Dissemination of scientific results in High Energy Physics: the CERN Document Server vision.

A. Pepe, J-Y. Le Meur, T.Šimko*
CERN, Geneva, Switzerland

Abstract

The traditional dissemination channels of research results, via article publishing in scientific journals, are facing a profound metamorphosis driven by the advent of the Internet and broader access to electronic resources. This change is naturally leading away from the traditional publishing paradigm towards an archive-based approach in which institutional libraries organize, manage and disseminate the research output.

Within this context, CERN has been committed since its early beginnings to the open divulgation of scientific results. The dissemination started by free paper distribution of preprints by CERN Library and continued electronically via FTP bulletin boards and the World Wide Web to the current OAI-compliant institutional repository, the CERN Document Server (CDS). By enforcing interoperability with peer repositories, like arXiv and KEK, CDS manages over 500 collections of data, consisting of over 800,000 bibliographic records in the field of particle physics and related areas, covering preprints, articles, books, journals, photographs and more.

In this paper we discuss how the CERN Document Server is becoming a solid base for the collection and propagation of research results in high-energy physics by implementing a range of innovative library management services. In particular, we focus on metadata extraction to create information-rich library objects and groupware and collaborative features that allow users to comment and review records in the repository. Moreover, we explain how the existing document ranking techniques, based on usage and citation statistics, may provide original insights on the impact of selected scholarly output.

INTRODUCTION

The divulgation and long-term preservation of research results are two crucial objectives for the management of scholarly communication. Historically, these tasks have been accomplished via publishing: distribution of paper copies allows the propagation of the research output and their storage at libraries ensures retention and availability over time. The advent and large-scale spreading of the Internet, combined with a broader access to resources, are naturally shifting the traditional publishing paradigm towards an electronic archive-based approach.

Within this context, a number of software technologies, protocols and applications emerged in the past few years to allow libraries to set up their own electronic repositories. At present, a large (and growing) number of academic and research institutions worldwide have established their own repositories and thus store, manage and share documents and multimedia material in electronic form. In this archive-based approach, document preservation and divulgation are maintained via compliance to widely recognised storage formats (e.g. PDF[1]) and interoperability protocols (e.g. OAI-PMH[2]) that permit to share document copies among repositories.

Very importantly, the recent intensification and growth of digital archives is naturally driving a widespread movement aimed at making literature open access (OA), i.e. "digital, online, free of charge, and free of most copyright and licensing restrictions" [1]. The two primary vehicles to openly deliver research articles are OA repositories, set up by libraries, institutions or individuals, and OA journals, that offer peer-review and free of charge access to their content. In this perspective, libraries — naturally prone to give free access to scholarly material to anyone — find themselves in the best position to lead the metamorphosis to "more open" publishing solutions. On the other hand, publishing houses, funded by costly subscriptions by libraries and individuals, are hesitating to adapt their policies to the new tendency for a number of reasons, partially because of the momentum of the traditional model, partially because of unsure monetary issues since "open access is a kind of access, not a kind of business model" [1]. The open access initiative is still the subject of very much enthusiasm as well as controversy.

In the field of high-energy physics (HEP) and related areas, subject repositories such as arXiv.org [2] have been successfully in place for many years and they act as an efficient means for the free distribution of electronic preprints of scientific papers. CERN's institutional repository, the CERN Document Server (CDS) [3], is in a way "orthogonal" to arXiv.org as it hosts documents in the same field, but at the same time, it offers a range of library management services such as a user-friendly interface, powerful search functions and collaborative features.

* [Alberto.Pepe, Jean-Yves.Le.Meur, Tibor.Simko] @cern.ch

[1] Portable Document Format
[2] Open Archive Initiative - Protocol for Metadata Harvesting

DISSEMINATION OF SCIENTIFIC RESULTS AT CERN

A bit of history

The dissemination, management and preservation of the results coming from high-energy physics experiments has been at the core of CERN's policy, since its creation.

