Observations on the Future of Grey Literature
Joachim Schöpfel

To cite this version:

HAL Id: sic_00168998
https://archivesic.ccsd.cnrs.fr/sic_00168998
Submitted on 31 Aug 2007

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Observations on the Future of Grey Literature

Joachim Schöpfel

ABSTRACT

The article provides an overview on the definition and evolution of grey literature (GL) in the emerging environment of online resources and open access to scientific and technical information. First, it gives some empirical evidence on the importance of GL in scientific publications from different domains, especially education, based on citation analysis. Other topics: the impact of Internet on the production of GL, the place of grey resources in open archives and institutional repositories, the development of bibliographic control and standardization of GL and the difficulties of identification and accessing of grey documents. The article ends with some prognostics on the future of GL and open questions for research in library and information sciences on GL.

LANDMARKS

Cartographers everywhere, even those who map the archipelagos of knowledge (Baltz, 2003), need a few landmarks to guide their sketches and explorations. There are several definitions of grey literature, the most common being the so-called ‘Luxemburg definition’, which was discussed and approved during the 3rd international conference on grey literature in 1997: “(Grey literature is) that which is produced on all levels of government, academics, business and industry in print and electronic formats, but which is not controlled by commercial publishers”.

This definition in itself contains two of the main characteristics of “grey” resources: On the one hand they are universal and ubiquitous, but on the other hand, they are difficult to identify and to obtain through conventional publishing circuits. The Luxemburg definition is also vague enough to reflect the problem of determining exactly what a type of literature variously described as «underground», «ephemeral» or «non-conventional» really means. To quote two experts from the British Library, “grey literature is difficult to define” (Wood and Smith, 1993).

In fact, the term traditionally covers three categories of documents – conference proceedings, reports and doctoral theses - that are often printed in short runs. Nevertheless, the borderline with “white” or “conventional” literature is permeable, since some conference proceedings are published by commercial publishers, as monographs in serial publications or journals, and the same is true for some reports. As for doctoral theses, especially in the humanities and social sciences, some are also found on the commercial publishing market.

However, regarding all the other documents that circulate outside conventional publishing circuits, the lack of “commercial oversight” raises real problems for teachers and

1 Based on an article first published in French in Perspectives Documentairses en Education 2006, n° 62.
Researchers as well as for information professionals when it comes to locating and acquiring them. The lack of “commercial oversight” and promotion (advertising) also often implies a lack of “bibliographic oversight”. In other words, these documents are often inadequately referenced in catalogues and databases, so that searches through this category of scientific information requires specialised knowledge on sources and circuits – as evidenced by the INRP’s *Observatoire des thèses concernant l’éducation* (Observatory on theses in education sciences).

Information professionals – including archivists, librarians, researchers and teachers – have been contributing to studies on grey literature for nearly 30 years now, compiling a rich corpus of articles and, since 1993, international conference papers on grey literature. In 1985, several European countries founded the EAGLE association to identify and disseminate grey literature. In the 1990s, national and inter-ministerial initiatives in France (including GRISÉLI and LIGRIA, see Comberousse, 1995 and Desmichel, 1998) resulted in particular in the establishment of two national “one-stop shops” for accessing reports (La Documentation Française and INIST-CNRS). Other countries, including Britain and the US, have designated “deposit and distribution centres” (such as the British Library) or established portals for scientific reports (such as GrayLit for US Federal Agencies).

Rather than attempting to summarise the full range of this abundant literature, our study offers observations on its importance and how it is evolving in the digital environment. We also include comments on the problem of descriptive referencing and distribution, to conclude with a series of open questions.

**ON THE (RELATIVE) IMPORTANCE OF GREY LITERATURE**

Grey literature has a role of its own as a means of distributing scientific and technical information (see Sondergaard et al., 2003), and professionals insist on its importance for two main reasons: research results are often more detailed in reports, doctoral theses and conference proceedings than in journals, and they are distributed in these forms up to 12 or even 18 months before being published elsewhere (see Wood and Smith, 1993).

