Information Services & Use 33 (2013) 93–101
DOI 10.3233/ISU-130704
IOS Press

PlosOpenR – Exploring FP7 funded PLOS publications

Najko Jahn [a,*], Martin Fenner [b,c] and Jochen Schirrwagen [a]

[a] *Bielefeld University Library, Bielefeld University, Bielefeld, Germany*

[b] *Hannover Medical School, Hannover, Germany*

[c] *Technical Lead PLOS Article Level Metrics, Public Library of Science, San Francisco, CA, USA*

Abstract. This case study explores alternative science metrics on grant-supported research publications. The study is based on plosOpenR, a software package for the statistical computing environment R. plosOpenR facilitates access to the application programming interfaces (API) provided by Open Access publisher Public Library of Science (PLOS) and OpenAIRE – Open Access Infrastructure for Research in Europe.

We report 1,166 PLOS articles that acknowledge grant support from 624 different research projects funded by the European Union's 7th Framework Programme (FP7). plosOpenR allows the exploration of PLOS Article-Level Metrics (PLOS ALM), including citations, usage and social media events as well as collaboration patterns on these articles. Findings reveal the potential of reusing data, that are made openly and automatically available by publishers, funders and the repository community.

Keywords: Research evaluation, article-level metrics, statistical computing, R, PLOS, open access, OpenAIRE, 7th Framework Programme

1. Introduction and motivation

With the growing number of openly available research services, new opportunities arise for measuring performance and impact of research publications. The quantitative study of scholarly communication no longer solely depends on traditional citation analysis, but is complemented by 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)[1] has been a strong advocate of alternative ways to publish and measure research. On every article, PLOS displays a suite of indicators including citations, information on usage and social media activity.[2] In November 2012, PLOS also started the Altmetrics collection inviting research articles on alternative science metrics [8].

It is promising to apply PLOS Article-Level Metrics (PLOS ALM) in broader contexts, as for instance on a particular funding programme. OpenAIRE – Open Access Infrastructure for Research in Europe[3] gives access to results funded by the European Union's 7th Framework Programme (FP7) in general and FP7 grants subject to the Open Access pilot of the European Commission under the Special Clause

[*] Corresponding author: Najko Jahn, Bielefeld University Library, Bielefeld University, Universitaetsstrasse 25, 33602 Bielefeld, Germany. E-mail: najko.jahn@uni-bielefeld.de.

[1] Public Library of Science (PLOS), http://www.plos.org/.

[2] PLOS Article-Level-Metrics (ALM), http://article-level-metrics.plos.org/.

[3] OpenAIRE, http://www.openaire.eu/.

39 (SC39) in particular. OpenAIRE features a dedicated work package for exploiting usage metrics on Open Access publications as a supplement to conventional citation analysis. To this end, OpenAIRE has started to aggregate data from federated usage data providers [7].

In this case study, we stress the potential of PLOS ALM information on exploring FP7 grant-supported research publications. Moreover, our considerations reflect a growing community of scientists that collaborate on statistical tools to access and analyse research output [2,3]. The case study, therefore, derives from a set of tools for the statistical computing environment R [9] – plosOpenR.[4]

After introducing the APIs and existing software solutions used in the next section, we demonstrate plosOpenR on the basis of 1,166 PLOS research articles acknowledging 624 FP7 funded projects. Finally, we discuss the potential benefits and limits from the perspectives of research infrastructure and quantitative science studies.

The dataset of our case study is available at doi:10.5281/ZENODO.1239.

2. Background and data

PLOS offers two public available APIs that plosOpenR uses. The PLOS Search API[5] gives developers access to the fulltext-corpus of all PLOS articles published. The search fields correspond to article sections. For our study on information about FP7 funded research published in PLOS, we mainly rely on the fields listed in Table 1.

The PLOS ALM API is used to retrieve metrics on PLOS articles.[6] PLOS releases the underlying software under an Open Source licence.[7] For our work, we analysed the following PLOS ALM providers listed in Table 2.

