Linking full-text grey literature to underlying research and post-publication data: An Enhanced Publications Project 2011-2012

Dominic J. Farace and Jerry Frantzen, GreyNet International, Netherlands;
Christiane Stock, INIST-CNRS, France
Laurents Sesink, DANS, Data Archiving and Networked Services, Netherlands
Debbie Rabina, Pratt Institute, School of Information and Library Science, United States

Abstract

This project seeks to circumvent the data vs. documents camp in the grey literature community by way of a middle ground provided through enhanced publications. Enhanced publications allow for a fuller understanding of the process in which data and information are used and applied in the generation of knowledge. The enhanced publication of grey literature precludes the idea of a random selection of data and information, and instead focuses on the human intervention in data-rich environments. The definition of an enhanced publication is borrowed from the DRIVER-II project, "a publication that is enhanced with three categories of information: research data, extra materials, and post-publication data". Enhanced publications combine textual resources i.e. documents intended to be read by human beings, which contain an interpretation or analysis of primary data. Enhanced publications 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.

Introduction

This enhanced publications project seeks to circumvent the data vs. documents camp in the grey literature community by way of a middle ground. Enhanced publications allow for a fuller understanding of the process in which data and information are used and applied in the generation of knowledge. The enhanced publication of grey literature precludes the idea of a random selection of data and information, and instead focuses on the human intervention in data-rich environments. The definition of an enhanced publication is borrowed from the DRIVER-II project, "a publication that is enhanced with three categories of information: 1) research data, 2) extra materials, and 3) post-publication data".[1] Enhanced publications combine textual resources i.e. documents intended to be read by human beings, containing an interpretation or analysis of primary data. Enhanced publications 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.

The goal of this project is threefold: to enhance GreyNet's existing collection of conference preprints by adding corresponding links to research data, to include commentaries i.e. post-publication data on GreyNet's existing conference preprints in metadata records, and to establish a workflow for future GreyNet enhanced publications based on the results of this project, where permanent access to full-texts and their data enriched components are made available via persistent identifiers accessible in trusted repositories.

Each of the four partnering organizations is brought together based on their expertise and tasks they will execute during the course of the project. GreyNet works together with INIST-CNRS to devise a questionnaire and carry-out a survey among its author base in the acquisition of research data linked to conference preprints. GreyNet will

facilitate data entry in the DANS Easy Repository with link backs to corresponding metadata records in the OpenGrey Repository. And, GreyNet cooperates with Pratt Institute to establish basic criteria upon which commentaries by LIS students will be compiled and added to existing metadata records.

Not all grey literature is based on research data, and this holds for GreyNet's collection of conference preprints. While it is anticipated that GreyNet's contributing authors will be inclined to make their research data available, some data from previous years will not be retrievable. On the other hand, since student commentaries are related to academic credit, this portion of the project will harvest optimal results. And, the combined results of the project will contribute to a future workflow not only for GreyNet's enhanced publications but also for other grey literature communities.

A research project in progress

The international grey literature community is both diverse in scale and foci of research. The HEP (High Energy Physics) community[2] depicted at CERN in Geneva and the Karst communities such as the one at Florida State University[3] are two such examples. In fact, GreyNet's own community[4] of researchers in the field library and information science is yet another example – one in which this study is focused.

This project carries with it the connotation of repurposing grey literature in that it not only has a retrospective element but seeks also to enrich already existing records in the collection. Through the years, GreyNet has been involved in providing retrospective access to its collections of conference preprints, which first started with bibliographic records in the SIGLE database, later their full-text records in the OpenSIGLE repository, and now in the newly launched OpenGrey Repository. The notion of enhancing these metadata records with related research data as well as post-publication data such as commentaries explains what is meant by 'repurposing grey literature'.

