizations-using-the-plos-search-and-alm-apis/

To allow a broader query and better match of FP7 funding acknowledgement visible in PLOS articles, another processing step outside of R is implemented for this case-study. It reuses a text mining/rule-based named entity recognition approach that has been developed in OpenAIRE for information space curation and enrichment.[6]

3. Results

3.1. FP7 Contribution in the PLOS Domain

Funding information for articles published in PLOS is provided as free-form text in the financial disclosure section. The openly available PLOS Search API allows specific queries of this section. Querying the search field `financial_disclosure:`

```
((europ* AND (union OR commission)) OR fp7) OR ((seventh OR 7th)
AND framework) OR (ERC OR (European Research Council)) OR ((EU OR EC)
AND project)
```

we obtained 2,562 candidate publications on 19 July 2012. In total, we matched 1,166 PLOS articles that acknowledged at least one FP7 research project. The FP7 acknowledgement by PLOS journal and publishing year show a moderate growth in most journals but a strong growth in PLOS ONE (Table 3). 77.78% of FP7 acknowledgement was detected in PLOS ONE contributions.

Journal / Publishing Year	2008	2009	2010	2011	2012*	Σ
PLoS ONE	8	36	132	358	335	869
PLoS Computational Biology	1	10	16	21	16	64
PLoS Genetics		10	17	20	26	73
PLoS Biology	1	5	6	12	8	32
PLoS Pathogens	1	8	15	41	26	91
PLoS Medicine			3	3	6	12
PLoS Neglected Tropical Diseases		1	5	7	12	25
Σ	11	70	194	462	429	1166

Table 3. FP7 funding acknowledgement in PLOS journals 2008–2012 (*until 19 July 2012)

On this basis, the compound annual growth rate can be calculated for PLOS ONE. For the period from 2009 to 2011 the compound annual growth rate for PLOS ONE articles is calculated as being 215.35 %. This number is consistent with the overall fast growth of the journal.

We identified 624 FP7 projects that have published in PLOS journals until July 19 2012. Table 4 on the next page presents the distribution over projects by its number of publications in PLOS journals.

The figure shows a positively skewed distribution of PLOS FP7 contributions with 57.53 % of the FP7 projects to be acknowledged once, while 1.92 % more than 8 times in PLOS journals. With 30 contributions, the FP7 research project *European Network for Genetic and Genomic Epidemiology* (ENGAGE) has published the highest number of PLOS articles.

PLOS per FP7	Frequency	Relative Frequency (in %)
1	359	57.53
2	129	20.68
3 – 8	124	19.87
9 – 30	12	1.92
	624	100

Table 4. PLOS contributions by EC funded research projects

The OpenAIRE OAI-PMH interface[13] provides further funding information for each FP7 research project. At the time of our study, 17,736 FP7 research projects were exposed which are distributed over 23 funding programmes. To determine the visibility of FP7 funding programmes in the PLOS domain, we examined the overall share of funding programme membership for every FP7 project and compared them with our PLOS sample (Table 5 on the following page). Like the distribution over FP7 research projects, funding acknowledgement in PLOS articles were also unequally distributed over funding programmes; whereas 27.96 % projects in the funding programme *Health Research* (HEALTH) contributed at least once in PLOS journals, no funding acknowledgement could be detected for 8 funding programmes (34.78 %), e.g. *Transport (including Aeronautics)* (TPT).

In total, 3.52 % of FP7 funded research projects and 65.22 % FP7 funding programmes are acknowledged in at least one PLOS article. The distribution of funding programmes is consistent with the PLOS focus on biomedical research and related fields.

In contrast, taking FP7 projects under the SC39 clause into account, we revealed that the proportion of SC39 funded research in the PLOS domain (94 out of 624 FP7 publications, 7.34 %) is higher than its proportion in the FP7 funding scheme (530 out of 17,736 FP7 projects, 3.22 %).

3.2. Article-Level Metrics

With its PLOS Article-Level Metrics (PLOS-ALM) API, PLOS provides information about received citation, usage data and dedicated social media events. Figure 1 on page 7 shows the PLOS proportion of FP7 articles covered by ALM source until 3 September 2012.

For every PLOS contribution in our sample, we were able to collect usage data from both the PLOS journal website (counter) and from PubMed Central (pmc).