OpenAIRE exposes FP7 funding information via its OAI-PMH interface[8] which reflects the OpenAIRE data model [4]. Detailed project information are listed in the set `project`. Fields that we have used to identify and contextualize projects in correspondence with acknowledgement in PLOS articles are `GrantID, acronym, title, call id, fundedby, fundedhow, sc39`.

plosOpenR follows three steps to explore PLOS ALM on FP7 grant-supported research publications:

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

Table 1

PLOS search API fields

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

[4] plosOpenR repository on GitHub, https://github.com/articlemetrics/plosOpenR.

[5] PLOS Search Fields, http://api.plos.org/solr/search-fields/.

[6] PLOS ALM FAQ, http://api.plos.org/alm/faq/.

[7] ALM software repository: https://github.com/articlemetrics/alm/.

[8] OpenAIRE BaseURL, http://api.openaire.research-infrastructures.eu:8280/is/mvc/openaireOAI/oai.do.

Table 2

PLOS ALM source fields examined

Family/provider	ALM-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 cited 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

For this purpose, plosOpenR reuses already existing tools to query and analyse PLOS research output that belong to the rplos package. rplos is developed by rOpenSci, a collaborative effort to provide R-based applications for facilitating Open Science.[9]

After querying the PLOS APIs, plosOpenR transforms the so retrieved JSON and XML outputs into `data.frame` structures in order to allow easier statistical analysis within R.

PLOS ALM are much easier to understand through visualizations. R provides powerful graphic devices often used in statistics. plosOpenR demonstrates different visualisation techniques which are documented on the PLOS API webpage.[10] For our case study, we focus on alternative scatterplots to explore ALM distributions, on network visualisations to examine collaboration patterns [6] and on choropleth maps displaying author's country of affiliation. For the latter, plosOpenR uses the Thematic Mapping API.[11]

To allow a broader query and more reliable match of FP7 funding acknowledgement visible in PLOS articles, a processing step outside of R applies OpenAIRE's text mining, rule-based named entity recognition approach to identify FP7 project Ref. [5].

3. Results

3.1. FP7 contribution in the PLOS domain

PLOS gives access to grant information in the financial disclosure section. The openly available PLOS Search API allows specific queries of this section. On 19 July 2012, we obtained 2,562 candidate publications after querying the search field `financial_disclosure`:

[9] rOpenSci: http://ropensci.org/.

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

[11] Thematic Mapping, http://thematicmapping.org/.

```
  ((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)
```

In total, we matched 1,166 PLOS articles that referenced 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). This journal represented 77.78% of FP7-supported research publications.

On this basis, we calculated the compound annual growth rate for PLOS ONE as being 215.35% over the two-year period from 2009 to 2011. This number is consistent with the overall fast growth of the journal.

We identified 624 FP7 projects that were acknowledged in PLOS-journal articles. Table 4 presents the distribution over projects by its number of publications in PLOS.

The figures indicate a positively skewed distribution of FP7 contributions in PLOS with 57.53% of the FP7 projects referenced once, while 1.92% published more than 8 times with PLOS. With 30 contributions, the FP7 research project *European Network for Genetic and Genomic Epidemiology* (ENGAGE)[12] published the highest number of PLOS articles.

The OpenAIRE OAI-PMH interface provides more detailed access to FP7 funding information. At the time of our study, 17,736 FP7 research projects were exposed which were distributed over 23 funding programmes. On this basis, we determined the visibility of FP7 funding programmes in the PLOS domain (Table 5). Similar to the distribution over FP7 research projects, FP7 grant acknowledgements were unequally distributed over funding programmes which coheres PLOS' focus on biomedical research and related fields: We found that 27.96% of the projects funded within the research programme *Health Re-*

Table 3

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

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

Table 4

PLOS contributions by EC funded research projects

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

[12] ENGAGE project information, http://cordis.europa.eu/projects/201413.

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

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	17,736	624	3.52

search (HEALTH) published at least once in PLOS, but we could not detect references to eight funding programmes, e.g. *Transport (including Aeronautics)* (TPT).