The Enhanced Publications Project (EPP) is being carried out in six phases, some of which overlap depending on the partnering organizations responsible for their implementation. The initial three phases comprise the first part of the project carried out in 2011 and the final three phases account for the second part of the project that will be carried out in 2012:

1. Project Proposal and Formation of the Team
2. Design of the Questionnaire and Author Survey
3. Criteria for commentaries and the selection of eligible conference preprints
4. Acquisition and Submission of research data
5. Data upload and cross-linking
6. Draft of enhanced publication guidelines and the design of a future workflow

Phase 1: Project Proposal and Formation of the Team

The project proposal was sparked by a paper written by Carroll el al[5] delivered at the Twelfth International Conference on Grey Literature entitled Scientific Data: Increasing Transparency and Reducing the Grey. In order to carry out the project, GreyNet would need to team-up with other partners and to this end, three organizations were contacted. INIST-CNRS who is the data provider for the OpenGrey Repository[6] in which GreyNet's publications are housed. The DANS Easy Repository[7] that would house the research data with crosslinks to corresponding the metadata records in the OpenGrey Repository. And, Pratt Institute's School of Information and Library Science[8] that would involve LIS graduate students in writing-up the commentaries of existing full-text papers in GreyNet's collection.

Phase 2: Design of the Questionnaire and Author Survey

The population of the survey was selected from among the 286 authors and co-authors in the International Conference Series on Grey Literature. It was decided that only first authors would receive the questionnaire, which narrowed the potential population of the survey to 162 authors of which only 95 were actually sent the online questionnaire. The reason the other 67 first authors were not included in the final survey population was due to a number of factors such as no current email address, retired, deceased, etc.

Number of (co)authors in the GL-Series	Number of first authors in the GL-Series	Number of EPP Survey Recipients	Number of EPP Survey Respondents
286	**162**	**95**	**50**

The 95 authors were sent a personalized email with a standardized text inviting them to participate in the survey by completing the online questionnaire. The survey was carried out using the freeware 'Survey Monkey' and the questionnaire contained 10 items, three of which were open-ended. Subheadings were also inserted in the questionnaire set off by quotation marks. These subheadings preceded each odd numbered question and were deemed relevant in achieving informed responses. The final results are based on the response of 50 of the 95 survey recipients, which amounts to roughly a 53% response rate.

Survey Results

While maintaining the anonymity of the individual respondent, it was possible to determine their geographic region (see Table 1, below). This was based on the response to item ten on the questionnaire in which the respondent was asked to enter his/her name and email address.

Table 1: Geographic Region of Respondents

	Percentage	Number
Asia	14,3%	6
Europe	45,3%	19
North America	33,3%	14
Non-Applicable	7,1%	3
	100,0%	42

While only 42 of the 50 survey respondents completed this item on the questionnaire, we were able to determine that authors from some 14 countries responded to the survey, and that the difference in the number of male and female respondents is negligible.

Table 2: Affiliation to the GL-Conference Series

	Percentage	Number
< 1 Year	42,9%	18
2 - 4 Years	16,7%	7
5 > Years	33,3%	14
Non-Applicable	7,1%	3
	100%	42

The respondents to item ten could also be classified into three groups based on their prior affiliation to the GL-Conference Series (see Table 2, above). Results show that about 43% were involved over the past year, about 17% were involved in the conference series 2 to 4 years ago, and some 33% of the respondents were involved as authors in the conference series 5 or more years ago. It was quite interesting to find that twice as many authors who contributed 5 or more years ago to the conference series chose to respond to the survey than authors distanced 2-4 years from the conference series.

"Data exchange is becoming the norm in open access communities"

Question 1: Does one or more of your conference papers in the GL-Series base its findings on empirical or statistical data?

	Percentage	Number
Yes	60%	30
No	40%	20
	100%	50

While it is generally known that not all of the conference papers in the GL-Series are based on statistical data, it still has yet to be determined what percentage of the collection is. However, the fact that 60% of the respondents' state that their work relies on empirical research data provides a clear indication of the relevance that this could have for the project.

Question 2: If so, would these data and/or datasets still be available in part or whole for archiving purposes?

	Percentage	Number
Yes	54,1%	20
No	45,9%	17
	100%	37

Some 54% of the respondents to this question indicate that they still have research data available from their former studies. This is no doubt indicative of the value they place on their work.

