Open Data engages Citation and Reuse: A Follow-up Study on Enhanced Publication

Dominic Farace, Jerry Frantzen, Joachim Schöpfel

Open Data engages Citation and Reuse:
A Follow-up Study on Enhanced Publication

Dominic Farace and Jerry Frantzen, GreyNet International, Netherlands
Joachim Schöpfel, University of Lille, France

Introduction

In 2011, GreyNet embarked on an Enhanced Publications Project (EPP) in order to link its collection of full text conference papers with accompanying research data. The initial phase in the study dealt with the design and implementation of an online questionnaire among authors, who were published in the International Conference Series on Grey Literature[1]. From 2012 onwards, subsequent phases in the project dealt with the acquisition, submission, indexing, and archiving of GreyNet's collection of published datasets now housed in the DANS EASY[2] data archive.

In 2017, GreyNet's Enhanced Publications Project was further broadened to include a Data Papers Project[3], where emphasis focused on describing the data and methods applied in gathering it rather than analyzing it. As such, the data paper signals data sharing and in this way promotes both data citation and the potential reuse of research data including any limitations on its potential reuse. This is in line with the FAIR Guiding Principles[4] for scientific data management and stewardship.

Available results from the Data Papers Project presented last year at GL19 concludes where this study commences. Here, we now seek to demonstrate the reuse of survey data collected in 2011 combined with survey data that was newly collected via an online questionnaire carried in May/June of 2018. The results of this study were expected to demonstrate an increased willingness among GreyNet authors to share their research data – this in part due to GreyNet's program of enhanced publication embedded in its workflow over the past seven years. The study sought to provide an example of the reuse and further comparison of the results of survey data, which can be incorporated in GreyNet's program of training and instruction. However, statistics on data citation and referencing are less likely expected to provide as yet indicative results.

Enhanced Publication

In 2011 the guiding principle for enhanced publications was that they inherently contribute to the review process of grey literature as well as the replication of research and improved visibility of research results in the scholarly communication chain. However, in 2018 and in light of the FAIR Principles there is one modification. Emphasis moves from replication to reuse. Further, GreyNet's original project on enhanced publications in 2011 as with this follow-up study in 2018 is intended not to be seen solely as a descriptive case study, but rather as a use case – one which can serve other communities of practice in the field of grey literature.

This study may no doubt be of interest to those who contributed over the years their full texts, research data, and metadata to GreyNet's collection of enhanced publications. However, to show the value of this use case beyond our own community, we are now required to demonstrate how GreyNet's open data engages citation and reuse.

FAIR Data Principles

In follow-up to the initial implementation of the Enhanced Publications Project in 2012, emphasis now has come to rest on the publication process of grey literature in which the individual enhanced publication becomes the beneficiary.

Two current developments in the field of information have considerable impact on grey literature – one being the FAIR Data principles in which data is to be findable, accessible, interoperable, and reusable; and the other, which deals with the publication of data papers defined as "Scholarly publications of a searchable metadata document describing a particular online accessible dataset or a group of datasets published in accordance to standard academic practices. As such, data papers represent a scholarly communication approach to data sharing"[5]. It is by way of a data paper that the FAIR Data principles are implemented. The FAIR Data Principles formulated in 2014 are related to the data and/or datasets deposited in archives. Prior to FAIR was the Data Seal of Approval[6], which was conferred upon the archive and not the individual data or dataset.

Now, in demonstrating how the data from GreyNet's collection of conference papers are findable, we can say that they have been deposited and are preserved in a national data archive. In demonstrating that GreyNet's research data is openly accessible, we can point out that the creators have waived their rights via Creative Commons Zero (CC0) thus allowing optimal access. And, in demonstrating that GreyNet's data are interoperable, we can refer to the rich metadata attributed and linked to the data and datasets including ORCiD and DOI persistent identifiers. However, to demonstrate the potential for how GreyNet's deposited data and datasets can be reusable, research was needed. And this project was born.