We also took projects under the SC39 clause into account and revealed that the proportion of SC39 funded research in the PLOS domain (94 out of 624 FP7 projects, 7.34%) was higher than the FP7 funding scheme share (530 out of 17,736 FP7 projects, 3.22%).

3.2. Article-level metrics

The PLOS ALM API provides information on citations, usage and dedicated social media events. Figure 1 shows the coverage of FP7 articles by ALM source field 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: between 43% (pubmed) and 63% (crossref) of the articles were cited. Social media services mentioned between 8% (comments on PLOS articles) and 81% (Mendeley readerships) articles granting FP7 support. Note that the collection of Twitter mentions within PLOS ALM started on June 1st, 2012.

Data from the PLOS ALM API furthermore allows us to compare the occurrences of ALM event types for every day since publication. Figure 2 depicts article age (in days since publication) and total views on the PLOS website. As a third variable we compared Scopus citations and Facebook shares

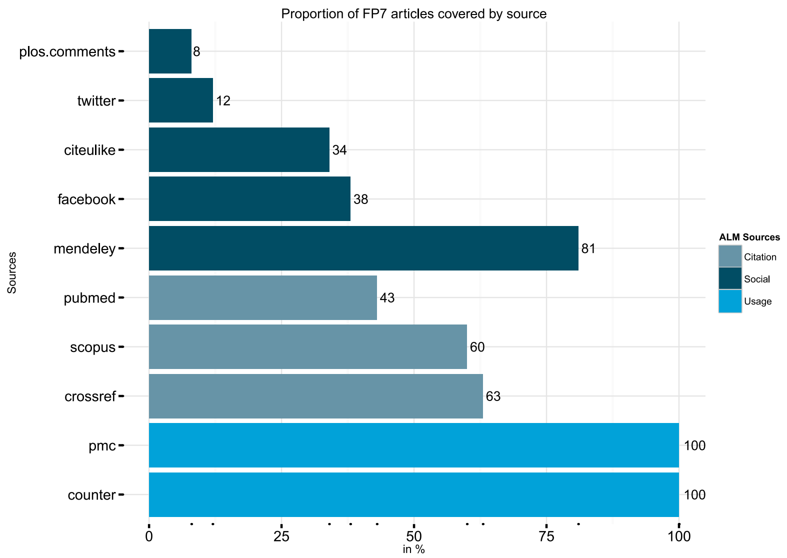
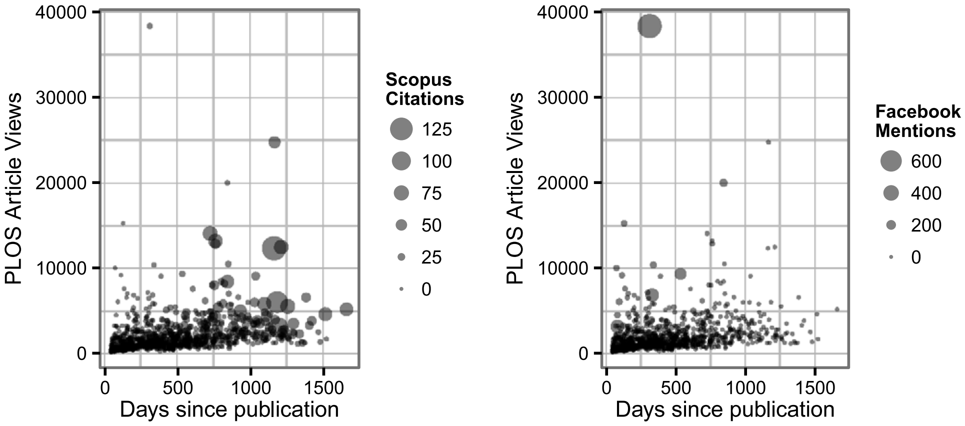


Fig. 1. PLOS proportion of FP7 articles covered by ALM source field. (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/ISU-130704.)



