« Open Archives (OA), the expectations of the scientific communities »

Open archives in all disciplines: the Hal project

CERN, OAI-5 meeting (20th of April 2007)

Franck Laloë, laboratoire Kastler Brossel (ENS, Paris)
Muriel Foulonneau (CCSD, Lyon)
Marco Picco (LPTHE, Paris), Daniel Charnay (CCSD), CCSD team.
Outline

• OA: a communication channel that comes in parallel and complements peer-reviewed journals

• What scientific communities expect from OA

• Additional requirements introduced by interdisciplinarity

• The Hal project: objectives

• The Hal project, present state
Open Access and Open Archives
(OA vs. OA, in particular OAP)

Open access includes two very different concepts:

- Open archives, which are (open, open)

- Open Access Publishing (OAP), where the traditional scheme (open, pays) is replaced by (pays, open)

« author/institution/consortium pays » financial model; see Jens Vigen’s talk

I will discuss only open archives/self-archiving, « direct scientific communication » (DSC)
Direct scientific communication (DSC): a revival

DSC is a revival of the old tradition of direct exchange of (hand written) letters between « natural philosophers »

Peer-reviewed journals, as we know them today, started around 1940-50. Peer review is very useful (selection, improvements suggested by the referees, etc.); it also has a cost. Someone has to pay for the process, either from the author of reader side.

Open archives (self-archiving) are about 15 years old (ArXiv). They have other complementary features that makes them very useful also. Research needs them too!

The two channels seem to coexist rather peacefully.
Useful features of OA

• Instantaneous free access to all the scientific information in a field of research in one place

• Good for research in developing countries

• More direct control of the author(s) onto the text

• Versioning (OA are no longer only preprint servers!)

• etc.

• This is all well-known. But aren’t there other features that the scientific communities expect from OA?
The expectations of scientific communities (1)

• Garanteed free access to full text; garanteed file readability

• Homogeneity of the scientific content

• Connected to international bases (ArXiv, PMC, etc.)

• Stability of http addresses: essential to be able to cite the work; no withdrawal of documents (if only to avoid fraud), but versioning possible (as errata in journals)

• Long term stability: migrations should be performed by specialized centers in time (hard and soft).
The expectations of scientific communities (2)

• Auxiliary services: automatic extraction of lists of publications (authors, labs), of the scientific production of an institution, etc..

• A single deposit, even if the article results from a collaboration, or if the lab belongs to several institutions. In many fields of research, it is rare that an article « belongs » to a single institution.
Additional requirements from pluridisciplinarity

• Some disciplins have their own requirements: MESH compatibility for life sciences, ADS bibcode for astrophysics, REPEC for economy, etc..

• They may also wish to have their own view of the generic archive (e.g. humanities)

A pluridisciplinary archive must treat the metadata, the protocols, etc.. in a disciplin-dependent way.
What scale is appropriate for an OA?

Many choices are possible: personal sites, lab site, department, institution, consortia of institutions, country,… planet, galaxy. Yes, small is beautiful.. but with a short lifetime!

Institutional level: a possibility has advantages, but:

- not cost/labour effective
- not easy to implement single deposit
- danger: Babel tower (contents of metadata)
- danger: mixing various types of documents (reports, data, etc.)
- too « provincial » (narrow) for scientists!
The Hal project (1)

In view of all requirements, a consortium of institutions seemed to be the best solution (for France). The Hal project started in 2000.

Objectives:

• offer to all disciplines the tools that are so useful in physics (ArXiv model)

• collecting full text scientific documents/articles and PHD theses

• be useful to scientists by guaranteeing a reasonably homogeneous scientific quality (independently of the editorial destiny of a document)
The Hal project (2)

• Not restricted to French scientists!

• Research oriented. Users = researchers

• Covers virtually all disciplines

• Imports metadata from PM and ADS, exports full text documents to ArXiv and PMC, etc.

• Exports metadata through many protocols (OAI, Repec, etc.)

• Also provides auxiliary administrative services
The Hal project (3)

• Hal is part of the DRIVER initiative

• Combines institutional and scientific perspectives through institutional portals built inside Hal, with a large flexibility (provided a core of rules is obeyed); institutions may have different policies towards AO within Hal (none, recommendation, mandate, ..)

• Validations « stamps »

• Webservices

• Many other features
Scientific domains contained in Hal
Hal institutional portals

DEPOSIT

Generic interface

INSERM

INRIA

HEP

HAL

READ

Generic interface

INSERM

INRIA

HEP

PMC, ArXiv, REPEC, ..
Stamps provide a kind of validation/authentification: the owner of a stamp can stamp (or unstamp) any document contained in Hal, to create his/her own extraction, a « stamped collection »

Typically, a lab (or institution) manager will stamp the documents that belong officially to its production, and use the link to this collection on the site of the institution.

A given document may be belong to an arbitrary number of collections. Stamping is made version by version.

Should be useful to create « epijournals » (overlay journals)
about 15% of documents are PHD theses
The French scientific output is about 10 000 articles per month
Two extrapolations

Dépôts sur Hal, deux extrapolations

- Fit exponentiel : $\exp(\text{années}/1.526)$
- Fit puissance : $\text{années}^{3.4}$
French contribution to ArXiv

Change from 2000 to now:

General: from 5 to 7%
Maths: from 9 to 12%
Computer Science: from 5% to 17%