Institutional archives for research: Experiences and programs in open access
Rome 30 November – 1 December 2006

Support of Open Archives at National Level. The HAL Experience

ANDRE Francis (a) CHARNAY Daniel (b)

(a) Institut d'Information Scientifique et Technique INIST-CNRS
2, allée du parc de Brabois
CS 10310
54519 Vandoeuvre-les-Nancy
France

(b) Centre pour la Communication Scientifique Directe CCSD-CNRS
Domaine scientifique de la Doua
12-14 boulevard Niels Bohr
66229 Villeurbanne cedex
France

Abstract

The French research institutions have recently signed a memorandum of understanding for the joint deployment of open archives based on the HAL platform (Hyper Articles on Line) developed by CCSD-CNRS. This unprecedented commitment pools all universities and Grandes Ecoles through their respective Conferences, and research organizations like CNRS, Inserm, INRA, INRIA, CEMAGREF, CIRAD, IRD and Institut Pasteur, and represents almost all researchers and academics in the French government sector. Some other public research organisations CEA, INERIS, INRETS and IFREMER recently joined the movement.

This decision to take on a common platform for depositing and valorizing scientific output comes as a result from a far-reaching process started up in 2000 when the CNRS founded its Center for Direct Scientific Communication (Centre pour la Communication Scientifique Directe) with the aim like arXiv to provide scientists with the capacity to disseminate freely their scientific output. The HAL service was launched in 2001.
Additionally, CNRS by committing themselves to the international Open Access movement backed the Max Planck Gesellschaft in the Berlin Declaration signed in October 2003. The Berlin Declaration has been signed by over 160 institutions now – including a great many Italian universities! – and is considered the starting point of institutional commitment to Open Access. In March 2005, in a joint press release, the four largest French research institutions (CNRS, INRA, INRIA, Inserm) announced their agreement to develop interconnected institutional open access repositories.
This decision provided ground to the HAL platform that became the repository supported by national-level research institutions. At the time, the platform was moving towards a repository collecting both doctoral dissertations and scientific papers in a wide range of fields, thereby providing various subject communities with specific deposit and retrieval interfaces.
The national agreement that has been recently signed now brings another significant challenge to the fore: the need to set up technical and organizational rules for circulating formatted data flows between the organizations' internal information systems and the HAL platform. This is the only means to make the

platform a reliable tool for the valorization of French scientific output. This challenge is all the more real as HAL is the French member of European DRIVER project aimed at interconnecting Europe's institutional repositories.

E-mail of corresponding authors: francis.andre@inist.fr charnay@in2p3.fr
1

Introduction

The signing in 2006 of a memorandum of understanding for a coordinated development, at the national level, of a common open archive platform was a milestone in the French research institutions' commitment to the creation of open archives. In a previously unheard of example of cooperation, the signatories of the memorandum represented almost all research institutions in France: universities, specialized higher education establishments (the so-called *Grandes Ecoles)* via their respective boards, as well as French public research institutions (CNRS[1], INSERM[2], INRA[3], INRIA[4], CEMAGREF[5], CIRAD[6], IRD[7], Institut Pasteur, CEA[8], IFREMER[9], INRETS[10], INERIS[11]) i.e. the great majority of researchers involved in public research projects. But before going into the details of the scope of this memorandum and its impact on the development of open archives in France, it might be useful to briefly sketch the background of French public research to back up the discussion of the technical and structural elements that were selected for the HAL open archive.

The Structure of French Public Research (1)

French public research is divided into three distinct institutional and complementary sectors:

- Government bodies: Public research institutions and related government departments;
- Higher Education: Universities, specialized higher education establishments (*Grandes Ecoles*), university hospitals and medical centers;
- Non-profit associations and foundations.

