# CERN Document Server. Institutional Repository and Service. CERN, Geneva, Switzerland: Critical mass gained! Depositing and reclaiming particle physics content

**CERN Document Server: Institutional Repository and Service. CERN, Geneva, Switzerland**[1]

**Critical mass gained! Depositing and reclaiming particle physics content.**

http://cdsweb.cern.ch/

**Background**

CERN is a unique organisation in the Institutional Repository (IR) community from a number of stand-points. CERN is an international research organisation which serves a specific subject community, i.e. that of particle physics with a longer tradition in self-archiving since 1991. Approximately 7,000 authors and 25,000 readers, just over a third of whom are mobile and working off-site, are served by the library. It is the library which manages the CERN IR Document Server. CERN has two pillars to its strategy for open access, one being self-archiving of its research results, and the other promoting open access publishing.

*Self-archiving*

Since CERN's establishment in 1953 and as part of it's constitution, all CERN's research results should be published or made otherwise publicly available ("an early concept of OA", says Jens Vigen, Group Leader Scientific Information Service and Head of the CERN IR). This has meant that preprints and reprints have been deposited via the library since day one. CERN aims for 100% coverage of its current research. It is well on its way reporting 79% coverage in 2005. CERN aggregates academic output not only from its employees, but also from those using CERN facilities. It also serves as an archive and access point for CERN management information.

CERN has tried to give its IR a place within its research community. It has expanded its IR by developing a global information service on particle physics, bringing together world content with its own output to obtain a critical mass of content. CERN hopes to encourage further local IR deposit by doing this. CERN thereby not only focuses on the infrastructure it maintains, i.e. its IR, but on a service which aims to further serve its authors and readers in information retrieval. This is a difficult aim to fulfil considering the other services which exist containing similar content and serving the same types of users. arXiv.org and SLAC SPIRES-HEP have developed services which are well-embedded in the research traditions of the particle physics community for example. [2][3]

*Open access publishing*

However, it seems to be the open access (OA) golden road rather than the green self-archiving one which sees CERN's preference. A place where CERN hopes that all of its future publications will be made OA using the author pays model.[4] For this reason,

[1] This case study write-up was executed as part of the *Stimulating the Population of Repositories* research project which was carried out as part of the European DRIVER project http://www.driver-community.eu/ It was conducted in 2006 and 2007. See http://dare.uva.nl/aup/nl/record/260224 for the publication.

[2] arXiv.org e-print archive: http://arxiv.org/

[3] SPIRES http://www.slac.stanford.edu/spires/

[4] Harnad, S., Brody, T., Vallieres, F., Carr, L., Hitchcock, S., Gingras, Y, Oppenheim, C., Stamerjohanns, H., & Hilf, E. (2004) The Access/Impact Problem and the Green and Gold Roads to Open Access. http://dx.doi.org/10.1016/j.serrev.2004.09.013 *Serials Review* 30 (4) 2004

the CERN library - with CERN management support - is very much focussing on open access publishing. Considerable time has been invested to prepare the road for the OA publishing of all of its CERN research. CERN has interestingly used the launch of its next new physics machine the LHC to gain some bargaining power with its publishers to implement its OA publishing strategy. CERN will insist that future publications can only be open access at a time where publishers are keen and competing for these research results.

**Policy**

Since CERN's establishment, all CERN researchers have been required to publish and/or make their results otherwise publicly available. However, in answer to developments in the OA community, the library formally reinforced this policy by publishing an official mandate and legal document endorsed by the library committee and CERN Director General in 2001. This was then revised and summarised in 2003 to adjust to further OA developments and stimulate active deposit, once again endorsed by the CERN Director-General. This has 2 key principles:1) to promote OA publishing and 2) to direct research departments to submit research results produced at CERN or its facilities to the library. This case study report will focus on the second principle.

It is vital for CERN's library to achieve 100% coverage of CERN"s current and recent research output. According to Vigen, "This is what it must strive for to get onto the golden road of open access publishing" to accurately determine the future costs of OA author-pays publishing. However, when asked how important a mandate was to their IR's population, Vigen interestingly answers, "I do not believe that mandating is critical, although I am glad to have it in place." "It is not critical as in the end you need to convince authors that OA publishing is in *their* best interest to deposit. However, as a librarian I feel more comfortable to pursue reluctant authors with a mandate as a back up." Although policies are in place, and stimulate deposit (approx. 50% of its deposits can be seen as a result of them) the library still has challenges to retrieve the remaining material.

