



Behind the scenes of Open Science with Gigi

Salvatore Mele, CERN
Pisa, Nov 7th, 2025

CERN Scientific Information Policy Board



- Evolution of CERN's original "library committee"
- Reports to CERN DG (later Director for Research)
- "Deals with any matters related to policies and strategies of Scientific Information Services to the HEP community"
- Practically: guidance and advice to library team

Transformational projects under Gigi's tenure

- Open Access
- INSPIRE
- Open Data

Chairs:

John Ellis (TH): 1990–1994

Walter Blum (EP): 1994–1997

Gabriele Veneziano (TH): 1997–1998

Rudiger Voss (EP): 1998–2002

Guido Altarelli (PH/TH): 2002–2006

Gigi Rolandi (PH/EP): 2006–2017

Michelangelo Mangano (TH): 2017–

Open Access

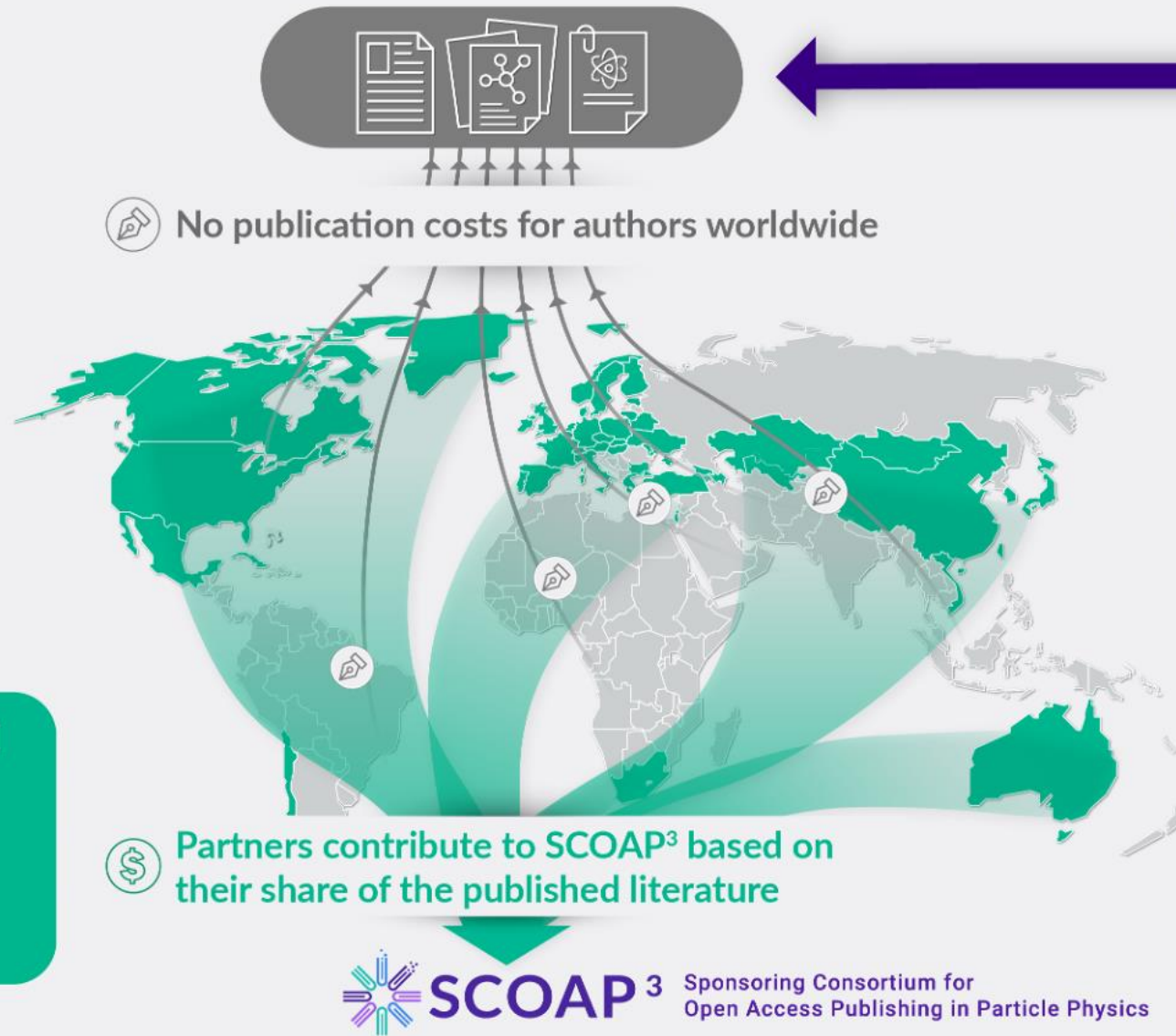
- Very new concept in 2006, emerging in life sciences:
 - Journals experimenting with 'pay to publish', thousands of CHF/EUR/USD
 - Universities running 'repositories' of 'preprints'
 - Considerable amount of polarization, confusion, attrition, risk
- Very special conditions in HEP:
 - Everything on arXiv.org BUT library purchase expensive (paper!) journals
 - Risks to 'leave behind' some parts of our community
 - The (many!) agencies supporting LHC considering (multiple!) requirements
 - Opportunity to leverage HEP's collaborative tradition and CERN strengths
- A disruptive solution for (HEP and) CERN theory and HEP experimental papers
SCOAP3
- A new approach for non-HEP papers at CERN?
A policy. Agreements. Funding.