"A data policy should be in place within research communities and organizations"

Question 3: Are you aware of any existing data archives or data initiatives in your country related to grey literature or other scientific publications?

	Percentage	Number
Yes	43,5%	20
No	56,5%	26
	100%	46

The response to this question shows that over 56% of the respondents are unaware of data initiatives related to grey literature in their own country. Unfortunately, we do not yet have cross tabulations by country. The incentive to further explore this finding takes on the implication that our project could contribute to an increased awareness of such initiatives within the grey literature community. It is also safe to assume that many if not most repositories are insufficiently robust to house and store statistical and other research data - this being the case with the OpenGrey repository. And, it is for this reason that GreyNet sought to partner with DANS in carrying out that phase of the project.

Question 4: If so, please provide the name(s) and corresponding URL(s) here?

	Percentage	Number
Specific	72,2%	13
General	11,1%	2
Non-Applicable	16,7%	3
	100%	18

Responses to this open-ended question were categorized into one of three clusters depending on how specific, general, or non-applicable the responses were. It can be noted that the number of respondents to this question was significantly less than to other items on the questionnaire. Nevertheless, if we look at the number of those who did provide specific names and/or URLs, we are able to compile a short list of archives housing research data on grey literature as show below.

ADP (SI), DANS (NL), IATUL (INT), ICPSR (US), IOP (INT)	ISS (IT), METIS (NL), Morphbank (US), NASA (US), NIH (US)	NOAA (US), NSF (US), NSIDC (US), NUSL (CZ), ORNL (US)	PLEIADI (IT), SIDR (FR), SYNABA (PL), TRAIL/CRL (US),

Of the 19 archives/portals named above, 8 are European, 9 are North American, and 2 are listed as international.

"Data counts as science output and should be recognized in references and citations"

Question 5: Would you be willing to submit data, datasets, or subsets to DANS (Data Archiving and Networked Services) that would in turn be linked to their existing metadata records in the OpenGrey Repository?

	Percentage	Number
Yes	48,9%	22
No	6,7%	3
Uncertain	44,4%	20
	100%	45

Almost half of the respondents appear willing to submit their data to the DANS Archive, while 44% express some uncertainty. The results of this question prove challenging for our project in that it will first be necessary to address the reasons for the respondents' hesitance, before embarking on a campaign to solicit the research data. Perhaps, by underscoring the advantages authors gain through increased referencing and citing of their work, as well as by providing them with ready guidelines for data entry, these authors/researchers would be more than willing to contribute their data to the project.

Question 6: If so, would you prefer that GreyNet entered your retrospective data and/or datasets in DANS, or would you prefer to do this directly?

	Percentage	Number
GreyNet	44,7%	17
Myself	18,4%	7
No Preference	36,9%	14
	100%	38

The 'no preference' response to this question can be interpreted not only as encouraging for acquiring retrospective data but also for the acquisition of ongoing research data that would be added to the DANS Repository and cross-linked to the corresponding metadata records in OpenGrey. While the initial scope of this project is geared to retrospective input, it is to be understood that empirical and statistical data underlying future conference preprints will be directly entered in the DANS Repository by authors themselves.

"Research data should be preserved and accessible in order to enhance scholarly communication"

Question 7: Do you agree that both the data producer as well as the data user stand to benefit by submitting data, datasets, or subsets for this Enhanced Publications Project?

	Percentage	Number
Yes	84,4%	38
No	0%	0
Uncertain	15,6%	7
	100%	45

Over 84% of the respondents agree that both the researcher/author as well as the potential data user would stand to gain from the enhancement of conference preprints with related data, datasets, and/or subsets. While it is long understood that researchers and authors are at the same time information users, and while it has been demonstrated that specific communities of researchers are more likely to first use the sources/resources produced within their own community before searching beyond[9], we are now seeing a new development in which wider audiences (i.e. net users) now have open access to underlying research data, whereby this type of grey literature becomes more transparent and accessible worldwide.

Question 8: Do you think that guidelines for data entry should be available for future conference papers and other types of grey literature?