- **1954**: a policy for the dissemination of research results is already present in the CERN Convention: "..the results of its [CERN's] experimental and theoretical work shall be published or otherwise made generally available" [4].
- **1954-1960s**: from the very beginning of the existence of the research institution, the CERN Library operates a document archive and free paper distribution of preprints.
- **1965**: the CERN Library introduces its first computers to facilitate document classification.
- **1990-1993**: with the advent of the Internet, the preprint distribution continues electronically via FTP bulletin boards.
- **1993**: the first institutional repository, the CERN Preprint Server, starts its life on the web with two original collections: CERN preprints and SCAN series, composed of physics papers received from external institutions.
- **1996**: the Preprint Server becomes WebLib, the CERN Library server, using the same underlying software to provide access to periodicals, books and most of the material kept in the library.
- **2000**: multimedia material such as photos, posters, brochures and videos produced at CERN are integrated into the repository, that is renamed CERN Document Server (CDS). The software that powers the archive is packaged for the first time, under the name of CDSware.
- **2002**: CDS adopts the Open Archive Protocol [5] to expose its metadata in the OAI-PMH format to thusly enhance and facilitate remote access to content.
- **2003**: a policy document [6] is issued in order to reinforce the habit of self-archiving.
- **2004**: CERN signs the Berlin Declaration and the Organization is officially committed to open access.
- **2005**: CERN's Executive Board approves the new CERN policy on Open Access [7]. CERN moves "a step forward for open access publishing" [8] by bringing together high representatives from physics publishers, laboratories and funding agencies to promote open access and demand the creation of a task force.

Access to scholarly output: the physics scenario

When it comes to Open Access to literature the physics community has an excellent record of conduct. Although physicists still submit their articles to traditional subscription journals, they tend, in most cases, to also upload their preprint "to the open", i.e. to a subject repository such as arXiv.org or to their institutions' repositories, like CDS for CERN. In this way, although the final, peer-reviewed, version is only available in the published (or electronic) journal, its preprint (and, in most cases, the corrigenda[3]) uploaded directly by the author, are available for free to anyone.

By operating in this fashion, the HEP community has been succeeding in quietly lowering the barriers set out by costly journal subscription. In the light of this, it can be said that particle physicists are definitely in the best position to "lead the way in a paradigm shift in scientific publishing to give everyone free access to research results" [9].

One could ask herself why despite such popularity of self-archiving among physicists, submission of articles to traditional subscription journals continues. The main reason behind this boils up to author recognition: the peer-review mechanism operated by the journal guarantees the quality of its articles and rewards the authors with the prestige of having published in the journal.

Thanks to the growing influence of the open access initiative, the well-established traditional publishing model is experiencing some changes, the proof being the multiplication of "non-profit" journals, i.e. journals with low or no subscription fee and the increasing interest towards open solutions from commercial parties (e.g. Springer Open Choice [10]). A remarkable example of the former is the Journal of High Energy Physics, an open access peer-reviewed low-cost journal, that has recently reported the highest impact factor among other long-standing particle physics journals [11].

CDS VISION: ARCHIVE-BASED ACCESS AND DIVULGATION

At present, traditional publishing schemes are said to be challenged by an increasing freedom in the provision and access to information. Initiatives like Google Scholar [12], the Scirus project [13] and Wikipedia [14] are catching popular attention and causing some controversial rumors. Despite some widespread skepticism, one thing can be taken for granted: these initiatives are there to stay. In fact, they can only grow bigger and trigger the fostering of other initiatives, since the web is *de facto* a global information space predisposed towards open information access and sharing.

We feel that the position of institutional and subject repositories within this context is not to oppose traditional publishing — still a fundamental means of propagation of serious scientific research — but rather to offer valid services that mimic and complement the functions of publishers in order to drive the transition towards a "more open" access to scholarly material.

