# Usage of grey literature in open archives: state of the art and empirical results

Joachim Schöpfel, Hélène Prost

# Usage of grey literature in open archives: state of the art and empirical results[1]

Joachim Schöpfel

*Charles de Gaulle University of Lille 3*

Hélène Prost

*Institute for Scientific and Technical Information (INIST-CNRS)*

## Abstract

The purpose of our communication is to present first results from a current research on the development and usage of open archives in France. This study aims to gain empirical insight in usage patterns of freely available scientific items deposited in open archives, especially of non-commercial material, e.g. grey literature, mostly not distributed through other channels.

We present a state of the art of published empirical data, standardization, research projects etc., together with a survey on the development and the usage of French open archives, based on open source methods and investigation.

The usage of grey literature in open archives is a recent field of professional and scientific interest. So far, little has been published on usage of open archives, and even less, on usage of deposited grey literature. Nevertheless, there are some promising new initiatives and projects and first empirical data. Our communication will combine review, quantitative and qualitative survey data and case study in order to provide a realistic insight into this emerging field.

Expected results: Empirical data allowing for first comparison between

[1] Acknowledgement to C. Boukacem-Zeghmouri for her helpful advice and contribution to data collection and analysis.

different archives and document types. Awareness on the scientific but also professional and economic interest of these data. A contribution to standardization (recommendations on data production, delivery and analysis).

## Notes on the authors

**Joachim Schöpfel** is senior lecturer in information and communication sciences at the Charles de Gaulle University of Lille 3 and is member of the GERiiCO laboratory, of GreyNet and EuroCRIS. He published on grey literature, scientific publishing, document delivery, digital libraries, usage statistics and professional development.

*Université Lille 3, UFR IDIST, BP 60149, F-59653 Villeneuve d'Ascq Cedex.*

joachim.schopfel@univ-lille3.fr

**Hélène Prost** works since 1995 as a librarian at INIST-CNRS and is specialised in the evaluation of collections and document supply. Actually she is preparing a new SS&H database as a part of the FRANCIS database. She obtained a Master in History in 1991 and a Master in Scientific Information in 1993. She published on statistical and bibliometric analysis of information.

*INIST-CNRS, 2 allée du Parc de Brabois, F-54519 Vandoeuvre-lès-Nancy Cedex.*

prost@inist.fr

## 1. Introduction

Grey literature represents a substantial part of the scientific production (Schöpfel & Farace, 2009). Since the Seventh International Conference on Grey Literature at Nancy in 2006, the GreyNet community intensified its research activities on the impact of the open access movement on the grey literature.

The purpose of this communication is to provide a follow-up study to our 2008 evaluation on the integration of grey literature in French open archives (Schöpfel & Stock, 2009) that described "a landscape in movement", with a significant increase of university institutional repositories supported by the academic consortium COUPERIN.

We considered that "the impact of grey material (…) in open archives is real and will stay", with an overall part of 17% of the deposited items. On the other hand, our survey revealed three major problems:

"(1) Policy statements need improvement. Often, the strategy and positioning of repositories are not explicit or simply missing.

(2) Especially grey items in open archives need improved bibliographic control. Compared to traditional cataloguing standards, metadata for grey material are less specific or again, simply missing. This is a problem for referencing, efficient search strategies and evaluation.

(3) Mostly wanted are detailed usage statistics on access and download of documents and other items in open archives."

The 2009 follow-up study surveys data on the development of the archives, e.g. evolution of deposits, and investigates usage statistics. The GL11 communication provides preliminary results from ongoing empirical and statistical analyses.

The study was funded by the Charles de Gaulle University of Lille 3. Special thanks to the professional team of the academic library of the Lille 1 campus, partner of the research project, and to Chérifa Boukacem-Zeghmouri for her contribution and advice.

## 2. Methodology

We basically applied the same approach as last year. The 2009 survey includes 150 representative (e.g. registered either with a dedicated

platform or as data provider for harvesting) French digital repositories. The different archives were selected through 19 French and international registries of open access repositories or service providers (see appendix), between May and June 2009, and followed a defined set of criteria (located/hosted in France, living archive, size>0).

Each registered archive (URL) was checked; errors (incorrect URLs etc.) and duplicates were eliminated. Information about the remaining archives were incorporated into a spreadsheet with 58 data columns in 5 categories:

1. General (background) information about the archive.
2. Specific information about the archive.
3. Content information.
4. Qualitative data.
5. Comments.

Usage data were collected separately.

## 3. Results

Our communication will concentrate on three empirical results:

- The development of French open archives, e.g. evolution of size and content.
- The development of grey items in these archives.
- The dissemination and content of usage data.

