The Centre for Direct Scientific Communication

Daniel Charnay

▶ **To cite this version:**

Daniel Charnay. The Centre for Direct Scientific Communication. Open access to scientific and technical information : state of the art and future trends, ICSTI/INSERM/INIST, Jan 2003, Paris, France. pp.133-137. hal-01279410

HAL Id: hal-01279410
https://hal.science/hal-01279410

Submitted on 26 Feb 2016

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.

Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

Information Services & Use 23 (2003) 133–137
IOS Press

The Centre for Direct Scientific Communication

Daniel Charnay
Deputy Director, Centre for Direct Scientific Communication, CNRS, France

Abstract. This paper describes the Centre for Direct Scientific Communication (CCSD) of the French National Centre for Scientific Research (CNRS) and focuses more on the "container" rather that on the "content". Dedicated to open access and the archiving of scientific publications, the centre has a threefold mission: creation of full-text international databases, self-archiving by scientists, long-term preservation. To achieve this mission, the centre relies on two servers, HAL for articles and TEL for theses, and on its mirror sites for ArXiv and PhysNet. The underlying technology used is the Eprint technology adapted to French. The centre also relies strongly on its collaboration with ArXiv at Cornell, CERN in Geneva and INIST and the MathDoc network in France.

1. Introduction

I will be providing a slightly different view from what has already been presented. That is, I will be discussing the "container" rather than the "content" itself.

2. The Centre for Direct Scientific Communication

The Centre for Direct Scientific Communication (CCSD) is a structure created by the CNRS for the archiving of scientific documents. It is still a small unit, made up of 2.5 engineers.

2.1. Objectives

It has three main objectives. First, to create international databases for use by the outside world and not just the CNRS. Second, the underlying principle is one of self-archiving, that is, researchers archive their own documents in these databases. Third, it is concerned with long term archiving. This is why the CCSD is hosted by the IN2P3 laboratory Computing Centre that is normally concerned with the migration of data.

The metadata collected from the material submitted by the researchers enable us to build up managing tools for bibliometric studies.

2.2. Services

The CCSD provides two main services. It has two general servers: HAL (Hyper Articles on Line) and TEL (Thesis on Line). It also has three specific servers: Jean Nicod (cognitive sciences), ArchiveSIC (information sciences), and Prélude (Education). In addition, it has mirrors for ArXiv and PhysNet.

0167-5265/03/$8.00 © 2003 – IOS Press. All rights reserved

2.3. Partners

In order to set up the CCSD, our main collaboration was with Paul Ginsparg's team, using the ArXiv model, which is well known in physics. We also had a very close collaboration with the MathDoc Unit, presented earlier today. We also work with CERN in relation with submission of documents to ArXiv. All of the data and metadata we collect are shared with other databases such as those held by INIST and CERN.

2.4. Hyper Documents on Line (HAL)

The most significant aspect of the Centre is the HAL system. Initially, the aim was to obtain documents from French and overseas institutes and place them on ArXiv. This was the case for all documents that corresponded to the disciplines stocked on ArXiv. We used the same formats and rules as developed by Ginsparg, although we tried to recover more metadata than was done with ArXiv. We wanted the archive to be multidisciplinary.

Our metadata is thus more extensive, and we created an extension that allows the recovery of metadata on its own. While our customers were pleased with the system, they wanted private views of all that is stored. Buffers are thus used to allow people to obtain private views on their own web servers. The system was launched in January 2003 and we does not, as yet, include many articles.

HAL is a standard database, using Eprint-type technology. It provides access to full texts and to metadata. Our main partner is ArXiv, and all documents that are stocked at the Centre will be almost simultaneously available on ArXiv. Our other major partner is going to be INIST, which is also fed by the

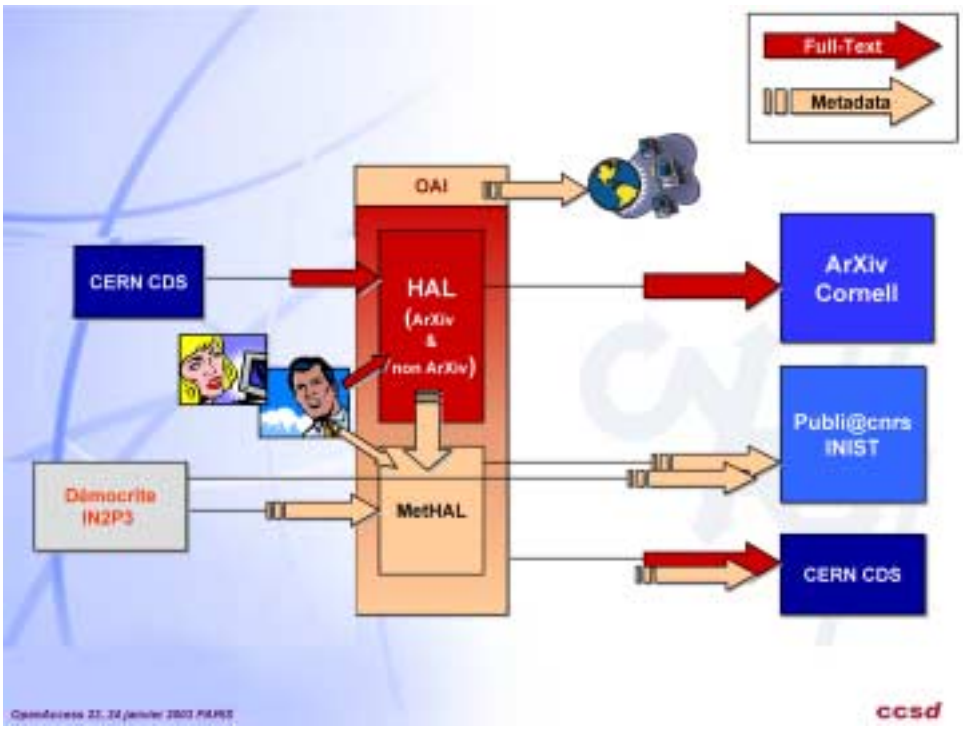

Fig. 1. The partners in HAL.