With this aim, the CERN Document Server (CDS), and its underlying software, CDSware, have grown in parallel for the last decade. Driven by requests of users and

[3] the differences between the preprint and the postprint

librarians, CDS currently holds more than 800,000 bibliographic records in the field of particle physics and related areas, including 400,000 fulltext documents, ranging from preprints, books, journals to photo, audio and video content. With 20,000 unique visitors performing 200,000 searches per month on average, the CDS portal has grown into one of the largest access points for high energy physics related material.

Besides archiving, classification and powerful searching, CDS is introducing more advanced user and library-oriented functionalities in an attempt to technologically support the shift towards a repository-based approach and thus make the transition more viable. The features described below are either already in place or are currently being developed and planned to be operational in the near future.

Information-rich library objects

In order to ensure long-term preservation and readability of documents, the CERN Document Server adopts worldwide recognised formats to store and organise records.

Currently, the CDS document archive is fed by direct author submission, OAI harvesting from fellow repositories as well as through many other data acquisition channels. In all cases, metadata is converted into its internal library standard format (MARCXML) whereas fulltexts are converted into PDF. Automated and manual procedures to assure the quality and correctness of the metadata gathered are applied then. Before metadata and fulltext are uploaded into the bibliographic and document servers, a final step in the data acquisition workflow takes place: the fulltext is analysed and important information such as citations and keywords is extracted to enrich the metadata.

A detailed account of the information extraction process is certainly outside the scope of this paper; however, it is interesting to outline the fact that both the keyword and the citation extraction is performed upon subject knowledge-bases in order to control and validate the accuracy of results. The retrieval of such information participates in creating content-rich, well-documented library objects. In particular, automatically extracted keywords can aid indexing and classification whereas citations can prove as a valid indicator to define the impact of a document (see subsection *Citation and usage statistics*).

Collaborative and groupware features

A clear advantage of digital repositories over traditional publishing channels is that active discussion and reviewing of published content can be conveniently done online. This type of review mechanism, referred to as "open-review" [15], is often compared with the peer review operated by most journals to select and to screen article submissions. Yet, these two means of debating the validity of a document are fundamentally different for a number of reasons, one above all being the fact that peer review is performed by peers, academics and scholarly professionals, whereas open reviewing is generally open to anyone (e.g. Amazon-style commenting and reviewing of books). Moreover, peer review is generally performed in order to filter submitted material — it is done *a priori* — whereas open reviewing is generally triggered *a posteriori*, after the submission of the material into a repository.

An open reviewing system adopted on top of a document repository can have the following advantages:

- it is fast — response from readers is immediate
- any user is allowed to express her opinion on a document — reviewers are not appointed by an external body
- reviews may be signed — reviewers are more careful about what they say
- comments and reviews can be made public or private, according to the situation
- users can join private and public groups — different access privileges and restrictions can apply to different groups
- users can privately interact with each other using an email-like web messaging system
- documents, as well as users and reviews, can be rated

On the other hand, open reviewing is still subject to a lot of controversy, especially due to the difficulty of moderation, correctness of information and the threat of vandalism[4]. Despite the controversy, we believe that open reviewing can be certainly regarded as an added-value to the peer-review process and, in the absence of peer-review, as an interesting alternative.

The CERN Document Server will shortly adopt a comprehensive system of commenting, reviewing and messaging that will allow users and groups to discuss content and share knowledge privately and publicly.

Citation and usage statistics

The adoption of citation and usage data as impact indicators has become increasingly more popular and accepted. A concrete example of how the traditional academic world inspired the development of global information sources might be the successful Google PageRank algorithm [16] that ranks web pages due to the numbers of pages linking to them, similarly to the citation techniques of the academic world.

Quantitative study of citations has a long history in the field of bibliometrics, a subfield of library and information science [17]. It has been used for over fifty years now, as an effective means of measuring the impact of selected authors/articles/journals. Nowadays, several projects aimed at the extraction and analysis of citation data exist (e.g. citebase [18]). At CDS, a citation index is generated through the extraction of references from fulltext. The index is then used to rank documents according to the number of times it is cited by or co-cited with other papers. The

[4] These are all recurring problems common to most collaborative writing technologies, such as wikis

system can be used to provide interesting insights on the impact of scholarly material, e.g. comparing article impact factors vs. the averaged journal impact factors.