The IR's management structure consists of a Scientific Policy Board (type of academic library committee) which advises the library. It is responsible for policy development and operational issues. New developments, statistics, updates and challenges are presented to this Board several times a year by the library.

CERN has a clear policy to continue working on international levels for both developing added value services / archives for its particle physics community as well as for the further development of its own library. Networking is important to keep track of relevant developments in publishing and information services in the particle physics field. Knowledge exchange is important as it plays an important role in library personal development both on national and international levels, e.g. with its Swiss Library Science Talks, and with its CERN OAI conference/workshops held every 18 months for the OAI community.[5]

In the 2005 CERN Annual Report, the Director-General made the following statement, "The laboratory's commitment to open access publishing, further emphasises its mission to communicate to all members of society throughout the world."[6] This statement, coupled with the reinforced mandate of 2003, and the

[5] Library Science Talks: http://librarysciencetalks.web.cern.ch/librarysciencetalks/
[6] Message from the Director-General, CERN Annual Report 2005, pg 5

investment in its IR and associated services, demonstrates CERN's continuing support through policy for the principles and practice of open access.

**Establishing the IR / service and sustainability**

The CERN IR is managed by the library. The archiving of CERN academic output has been fully embedded in the library since the 1950's. However, historical milestones brought changes as things went digital when the 80's saw the electronic library catalogue take form and the advent of the digital library. In addition, the first web server at Stanford University Library, invented by a CERN researcher in fact, was the forerunning technology of CERN's preprint server set up in 1994.

This server was crucial for setting up CERN's IR in 1996, which brought a simple interface to global and CERN particle physics research. This became the CERN Document Server (CDS) IR global information particle physics portal. CERN developed home-grown IR software (formerly CDSWare, now Invenio) hosted locally using the OAI-PMH protocol and MARC, MARCXML and NLM metadata standards. The IR software was developed by the library together with an independent IT department. CERN hoped to gain extra revenue for IT staff by providing CDSWare support and development to the broader OA community but its ambitious hopes have proved to be a little far-reaching; this activity is therefore under review.

Another key milestone in the development of the IR and the CERN policy was the Southampton OAI Conference in 2005. It was here that CERN management became more aware of the arguments for OA and its significance for CERN at a CERN-hosted event in autumn 2005, entitled *Changing the Publishing Model*. CERN's Director General gave a speech, with 100 researchers and many journal editors in the audience. More hearts and minds were won for both the IR and for the principles of open access.

As a follow up to this meeting, a tripartite meeting and task-force was set up between the key stakeholders in particle physics publishing to identify the elements in place and necessary to convert the existing journal model into OA. This task force consists of particle physics publishers (those supportive or at least sympathetic of OA), funders and researchers. The Chair of the CERN Scientific Policy Committee consequently wrote an article supporting OA, afterwards claiming that he had never had so much positive response to a paper. CERN's investment in OA publishing has also been made to better guarantee the smooth deposit of future CERN output into its IR without publisher obstruction.

Departments are responsible for depositing content into the system mainly on behalf of its authors. This organisational process has been in operation since the IR's establishment. Content not deposited by CERN researchers is harvested by the library from other information providers such as arXiv.org (see *Populating the IR* below). IR activities are maintained by the library with almost 4 full-time staff from the collection development team, a full-time programme head and IR manager working on the IR and its information service, and four technical employees continue developing services. This allows the continued acquisition of new material, completing the collection of CERN output from previous years as well as the development of added value services to encourage further population and to attract the arXiv.org faithful.

*Services*

CERN has a developed service layer on its IR. Various services have been developed to encourage research output deposit as well as to provide digital library user services.

The global CERN database (CDS) has already been mentioned, partly developed for information discovery. It is integrated into the library catalogue. Other related activities include hyperlinked citations and other authors who consulted this document consulted ….. Publication lists are generated automatically from repository input, and reference lists can be exported. These services have been developed to save time on administrative activities. Research development is stimulated by a so-called "basket function" where documents can be stored and shared with colleagues from certain communities where comments can be added to documents. In 2007 and beyond, CERN will aim to create author tools for writing documents and converting them to xml, and tagging them more effectively as a new service.

