Chapter 10

The driving and evolving Role of Grey Literature in High-Energy Physics

Anne Gentil-Beccot, CERN, Switzerland

10.1 Introduction

If grey literature (GL) is often seen as a marginal part of the scientific information landscape, this is absolutely not true for high-energy physicists (HEP) who developed, decades ago, their own scientific communication scheme using this allegedly “darker” fraction of the literature. Today, grey literature remains a living and indispensable resource for this discipline. What is more, grey literature has become a driving force, motivating many evolutions in the HEP information landscape.

At the same time, publication in journals continues to be essential for the scientific community, most of the preprints being eventually published. But new challenges in scientific publication, such as Open Access publishing, are under discussion nowadays. Furthermore, the information landscape is becoming increasingly complex. In addition to the tools developed by the community, scientists can use many information products such as commercial databases or search engines like Google or Google Scholar - users can now access, easily or not, huge amounts of varied information. In addition, needs are changing, new technologies appear every day and new ways of interacting with users evolve. In such a context, it is interesting to analyse the current role of grey literature in HEP, and try to understand what the future evolution might be.

After describing in detail the pervasive role GL has taken in HEP for several decades and how the community has developed dedicated tools adapted to these specific resources, we will show that HEP scholars continue to rely on this literature for accessing information because it meets their need for fast communication. Finally, we will show how the community is continuously adapting its information tools, evolving with the needs of the users and the fast-changing technologies.

10.2 The pervasive “preprint culture”

High-energy physics aims at discovering the constituents of matter and understanding their interactions [1]. This is a small and cohesive community, counting around 30,000 scientists, with a strong collaborative spirit. Because of its special characteristics, this community has developed its own scientific communication scheme, mostly based on grey literature.

10.2.1 Some historical perspectives

Fifty years ago, long before the birth of the online world, the delay between the submission of a scientific paper and the time it reached the reader appeared unacceptable to HEP scientists, who therefore adopted preprints as their main communication channel. The community was already composed of two sub-groups who needed to communicate both internally and externally: experimental physicists working at accelerators of ever-increasing energy, regularly witnessing new discoveries during the early stages of the discipline; and theoretical physicists interpreting these results, improving their theories and suggesting new projects. It was simply out of the question to accept months of delay in communication - the average turnaround time of ideas and experimental results was no more than several months - grey literature was the solution!

For decades, theoretical physicists and experimental collaborations, wanting to disseminate their results in a way faster than the distribution of conventional scholarly publications, printed and mailed to all major HEP institutes copies of their manuscripts at the same time as submitting them to peer-reviewed journals [2,3,4]. The same institutes financially supported the dissemination of the scientific results of their researchers, and this implied high costs¹. Libraries spent also resources indexing all these preprints, working papers and reports, making them accessible to the institute’s researchers.

In a sense, this “preprint culture” in high-energy physics pioneered the “open access” distribution of scientific results. This form of “institute-pays” Open Access ensured the fastest and broadest possible dissemination of scientific results. It is worth noting that this process favoured scientists working in well off institutions. These could pay for the mass mailing and were most likely to receive copies of preprints from other scientists seeking recognition. Smaller and less well off institutions had therefore less chance to disseminate their results and become aware of the research of other scientists.

1 In 1990, CERN used to spend over 1 Million Swiss francs a year for printing and mailing expenses.

10.2.2 Development of dedicated tools

With the increasing use of the Internet, the process continued electronically and the cards could be redistributed more equally. Indeed, the community launched its own tools to manage grey literature: in 1991, even before the web was invented, Paul Ginsparg, at the Los Alamos laboratory, launched arXiv [5], the first physics preprint repository. This new tool ensured the transition from an old preprint world to a new electronic system, offering all scientists an easy and less-restricted way to access and disseminate information, by removing the cost-barrier of mass-mailing preprints all over the world.

With more than 500,000 articles, arXiv has today grown beyond the field of HEP, becoming the major repository for many other disciplines [6].

Figure 1. Fraction of articles published in the main peer-reviewed HEP journals which also appeared, in some version, on arXiv.org.