Usage statistics is a somehow more recent impact indicator, as it is based on the analysis of the user habits when accessing metadata and fulltexts on the web. Yet, it has already suscitated a considerable activity and research (e.g. the Los Alamos initiative [19]) and attempts for multinational collaboration (e.g. the European initiative on alternative metrics [20]). At CDS, ranking of documents based on the number of downloads can be performed via analysis of access logs, that are anonymized to respect the viewers privacy. Based on average access stats, a useful information such as "people who viewed this document have also viewed" is proposed, permitting the end users to discover new sources of information related to their interests. (This is akin to Amazon.com's "people who bought this book have also bought" functionality.)

The analysis of usage and citation statistics can prove a valid means to give authors recognition and prestige. For example, it has been shown that the Open Access publishing usually provide higher impact of the scientific work on the community when compared to the traditional non-OA publishing [21, 22].

CONCLUSIONS

The traditional publishing paradigm is undertaking a profound change, impelled by the proliferation of subject and institutional repositories and a broader availability of electronic resources and tools.

In the field of particle physics and related scientific disciplines, subject repositories such as arXiv.org have successfully been in place for some time, offering access to preprints and proving the real access points for acquisition and discussion of latest research results.

CERN's institutional repository, CDS, has operated for over a decade now, providing not only open access to a vast amount of scholarly material in the field of particle physics and related areas, but also a range of advanced user and library-oriented services aimed at supporting the proliferation of open, interoperable, archives.

In this paper, we have briefly discussed the technology of such features, namely information-richness of library objects, social tools and alternative impact indicators, and the benefits that could originate from their widespread usage.

REFERENCES

[1] Peter Suber. Open Access Overview. http://www.earlham.edu/ peters/fos/overview.htm

[2] arXiv.org. http://www.arxiv.org/

[3] CERN Document Server. http://cdsweb.cern.ch/

[4] CERN Convention of Member States. http://doc.cern.ch/archive/electronic/other/preprints//CM-P/cm-p00046871.pdf

[5] Open Archives Initiative. http://www.openarchives.org/

[6] An electronic publishing policy for CERN. http://library.cern.ch/cern_publications/SIPBPubPol.17.11.03.htm

[7] Continuing CERN action on Open Access. http://library.cern.ch/cern_publications/CERN_exec_board_23.03.05.html

[8] A step forward for open access publishing. http://press.web.cern.ch/press/PressReleases/Releases2005/PR18.05E.html

[9] K. Peach, Join the open-access revolution, CERN Courier. June 2005.

[10] Springer Open Choice. http://www.springer.com/sgw/cda/frontpage/0,,1-40359-0-0-0,00.html

[11] 2001 JCR Science Edition. Institute for Scientific Information (ISI)

[12] http://scholar.google.com/

[13] http://www.scirus.com/

[14] http://wikipedia.org/

[15] F. Godlee. Making Reviewers Visible. http://jama.ama-assn.org/cgi/content/full/287/21/2762

[16] Google Technology. http://www.google.com/technology/

[17] S. Redner. Citation Statistics From More Than a Century of Physical Review

[18] http://www.citebase.org/

[19] J. Bollen, H. Van de Sompel. A framework for assessing the impact of units of scholarly communication based on OAI-PMH harvesting of usage information. http://library.lanl.gov/cgi-bin/getfile?LA-UR-05-8420.pdf

[20] International Workshop on institutional repositories and enhanced and alternative metrics of publication impact. http://www.dini.de/veranstaltung/workshop/oaimpact/

[21] S. Harnad, T. Brody. Comparing the Impact of Open Access (OA) vs. Non-OA Articles in the Same Journals. http://www.dlib.org/dlib/june04/harnad/06harnad.html

[22] T. Brody et al. The effect of Open Access on citation impact. http://opcit.eprints.org/feb19oa/brody-impact.pdf