[1] Centre National de la Recherche Scientifique – National Center for Scientific Research
[2] Institut National de la Santé et de la Recherche Médicale – National Institute for Health and Medical Research
[3] Institut National de la Recherche Agronomique – National Center for Agricultural Research
[4] Institut National de la Recherche en Informatique et en Automatique – National Institute for Research in Computer Science and Control
[5] Agricultural and Environmental Engineering Research
[6] Agricultural Research for Developing Countries
[7] Institut de Recherche pour le Développement – Research Institute for Development
[8] Commissariat à l'Energie Atomique – Atomic Energy Commission
[9] Institut Français de Recherche pour l'Exploitation durable de la Mer – French Research Institute for Exploitation of the Sea
[10] Institut National de Recherche sur les Transports et leur Sécurité – French National Institute for Transport and Safety Research
[11] Institut National de l'Environnement Industriel et des Risques – National Institute for industrial Environment and Risks management

These three sectors account respectively for 58%, 38% and 4% of the government's public research efforts.

In 2004, public research institutions had a total research staff of 41 000 researchers and research engineers (out of an overall staff of over 80 000) divided for the most part among public scientific and technical institutions (EPSTs), industrial and commercial institutions (EPICs) and to a lesser extent public administrative institutions (EPAs).
The major research structures are in decreasing order of size:

- CNRS, National Center for Scientific research, a multidisciplinary institution,
- INSERM, National Institute for Health and Medical Research,
- INRA, National Institute for Agricultural Research,
- INRIA, National Institute for Research in Computer Science and Control,
- CEA, Atomic Energy Commission
- CNES, National Center for Space Research,
- IFREMER, French Research Institute for Exploitation of the Sea.

Higher education with its 81 universities and dozens of *Grandes Ecoles* accounts for 50 000 research faculty members and research engineers out of a staff of over 70 000. Non-profit associations and foundations, for their part, account for about 3 000 researchers.

Overall budget for these public research institutions amounted to over 13 billion euros in 2004 compared to 22 billion euros for French industrial research and development.

Ninety-four thousands researchers and research engineers contribute to knowledge growth through publications such as articles, doctoral dissertations and research reports. French scientific publications are estimated to account for 4.7% of global output and for 13.6% of European journal article output, with distribution variations according to disciplines (2):

- Biology: 4.8%
- Medical Research: 4.5%
- Applied Biology: 3.7%
- Chemistry: 4.5%
- Physics: 5.2%
- Earth and Space Sciences: 5.0%
- Engineering: 4.3%
- Mathematics: 7.1%

Given these estimates and if they were extended to all types of scientific publications, including preprints and working papers, valorizing the national scientific output would mean collecting between 70 000 and 90 000 electronic documents per year or close to 7 000 monthly electronic deposits. Therefore, it could only be achieved if all the research institutions and all the universities across the country participated in a common effort to develop open archives. Thus the challenge was to provide a platform enabling researchers to valorize their scientific output while respecting the specific disciplines of each institution.

The HAL Archive: A Multidisciplinary Platform for Researchers

The Center for Direct Scientific Communication (*Centre pour la communication scientifique directe – CCSD)*, the CNRS unit that hosts the HAL platform, was created in 2000 at the initiative of the physicist Franck Laloë who had the arXiv model in mind. The initial objective, similar to Paul Ginsparg's objective, was to provide researchers with a self-archiving tool, open to all scientists whatever their institution or nationality, but to extend coverage beyond physics and mathematics. The HAL repository was designed to

provide access to a database of full text of documents build by researchers for researchers, giving enhanced visibility and dissemination, and guarantying long term availability. HAL was designed from the beginning to support the new practices of research work dissemination. Today, the archives contain over 41 000 full text documents, with an average monthly deposit rate of 1200 documents (fig 1).