The SCOAP³ Model: How It Works

SCOAP³ centrally underwrites Open Access to research in high-energy physics, enabling free publishing, global access, and re-use



Everyone around the world can access and reuse any SCOAP³ article for free



SCOAP3 2014-2025: 79'000 articles, 13 journals, 1 00K+ authors
3000+ partners in 37 countries (agencies, labs, libraries, universities)

List price APCs of comparable journals

Phys.Rev.Letters (APS)

EPJA (Springer)

Phys.Rev. D (APS)

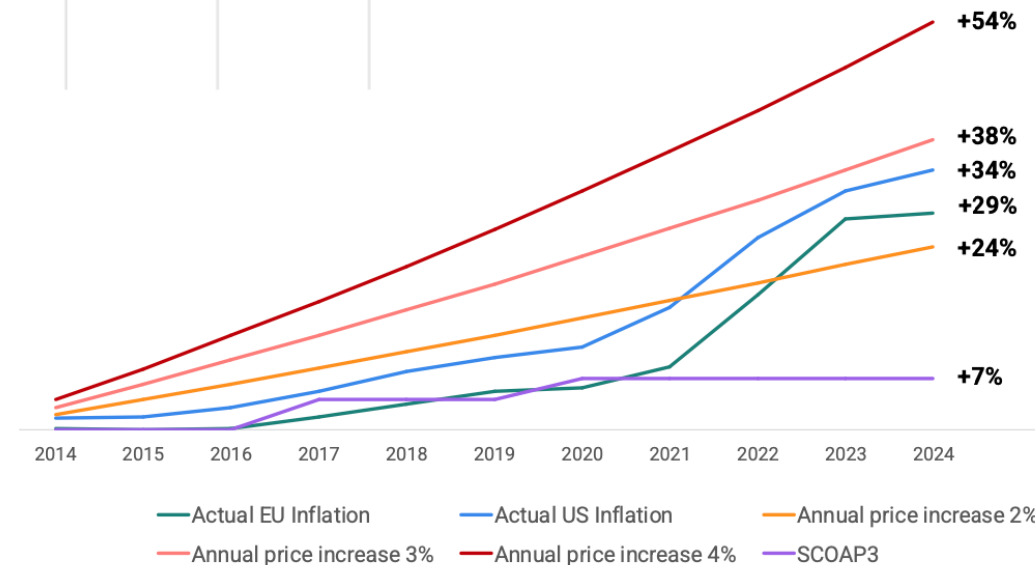
JCAP (IOPp/SISSA)