The following discussion includes a short state of art of international projects of standard usage data in open repositories. The communication ends with some recommendations for usage assessment and gives an overview of future research.

### 3.1. Development of open archives in France 2008-2009


**Figure 1: Number of items 2008 and 2009**

The total number of referenced open archives in France has increased since last year. While we identified 56 repositories in 2008, their number attained 150 in 2009 (+168%).

In the same time, the total number of deposited items increased from 703,178 to 1,878,520 (see Figure 1). Especially the number of small and medium-sized repositories increased rapidly, with an item number between 100 and 10,000.

Nearly 60% of these archives are institutional repositories, mostly hosted by universities and other HE structures. About 15% are thematic archives (one discipline or one subject), another 15% contain only one category of documents (mostly electronic theses or dissertations, ETDs). The rest are mixed or heritage repositories.

On 103 sites the author or his laboratory can deposit his documents online. Other referenced repositories seem to be rather normal websites, digital libraries or portals, without any characteristics related to the open access movement.

Roughly one third of the documents are journal articles (pre- or post-print), another third are grey material or datasets. The rest are different types of documents, for instance heritage items, or cannot be correctly

identified. Following these figures and information found on the web sites, we can estimate that more than 60% of the repositories have some kind of quality control or other validation procedure – significant more than last year.

### 3.2. Grey literature in French open archives

74% of all referenced repositories contain grey literature. In fact, nearly all institutional repositories and most of the document-specific archives contain grey material.

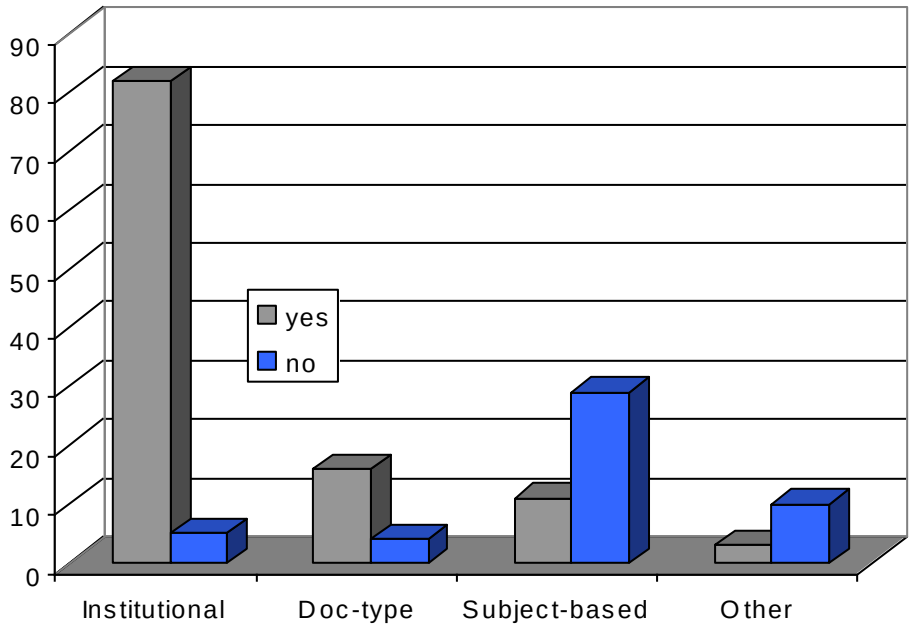


**Figure 2: Grey literature in different kinds of archives**

A deeper analysis of this material reveals the following figures (Table 1).

| Type of grey literature | Number of items |
|---|---|
| Theses and dissertations | 70,488 |
| Reports | 36,186 |

| | |
|---|---|
| Conferences | 157,257 |
| Working papers | 6,637 |
| Courseware | 2,875 |
| Other | 43,308 |

**Table 1: Grey literature in repositories**

The most important part of grey literature are conferences (proceedings, communications etc.), followed by theses and dissertations and, at a lower level, by reports. Nevertheless, an important part cannot be correctly identified ("unpublished work" etc.).

The reason is twofold. On the one hand, not all open sites allow for an advanced search on document types. Another reason are missing and/or non standard (non comparable) metadata. Without metadata, the survey becomes tough – one should access each item in order to define or guess its category. Mission impossible for more important repositories.

In fact, nearly 70% of the repositories – a little bit more than in 2008 – provide specific metadata for grey literature. For most of them, this means in particular some specific data for theses (for instance, the university) or for conferences (location and date of the conference).