	Percentage	Number
Yes	91,1%	41
No	0%	0
Uncertain	8,9%	4
	100%	45

Over 90% of the respondents to this question are of the opinion that a set of guidelines should be made available for data entry. And, such guidelines will be addressed in the final phase of this Enhanced Publications Project.

"Data is disciplinary or subject based and this accounts for differences in formats used to acquire it"

Question 9: What kind of data and data formats have you used/are using in your research?

	Percentage	Number
Specific	44%	15
General	38%	13
Non-Applicable	18%	6
	100%	34

Responses to this open-ended question were categorized into one of three clusters depending on how specific, general, or non-applicable the responses were. If for example a respondent replied that they were 'no longer engaged in research' or 'not engaged in research at the moment', then such responses were categorized as non-applicable. If the respondents replied to this question in broad terms such as tables, charts, times series, etc., these responses were categorized as general. And, those respondents who actually named particular software or statistical packages such as SPSS, Mekov, Excel, Minitab, MS Access, etc. were identified as having provided a specific response.

Question 10: Please enter your name, email address, and any other comments or recommendations for this Enhanced Publications Project?

As indicated earlier in the paper, the response to this open-ended question allowed us to specify the geographic region of the author as well as his/her affiliation over time in the GL-Conference Series. Comments by four of the survey respondents are of particular interest in that they are assumed to be shared by other authors/researchers. These comments are recorded as follows:

> **"**I'm a firm believer that not all data is worth archiving.**"**
> **"**Will your system support the preservation and migration to new platforms?**"**
> **"**For many, it would need to be a local activity linked to our own sites.**"**
> **"**I share data with my colleagues and research teams, but I'm not sure if I would be willing to share them with anybody else at the moment?**"**

It is believed that the tenor of these comments have much to do with the uncertainty expressed in response to Question 5, and must likewise be kept in mind during the acquisition phase of this project.

Phase 3: Criteria for commentaries and the selection of eligible conference preprints

Graduate students from Pratt Institute's School of Information and Library Science were engaged in the selection of GreyNet's conference preprints currently accessible via the OpenGrey Repository. In their initial selection and review of the full-texts, they first determined whether the content would be of value in the research chain and whether the manuscripts were clearly written given the fact that English is not always the first language of the authors in the International Conference Series on Grey Literature.

Standardized Format

In order to facilitate their work, the students developed a standardized format used in drafting the commentaries. Each commentary comprises a brief summary, the strengths of the research, any noticeable limitations, and the takeaway *i.e.* what the student considered the most salient aspect in the study. A sample commentary is provided below.

http://hdl.handle.net/10068/697760 GL8_Anderson_et_al_2007_Commentary_by_Pratt_Institute.pdf

Harnessing NASA Goddard's Grey Literature: The Power of a Repository Framework

Summary:
In an organization like NASA, where researchers come and go depending on the project, it is especially important to have a central repository for the valuable scientific grey literature that is produced during the lifetime of any given mission. This paper details the steps that went into NASA Goddard Library's development of a Digital Asset System (DAS) capable of managing the wide range of items (multimedia objects as well as text documents) and multiple vocabularies used in its various projects over its long history. The authors developed their own extension of a Dublin Core Metadata scheme, and also created oral and video histories of many projects that had concluded.

Strengths:
- The description of the process the authors used to create and improve upon the DAS is clear and thorough.

Limitations:
- The paper might benefit from more in the way of concrete examples.

Takeaway:
This is a fascinating look at the development of a repository for a fabled government agency, the problems that arose along the way, and the solutions the authors devised. It is a valuable working document for anyone undertaking the creation of a digital repository for grey literature.

Reviewer Eloise Flood, Pratt SILS 2011

Currently 205 commentaries on GreyNet's conference preprints now are available in the OpenGrey Repository. This covers 79% of the total collection. More on this phase of the project was presented by a group of Pratt students during the GL13 Conference Poster Session.[10]

Project Continuation

The final three phases of this project will be undertaken in 2012 and involve the acquisition of research data and their subsequent upload and cross-linking between the DANS Repository and the OpenGrey Repository. This will allow for open access to research data linked to underlying full-texts, to extra materials such as PowerPoints, abstracts, and biographical notes already available via the metadata records, as well as the post-publication commentaries. A set of guidelines will be drafted and used in the future workflow of GreyNet's enhanced publications. In this last phase of the project, existing guidelines for the re-use, verification, and preservation of data[11] will be employed. Furthermore, in line with GreyNet's policy on information access and retrieval, neither the submission of the research data nor the format will be mandated.