Nucl.Phys.A (Elsevier)

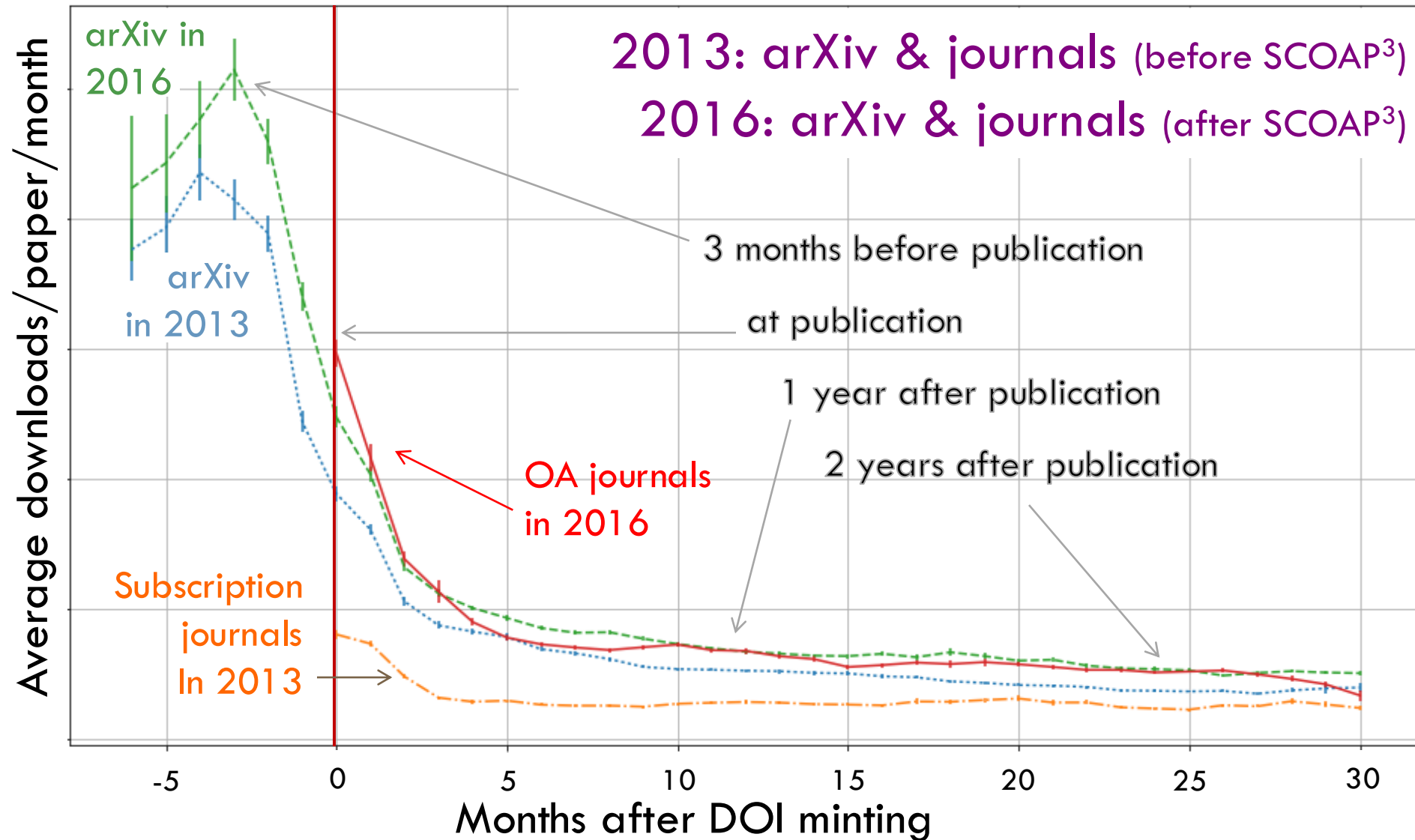
Effective cost per article for SCOAP³

SCOAP3 average all phases

- Low costs vs. pay-to-play Open Access
- >75% of funds from subscription redirection
- Managing inflationary pressure vs. increase in subscriptions