However, coverage within the ALM categories of citation and social media events is more heterogeneous: citing articles referred to between 43 % (pubmed) and 63 % (crossref) of the articles and between 8 % (comments on PLOS articles) and 81 % (Mendeley readerships) were mentioned in social media services. Note that the collection of Twitter mentions within PLOS ALM started June 1st, 2012.

The inclusion of the date of publication in the PLOS ALM API allows us to compare the occurrences of ALM event types for every day since publication. Figure 2 on page 7 depicts article age (in days since publication) and total views on the PLOS website. As

[13] Base URL http://api.openaire.research-infrastructures.eu:8280/is/mvc/openaireOAI/oai.do

EC Funding Programme	Projects funded	Projects acknowledged in PLOS	Ratio (in %)
HEALTH	769	215	27.96
KBBE	421	46	10.93
GA	25	2	8
INFRA	311	16	5.14
ENV	406	20	4.93
REGPOT	169	8	4.73
ERC	2909	122	4.19
ICT	1731	67	3.87
PEOPLE	7878	115	1.46
NMP	584	8	1.37
Fission	101	1	0.99
SiS	142	1	0.70
SEC	198	1	0.51
ENERGY	303	1	0.33
SME	694	1	0.14
Fusion	3	0	0
COH	23	0	0
INCO	126	0	0
CIP-EIP	15	0	0
TPT	521	0	0
REGIONS	65	0	0
SPA	162	0	0
SSH	180	0	0
Total	17736	624	3.52

Table 5. Comparing the proportion of FP7 funded projects and their acknowledgement in PLOS articles by funding programmes

a third variable we compared Scopus citations and Facebook shares received for each FP7 funded research article in PLOS journals (mapped as point size). Citation rates are time-dependent with few citations in the first 12 months of publication, and the article set contains many articles published in 2012 (Figure 3.2 on the next page). With 38,361 total views, best represented in our sample is the publication:

> Vitali S, Glattfelder JB, Battiston S (2011) The Network of Global Corporate Control. PLoS ONE 6(10): e25995. doi:10.1371/journal.pone25995 [published 2011-10-26]

However, while this article funded by *Forecasting Financial Crisis FOC-II* is also best ranked in the event type category Facebook shares, with 10 times cited according to Scopus the article is ranked in the 87th percentile.

According to the blog of the FOC project, the article has gained broad public media attention.[14] In this post, the authors warn to misinterpret and their findings as it seems to

[14] http://forecastingcrises.wordpress.com/2011/10/27/the-network-of-global-corporate-control-2/

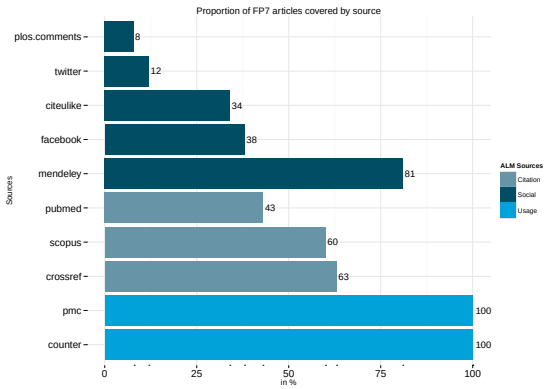
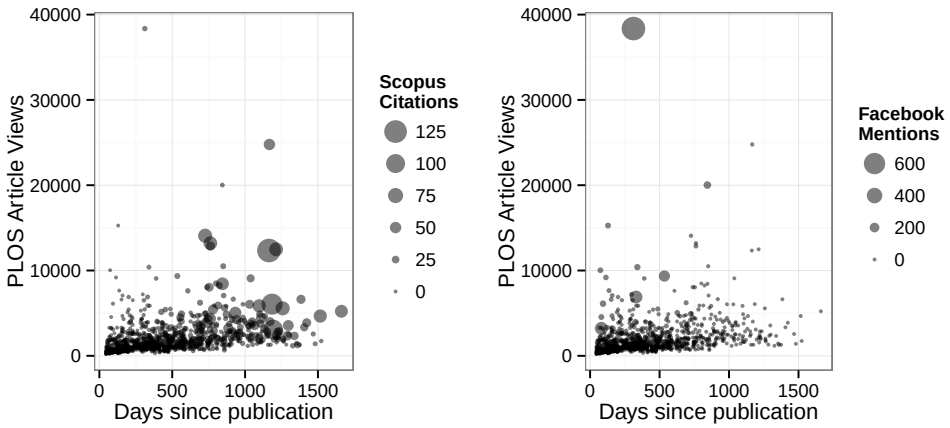