The SPIRES database [7], the first grey literature electronic catalogue [8,9], was born at SLAC (Stanford Linear Accelerator Center) HEP laboratory in Stanford, California, in 1974, and was developed in collaboration with DESY, in Hamburg, Germany and Fermilab, Chicago. It listed preprints, reports, journal articles, theses, conference talks and books, and it now contains metadata for about 760,000 HEP articles. This tool took advantage of the invention of the web, and became the first web server [10] in the U.S. In summer 1992, SPIRES linked to the arXiv for full-texts, starting a close partnership, and bringing preprints onto the web, accessible through detailed indexing including reference to the published versions, when available.

The community produces around 5,000 journal articles per year. The large majority of these articles are published in just six peer-reviewed journals from

four publishers [11]. In figure 1, we see that 90 to 100% of the articles published in these six journals are also submitted to arXiv; we see also that this situation is stable and has lasted for ten years already. In addition, it is important to say that many HEP scientists routinely upload to arXiv a revised version of their preprint, which matches the final peer-reviewed version, including the corrections introduced during the publication process.

Even in the era of electronic journals, grey literature fully retains its importance in the discipline; arXiv today contains the vast majority of preprints (in most of the cases in their peer-reviewed version) in the field, this means that almost the entire literature in the discipline is freely accessible on the web. It is important to say this situation happened without any debate or mandate, driven by the specific needs of the scientists, as we shall see in the next section.

10.3 GL enabling fast and immediate communication

Almost the whole literature produced in HEP is available on arXiv. But why is this a reality in high-energy physics, while in many other disciplines repositories are hardly filled?

10.3.1 A direct benefit for the community

A study [12] has been carried out in 2009 on the actual usage of information in the community using citations and logs analysis.

We see in Figure 2 one of the findings of this study. Articles published in two leading HEP² journals over 10 years were split into two samples, those which were submitted to arXiv (96.4% of the total) and those which were published without appearing on arXiv (3.6% of the total); we see clearly that articles submitted to arXiv begin accumulating citations long before publication. This shows that in HEP the scientific discourse happens when the literature is in its “grey” stage. Citation begins well before publication, because authors read the preprint earlier. It is worth adding that, in this graph, citations from preprints have been taken into account, this explains the peak of the top line appearing almost at the publication date, since no publication delays (either from the citing paper, or from the cited paper) are taken into account, and this also explains the very steep rise of the bottom line, as here again the citing paper, which might be in a preprint form, is citing immediately after publication of the cited paper. This demonstrates twice that scientific discussion starts much before publication in a journal.

Hence, there is an immense incentive for scientists to use grey literature: the speed of information dissemination. And, in the same way that in the 1960s scien-

2 Physical Review D (published by APS) and Journal of High Energy Physics (published by IOP/SISSA till December 2009 and by Springer/SISSA as from January 2010)

tists mass-mailed preprints to disseminate their research results as fast as possible, today, they use arXiv for the very same reason, with, obviously, much greater efficiency.

Figure 2. Average number of citations per article per month as a function of the time of the citation relative to the time of publication. Data is from 26,741 articles from the Journal of High Energy Physics and Physical Review D over the period from 1998 to 2007.

While the above citation analysis gives an understanding of the speed and manner of the scientific discourse in HEP, the analysis of click streams provided by SPIRES gives even more information about the actual reading habits of HEP scientists.

10.3.2 The predominance of grey literature even for peer-reviewed information

A survey [13] performed in 2007³ demonstrated that about 50% of HEP scientists use SPIRES for a bibliographic search. Therefore, the analysis of click streams of SPIRES users once an article has been identified and can be accessed, gives a

³ From April 30 to June 11 2007, a survey was jointly conducted by several High Energy Physics institutes: CERN (European Organization for Nuclear Research), SLAC (Stanford Linear Accelerator Center), DESY (German Electron Synchrotron) and Fermilab (Fermi National Accelerator Laboratory). The aims of the survey were to understand the users' perceptions of current HEP information systems, to assess user requirements and preferences, and to define future needs. During the survey period, 2110 answers were received.

clear representation of the reading habits of the community. SPIRES click streams collected during October 2008 have been analysed (30,000 clicks) [12]. The study was restricted to clicks that occurred from records displaying both a link to arXiv and to a publisher website. We discovered that in 82% of cases, arXiv is preferred. The survey showed also that 40% of the scientists go directly to arXiv to access information, they are therefore not counted in the data mentioned here. The advantage of arXiv over the published version might thus be much higher than what is shown by the click streams analysis.