SCOAP3 tripled downloads of journal articles



Downloads 3Q13, 4Q13, 1Q16, 2Q16 on arXiv.org and publishers' platforms

50k non-Open Access articles and 8k Open Access articles

Elsevier: *Phys.Lett.B*, *Nucl.Phys.B*; Springer: *Eur. Phys. J. C*, *JHEP*

CERN's first Open Access policy and its long-term results

Approved by the Director-General - 16.10.2014

Open Access Policy for CERN Physics Publications

CERN is committed to Open Access, it represents values written in our Convention sixty years ago, and it is increasingly important for our Member States. For over twenty years, most of our Physics publications have been immediately available online as CERN-PH-TH or CERN-PH-EP preprints before publication in journals. This is generally called “Green” Open Access.

“Gold” Open Access journals go a step further and make articles available to readers free of charge. These journals are financed by a per-article fee, called Article-Processing Charge. Through 2013, the Scientific Information Service has centrally covered such fees for certain journals. In addition, thanks to partnerships with most publishers, all LHC experimental articles have been published as Gold Open Access since 2010.

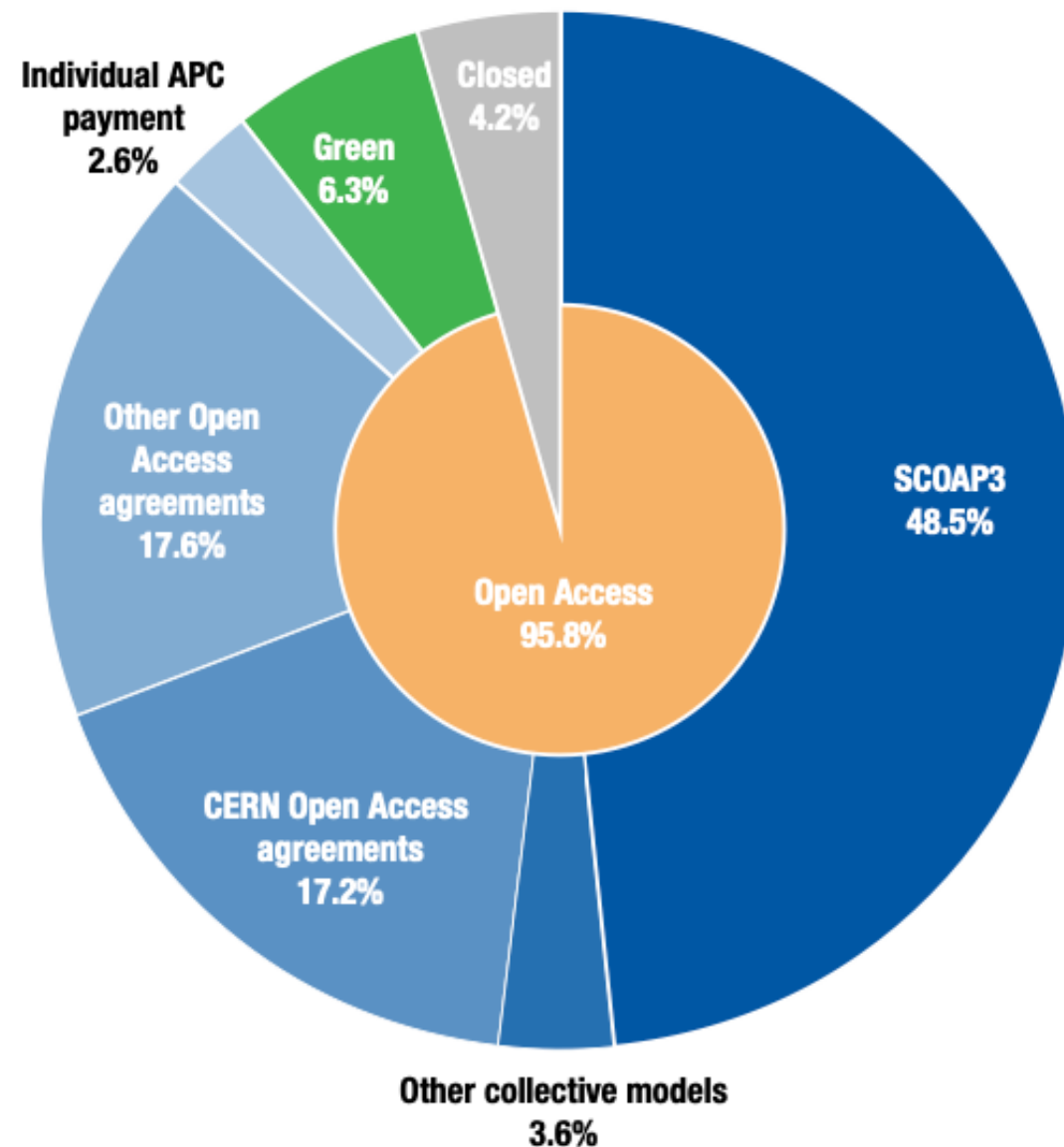
As of January 2014, CERN and international partners in over 38 countries launched the SCOAP3 initiative, which has converted to Gold Open Access most journals in the field at no cost for any author worldwide. Recently, CERN and the American Physical Society (APS) announced a partnership¹ to publish all CERN articles in APS journals Gold Open Access in 2015 and 2016.