Figure 3 shows the development 2008-2009 of some more important and well-known French repositories. Except PERSEE – the national repository for the back files of French SS&H journals – all important repositories contain grey literature, and for all we observe a relative more important increase of the part of grey literature than for the overall number of items. The future may show if this is a stable or a transitory evolution.

For instance, HAL the central open archive for academic publishing[1] arrives in 2009 at about 60,000 grey items. The institutional repository of the national centre for agronomic research INRA holds between 40 and 50,000 grey items – significantly more than in our 2008 study.

[1] HAL ranks second in the list of world repositories *Webometrics* http://repositories.webometrics.info/top400_rep_inst.asp

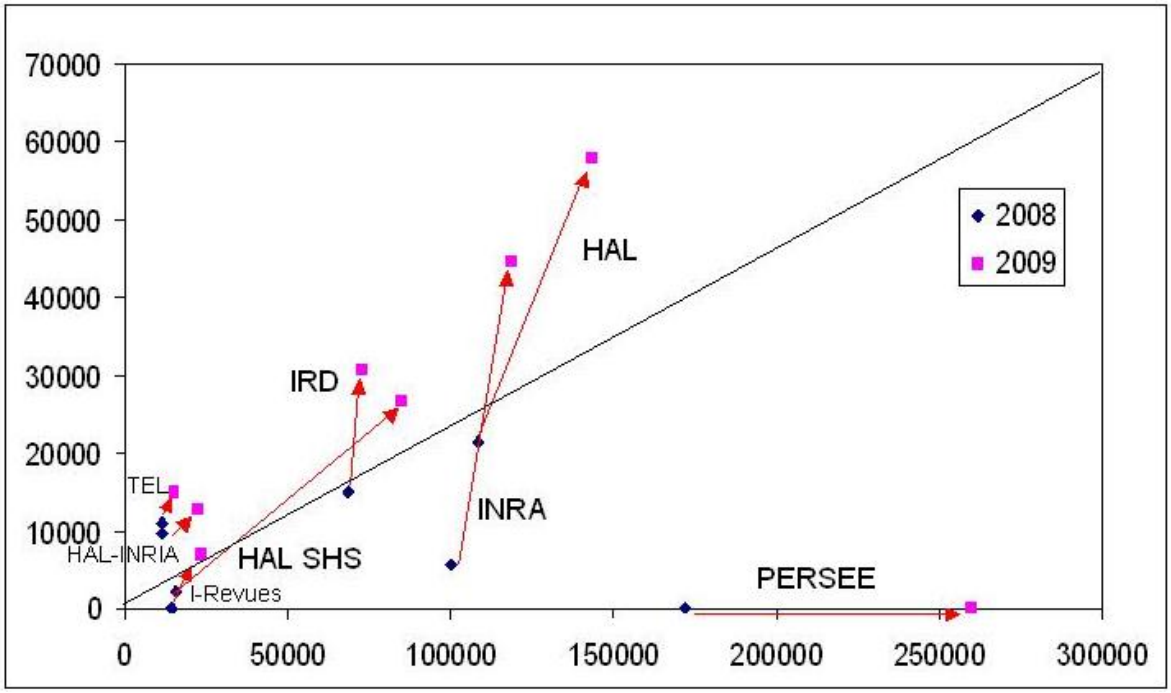


**Figure 3: Development of some selected repositories (left side: grey items, below: overall number of items)**

Even the CNRS plat-form for open e-journals, I-Revues, began to upload non-commercial material, e.g. conference proceedings.

### 3.3. Usage statistics

Our survey identified only a small number (7%) of repositories that publish usage statistics. Sometimes "usage" is misinterpreted in terms of upload figures, instead of access and downloads statistics. In the following we provide a short overview on some statistics related to grey literature.

The statistics of the University of Toulouse repository OATAO distinguish between published articles and electronic theses. The difference is significant: the average downloads number of articles is 49 per item while it is 109 for theses, e.g. the grey (unpublished) material is around 2,2 more often requested than journal articles (Malotaux, 2009).

The IFREMER repository allows for comparison between monthly downloads of articles, reports, conferences and theses (see table 2).

| | 2007 | 2009 |
|---|---|---|
| **Articles** | 10 | 7 |
| **Theses** | 70 | 33 |
| **Reports** | 30 | 10 |
| **Conferences** | n.d. | 9 |

**Table 2: Average monthly downloads (IFREMER repository)**

All types of grey documents are more often requested than published articles. And at least for this repository, the most interesting materials for visitors are the electronic theses that can't be accessed elsewhere (Merceur, 2007 and 2009).