This brings us to the fact that HEP scientists prefer to read the arXiv version of published papers, giving grey literature an even more important place. Several typical characteristics of the community help to explain this result. One of the main reasons is that, in most of the cases, the author resubmits a revised version of the preprint including the corrections brought by the peer-review process. Besides, arXiv provides free access to its content, whereas the published version on the journal website is often under subscription, restricting access. It confirms a result from the 2007 survey: respondents were asked which system they used the most when looking for preprints or published articles in different search situations. It became clear that the overall landscape does not change substantially between preprint and article searching. This is extremely relevant in the interplay between grey literature and “conventional” literature in the field: when HEP scholars need to access “conventional” literature they still use the systems that were initially conceived to index and curate the grey literature in the field! Solutions invented for grey literature are therefore mainstream in this community.

10.3.3 A redistribution of the roles in the HEP information landscape

As demonstrated, peer-reviewed journals have lost their role as providers of information and as a means of scientific communication, which has effectively moved to the grey literature. However, HEP peer-reviewed journals continue to play an indispensable role, providing independent quality control, which is necessary in this field as in the entire academic community. This situation, far from increasing the gap between GL and published literature, allows the clarification of their mutual role which leads to their separate but synergic evolution. There is no space here for further discussion of this issue, but the SCOAP3 project [14,15] aims to convert all HEP journals to Open Access, according to a model where the peer-review role of journals, rather than the dissemination, would be financed by the community.

In addition, the success of GL in HEP is due to the fact that GL in this field has taken advantage of the new technologies, adapting and shaping them, rather than resisting and retrenching. This is the topic of the next section.

10.4 Evolving needs, evolving tools

The survey performed in 2007 [13], mentioned above, aimed to assess the current usage of information resources by HEP scholars as well as their future needs. This section will discuss more specifically some of these results.

10.4.1 The predominance of “community-made” tools: today and tomorrow

Figure 3: Favourite information resources for HEP scholars.

Thanks to the association of arXiv and SPIRES, scientists have access to the whole HEP literature, either in the preprint or the published version, SPIRES providing detailed metadata and publication information, and arXiv providing full-text preprint versions of nearly all journal articles. As we mentioned in the previous section and as shown in figure 3, the survey demonstrated that arXiv and SPIRES were the two main tools used by the community. 91.4% of the respondents prefer community-based services⁴. This is not a surprising result since most of these community-based systems were created to meet the needs of HEP scholars, and these tools tend to be tailored specially to the evolving needs of the HEP community. These systems have been user-driven for decades.

On the other hand, 9% of the respondents claimed to use Google or Google Scholar. The survey also showed that the use of Google tends to increase as the

⁴ CDS (CERN Document Server)[16] and NASA ADS [17] are also tools created by the community.

age of the respondents decreases. This is a reflection of the integration into the HEP community of scholars belonging to an age group that has been exposed to Internet search before their academic career, as opposed to scholars who first used Internet search engines during their professional activity. It must be noted that Google might actually be only a gateway to other sources, as it indexes material from arXiv and SPIRES. Effectively, those using Google ultimately use some of the other information services to access the document. This result cannot therefore be seen as proof that Google is becoming more important than the community tools, it only shows the increasing need for users to have a single entry to their information resources.

In summary, the survey demonstrated that community-based systems largely dominate the landscape, even if Google takes a non-negligible part. The choice of these information systems corresponds to the need of scholars for easy access to full-text and a wide coverage of all literature in HEP, which is exactly what is currently offered by the combination arXiv.org & SPIRES.