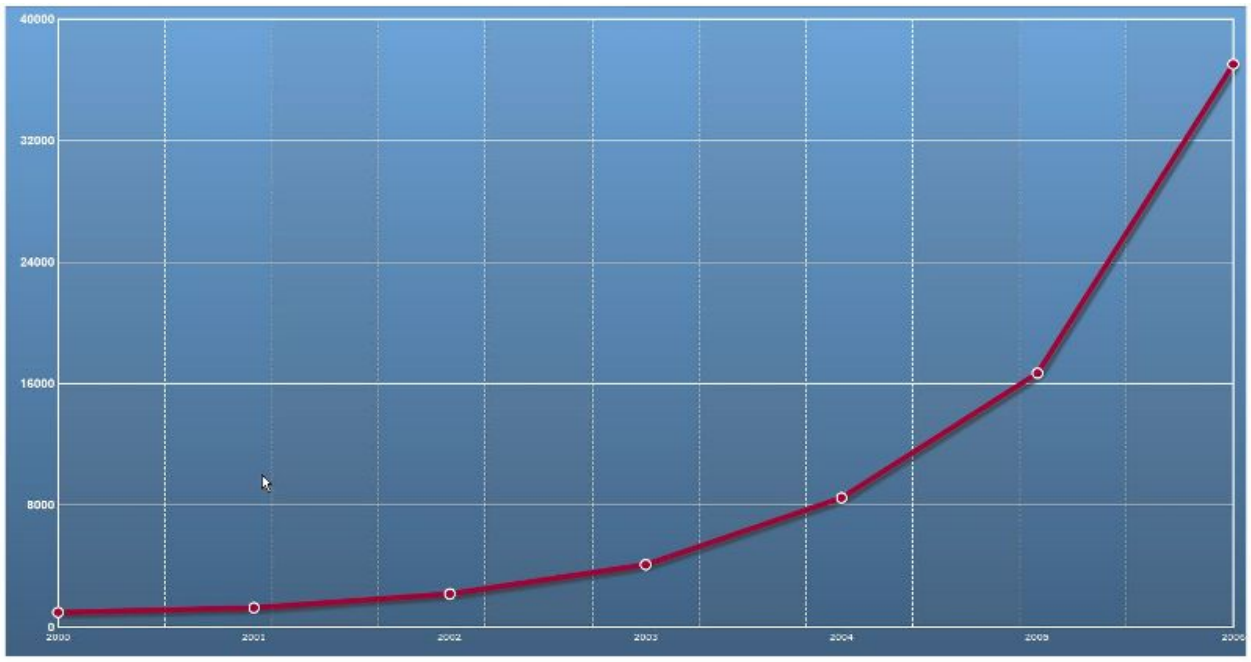

Figure 1 : HAL depositing curve

A single platform for multiple repositories

HAL, a software platform, is used by researchers belonging to a variety of research institutions and a great variety of disciplines ranging from Life Sciences, Engineering and Earth Sciences to Humanities and Social Sciences. Scientists strongly expressed the wish for a document deposit interface and access facilities that matched their individual needs. Indeed, wide variations between publication practices exist according to the disciplines involved such as variations between the importance given to doctoral dissertations, to research reports, to proceedings, to preprints, as well as in the ways scientific findings are used. Therefore, it was important to respect these different practices while striving to find a common model to describe the archived data. To achieve this objective, a core of common metadata was defined to ensure quality and homogeneity. It included a controlled description of scientific disciplines, of the periodical titles in which the deposited documents could be published and of the names of the laboratories and institutions to which the depositors were affiliated. These controlled data call upon authority files currently maintained in the HAL environment but which are to be replaced at some point by external sources validated at the national level.

A number of common operating rules were decided upon:

- The types of documents to be accepted for deposit: research articles, already published or to be submitted to a journal, doctoral dissertations that have been defended, book chapters, or more generally any document liable to be considered for publication by the peer-review committee of a journal;
- An access policy whereby documents can either be readily visible as soon as they are deposited or after an adjustable time lapse.

A deposit in the HAL repository is basically a four-step procedure with always the possibility of going back to the previous step:

- Entry of the document's descriptive metadata;
- Entry of the link between the author(s) and their affiliation(s);
- Transfer of the full text of the document;
- Validation and confirmation of the deposit.

Although the deposited documents are located in a single repository, they can be approached from various angles so that HAL can be seen as a collection of specific repositories based on various criteria:

- Scientific field: the HALSHS repository (http://halshs.archives-ouvertes.fr/) , for example, can be used to access and/or deposit documents in the field of Humanities and Social Sciences;
- Document type: TEL (Thesis on line - http://tel.archives-ouvertes.fr/) is used to access and deposit doctoral dissertations;
- Institutions: HAL INRIA (http://hal.inria.fr/) and HAL INSERM (http://www.hal.inserm.fr/) for example are specific respectively to the National Institute for Research in Computer Science and Control and to the National Institute for Health and Medical Research.