Figure 1. PLOS proportion of FP7 articles covered by ALM source



Scopus citations mapped to point size Facebook shares mapped to point size

Figure 2. Scatterplots of total PLOS views for days since publication with (a) Scopus citations and (b) Facebook shares mapped to the size of points

be done in public media.

3.3. Collaboration Patterns

Our dataset can also be used to explore the collaboration patterns between FP7 research projects in joint PLOS publications. We found that 9.52 % of the PLOS publications under consideration acknowledged more than one FP7 research funding project (Table 6 on the following page). From 624 FP7 projects identified, 26.28 % were acknowledged together with another FP7 projects. Figure 3 on the next page explores the relations between single FP7 research projects. The edges represent the number of joint PLOS contributions. In total, we detected 57 components that consist of 164 projects.

The FP7 research project *European Network for Genetic and Genomic Epidemiology* (ENGAGE) is most frequently represented in our sample again, counting for both the most PLOS contributions and the highest number of direct links in the network (20).

FP7 Projects per paper	Frequency	Relative Frequency (in %)
1	1,055	90.48
2	93	7.98
3	14	1.20
4	4	0.34
Σ	1,166	100

Table 6. FP7 funding acknowledgement per PLOS article

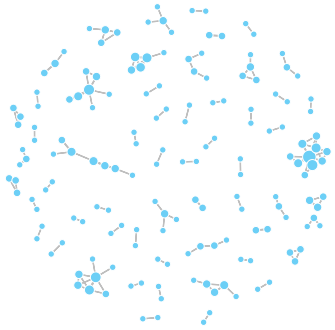


Figure 3. Collaboration network FP7 projects in PLOS journals. Edge width show the number of joint articles of a collaborating institutional pair, vertice size represent the degree centrality.

Furthermore, we used the PLOS Search API to retrieve the affiliations of the authors. In total, we obtained 6,090 author addresses distributed over 1,166 publications. 6,049 correctly formatted country names could be extracted that are distributed over 95 countries (Figure 4 on the facing page). Affiliations listed originated most frequently from United Kingdom (12.25 %), Germany (11.77 %) and the United States (11.69 %).

If compared to World Bank regions,[15] more than every third affiliation listed originated from the region of Western Europe. Regarding the distribution of PLOS publications over FP7 projects and countries, this distribution is positively skewed with the regions of Western Europe, Northern Europe, Southern Europe and Northern America accounting for 92.97 % of all affiliations detected.

4. Discussion & Conclusion

In this study, we report 1,166 research contributions in PLOS journals acknowledging FP7 funding from 624 different research projects. With plosOpenR, the sources used

[15] http://www.worldbank.org/countries

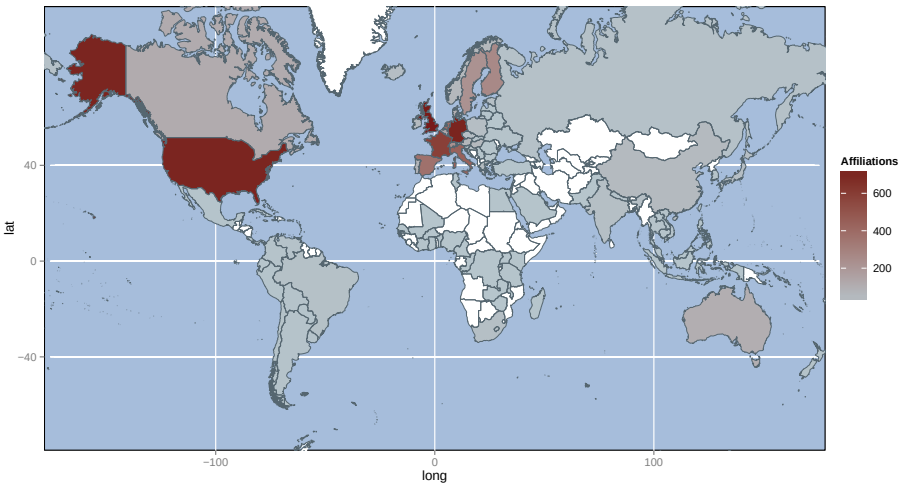


Figure 4. Affiliation world map of FP7 contributions in PLOS journals

for this report are collected in a public GitHub repository to be reused in the statistical computing environment R.