Author Survey on Open Data

This study sought to demonstrate the reuse of survey data collected in 2011 combined with survey data that was newly collected via an online questionnaire. In order to do so, a selection of questions from the 2011 Survey was joined with newly formulated questions in constructing the 2018 Questionnaire**.**

Survey Population

The selection used to define the population of the 2018 survey is much in line with the selection used in 2011. It was assumed that in this way that the data collected would allow for insight into changing attitudes and practices within GreyNet's research community and as such would be of more interest to other grey literature communities.

Survey Population	First Authors	Survey Recipients	Survey Respondents	Survey Results %
2011	162	95	50	52,6%
2018	115	94	44	46,8%

As shown in the chart above, the population of the 2018 survey was selected from among 115 first authors in the International Conference Series on Grey Literature. The selection comprised the respondents to GreyNet's 2011 Survey on Enhanced Publications along with first authors in the GL-Conference series from 2012 to 2018. Once the survey population was further reduced, either because an author's email address was currently unavailable, the author had retired or had since

moved to another field, the questionnaire was then sent out to the remaining 94 authors/researchers via personalized emails. The final results of this study rest on the responses of 44 survey respondents, which accounts for a near 47% response rate.

Survey Questionnaire

The data collected based on the responses of those forty-four authors/researchers was to ten questions, two of which were open-ended. Six of the questions were taken from the questionnaire carried out in 2011, which dealt with the author's own empirical research data, its availability, the formats in which it appears, and the author's willingness to archive it and make it openly accessible. The four additional questions deal with the respondent's citation and reference to data, their use of data journals in carrying out search and retrieval, and whether they (co)authored a data paper or data article. The research data – long tail[7] in contrast to big data – was collected via SurveyMonkey between May 18th and June 15th 2018, where it remains stored along with a copy in .ods format[8] in the DANS EASY Archive.

Comparison of Survey Results 2011-2018

In the following tables, responses to six of the 2018 survey questions are shown compared with the responses to the same questions that appeared in the 2011 questionnaire. It should be noted that the numerical order of the questions follows that of the 2018 questionnaire.

Q1. Does one or more of your conference papers in the GL-Series base its findings on empirical or statistical data?

2011	Question 1	2018	Question 1
Yes	60%	Yes	70%
No	40%	No	30%
	100%		100%

Q2. If so, would these data and/or datasets still be available in part or whole for archiving purposes?

2011	Question 2	2018	Question 2
Yes	54,1%	Yes	60%
No	45,9%	No	20%
Other	0%	Other	20%
	100%		100%

Q3. Would you be willing to submit data, datasets, or subsets to DANS that would in turn be linked to their existing metadata records?

2011	Question 5	2018	Question 3
Yes	48,9%	Yes	51,2%
No	6,7%	No	9,3%
	Uncertain		**Other**
	44,4%		39,5%
	100%		100%

Q4. If so, would you prefer that GreyNet entered your (retrospective) data and/or datasets in DANS, or would you prefer to do this directly?

2011	Question 6	2018	Question 4
GreyNet	44,7%	GreyNet	41,5%
Self	18,4%	Self	41,5%
	No Preference		**Other**
	36,9%		17%
	100%		100%

Q5. What kind of data and data formats have you used/are using in your research?

2011	Question 9	2018	Question 5
	Specific		**Specific**
	44%		75,9%
	General		**General**
	38%		18,9%
	N/A		**N/A**
	18%		5,2%
	100%		100%

Q10. Please enter your name, email address, and any other comments or recommendations that would be of benefit to this survey

2011	Question 10	2018	Question 10
Answered	84%	Answered	73%
Skipped	16%	Skipped	27%
Total 50 Respondents	100%	Total 44 Respondents	100%

A comparison of the results of the six questions that were repeated in the two questionnaires held seven years apart indicate that papers in the GL-Conference Series have since increased in the amount of empirical data they contain and that these data are available for archiving purposes. However, the results indicate that only a small increase is shown in the number of respondents willing to submit their data for archiving purposes. A marked increase shows that the respondents would prefer to archive their own data rather than that GreyNet do so on their behalf. The kinds of data formats, which the respondents use in their research has significantly increased since the earlier survey. And, finally it can be observed that the number of respondents willing to provide contact details has decreased since the 2011 questionnaire.