Each specific repository has its own interface with its own design and its own search criteria; it also has a deposit interface adapted to its own needs where additional metadata related to the scientific field or the institution can be entered. For example, the MESH thesaurus is linked to the deposit of the publications of INSERM researchers.
Document stamping is an original feature that can be used for various actions:

- To confirm the validity of a document deposit and authorize putting it online; this is usually done by the authority in charge of the repository.
- To separate documents according to thematic or institutional criteria in order to create virtual collections maintained by a member of the scientific community involved. Currently there is around 150 active virtual collections such as for example the virtual collection of the cognitive sciences network (http://hal.archives-ouvertes.fr/ACI_COGNITIC/fr/).

Thus, HAL is at the same time a single centralized repository, accessible through a generic interface used both for depositing and searching documents (http://hal.archives-ouvertes.fr) and a collection of subject-oriented or institutional repositories, with specific deposit and search interfaces and with validation rules specific to a community. Only an extremely modular and open design of the platform could accommodate the variety of constraints expressed by the different scientific communities.

Ensuring the international visibility of deposits

At the beginning, the activities of the CCSD naturally led at first to the definition of links with the physicists' community through the hosting of an arXiv mirror site and the implementation of a software connection between the HAL and arXiv repositories. French researchers archiving in HAL can have their deposited documents automatically submitted to arXiv, therefore ensuring, through a single deposit act, their visibility in the international repository of their discipline.

Later on, this discipline element was extended to other disciplines. Researchers in Life Sciences can also deposit their publications in PubMed Central, the repository maintained by the US National Library of Medicine (NLM), provided their publication has a PubMed identifier. Similarly, there is an enabled link to the ADS (Astrophysics Data System) repository maintained by NASA. Other data export formats/protocols are also available:

- The REDIF format, used by the economists' community, to link to the REPEC international repository;
- And the seminal OAI-PMH protocol with several XML schemas.

The mere act of collecting at an international level the researchers' scientific output could be considered as an objective in itself. However, because of today's international contextualization of most scientific disciplines, efficient dissemination of these outputs can only be ensured by exposure in international subject-oriented repositories accredited by the concerned scientific communities. This is why links to external repositories were and will be implemented.

A constantly evolving technical frame

HAL is a platform that is totally developed and maintained by the CCSD. Based on open technologies, it is constantly evolving to accommodate the wishes of researchers and their institutions:

- Linux, Apache, MySQL and PHP are HAL's core technologies,
- The deposit interfaces, constantly evolving, now use the AJAX metalanguage, thus offering a better ergonomics,
- Interfaces include the REDIF, OAI-PMH, SOAP and RSS open protocols.

The basic philosophy is to have a tool offering gateways to external applications. Currently, HAL's role is not to provide a single tool to collect and archive any type of publications but rather to act as a relay towards international subject-oriented repositories such as REPEC, ArXiv, and PubMed Central.

Why deposit in HAL? What can researchers expect in return?

Researchers are in charge of depositing their documents in the HAL repository. It is not always easy to convince them of the usefulness of auto-archiving solely for the sake of the increased visibility their publications can achieve through open access. Therefore, to make HAL more attractive to researchers, several features were developed for their individual needs and among others:

- Creation of personal pages;
- Export in various formats (XML, LateX, RTF,…) of the bibliographic references of their publications;
- Selection of deposited documents by researcher, by laboratory, by research institutions'
- RSS feeds based on various criteria to keep up with recently deposited articles;
- Access to statistics of their publications' downloads.

With these features, researchers can dynamically obtain on request the fulltext of their publications and lists of references useful for grant requests or assessment files.

A National Memorandum of Understanding for an Open Archive Platform

During the 2006 Summer, French research institutions started a collaborative process previously unheard of at the national level. The aim was the joint development and operation of a shared open archive platform. A Memorandum of Understanding "for a coordinated approach on a national level to open

archiving of scientific output" was signed by all Universities and all *Grandes Ecoles* through their respective Boards and research institutions such as CNRS, INSERM, INRA, INRIA, CEMAGREF, CIRAD, IRD, Institut Pasteur, recently joined by CEA, INERIS, INRETS and IFREMER, thus representing almost all public French researchers (3).