CERN authors² are now requested to publish all their results Gold Open Access.

While all LHC physics articles are Gold Open Access, only 60% of CERN physics results were published as Gold Open Access in 2013. Thanks to the SCOAP3 initiative and the additional partnership with APS this figure will rise to 90-95% in 2015. CERN now aims to reach 100% Gold Open Access for all its original High-Energy Physics results³, experimental and theoretical, by the end of 2016.

The easiest way to publish Gold Open Access is to prefer journals covered by SCOAP3⁴ [<http://scoap3.org/scoap3journals>] or through Open Access arrangements where no additional fee should be paid by CERN. The Scientific Information Service can help to understand the situation.

When circumstances require publication in any other journal, Open Access fees can be paid with funds external to CERN, e.g. E.U. projects. Where this is not possible, authors are requested to get permission from their Department Head.



*The open access share of CERN published articles in 2023/2024
Number of articles: 2227 (not including conference papers)*

[Institutions](#)[Experiments](#)[Journals](#)

Discover High-Energy Physics Conte

INSPIRE is a trusted community hub that helps researchers to share and find accurate scholarly information

literature ▾

[How to search?](#)

BROUGHT TO YOU BY

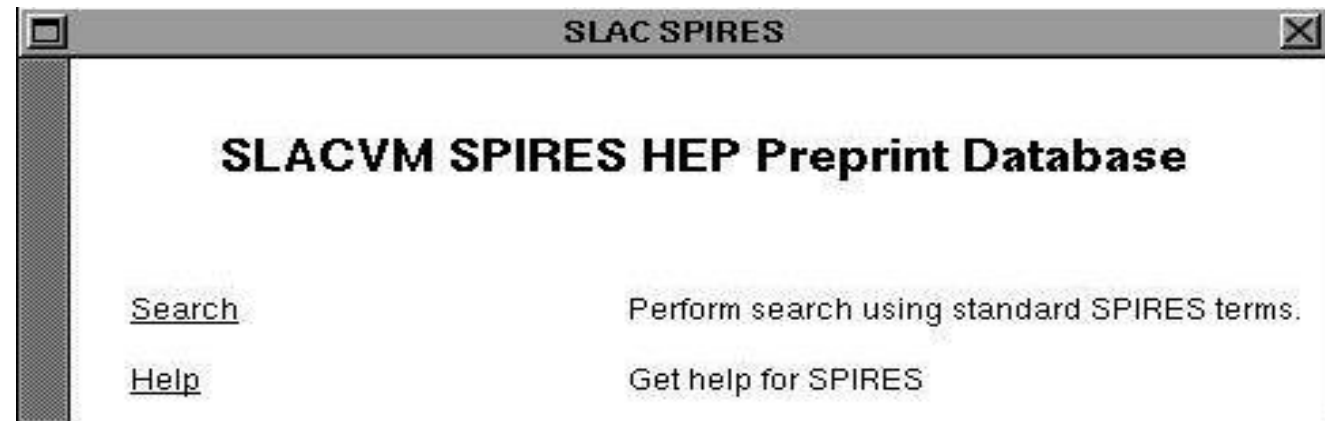


AUTHOR Parker, Edwin B.
TITLE SPIRES (Stanford Public Information Retrieval System)
1970-71 Annual Report.
INSTITUTION Stanford Univ., Calif. Inst. for Communication
Research.
SPONS AGENCY National Science Foundation, Washington, D.C. Office
of Science Information Services.
PUB DATE Dec 71
NOTE 154p.; (23 References)

EDRS PRICE MF-\$0.65 HC-\$6.58
DESCRIPTORS *Computer Programs; *Information Retrieval;
*Information Storage; *Information Systems; *On Line
Systems
IDENTIFIERS Computer Software; SPIRES; *Stanford Public
Information Retrieval System

ABSTRACT