But how do researchers actually use grey literature? One way of evaluating this is to analyse the citations given in their publications. In order to obtain a statistical evaluation, we used two different sources: a Franco-Dutch study launched in 2004 and the journal entitled *Perspectives Documentaires en Education*.

**Scientometric studies (citation analyses)**

We look first at some results from the Franco-Dutch study (Schöpfel et al. 2005, Farace et al. 2006), which analysed 64 scientometric articles published between 1987 and 2005 and citing several thousands references altogether. The table below shows the proportion of grey literature cited in publications from different scientific disciplines:

<table>
<thead>
<tr>
<th>Field</th>
<th>Grey literature citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soil science</td>
<td>14%</td>
</tr>
<tr>
<td>Biology</td>
<td>5-13%</td>
</tr>
</tbody>
</table>

---

2 INRP: French National Institute on Research in Education Sciences
3 http://www.textrelease.com
Veterinary medicine | 6%
---|---
Psychiatry (addiction) | 1%
Psychology | 3%
Engineering Sciences | 39-42%
Economic Sciences | 9-17%
Sociology | 7-9%
Education Sciences | 14-19%

Table 1: Proportions of grey literature cited in scientific publications

These analyses show that the relative importance of grey literature is largely dependent on research disciplines and subjects, on methodological approaches and on sources used. In some fields, especially the life sciences and medical sciences, there has been a traditional preference for conventional distribution media (journals), while in others, such as agriculture, aeronautics and the engineering sciences in general, grey literature resources tend to predominate.

In particular, public administrations and public and industrial research laboratories produce a great deal of “grey” material, often for internal or “restricted” circulation (see Ullah et al., 2004, for example).

The journal Perspectives Documentaires en Education (case study)

For a more accurate idea of the way grey literature is used in education sciences, we analysed several recently published articles from Perspectives Documentaires en Education. The results vary widely.

We looked first at four studies published in 2003 in issue n° 59, quoting 372 bibliographic references altogether. The proportion of citations referring to “grey” documents varied from 12% (Lüdke 2003) to 37% (Duarte 2003) to 45% (Landesmann 2003) to 56% (Ndoye 2003). The latter – a study summarising documentation on school performance in Africa – is symptomatic: the majority of the sources used by Ndoye are documents produced by national or international bodies (ministries, universities, UN, World Bank etc.), which have never been distributed commercially and, three years later, have probably become very difficult to locate and obtain.

In another example, Auricombe (2001), in his study on searches and uses of documentation by researchers and teachers studying CNAM distance learning courses, highlights the importance of grey literature, from which 21% of his own references are taken (e.g. INTD and ENSSIB theses).

Finally, we analysed the different studies and notes in issue n° 60 of Perspectives Documentaires en Education (2003). 15 of the 198 sources quoted by the authors (= 7%) refer to conferences, seminars, legal texts and doctoral theses. The other citations, such as those in the same issue’s “current bibliography”, refer to books and journal articles.

A comparison was made with two studies in education sciences published in other journals: Okiy (2003), analysing 4012 references to 70 theses submitted in Nigeria, found that 14% of the sources were from grey literature (theses, reports, conferences, etc.). An analysis of 1842 references from theses submitted in the US (Beile et al., 2004) produced a figure of over 19% of “grey” document sources.
To sum up, the relative proportion of grey literature in scientific information varies widely with scientific fields but also with subjects, methodologies and even geographic origin. Grey literature plays a considerable part in the education sciences, accounting on average for 10-20% of all sources used.⁴

**TYPOLOGY: FROM PAPER TO DIGITAL**

To return now to the definition of grey literature, we indicated earlier that the term traditionally refers to reports, conference proceedings and doctoral theses. We will now take a closer look at what these cover in reality. SIGLE, the European database of grey literature, has been supplied since the 1980s with information from organisations in several EU countries. Its 800,000+ references are distributed as follows:

<table>
<thead>
<tr>
<th>Document types</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reports</td>
<td>62.7%</td>
</tr>
<tr>
<td>Theses</td>
<td>31.7%</td>
</tr>
<tr>
<td>Conferences</td>
<td>2.3%</td>
</tr>
<tr>
<td>Data files</td>
<td>2.1%</td>
</tr>
<tr>
<td>Translations</td>
<td>0.9%</td>
</tr>
<tr>
<td>Other</td>
<td>0.3%</td>
</tr>
</tbody>
</table>

*Table 2: Distribution of different document types in the SIGLE database*

Reports are the most numerous by far among the different types of grey literature. But the ‘reports’ category covers a wide variety of very different documents: institutional reports, annual or activity reports, project or study reports, technical reports, reports published by ministries, laboratories or research teams, etc. Some are distributed by public bodies (ADEME, CERTU, Documentation Française, European Commission, for example), others are confidential, protected or circulated to a restricted readership, such as technical reports or those from industrial laboratories. Some are voluminous, with statistical appendices, while others are only a few pages long.

In the other categories, scientometric studies (see Farace et al., 2006) offer a tremendous range of grey resources: besides theses and conference proceedings, they also include unpublished manuscripts, news bulletins, recommendations and standards, patents, technical notes, data and statistics, presentations, personal communications, working papers, laboratory research books, preprints, university studies, lecture notes, and so on.

However diverse, these documents all have one point in common: they contain unique and significant scientific and technical information that is often never published elsewhere. The lack of descriptive referencing and adequate circulation is therefore, as we have said, a real problem for scientific communication.

However, the Internet is now changing the whole picture. Not only because of changing user behaviour (see for example a recent study by Le Roux, 2005, on document specialists in the teaching professions), but also, and especially, because more and more grey

⁴ By way of note to the reader, this present study contains 30% “grey” references.
literature is being published on the Web. Even so, as a study from the German centre for information in the social sciences has pointed out (Artus, 2005), the emergence of a new medium for publications does not necessarily mean that more grey literature is appearing. But although the switch from paper to digital has not increased the number of publications, it has radically changed access and distribution methods and, especially, accentuated the ephemeral and volatile nature of grey literature.

FROM PREPRINTS TO OPEN ARCHIVES

There is no need here to describe the movement towards open access to scientific information, which has been crystallising since 1994 around various initiatives to promote open archives. For an introduction to the topic, we refer readers to a number of recent resources (Chanier, 2004, Fily, 2005, Aubry and Janik, 2005, a guide to the URFIST Paris site by A. Mahé\(^5\), or the INIST current awareness site\(^6\)).

The case of the first preprint server, ArXiv, set up by P. Ginsparg at Los Alamos in 1991, is nonetheless of interest for the purposes of this study\(^7\). This was a means for distributing research results organised by and for physicists, with no intermediaries, which was entirely independent from any commercial circuit for scientific publications. In this sense, the server (which contains over 350,000 documents today) corresponded exactly to the definition of grey literature.

However, the case is more complex than that. “Preprints” would not have existed without “print” - in other words, ArXiv would not have existed without scientific journals. Ginsparg’s aim was simply to circulate results quickly and immediately. The point was not to bring commercial publishing into question, since Ginsparg did not offer any alternative to the peer review system, which was still run by the conventional publishing circuit. ArXiv was in fact creating a kind of symbiosis between grey literature and traditional publishing, which was linked to the highly specific organisation of the nuclear physics community.

The creation of the first open archive within the CNRS in 1998, by Franck Laloë, a physicist with the ENS, obeyed a broadly similar logic\(^8\), advocating direct, fast, free scientific communication between researchers in the same field.

But the crisis that has hit scientific journals (see Chartron, 2002 and Keller, 2001) and the appropriation of the NICTs by information professionals have helped to turn the new means of distribution into an alternative model of scientific publishing. In France, the CCSD’s HAL is fast becoming the main countrywide institutional archive. This has two objectives:

- Economic: to offer a cheaper alternative to scientific publications by exercising increasing pressure on STM publishers.
- Administrative: to facilitate oversight over scientific production from the various research organisations (EPST, universities) by identifying and evaluating researcher publications via a central database.