Results based on Responses to the new Questions in the Survey

Question 6

Have you ever cited or referenced data(sets) in one or more of your publications?

Answered: 43 | Skipped: 1
Yes 44%
No 44%
Other 12%

Question 7

When doing research, have you or a colleague ever reused data

Answered: 40 | Skipped: 4
Yes 50%
No 50%
Other -

Question 8

Do you at times include data papers or data journals in your browse and search strategy?

Answered: 43 | Skipped: 1
Yes 39%
No 56%
Other 5%

Question 9

Have you ever cited or (co)authored a data paper or data article?

Answered: 42 | Skipped: 2
Yes 24%
No 67%
Other 9%

The results of the two questions dealing with the citation and reuse of data by the respondents clearly indicate an even yes-no response rate to both questions. However, the responses to the questions as to whether data papers are included in the authors browse and search strategy and whether they have authored or coauthored a data paper/article are significantly non-affirmative.

Analysis of the Survey Data 2011-2018

An analysis of the survey data demonstrate significant change in the responses to three of the questions in 2011 compared with the same questions repeated in 2018. More data and/or datasets are available in part or whole for archiving purposes 54% → 60% (p =.005), more authors prefer entering their data and/or datasets directly in the DANS Archive 18% → 42% (p =.05), and the specificity of the data formats listed by the respondents increased significantly 44% → 76% (p=.1). As to the other questions repeated in the survey, little or no change was observed except in the final open-ended question requesting contact details.

Some Conclusions and Further Comments

This follow-up study demonstrates a community of practice moving further to open data. There is evidence of more data awareness and data literacy among GreyNet's authors and researchers. However, one must not overlook the fact that not all papers in the GL-Conference Series are based on empirical data and not all data can be shared for reasons of confidentiality, embargo, licensing, (lack of) policy directives, or sheer hesitance by the author/researcher.

Furthermore, this study demonstrates that research data can be published prior to the research paper, allowing for more immediate citation, reuse, and usage statistics. Also, a data paper[9] focusing on the research data that includes persistent identifiers such as ORCiDs, DOI's and other hyperlinks, can further add to the increase in citation, reuse, and usage statistics. Finally, the results of this study have already been incorporated in GreyNet's series of workshops[10] and training on data and data papers offered both within and outside its community of practice.

References

[1] Farace, D. et al. (2012). Linking full-text Grey Literature to underlying research and post-publication data: An Enhanced Publications Project 2011-2012. – In: The Grey Journal, Volume 8, Issue 3, 2012. – pp. 181-189. – ISSN 1574-1796

[2] GreyNet's collection of published datasets in the DANS EASY data archive, https://easy.dans.knaw.nl/ui/?wicket:bookmarkablePage=:nl.knaw.dans.easy.web.search.pages.PublicSearchResultPage&q=greynet

[3] Farace, D., Frantzen, J. and Smith, P.L. (2018). Data Papers are Witness to Trusted Resources in Grey Literature: A Project Use Case. – In: The Grey Journal, Volume 14, Issue 1, 2018. – pp. 31-36. – ISSN 1574-1796

[4] FAIR-Data Principles https://www.force11.org/group/fairgroup/fairprinciples

[5] https://en.wikipedia.org/wiki/Data_publishing#Paper

[6] https://www.datasealofapproval.org/en/

[7] Long tail of research data https://www.radar-projekt.org/display/RE/Glossar#Glossar-Longtailofresearchdata

[8] https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:110917/tab/2

[9] Farace, D. and Schöpfel, J. (2018). Data from "Open Data engages Citation and Reuse: A Follow-up Study on Enhanced Publication". – In: The Grey Journal, Volume 14, Issue 3, 2018. – pp. 149-150. – ISSN 1574-1796

[10] http://greynet.org/greyforumseries.html