The statistics of the University of West Britanny (Brest) confirm that the electronic theses are more often downloaded than other items. 17 of the 25 most requested documents are theses (Bertignac & Gac, 2009).

Other statistics include the list of the requested theses (INP Toulouse) or provide information on activity metrics, information seeking and user characteristics without details on grey literature (ParisTech).

## 4. Discussion

The 2009 follow-up study encountered the same kind of methodological problems than the 2008 survey (see Schöpfel & Stock, 2009) - the overall number of items in some repositories is difficult to define, and identifying grey content remains a challenge. Also the architecture of the HAL repository (or HAL system) is complex, and one part of deposits is stored in two or three of the HAL archives. Nevertheless, our survey data allows for some more general statements.

### 4.1. The changing nature of open archives

Open repositories are meant for publications. Yet, 16% of all items in French repositories are metadata (records) without full text, and 25% of the repositories in our survey contain at least 50% simple records. A comparison with US, UK or German repositories shows that this is not particular for French open archives. Paradoxically, the reason for this

evolution may be linked to the success of institutional repositories: as they are increasingly integrated into evaluation of institutions and scientists, hosting structures (libraries etc.) started to upload metadata even if the documents, for legal or other reasons, are not available.

Another evolution non-conform with the initial goals of the OA movement is the important part of scientific and/or national heritage items in institutional and other repositories. At the beginning, open repositories were created to foster and speed up direct scientific communication (cf. the role arXiv always plays for the high energy physics community). 30 sites (20%) in our survey hold such kind of content, with some of the most important archives (PERSEE for SS&H journal back files, the French national library's GALLICA and the history of art repository of INHA). All these sites are referenced as open archives. Are they, really? How should we fix the difference, for instance, between HAL with current publications and NUMDAM with back files in Mathematics? In 2009, around 46% of the deposits seem to be heritage items. Should we revise the definition of open repositories?

A third problem is the restricted access to some repositories. Basically, open archives stand for general and unrestricted access to scientific information. In reality, 12% of the repositories limit access to their content in some way or other. For some sites, visitors have to register; for others, access is restricted for non-institutional user.

## 4.2. Metadata and usage statistics

The part of repositories with specific GL metadata slightly increased (68%, +2% compared to 2008). But at least 15 open archives with grey materials lack specific metadata. Identifying different types of grey documents is often difficult. Sometimes, the only distinction made is between "published" and "unpublished" papers. In other archives, the GL categories appear to be rather ad hoc categories without any precise or standard definition.

Without metadata, a detailed and in-deep analysis of the repository content is impossible; also, without metadata, usage statistics cannot be linked to the content in terms of document types, institutions (affiliations), disciplines etc.

Nevertheless, the small number of (published) statistics confirms results

from international repositories such as RePEc[1] where grey documents – unpublished working papers – are consistently more requested than articles (see figure 3).

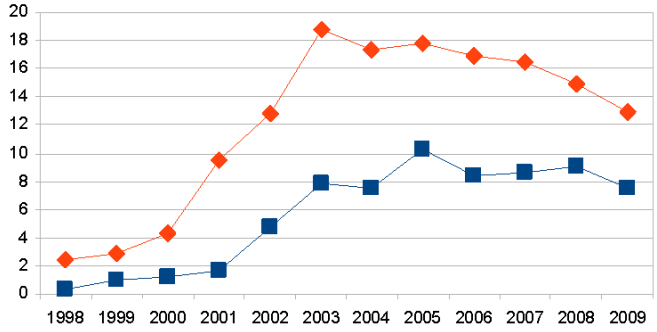


**Figure 4: Average annual downloads per document type in RePEc (red: working papers; blue: articles)**

Yet, more surveys are needed to confirm this statement, provide more detailed evidence and also help to understand the reason of this difference.

## 4.3. Recommendations

We already observed that only few repositories made their statistics available. And those who do so use different tools, methods, terms. In some way, this reminds the early period of digital libraries when publishers produced either non-standard or no usage statistics at all.

Nevertheless, as a recent report settled, "researchers want tangible, immediate benefits such as download statistics (…)" (Fry et al., 2009). In the following, based on our own survey and on other projects (JISC PIRUS[2], DINI OA-Statistik[3]) we suggest some recommendations meant to foster standard developments of open repositories.

1. Recipient: Usage statistics should be useful to all groups involved in the functioning of open repositories, e.g. usage data should be made available to authors, users (visitors), and institutions.