Given the methodological and hermeneutic diversity of the different scientific communities, one may well wonder whether a system that has operated for 15 years in physics will be equally effective in other fields, especially in view of the “constraints induced by research evaluation methods” (see Prévot, 2005, on the subject of Earth Sciences).

\(^5\) http://www.ext.umpc.fr/urfist/ArchiveOuverte/OA.html
\(^6\) http://www.inist.fr/openaccess
\(^7\) http://arxiv.org
\(^8\) http://ccsd.cnrs.fr
What is the part played by grey literature in this new environment? The first international ‘Directory of Open Access Repositories’ (OpenDOAR), established by the Universities of Nottingham and Lund, identifies 349 different sites, including 20 in France (February 2006). OpenDOAR indexes 12 categories of grey literature, including reports, theses, conference proceedings and preprints, but also working papers, learning objects (especially university lectures), PowerPoint presentations and student theses and research results (‘data sets’). The directory does not allow archive searches, so that there is no way of obtaining an exact figure for the numbers of documents involved. The table below therefore only shows the number of open archives that contain certain categories of grey literature:

<table>
<thead>
<tr>
<th>Document types</th>
<th>Number of archives</th>
<th>% of OpenDOAR sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>Theses</td>
<td>211</td>
<td>60%</td>
</tr>
<tr>
<td>Reports</td>
<td>146</td>
<td>42%</td>
</tr>
<tr>
<td>Conferences</td>
<td>146</td>
<td>42%</td>
</tr>
<tr>
<td>Preprints</td>
<td>89</td>
<td>26%</td>
</tr>
<tr>
<td>Student theses</td>
<td>72</td>
<td>21%</td>
</tr>
<tr>
<td>Working papers</td>
<td>66</td>
<td>19%</td>
</tr>
</tbody>
</table>

Table 3: Presence of different document types in the OpenDOAR archives

At first sight, these figures seem to suggest that grey literature is relatively well represented in open archives. 60% of recently identified sites contain doctoral theses, and over 40% of open archives contain conference proceedings and reports. However, the picture becomes less clear with an analysis of archive descriptions, which shows that the number of sites explicitly dedicated to grey literature is much smaller: only 45 sites are identified for doctoral theses (13%), 30 for reports (9%) and just 8 for conference proceedings (2%). Observations made at the 7th international conference on grey literature in 2005 indicate that these documents are often swamped within the sheer mass of documents that are deposited in archives and/or difficult to identify. The development of open archives does not therefore seem to have changed the situation of grey literature to any great extent.

We conclude this section with a few remarks on the situation in France. Even though 9 archives, including HAL, contain conference proceedings, these are referenced at the same level as articles (individual communications) without any specific search functions for this type of information (grouping by conference, etc.). The same is true for reports: although a number of archives, especially institutional ones (IFREMER, Institut Jean Nicod, MSH-Alpes), obviously contain reports, as far as we know there is only one project (LARA, run by INIST-CNRS, based on DSpace) that has been developed exclusively as a repository for scientific and technical reports (see Stock et al., 2006).

The situation is different for doctoral theses, where more advances have been made. This is probably accounted for by three factors: an international environment (NTLTD, ETD conferences), the national ABES system (SUDOC, the STAR project) and the scientific issues raised in doctoral theses (see Le Hénaff and Thiolon, 2005). Nevertheless, the situation in France is still characterised by the heterogeneity of institutional, thematic and thesis archives (e.g. Cyberthèses, PASTEL, TEL), which together account for only a small proportion of national production (see Pailllassard et al., 2005). Although this situation is unsatisfactory in

---

9 http://www.opendoar.org
comparison with other countries, it will undoubtedly evolve in the coming months towards more systematic identification and deposits of French doctoral theses in electronic format, through the ABES and CCSD sites.