metadata recovered. We are able to carry out massive inputs of data for institutes that require this service. We also work with CERN, providing direct access from their CDS interface to HAL. We are OAI compliant. Documents are exported in XML.

2.5. Theses On Line (TEL)

We tested a number of products and decided to launch a service of theses online for mathematics and physics, with underlying Eprint technology. This took the form of a collaboration with MathDoc. We have currently stocked about 500 theses, which are fed into the database directly by the author.

As for HAL, users can access personalised pages, viewed directly from their own web sites. They can also re-design the pages if they wish.

3. Submitting material

For both HAL and for TEL, any material submitted is controlled to ensure that it is what it says it is. However, we do not carry out a scientific evaluation. The system is compliant with the OAI 2.0 protocol.

3.1. Verification

Anyone is able to deposit documents in these databases. However, they need to be identified through the creation of an account prior to any deposit. The creation of an account does not involve a high level of verification. We simply verify the domain in which the author works. Buffers allow users to select and identify articles. Authors can deposit the same article in as many versions as they wish, and all versions are accessible to users.

3.2. Formats

Our aim was to make the architecture as simple as possible. We encourage authors to deposit documents in their original format. Given that physicians usually work in TeX and LaTeX, we will accept this format. If the original format is not available, we will accept PDF. The server adapts itself to the discipline, and can accept HTML, RTF, and even audio files, for example.

In relation to XML, we would have liked to store the document itself in XML, which would have made our task easier. However, we are unable to do so due to its incompatibility with mathematical formulae.

3.3. Technologies

HAL uses standard technologies, such as MySQL and PHP codes. TEL uses Eprint 1.1, as adapted by our engineers to make it more user friendly and French language specific. We will soon be moving to the second version of Eprint.

As I stated earlier, CCSD is hosted in a scientific computing centre. We thus have access to stocking and technological migration of data, if required. We can thus provide for long term saving of documents and technological migration, and offer our users personalised views of the relevant information. We are compliant with the relevant standards. Our aim is to increase both our national and international visibility.

4. Future steps

Our intention now is to move toward other disciplines, such as human and life sciences. We are also interested in retro-digitising older and more prestigious theses into our thesis database. We will aim for better coordination with classic scientific journals through cross-links and referencing. We believe that we can bring an impetus to the creation of new scientific "overlay journals". We have an international vision and do not want to be perceived as being a French organisation for the French speaking community.

We advise researchers not to necessarily set up individual archives in their own laboratories. This is technically and easily achievable. However, such archives are not sustainable in the long term as they often rely on the presence of a particular individual. Once that individual leaves the laboratory, the archive goes out of use. While we do not want to create a totally centralised system, an organisation such as the CCSD can provide a solution to researchers who want their data archived.

Relevant information on the CCSD can be consulted on our website: www.ccsd.cnrs.fr. And on the subsites:

- tel.ccsd.cnrs.fr
- hal.ccsd.cnrs.fr
- archiveSIC.ccsd.cnrs.fr
- JeanNicod.ccsd.cnrs.fr
- arxiv.ccsd.cnrs.fr
- www.ccsd.cnrs.fr/physnet

Discussion

From the floor

As a physicist, why should I send data to HAL rather than to ArXiv?

Daniel Charnay

We are aiming to obtain as much metadata as possible. For example, the CNRS physics and mathematics departments are not capable of measuring what is in the archive. HAL is transparent and the use of bibliometry tools is an advantage.

Franck Laloë, CCSD

I would add that ArXiv is a very good system. However, it is not very stable or technically viable, and needs to be improved and revamped. It is likely that, in the long term, the CCSD interfaces will be used for all physics databases in the world.

Bernard Lang

You stated that HAL is driven by international cooperation. The duplication of different archives in different continents, and even within countries, is crucial in this context. The issue you raised of individual researchers archiving material in their own laboratories can be resolved through the use of free open access software. Are these individual researchers able to work directly off your system? Can their work be integrated into the work of the Centre?

Daniel Charnay

The use of open access software should be possible in one year's time. Currently, the source codes are under development and are not yet available. We will have a stable version in one year's time. The

software will then be available on open access. In relation the issue of duplication, this is one of the underlying principles of ArXiv, which has a number of mirror sites around the world. Currently, the material stocked in the Centre that does not form part of the ArXiv disciplines is not duplicated.

Jean-Claude Guedon

In order to resolve the gap between highly centralised systems and decentralised systems, a system is being developed at Stanford: the LOCKSS system. It allows all computers in the network to automatically share and exchange documents. It can be characterised as a dynamic mirror system, and also ensures automatic migration of documents in systems that are evolving. The software is currently being finalised and exists as an open access code. No one in France has yet shown a desire to participate in the Stanford experiment.

Daniel Charnay

This is an extremely interesting piece of information.