Increased visibility is of importance to CERN and its authors both in subject-relevant search services and the more generic ones. Although encouraging arXiv.org to reciprocally harvest CERN material has failed so far, SLAC SPIRES-HEP has not. Google and Google Scholar remain important as well. CDS content is picked up by Google, and CERN has agreed on metadata specifications in order to deliver its content to Google Scholar more effectively.

Vigen still sees that there is work to be done. Apart from aiming for even wider distribution of CERN material, the CERN library will collaborate more with other service providers, and will develop more internal services. These include linking primary data to its own publications, offering a service to departmental web pages where publications can be generated automatically from the IR and carrying out data-mining. He also intends on acquiring more conference output for CDS which is thin on the ground.

A summit took place in spring 2007 at Stanford University to look at a) existing services in the area of particle physics and b) where information services will be 10 years from now. CERN is interested in seeing how SLAC SPIRES-HEP, the Particle Data Group, CERN and arXiv.org, for example, can work more efficiently together; Future roles may need defining, and learning new ways of how to better collaborate in the scientific arena may be called for.

As regards digital preservation, CERN is prepared to convert to new formats in the future, but beyond this sees this as a responsibility of national libraries.

To conclude, a mandate is indeed in place, and storing a copy of CERN's research has been in the organisation's constitution since its foundation. but services are additional means to reach the goal of 100% coverage of CERN-produced research.

*Costs and sustainability*

The CERN IR is fully embedded into the library's activities. The running costs of the IR and its service are: 1 Library Director of 0.02 FTE, 1 programme manager FTE, 1 IR manager FTE, 3.75 collection development staff FTEs, 4 programmers FTEs, and 0.2 PR support. Hidden costs are not included.
According to CERN, Since CERN has been developing its knowledgebase since the 50s, calculating development costs does not seem helpful.

## Populating the IR

Policies are well established and defined at CERN and one might therefore assume that population is no longer a challenge; this is incorrect as work still has to be done to persuade just under 50% of CERN researchers to deposit into its IR. Although CERN seems to constantly be achieving approximately 72% coverage of its research output since 2001, which is nothing short of admirable in comparison to other IRs, this is due to both deposit and harvesting from outside sources. For reasons of sustainability, compliance by the majority of CERN researchers to deposit material in the CERN IR is preferable to the library.

In contrast to most research areas, the particle physics self-archiving discipline has been running since the early 90's largely due to arXiv.org. So although the community does not need to be persuaded regarding the value of self-archiving, the challenge that CERN faces is that researchers are somewhat reluctant to deposit with its own IR in addition to a service such as arXiv.org which has served them well.

The organisational structure is in place to allow the self-archiving of CERN material in one archive, and functions well compared with other IRs (approx. 50% of CERN's output was deposited in 2005 in this way). Both metadata and full text is deposited via departmental secretaries who are key to the content acquisition process. Academic output is submitted to this departmental information coordinator or else authors submit themselves or via assistants. Preprints enter the system once given an ID number after having been submitted to a preprint series, which goes through an internal peer review process. The key reasons for researchers contributing to the IR are the mandate, that the work is done for them on their behalf, and increased visibility (user figures show that 70% who consult CERN metadata come from outside of CERN). In addition, the IR seems to make compiling their own publication lists much easier says Vigen.

Journal articles, books, working papers, conference proceedings and presentations, images, audio, visual files and teaching and learning materials, etc. are collected. The only material that the CERN IR has not yet targeted is primary data although this is under consideration. Most material is in the English language. Various versions are stored in the IR, from preprints to publisher PDFs, e.g. 1,500 PDF publisher formatted articles from the American Physical Society (APS). Some authors are sometimes reluctant to submit papers which are very similar in nature, however as online access can make duplicate papers more transparent to the research community. Conference papers therefore constitute a considerable extent of the IR's missing material, as do reports and embargoed material. CERN takes care to only make open access what is legally allowed to ensure publishers - as OA negotiating partners in the future. In 2007, CERN intends on concentrating on retrieving more conference papers as well as post-prints to increase the quality and access to revised versions.