**IMPROVING “BIBLIOGRAPHIC OVERSIGHT”**

We said earlier that the lack of “commercial oversight” of grey literature implies a lack of “bibliographic oversight”. In other words, “grey” documents are often inadequately referenced in catalogues and databases. This does not mean that there are no standards or recommendations for cataloguing reports, conference proceedings, theses and so on. However, contrary to the situation with journals and books, the absence of commercial stakes has contributed to the (very relative) “success” of the rules that have been set out.

The way the different types of grey literature are referenced still depends more on choices made by the bodies that produce, collect or distribute these documents than on any national or international standard (ISO, ISBD, AFNOR etc.). The failure of the international report numbering standard (ISRN) is symptomatic in this respect - having been the only country with an active ISRN agency (INIST) for several years, France eventually had to agree to the ISO abandoning the standard. The only remaining exception in France concerns doctoral theses submitted to French universities, for which the Téléthèses database and, later, SUDOC cataloguing, imposed a uniform bibliographic format.

At European level, input requirements for the SIGLE database forced the network’s member countries to attempt to harmonise their “grey” resources around a single SGML format. But input to SIGLE ceased in April 2005, and the network – the EAGLE association – has gone into liquidation. The predictable outcome is that each organisation will be returning to its own referencing methods and rules, outside any kind of uniform bibliographic oversight.

The rapid development of the Internet and its ever-multiplying on-line resources is affecting bibliographic oversight in two ways (see Artus, 2005). On the one hand, such “wild growth” is speeding a decline in the application of formal and qualitative standards, with the risk of grey literature becoming “even more greyish”. On the other hand, the same risk has also boosted awareness, in all countries, of the need to define a few minimal data – called metadata – to provide a framework for referencing digital documents.

Examples include: a project aiming to adapt the Dublin Core to reports (Jeffery et al., 2002) or doctoral theses; the French “Thèses Electroniques Françaises” (TEF) initiative, which is working on a set of metadata and a single XML schema, the “Text Encoding Initiative” (TEI) designed to develop and recommend, at international level, common tagging standards that are independent from IT upgrades, and the creation by the CNRS and INRIA, in 2005, of a TEI support centre for Europe, located in Nancy and built around the ATLIF, LORIA and INIST organisations, which will probably also address the matter of grey literature.

The problem of poor standards and lack of uniformity in referencing documents deposited in institutional archives was the reason why the JISC, in Britain, set up a committee in early 2006 to look into the interoperability of these archives and to describe their resources in order to facilitate their identification by end-users.

One final example: after the 7th international conference on grey literature in 2005, P. de Castro and S. Salinetti from the Istituto Superiore di Sanità in Rome initiated an

---

11 [http://www.tei-c.org](http://www.tei-c.org)
international committee on grey literature which would develop recommendations (the “Nancy Style”) for the production and distribution of scientific and technical reports. The first version of the document is now published on the Web¹² (GLISC, 2006), and translations into French and Italian are under way.

This example is symptomatic in some ways of the problem of bibliographic oversight of grey literature. This is an international issue to which national responses have only brought partial solutions, thereby actually increasing the diversity of data and the difficulties involved in identifying and locating documents. And because there are no commercial issues at stake, improved referencing will always be dependent on the initiatives and willingness of producer or distributor organisations and a few committed professionals.

ACCESS AND DISTRIBUTION

Identifying, locating and obtaining grey literature is generally not easy – and this is inherent to its nature. To get some idea of the problem, readers might attempt a search for the reports, studies, conference proceedings and doctoral theses cited by Ndoye (2003) or Landesmann (2003) in issue n° 59 of Perspectives Documentaires en Education...

For 20 years, the SIGLE database offered a solution at EU level, insofar as its partner organisations were under obligation to keep referenced documents at their end-users’ disposal via lending or document delivery services.

Users today are faced with a huge variety of sites, archives, catalogues and databases, which makes searches for “grey” information not only painstaking but sometimes prohibitive as well, not to mention the linguistic difficulties involved.