[1] http://repec.org/

[2] http://www.jisc.ac.uk/whatwedo/programmes/pals3/pirus.aspx

[3] http://www.dini.de/projekte/oa-statistik/

2. COUNTER principle: Usage statistics should be defined at different levels, with increasing complexity and a basic minimum level that corresponds to the COUNTER Journal Report 1.

3. Log files analysis: A selection of a minimum data set for a basic log files analysis should be defined. These data should cover the whole range of potential information (visitor, content, request type, date, unique identifier).

4. Terminology: The usual terms and wordings of usage statistics, log files and open repositories should be clearly defined and if possible, translated in French (access, downloading, visit, request, hit…).

5. Provision: Reports must be provided monthly following the COUNTER rules. Data must be updated within four weeks of the end of the reporting period. All of last calendar year's data and this calendar year's to date must be supplied.

6. Metadata: Usage of full text and metadata (records) must be clearly distinguished.

The usage data should be disseminated in en environment of added value services that take into account the specific needs of the different groups (authors, visitors, institutions).

"Value-added services such as download statistics, email alerts, etc would contribute to the perceived usefulness of repositories and would help them gain popularity in what is an increasingly competitive information landscape." (Fry et al., 2009).

The development of such services should include at least four key elements:[1]

- Modular statistics (collections, document types, time period etc.).
- Summary tables.
- Assistance-help online / FAQ.
- Link with other tools measuring the impact of deposited items (citations, tagging etc.).

[1] See for instance the metrics of the Public Library of Science PloS at http://article-level-metrics.plos.org/

These elements would help authors, visitors and institutions to evaluate the impact, popularity and quality of stored content.

## 5. Conclusion

Our communication presented preliminary results of an ongoing study on the development and usage of French open repositories.

The total number of referenced open archives in France has largely increased since 2008, from 56 to 150 in 2009 (+168%).

74% of all referenced open archives contain grey literature, especially institutional repositories. Yet, an important part of grey items cannot be correctly identified because of missing search options and/or metadata.

Only a small number of repositories publish usage statistics (7%). The available data confirm that grey documents are more often requested than articles.

We added some recommendations for the development of usage statistics in the open repository environment, following in particular the model of the project COUNTER.

In the next months, our research team will apply the log file analytical approach to the institutional repository of the university of Lille 1, IRIS[1] (the former *Grisemine* website, the first "grey" archive in France and presented at the Nancy conference), in order to illustrate the state of the art and the survey results. In the same time, we will elaborate a glossary of terms specific to usage statistics of open repositories. Probably, this part of work will be done together with the DINI team.

## 6. Bibliography

Bertignac C, Gac D (2009). La voie verte: les archives ouvertes. In: Open Access Week. 19 October 2009. Institut Universitaire Européen de la Mer.

http://www.eur-oceans.eu/WP3.1/OA/LaVoieVerte-Catherine\%20Bertignac.pdf

[1] https://iris.univ-lille1.fr/dspace/

Fry J, Oppenheim C, Probets S, Creaser C, Greenwood H, Spezi V, et al. (2009). PEER Behavioural Research: Authors and Users vis-à-vis Journals and Repositories. Baseline report. LISU Loughborough University.

http://www.peerproject.eu/fileadmin/media/reports/Final_revision_-_behavioural_baseline_report_-_20_01_10.pdf

Merceur F (2007). Gestion d'une archive et d'un moissonneur, l'exemple de l'IFREMER. In: RPIST 2007, 20 juin 2007.

http://rpist.inist.fr/sites/rpist/IMG/pdf/archimer-2.pdf

Merceur F (2009). Optimiser la visibilité de vos dépôts, au vu des statistiques d'Archimer. In: Open Access Week. 19 octobre 2009. Institut Universitaire Européen de la Mer.

http://www.eur-oceans.eu/WP3.1/OA/OptimisationVisibilite-FredMerceur.pdf

Schöpfel J, Farace DJ (2009). Grey Literature. In: Bates MJ, Maack MN eds. *Encyclopedia of Library and Information Sciences*. 3rd edition. Taylor & Francis.

Schöpfel J, Stock C (2009). Grey literature in French Digital Repositories: A Survey. *The Grey Journal*, 5(3), pp. 147-161.

Malotaux S (2009). OATAO Archive ouverte multi-établissements. Bilan après un an d'existence. In: Journées d'études sur les archives ouvertes. Consortium COUPERIN. 2 et 3 avril 2009.

http://journeesao.wordpress.com/2009/04/19/archives-ouvertes-et-regroupement-d\%E2\%80\%99etablissements-le-cas-des-ecoles-d\%E2\%80\%99enseignement-superieur-toulousaines-sandrine-malotaux/