CERN's IR also manages the content of the organisation, serving as a research management information system, including not only academic output, but formal memos, internal reports, CERN report series, lectures, etc. Most of this material is made publicly available except for some internal notes, reports, work in progress and some classified information.

However, a number of researchers do not deposit into the IR themselves nor submit their work for inclusion. It is the group of theorists at CERN who write the most, who submit to arXiv.org and generally not to CERN's IR for example. CERN realised that they needed to fit into the work processes of the researcher, and adapt to more

challenging groups and develop work-rounds to retrieve their content. In order to fill this gap in deposit, CERN further fills its CDS repository with material from outside – both capturing missing CERN material and other material of importance to the international community.

CERN has therefore created a global particle physics service, combining the contents of arXiv.org and about 100 other service providers with its CERN material. It does this by harvesting a mixture of metadata and full text from a number of services identified by a physicist who works for the library. Archives are analysed for relevance, and choices are made as to whether an entire set is to be harvested or a sub-set. CERN presently harvests 80-100 sets. Some are harvested from once a year to once a day as is the case with arXiv.org, and by agreement as long as they do not overload existing services.

"However, this method is not scalable for the future," says Vigen. "It is a good way of getting critical mass, and to be able to better verify what you have and what not. Quality control is still necessary on this mass of content using resources which could be used elsewhere." For this reason, tools have been developed such as an automated email system to alert researchers to deliver missing full text content which has been identified from reference lists for example, reminding him/her to deposit in the future. CERN also intends on using more advocacy to remind first authors to submit their papers.

Quality assurance measures are in place to check the identification of author names and their works when such data is imported into the system after utilising the CERN-developed harvester. Metadata is also enhanced with additional information with names of experiments, topics, etc, by the library and an automatic system checks to ensure that no PDFs are corrupted.

To summarise, CERN has gained a substantial amount of its output either by direct deposit or by reclaiming it back from other information providers. "Had we not aggregated material from outside we would have failed dramatically," concludes Vigen. This challenge is clearly relevant to most communities with an already established subject-community self-archiving tradition. As a consequence, the significance of the size of the repository, with an important core of the particle physics subject-community to hand, may well have an influence on the publishers of the future when talking to them about open access confirmed Vigen.

*Statistics*

The entire archive contains almost 900,000 records as of November 2006; giving access to approximately 360,000 full texts. 80-90% is automatically ingested from worldwide information resources.[7] 43,170 (12%) of those full text documents originate from CERN-authored work between 1952 and 2006. This includes post and pre-prints, conference papers, theses and reports and excludes CERN management documentation. Of the CERN material:

**Academic take-up**
51% of all academic staff contributed to the IR in 2005, 37% indirectly through arXiv.org, and 5% from certain publisher sites with permission.

**Annual academic output** in 2005: 2,210, i.e. 79% of a total of 2,674 documents. The average has been 72% since 2001.

**Total records** by type for 2005 (CERN only): 2,674 records, with 2,210 ft files including 722 conference contributions (33%), 426 articles (19%), 239 working papers/pre-prints (11%), 130 teaching and learning materials, 56 theses, 34 proceedings, 0 dissertations, and 600 other materials including 451 CERN research reports (20%), 98 CERN scientific committee papers, 40 press releases, 11 reports, nn books and nn chapters.
Total records by type as recorded in Feb. 2007 (CERN and external records): 84,330 records, with 54,477 object files including 22,476 working papers (incl. internal notes, scientific committee papers and pre-prints) (41%), 15,759 journal articles (29%), 10,512 conference contributions (19%), 1,078 teaching and learning materials, 522 conference proceedings, 343 chapters, and 138 books.

**Object deposit by year**
Document numbers by year: 2004: 2,546; 2005: 2,210 and 2006: 2,005.

**Freely and openly accessible** in 2005: 100% of ft.

Logs are gathered on the number of hits to the CERN Document Server, as well as to the downloads of full text material contained on total usage, geographical domain, article highest hits and IP numbers. However, logging is carried out at irregular intervals at present. This is fed back to the library management board and library staff.