This memorandum stipulates that signatories wished to acquire the necessary means to identify, disseminate, develop, promote and monitor the scientific output of their researchers and faculty members, within their research units and laboratories and, where applicable, of affiliated research teams. Joining forces to acquire a common platform for archiving scientific findings seemed the best way to maximize the chances that the project would succeed.

A clearly stated objective was the interoperability of this platform with other open archive repositories meeting the criteria for direct scientific communication such as arXiv and PubMed Central so that the visibility of French research could be enhanced within the international scientific community.

It was recommended that the HAL platform be used as a database to store and disseminate the findings of scientific works. However, it was stated that the common deposit database should be filled either through direct deposit whether or not through a specialized interface such as that defined for HAL-INRIA, or, where applicable, through indirect deposit from an institution's own information system properly interfaced with the HAL tool. This refers to the digital work environment currently under development in French Universities which should provide these universities with a local information system able to manage the totality of scientific outputs (publications but also learning objects).

This memorandum, concluded for a two-year period, also defined the structures that will preside over the realisation of the objectives. Thus, a Strategic Committee is responsible for defining and monitoring the strategic objectives, defining a work plan for the actions to be undertaken and defining, at the end of the Protocol, the most appropriate framework for perpetuating the shared platform. The Strategic Committee appoints the members of the Scientific and Technical Committee, consisting of scientific and technical representatives of the organisations involved in the actions to be carried out, will be responsible for implementing and carrying out the actions according to a work plan defined by the Strategic Committee

Notwithstanding the achievement of the stated objectives, this memorandum can already be considered as:

- An example of collaboration between higher education establishments and research institutions around the common objective of promoting scientific output. This collaboration should be achieved not only while respecting the mutual interests of each institution whether they are scientific or geographical but also while striving to maximize the use of resources allocated to scientific information management through controlled sharing.

- An opportunity to build a larger area of exchange and sharing. Open archives raise a number of issues such as how to ensure the management and the long term preservation of research data, how to ensure the widest possible access to these data, how to foster the appropriation of scientific communication technologies among the various scientific communities, how to have an efficient scientific steering system. This will call for the implementation of structures for skill sharing and even for the setting up of common structures.

Conclusion

The HAL platform is already six-years old. It has undergone numerous technical adaptations since its launch in 2001, especially to satisfy the needs of specific scientific communities whose dissemination

practices of research findings were quite different from those of the physicists that were at the origin of the project. Thus, Life Sciences researchers to whom the notion of preprint was quite foreign but for whom publications could not exist without a link with PubMed or without MESH indexing, could find in HAL a host in conformity with their requirements. Research institutions, to whom international visibility is increasingly important, appreciated the importance of supporting a platform that could adapt to their need for recognition. However, the spirit of the founding fathers did not disappear during these adaptations. HAL remains an open archive platform constituted by researchers and offering a range of services for researchers with as main objective to ensure maximum visibility of scientific works.

There was a real need for an original structure to collect and disseminate research results given the current structure of French research. Its research institutions are located all over the country and they are more and more interwoven with the university environment. The national representativeness of HAL makes the French repository an essential partner in the development of European research infrastructures and HAL is currently involved in the DRIVER, a European project aimed at interconnecting the institutional repositories of five European countries.

References

(1) Rapport sur les politiques nationales de recherche et de formations supérieures. Annexe au projet de loi de finances 2007. Ministère Délégué à l'Enseignement Supérieur et à la Recherche. Imprimerie Nationale, octobre 2006.

(2) Key figures on Science and Technology 2006. OST Economica; edited by Philippe Mustar and Laurence Esterlé ISBN 2-7178-5208-5

(3) Protocole d'accord en vue d'une approche coordonnée, au niveau national, pour l'archivage ouvert de la production scientifique.
Available at : http://hal.archives-ouvertes.fr/more/Communique_de_presse_11_10.pdf last visited 01/03/07