In Close

Final conclusions will have to await completion of the enhanced publications project. However, it is worthwhile to summarize the results of the first half pertaining to the survey and post-publication data. While it was known from the start of the project that not all of the conference preprints in GreyNet's Collection were based on research data, it was encouraging that 50 of the 95 authors surveyed completed the questionnaire, 30 of whom stated that their findings were based on empirical or statistical data, and 20 of whom still maintain these data, datasets, and/or subsets, which are still available for archiving purposes. While it was surprising to find that a little over half the respondents were unaware of existing data archives related to grey literature in their own country, it was considered worthwhile to have accumulated a short list of subject-based grey literature repositories that do house research data. This list can now be further expanded and even linked to other subject-based resources.[12]

In the second part of the questionnaire, which focused not only on retrospective data but also ongoing and future research, we found that very few respondents were unwilling to submit their data for input in the DANS Repository. However, on the other hand, nearly half voiced uncertainty. In the next phase of our project, it will be important to address the uncertainties before proceeding with a campaign for data acquisition. Finally, the overwhelming majority of respondents agree that both the data producer and user would benefit from enhanced publications, and they look forward to guidelines for future data submission.

In regard to post-publication data, it was anticipated at the start of the project that the LIS students would produce a significant number of commentaries given this phase of the project was linked to course credit. However, it was beyond expectation that almost 80% of the existing conference preprints would have been completed in the first part of our enhanced publications project. This wealth of post-publication data now provides the opportunity to assess the impact that these commentaries have on GreyNet's collection of conference preprints.

Acknowledgements

Special thanks to the students at Pratt Institute's School of Information and Library Science, who under the direction of Prof. Dr. Debbie Rabina carried out the post-publication phase of this project by enhancing GreyNet's Collection of conference preprints with commentaries. Likewise, a word of thanks to Nathalie Henrot at INIST-CNRS, who was responsible for integrating over 200 commentaries in the OpenGrey Repository.

References

[1] http://www.driver-repository.eu Enhanced Publications: State-of-the-Art (PDF), Enhanced Publications: Object Models and Functionalities (PDF)

[2] The driving and evolving role of grey literature in High-Energy Physics / Anne Gentil-Beccot http://www.reference-global.com/doi/abs/10.1515/9783598441493.2.155

[3] Grey Literature in Karst Research: The Evolution of the Karst Information Portal, KIP / Todd Chavez http://www.reference-global.com/doi/abs/10.1515/9783598441493.2.181

[4] Grey Literature Network Service http://www.greynet.org

[5] Scientific Data: Increasing Transparency and Reducing the Grey / Bonnie Carroll, June Crowe, and J.R. Candlish http://hdl.handle.net/10068/700004

[6] http://www.opengrey.eu/ OpenGrey Repository – System for Information on Grey Literature in Europe

[7] https://easy.dans.knaw.nl/ui/home DANS Easy - Data Archive and Networked Services

[8] http://www.pratt.edu/academics/information_and_library_sciences/ Pratt Institute; School of Information and Library Science, SILS

[9] See references 2 and 3 above.

[10] http://00215f8.netsolhost.com/images/GL13-PSS,_Edwards_et_al.pdf Shining a light on grey literature / Bethany Edwards, Eloise Flood, Thomas Keenan, and Ashley Rode

[11] http://www.dans.knaw.nl/en/content/categorieen/publicaties/dans-studies-digital-archiving-6 Selection of Research Data: Guidelines for appraising and selecting research data Heiko Tjalsma (DANS) and Jeroen Rombouts (3TU.Datacentrum), 2011.

[12] http://www.greynet.org/greysourceindex.html GreySource Index, A selection of classified web-based resources in grey literature