SPIRES (Stanford Public Information RETrieval System) is a computer information storage and retrieval system being developed at Stanford University with funding from the National Science Foundation. SPIRES has two major goals: to provide a user-oriented, interactive, on-line retrieval system for a variety of researchers at Stanford; and to support the automation efforts of the university libraries by developing and implementing common software. SPIRES I, a prototype system, was implemented at the Stanford Linear Accelerator Center (SLAC) in 1969, from a design based on a 1967 information study involving physicists at SLAC. Its primary data base is a high-energy-physics preprints file. Evaluation of SPIRES I resulted in the definition of a production information storage and retrieval system, SPIRES II. This system will be available daily, beginning in mid-1972, to faculty, staff, and students of the University. It is characterized by flexibility, simplicity, and economy. SPIRES II will operate on-line on an IBM 360/67 computer. This report summarizes the uses of the SPIRES I system over the past year and describes both the nature of SPIRES II and this system's development over the past year. (Author)



Date: Fri, 13 Dec 91 17:55:53 GMT+0100

From: timbl@nxoc01.cern.ch (Tim Berners-Lee)

Subject: WWW to SPIRES on SLACVM - Experimental

To: www-interest@cernvax.cern.ch, www-talk@cernvax.cern.ch

There is an experimental W3 server for the SPIRES **High energy Physics preprint database**, thanks to Terry Hung, Paul Kunz and Louise Addis of SLAC. It's only just been put up, so don't expect perfection. With the w3 line mode browser, follow a link to it from our home page,

- Tim

Paul Kunz wrote a few days ago:-

"The SLAC Library maintainer of SPIRES databases, Louise Addis, is absolutely delighted. She will ask for a permanent VM service machine and finish off the polishing. Things are really moving now."

By 2007, SPIRES based on a 40-year old DB model, un-maintainable code.

SLAC and DESY phasing out of HEP. Uncertain DoE support (incl. Fermilab)

A 'summit' of labs, publishers, services is convened at SLAC.

Gigi represents CERN and delivers an: 'Experimentalist view' talk

Information needs



- ◆ Changes in the information needs are influenced more by new opportunities driven by technology than by changes in the needs of researchers. **Today we can do more.....**
- ◆ In the last few years we became more globally informed, more used to get what we need in one or two clicks and much faster than we ever dreamed only 10 years ago.....