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

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

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. With 38,361 total views, the following article resulting from FP7 funded *Forecasting Financial Crisis* (FOC-II)[13] was most represented in our sample:

> 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)

[13] FOC-II project information, http://cordis.europa.eu/projects/255987.

However, while the article had received the highest number of Facebook shares in our study, the paper was only ranked in the 87th percentile according to Scopus citation count (10 times cited). A possible explanation gives a blog post from the FOC project. The project members describe that the article has gained broad media attention. However, the authors claim that media have misinterpreted their findings.[14]

3.3. Collaboration patterns

Our dataset can also be used to explore collaborations patterns between FP7 research projects by joint PLOS publications. We found that 9.52% of PLOS publications under consideration acknowledged more than one FP7 project (Table 6). From 624 FP7 projects identified, 26.28% were acknowledged together with another FP7 projects. Therefore, the relationships between projects and articles was furthermore explored as a scientific collaboration network (Fig. 3).

Figure 4 visualises the links between FP7 research projects that contributed to at least one PLOS publication together. In this figure, edge width represents the number of joint PLOS contributions. In total, we detected 57 components that consist of 164 projects. The FP7 project 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).

Lastly, we used the PLOS Search API to author affiliations. In total, we obtained 6090 author addresses distributed over 1166 publications. 6049 correctly formatted country names could be extracted

Table 6

FP7 funding acknowledgement per PLOS article

FP7 Projects per paper	Frequency	Relative frequency (in %)
1	1055	90.48
2	93	7.98
3	14	1.20
4	4	0.34
$\sum$	1166	100

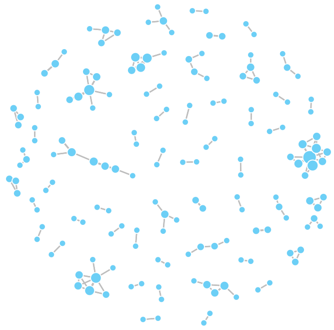


Fig. 3. Collaboration network of FP7 projects in PLOS journals. Edge width show the number of joint articles of a collaborating institutional pair, vertices size represent the degree centrality, i.e. the number of articles in PLOS. (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/ISU-130704.)