## Communication / advocacy

CERN did not need to preach to the converted, at least not on the principles of archiving and open accessibility. This has been due to the 10-year pre-print archiving tradition that arixiv.org has nurtured within its community. This has been an advantage. However, CERN has still had to invest time and effort in advocating the use of its own IR and services in a research community which largely uses another to deposit. Although CERN staff regularly present on advocacy at international events, it admits to spending very little time on active advocacy within CERN and more on retrieving external content, quality control and above all OA publishing to lay the foundations for seamless self-archival in the future.

[7] 888,954 as of 28 November 2006

CERN does not have a communication plan which could help organise and focus advocacy activities for the future, however, they do use some essential communication channels to inform its authors. IR issues appear regularly in the CERN e-bulletin bi-weekly newsletter. This raises significant awareness as most employees habitually read this; being probably the most effective PR vehicle according to Joanne Yeomans, Library Section Leader.

Face to face contact is essential for library relations. One to one meetings with individuals focus on acute library-related issues and are utilised by the library to raise awareness of IR activities. In future, the CERN library hopes to step up on advocacy by more structurally presenting at departmental meetings.

The IR's progress and related issues are consistently reported on at library policy board meetings. This proves to be very successful where higher-level CERN staff and researchers can be turned to for guidance.

CERN hopes to invest more in advocacy in the future in order to obtain its IR's missing content and better guarantee further take-up by CERN's research community. A mandate alone will not achieve this aim believes Yeomans.

**Legal issues**

Concerns regarding copyright have played a very minor role at CERN due to the tradition of the pre-print culture in the particle physics community. However, opportunities are often unknown amongst researchers as regards the storage of other versions in IRs. This is relevant for some researchers who might wish to post later versions such as post-prints, but are apprehensive in doing so. The CERN library raises awareness amongst new researchers at least of the opportunities open to them as regards storing publications in the IR and OA at induction days as a first step.

The library only knowingly gives access to post-prints or final articles where publishers allow. This is a strategic decision, seeing publishers as concrete partners for the future OA publishing of CERN research. For this reason, neither deposit licenses nor other licenses for OA distribution have been suggested so far says Vigen. It is important to add, however, that the principles of open access are vital both to the CERN scientific community, and the library, "We will respect publisher limitations but will continue to challenge them," concludes Vigen speaking for CERN as a whole.

**Critical success factors (CSFs) for populating your repository or service**

The following CSFs have been identified by CERN for populating a repository:

- Ally with stakeholders of the system, i.e. departmental administrators, members of the internal peer review process, to enable the efficient deployment of institutional content.
- In order to reach critical mass, although a mandate can boost the acquisition of content, in a community with a self-archiving tradition, it is necessary to negotiate and collaborate with the services which researchers deposit with to reach content acquisition goals, and thus it is essential to:
- Identify important information aggregators and establish agreements to ingest material into your IR (metadata *and* full text where possible) for the efficient acquisition of both new and missing content.
- Take an opportunistic approach – "just do it, don't spend too long thinking about it." Jens Vigen.

**The learning curve**

The following issues have been identified by CERN which can hamper the population of a repository:

- Competitive services exist where researchers have a long tradition in both archiving and utilising information retrieval services.
- A lack of awareness exists in the CERN research community of the value in submitting full text internally to the IR as opposed to other external information services.
- Copyright regulations prevent the aggregation of some full text which is of value to the researcher; i.e. publisher final PDF versions.
- Some authors are apprehensive about submitting several similar versions of one and the same paper and making this transparent through online dissemination via the IR.
- Certain authors underestimate the value of some of their papers, which prevents them from submitting them to the IR.

**Issues for possible further investigation**

1. How far are IR services taken up by the research population?
2. How effective are mandates to achieve the 100% aim?
3. Can successes be reaped by focussing on both open access publishing and self-archiving where resources are concentrated on the former? Is one more important than the other?
4. How cost-effective is the population of the CERN repository?
5. Would more advocacy contribute to an increase in take-up by CERN's research community?
6. Is there a maximum to visibility within specific communities or is there a point of saturation?

**Acknowledgements**

I would like to express my sincere and heartfelt thanks to members of the team behind CERN's repository. I would like to thank them for their precious time and engagement at both the time of interview and throughout the publication process. The persons I would particularly like to thank here are Jens Vigen, Joanne Yeomans and Ingrid Picchioli. These colleagues have been more than willing to share their knowledge to provide the information management community with an insight on their achievements. We are indebted to them.