- ◆ however the available on-line information has now increased by orders of magnitude and we face today a different problem: how to reduce the level of signal to noise when looking for what we need.... **I know that what I am looking for is there, but I do not know how to find it.**

- ◆ **How I am going to find THIS ?**

All instances of a scientific result



- ◆ A scientific result will appear in many preliminary forms: internal notes of collaborations; slides shown in internal meetings; conference contributions (both slides and proceedings). It will then become public in its final form in a seminar, possibly recorded in multi-media format, and as a short letter to a journal (appearing first as a pre-print and then as a published article).
- ◆ Will it be possible to retrieve all instances of the same result in one click ?

Looking at these and other needs, and with 'some negotiation',
INSPIRE was born: fusing SPIRES' data and CERN's software
Collaborative model to curate information, leveraging new tech.

Other things that Gigi saw in 2007... but needed wait for GPTs

Generic text search



- ♦ My student asks me how to compute the angular distribution of tau to mu gamma...I saw it few years ago in a paper but this is all what I know. How do I find it ?

Evolution in time of a “number”



- ♦ What was the top mass and its error in 2002 ?

Copilot ⓘ

The angular distribution for $\tau \rightarrow \mu \gamma$ is simple because it's a two-body decay. In the τ rest frame, the muon and photon are back-to-back, and the only nontrivial dependence comes from τ polarization. Using the dipole operator

$$\mathcal{L}_{\text{eff}} \sim \bar{\mu} \sigma^{\alpha\beta} (A_L P_L + A_R P_R) \tau F_{\alpha\beta},$$

the differential width is:

$$\frac{d\Gamma}{d\cos\theta} \propto 1 + \alpha_\mu P_\tau \cos\theta, \quad \alpha_\mu = \frac{|A_R|^2 - |A_L|^2}{|A_R|^2 + |A_L|^2},$$

where θ is the angle between τ spin and μ momentum. If $A_L = A_R$, the distribution is flat; if one chirality dominates, you get forward/backward asymmetry. For an unpolarized τ , the distribution is isotropic.

A good reference is:

A. Crivellin et al., “Lepton Flavor Violation in Tau Decays,” Phys. Rev. D 88, 076001 (2013), arXiv:1302.6467.

ChatGPT ▾

Get Plus ×

Share

...

In **2002**, the world average for the **top quark mass** was approximately:

$$m_t = 174.3 \pm 5.1 \text{ GeV}/c^2$$

This value came from combined measurements from the **CDF** and **DØ** experiments at **Fermilab's Tevatron collider**.

Context:

- The top quark had been discovered in 1995.
- By 2002, the precision was still limited mainly by statistical uncertainties and jet energy scale systematics.
- The **Particle Data Group (PDG 2002)** listed the world average as roughly **174.3 ± 5.1 GeV**.

Gigi's thoughts on data preservation with HEP characteristics (circa 2007)

THE PROBLEM of HEP data preservation



- ♦ The HEP data model is a **highly complex data model** (from the start difficult to export to OA a` la astronomy)
- ♦ Raw data -> calibrated data -> skimmed data -> high-level objects
- ♦ **Final results depend on all the grey-literature on constants, human knowledge, algorithms which are needed for each pass**
- ♦ Experiment lifetimes > computing environment lifetimes. Many migrations within the lifetime of an experiment (in this sense preservation is not an issue !)

The "Parallel way" to archiving and publishing data



- ♦ In addition to internal data models, elaborate a **parallel format** for useful and usable high-level objects
- ♦ Publish high-level objects behind each scientific paper (after a time lapse?)
- ♦ Publish all high-level objects after end of collaboration
- ♦ Address issues of accountability, reproducibility of results, "careless discovers", "careless measurements"

A possible R&D program



- ♦ **Use LEP or Tevatron or BABAR as a case study for information retrieval to better assess the different methods**
- ♦ **Define some high-level object to make a OA-based analysis possible for an "external" but "motivated" researcher of the field**
- ♦ **Propose strategies to define "parallel" high-level objects to be included in the LHC data model, that is not post-mortem but aim to make it part of the data-model designing process. This is very timely.**
- ♦ **Imagine solutions to expand digital-library records of experimental results to include the OA data behind the results**
- ♦ **Initiate a discussion on priority issues and time-delays in making these "parallel" high-level objects available. This is very timely.**