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

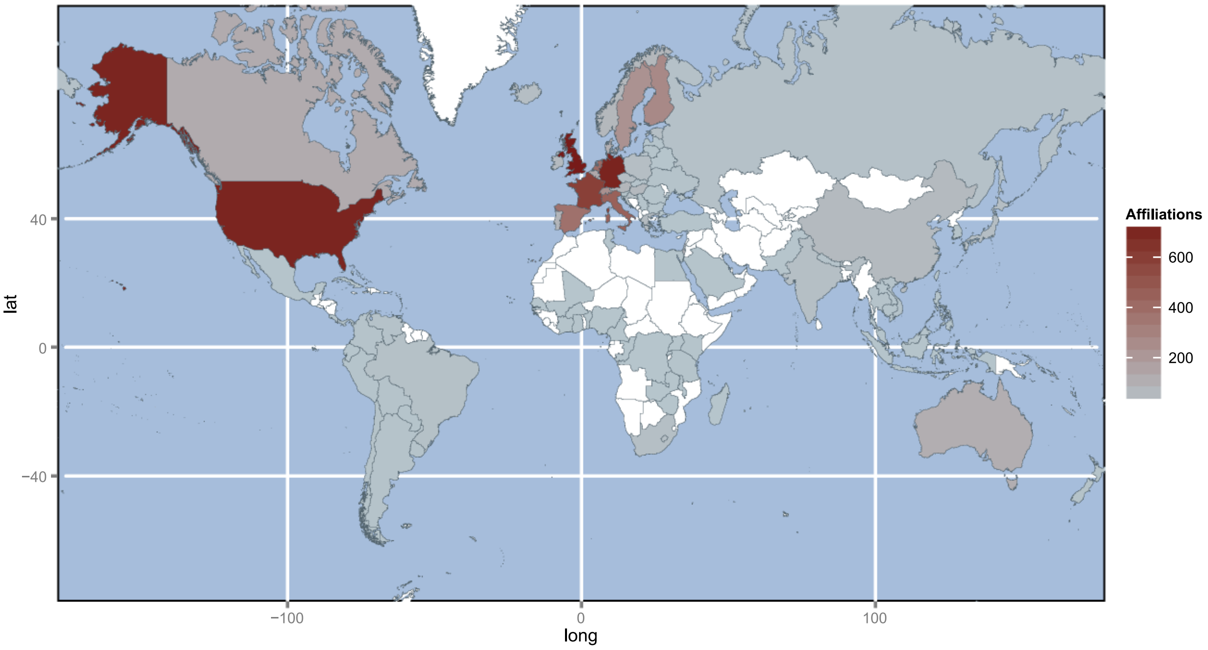


Fig. 4. Affiliation world map of FP7 contributions in PLOS journals. (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/ISU-130704.)

that are distributed over 95 countries (Fig. 4). 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 and conclusion

This case study presents first findings of plosOpenR, a software package for the statistical computing environment R. plosOpenR allows the exploration of grant-supported research publication in PLOS journals. We report 1166 articles that acknowledge 624 different research projects funded by the European Union 7th Framework Programme (FP7). Additionally, our case study presents metrics such as citations, usage and social media activity as well as collaboration structures between FP7 projects visible in joint publications.

From a research infrastructure and services point of view, our results highlight the importance of openly available research services. plosOpenR combines data sources from the Open Access publisher PLOS and OpenAIRE as well as reuses existing software packages from rOpenSci. While PLOS provides detailed information on articles, OpenAIRE enables a distinct view on FP7 funded research published in PLOS journals by exposing the relevant project information. Furthermore, we demonstrated the aggregation of alternative science metrics by the PLOS ALM API. For all research contributions under

[15] World Bank regions, http://www.worldbank.org/countries.

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).

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 publications in PLOS journals are to be expected in future. Secondly, the extreme positive skewness of most distribution under considerations demands careful analysis and interpretation. Especially, it has to be noted that our findings only partially cover all FP7 funding projects and programmes due to the disciplinary scope of PLOS.

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. In addition, 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, plosOpenR provides tools for easy on-time exploration of PLOS ALM in order to identify irregular patterns and motivate qualitative investigation.

Future work and studies on PLOS ALM will focus on main problem areas [1]. 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.

Acknowledgements

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] ALM workshop 2012 report, PLOS ALM, 2012, figshare, available at: http://dx.doi.org/10.6084/m9.figshare.98828.

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

[3] C. Boettiger, D.T. Lang and P.C. Wainwright, rfishbase: exploring, manipulating and visualizing FishBase data from R, *Journal of Fish Biology* **81**(6) (2012), 2030–2039.

[4] P. Manghi et al., OpenAIRE – data model specification, 2010, available at: http://puma.isti.cnr.it/dfdownload.php?ident=/cnr.isti/2010-EC-016.

[5] P. Manghi, L. Bolikowski, N. Manold, J. Schirrwagen and T. Smith, OpenAIREplus: The European scholarly communication data infrastructure, *D-Lib Magazine* **18**(9/10) (2012), available at: http://dx.doi.org/10.1045/september2012-manghi.

[6] M.E.J. Newman, The structure of scientific collaboration networks, *Proceedings of the National Academy of Sciences* **98**(2) (2001), 404–409.

[7] OpenAIRE, OpenAIRE guidelines for usage statistics v1.0, 2011, available at: http://www.openaire.eu/en/about-openaire/publications-presentations/publications/doc_ details/314-openaire-guidelines-for-usage-statistics-v10.

[8] J. Priem, P. Groth and D. Taraborelli, The altmetrics collection, *PLoS ONE* **7**(11) (2012), e48753.

[9] R Core Team, *R: A Language and Environment for Statistical Computing*, Vienna, Austria, 2012, available at: http://www.R-project.org.