In view of the rapidly changing face of STI within the digital environment, all the major traditional centres collecting and distributing grey literature, such as the British Library, the Canadian Institute for Scientific and Technical Information, the TIB in Hanover or INIST, have begun to develop free access services to these documents, especially for theses and preprints, but also for reports and so on (see Boukacem-Zeghmouri and Schöpfel, 2006).

However, there is still a notable absence – at least at national and European levels – of portals and search tools that are specifically dedicated to “grey” documents. Even Elsevier have started to index doctoral theses in electronic format in their Scirus search engine. The initiative from members of the EAGLE association to launch a metasearch engine dedicated to European grey literature collections is still in the draft stage (see Schöpfel, 2006). Meanwhile, users have no other choice but to conduct searches individually using whatever means they have to hand, or to keep using the search services provided by traditional organisations (SCD, BU, INIST).

THE FUTURE OF GREY LITERATURE

Grey literature will remain a challenge for information and documentation professionals as well as an interesting field for research and activity, in five areas at least (see Stock et Schöpfel, 2004):

The need for a new definition: The traditional definition of grey literature needs to be refined and supplemented through an accurate analysis of new means of access and

¹² http://www.glisc.info
distribution, in line with Mackenzie Owen’s observation (1997) that “Grey does not imply any qualification (but) is merely a characterization of the distribution mode”.

**The need for an economic model:** Despite the absence of “commercial oversight”, collecting, distributing and searching grey literature all come at a price, which may in fact be much higher than for article or book searches. To date, there is no economic model in this area and analysis is much needed in terms of investments, direct and indirect costs, acquisition prices and so on.

**The need to oversee archiving practice:** New technologies of information and communication facilitate resource archiving in general, and there is strong incentive from the “open access” movement. Nevertheless, the question of “who should archive what, where, when, and for how long” has remained largely unanswered to this day. Given the policy – and financial – aspects involved, answers are urgently needed, even for only part of grey literature resources.

**The need for a new ‘value chain’:** In the last few years, in the Netherlands, Rosendaal (2004) has been researching the process whereby universities have been reappropriating publications, and highlights the radical changes affecting the ‘value chain’ of scientific publications. Evaluations of scientific publications and their quality are set to become major issues in the context of emerging STI trends. The impact of new technologies of information and communication on non-commercial circuits is a complex matter that has been little analysed to date – and the potential field for research is vast.

**The need to clarify the legal aspects:** In our study, the issue of intellectual property rights in grey literature has been deliberately left aside. Nevertheless, the legal status of grey resources and rights in their use (deposit, archiving, distribution, etc.) is (another) major challenge for the future of this form of STI. The legal environment in France and internationally is evolving rapidly, and no documentary analyses have so far addressed the legal aspects and economic issues at stake in the field of grey literature.

To conclude, we offer some prognoses for further reflection.

It seems certain that:

- Grey literature will not disappear, but will continue to play a part alongside commercial publishing.
- The borderline between “grey” and “white” (commercial) literature will become increasingly indistinct, particularly in an environment that is moving towards free access to STI.
- The proportion of “grey” documents published on the Web will increase rapidly.
- The Internet will encourage greater diversity in “grey” resources (raw research results, notes and personal comments, lectures, etc.).

It seems likely that:

- Bibliographic oversight of grey literature will remain problematic despite the trend towards standardisation of digital documents.
- Open archives will offer more appropriate services and functions for some segments of grey literature, not only for preprints but also for doctoral theses and reports.
- Some organisations – especially public bodies but also in the private sector (Elsevier, Google, etc.) – will develop tools and services to aid more efficient exploitation of “grey” resources on the Web.
It seems unlikely that:

- Searching and collecting grey literature will become as easy as for journals and books from the traditional publishing sector.
- The new tools for collecting, depositing and archiving will make grey literature less ephemeral and volatile than in the past.

**BIBLIOGRAPHY**