10.4.2 Access to even more grey literature

The survey, however, showed a need from a large part of the users to have access to emerging forms of grey literature such as conference slides. Scientists go to conferences to present and discuss the latest research results. Slides shown during conference talks generated in digital format, constitute a new form of grey literature that other scientists want to access immediately, and often quote in their subsequent publications, without having to wait for a conference paper to be written and submitted as a preprint to a repository [18]. This is indeed the next frontier: capturing, storing and indexing the content. It is important since this information is not always organised: links to conference slides might get broken, as might the web sites of conferences. Thus, there is a need for a system to harvest and serve the content. Some projects aim to organize this information such as Indico [19], but not all conferences benefit from such tools and this is a new call to action for information providers. In addition, the survey underlined the need for a single interface to access all required information, whatever its type.

Another important field of improvement requested by the users is access to theses. Regarding theses, the survey shows that users access Google much more than for any other information type. This emphasizes that community-based systems do not yet cover the complete scientific information needs of the field, particularly in the area of grey literature represented by theses. While the vast majority of HEP literature can be found on arXiv and in a few major journals, conference proceedings and theses are distributed over a multitude of servers and are thus more difficult to collect. Also, commercial databases do not provide a better service in this regard. On the contrary, only Google has an advantage here since it indexes a lot of resources.

By way of conclusion, this then suggests further development of the community-based services with investment in the harvesting and preservation of theses as well as other non-peer-reviewed user-generated content.

This is once again proof of the growing importance of grey literature for the immediacy of communication in the discipline. Furthermore, many users see access to data and tables as important, yet another aspect of the future of GL in high-energy physics. Indeed, data such as tables behind graphs can also be considered a form of GL, since their publication is not yet standardised. The way these data are (will be) described, preserved and accessed constitutes the next frontier for community information systems, since they have the advantage of being close to the community, while publishers are only starting to think of how to provide these data for their users.

10.4.3 Integration of Web 2.0 technologies

Future needs imagined by users correspond primarily to a wish to have easier and wider access to content. But in a context where our daily communication channels include more and more Web 2.0 technologies, other needs appear, such as 'recommendation of articles' (almost 50% of the respondents think it is somewhat important).

Furthermore, a question in the survey tried to assess the potential for the implementation of Web 2.0 features to capture user-tagged content. We find that 63% of the respondents claimed they would be willing to spend between five minutes a day and an hour a week, showing that there is immense potential for user-tagged and user-curated content in the field of information provision in HEP. This question is essential because users need to retrieve correct and accurate information. This user-tagging and user-curating could help future information systems to provide accurate information to their users. We don't know yet how this will evolve, how far the community will use these new tools. But, this type of information, which can be considered as unstructured, will without any doubt be scientific content that will be put online and will have a value as such.

To summarize, many challenges appear when one starts analysing the wishes of the users, but they outline the best track to follow in order to build a system fully adapted to the community usage.

10.4.4 Towards a new system

The survey and discussions between the four leading HEP laboratories (CERN, DESY, Fermilab and SLAC), in synergy with other partners (notably arXiv) and in a continuous dialogue with major publishers in the field, led to the idea of the next generation HEP information system, merging the current SPIRES database and a modern platform, the Invenio open-source digital library software [20]. This

new information system, INSPIRE [21], is being developed by a collaboration of the four leading HEP laboratories mentioned above. It will integrate the content of present repositories and databases to host the entire set of metadata and the full-text of all open access publications, past and future, including conference material, becoming the single entry for the whole community to all HEP relevant material.

Grey literature has for decades been the driving information source for HEP scientists and will maintain this role even through the future evolutions of information systems.

10.5 Conclusions

With this last section, we conclude our snapshot of the HEP community scientific information practices. We saw that the community has always been anticipating and driving the evolutions. Grey literature became the most important communication channel because the community needed immediate access to information. This is still the case, and this is why the community develops new systems that make the access to information even more immediate. But needs evolve and the community must now go further: open access to full text is not enough anymore, scientists want improved access to greyer literature, such as conference proceedings, theses, or high-level data, not yet available anywhere. They need to interact more deeply with all these resources. This is one of the challenges the new HEP information system INSPIRE will meet.