From a research infrastructure and services point of view, our results highlight the importance of openly available research services that complement traditional citation analysis. In this report, we combine two such data sources, namely from the Open Access publisher PLOS and the OAI-PMH interface, an Open Access infrastructure for European funded research.

While PLOS provides us with detailed information on articles and their impact, OpenAIRE enables a distinct view on FP7 funded research published in PLOS journals by exposing the relevant project information. Furthermore, we demonstrate the aggregation of article-level metrics consisting of citation counts, usage and social media events. For all research contributions under investigation, we were able to retrieve usage data on a daily basis from the PLOS journal website and the disciplinary repository PubMed Central hosted by NIH (PMC). Normalized FP7 funding acknowledgement and address fields allow both to explore collaboration networks manifested in the research contributions.

When discussing the results in the light of quantitative science studies on performance and impact of research publications, it has to be noted that our study is limited in various ways. Firstly, we were only able to examine research contributions published in PLOS journals until 19 July 2012. The continuing increase in FP7-funded PLOS papers from January to June 2012 as well as the duration of the Seventh Framework programme suggests that potentially more FP7 funded research acknowledged in PLOS journals are to be expected in future. Secondly, the extreme positive skewness of most distribution under considerations demands careful analysis and interpretation. Even though we detected

more than three quarters of research articles in the multi-disciplinary journal PLOS ONE, it only partially cover FP7 funding projects and programmes.

Particular care needs to be taken if future studies rank research articles according to the different metrics in use and develop comparative indicators that rely on these data. For instance, our exploration of Scopus citation counts in comparison with the social media event type Facebook shares on the article level revealed that public media attention has effects on analysing and interpreting research publications. Whereas the majority of usage data and social web activity happens in the days and months after publication, citation data are accumulating much more slowly. The set of FP7-funded PLOS articles that we identified should therefore be reanalyzed at least two years after the last paper in the set has been published. However, with data sources and visualization methods suggested in this report, we provide tools for easy on-time exploration of article level metrics in order to identify irregular patterns that motivate qualitative investigation.

Future work and studies on PLOS article-level metrics will focus on main problem areas [7]. With the evolving European federation of usage-data providers, OpenAIRE has the potential to provide additional information about usage events and might complement PLOS ALM as PMC already does.

Acknowledgement

We thank Harry Dimitropoulos for applying his text mining/rule-based named entity recognition approach on our sample. Jochen Schirrwagen gratefully acknowledges support from the European Commission (FP7-INFRA-2007-1.2.1, Grant Agreement no. 246686).

References

[1] Priem J, Groth P, Taraborelli D. The Altmetrics Collection. PLoS ONE. 2012;7(11):e48753.

[2] OpenAIRE. OpenAIRE Guidelines for Usage Statistics v1.0; 2011. Available from: http://www.openaire.eu/en/about-openaire/publications-presentations/publications/doc_details/314-openaire-guidelines-for-usage-statistics-v10.

[3] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2012. Available from: http://www.R-project.org.

[4] Boettiger C, Lang DT, Wainwright PC. rfishbase: exploring, manipulating and visualizing FishBase data from R. Journal of Fish Biology. 2012;81(6):2030–2039.

[5] Boettiger C, Lang DT. Treebase: an R package for discovery, access and manipulation of online phylogenies. Methods in Ecology and Evolution. 2012;3(6):1060–1066.

[6] Manghi P, Bolikowski L, Manold N, Schirrwagen J, Smith T. OpenAIREplus: the European Scholarly Communication Data Infrastructure. D-Lib Magazine. 2012 Sep;18(9/10). Available from: http://www.dlib.org/dlib/september12/manghi/09manghi.html.

[7] ALM Workshop 2012 Report, PLOS ALM; 2012. Doi: 10.6084/m9.figshare.98828. figshare.