Another problem that will have to be addressed is the changing role of published literature, which no longer serves any communication purpose, even if its place is still vital for other scientific reasons.

But one constant remains in all these movements, the community itself takes the lead in all these changes, and that's why it is well placed to successfully achieve these evolutions, as they are driven by no more than its own needs.

10.6 References

- [1] CERN, public webpages
<http://public.web.cern.ch/public/en/Science/Science-en.html>
 [Last visited October 06, 2009]
- [2] Goldschmidt-Clermont L. (2002), Communication Patterns in High-Energy Physics, in High Energy Physics Libraries Webzine, 6 (2002)
<http://library.web.cern.ch/library/Webzine/6/papers/1/>
 [Last visited October 06, 2009]
- [3] Heuer, R.D., Holtkamp, A., Mele S. (2008), Innovation in Scholarly Communication: Vision and Projects from High-Energy Physics, in Information Services & Use 28 (2008) 83 – 96, arXiv:0805.2739.

- [4] Aymar R. (2009), Scholarly communication in High-Energy Physics : past, present and future innovations, in *European Review* 17 (2009) 33 – 51, CERN-OPEN-2008-015.
- [5] ArXiv website
<http://www.arXiv.org>
[Last visited October 06, 2009]
- [6] Ginsparg, P. (1994), First Steps Towards Electronic Research Communication, in *Computers in Physics* 8 (1994) 390.
- [7] SPIRES Website
<http://www.slac.stanford.edu/spires>
[Last visited October 06, 2009]
- [8] Addis, L. (1994), Brief and Biased History of Preprint and Database Activities at the SLAC Library, 1962-1994
<http://www.slac.stanford.edu/spires/papers/history.html>
[Last visited October 06, 2009]
- [9] Kreitz, P.A. and Brooks T.C. (2003), Subject Access through Community Partnerships: A Case Study, in *Science & Technology Libraries*, 24 (2003) 153, arXiv:physics/0309027
- [10] Kunz, P. et al, The Early World Wide Web at SLAC
<http://www.slac.stanford.edu/history/earlyweb/history.shtml>
[Last visited October 06, 2009]
- [11] Mele, S, Dallman, D, Vigen, J., Yeomans, J. (2006), Quantitative Analysis of the Publishing Landscape in High-Energy Physics, in *Journal of High Energy Physics* 12 (2006) S01, arXiv:cs.DL/0611130
- [12] Gentil-Beccot, A., Mele, S. and Brooks, T.C. (2009), Citing and Reading Behaviours in High-Energy Physics. How a Community Stopped Worrying about Journals and Learned to Love Repositories
arXiv:0906.5418
- [13] Gentil-Beccot, A., Mele, S., Holtkamp, A., O'Connell, H. and Brooks, T.C. (2009), Information Resources in High-Energy Physics: Surveying the Present Landscape and Charting the Future Course, in *Journal of the American Society for Information Science and Technology*, 60 (2009) 150–160, 12
arXiv:cs.DL/0804.2701v2
- [14] Bianco, S. et al., Report of the SCOAP3 Working Party, ISBN 978-92-9083-292-8
<http://scoap3.org/files/Scoap3WPReport.pdf>
[Last visited October 06, 2009]
- [15] SCOAP3 website
<http://www.scoap3.org/>
[Last visited October 06, 2009]
- [16] CERN Document Server
<http://cdsweb.cern.ch>
[Last visited October 06, 2009]
- [17] NASA Astrophysical Data System
<http://adswww.harvard.edu/>

[Last visited October 06, 2009]

[18] Gentil-Beccot, A. (2008), How do High-Energy Physics scholars discover scientific information? In TGJ, 4 (2008).

[19] Indico website

<http://cdsware.cern.ch/indico/>

[Last visited October 06, 2009]

[20] CDS Invenio website

<http://cdsware.cern.ch/invenio/index.html>

[Last visited October 06, 2009]

[21] Inspire website

<http://hep-inspire.net/>

[Last visited October 06, 2009]