Sometimes you get more than you wish for: LHC 2020 Open Data Policy

CERN Open Data Policy for the LHC Experiments
November, 2020

The CERN Open Data Policy reflects values that have been enshrined in the CERN Convention for more than sixty years that were reaffirmed in the European Strategy for Particle Physics (2020)¹, and aims to empower the LHC experiments to adopt a consistent approach towards the openness and preservation of experimental data. Making data available responsibly (applying FAIR standards²), at different levels of abstraction and at different points in time, allows the maximum realisation of their scientific potential and the fulfillment of the collective moral and fiduciary responsibility to member states and the broader global scientific community. CERN understands that in order to optimise reuse opportunities, immediate and continued resources are needed. The level of support that CERN and the experiments will be able to provide to external users will depend on available resources.

This policy relates to the data collected by the LHC experiments, for the main physics programme of the LHC — high-energy proton–proton and heavy-ion collision data. The foreseen use cases of the Open Data include reinterpretation and reanalysis of physics results, education and outreach, data analysis for technical and algorithmic developments and physics research. The Open Data will be released through the CERN Open Data Portal which will be supported by CERN for the lifetime of the data. The data will be tailored to the different uses, and will be made available in formats defined by each experiment that afford a range of opportunities for long-term use, reuse and preservation. In general, four levels of complexity of HEP data have been identified by the Data Preservation and Long Term Analysis in High Energy Physics (DPHEP) Study Group³, which serve varying audiences and imply a diversity of openness solutions and practices.

Published Results (Level 1) Policy: Peer-reviewed publications represent the primary scientific output from the experiments. In compliance with the CERN Open Access Policy, all such publications are available with Open Access, and so are available to the public. To maximise the scientific value of their publications, the experiments will make public additional information and data at the time of publication, stored in collaboration with portals such as HEPData,⁴ with selection routines stored in specialised tools. The data made available may include simplified or full binned likelihoods, as well as unbinned likelihoods based on datasets of event-level observables extracted by the analyses. Reinterpretation of published results is also made possible through analysis preservation and direct collaboration with external researchers.

Outreach and Education (Level 2) Policy: For the purposes of education and outreach, dedicated subsets of data are used, selected and formatted to provide rich samples to maximise their educational impact, and to facilitate the easy use of the data. These data are released with a schedule and scope determined by each experiment. The data are provided in simplified, portable and self-contained formats suitable for educational and public understanding purposes; but are not intended nor adequate for the publication of scientific results. Lightweight environments to allow the easy exploration of these

Published Results (Level 1) Policy: Peer-reviewed publications represent the primary scientific output from the experiments. In compliance with the CERN Open Access Policy, all such publications are available with Open Access, and so are available to the public. To maximise the scientific value of their publications, the experiments will make public additional information and data at the time of publication, stored in collaboration with portals such as HEPData,⁴ with selection routines stored in specialised tools. The data made available may include simplified or full binned likelihoods, as well as unbinned likelihoods based on datasets of event-level observables extracted by the analyses. Reinterpretation of published results is also made possible through analysis preservation and direct collaboration with external researchers.

Reconstructed Data (Level 3) Policy: The LHC experiments will release calibrated reconstructed data with the level of detail useful for algorithmic, performance and physics studies. The release of these data will be accompanied by provenance metadata, and by a concurrent release of appropriate simulated data samples, software, reproducible example analysis workflows, and documentation. Virtual computing environments that are compatible with the data and software will be made available. The information provided will be sufficient to allow high-quality analysis of the data including, where practical, application of the main correction factors and corresponding systematic uncertainties related to calibrations, detector reconstruction and identification. A limited level of support for users of the Level 3 Open Data will be provided on a best-effort basis by the collaborations.

Public data releases will occur periodically, following an appropriate latency period to allow thorough understanding of the data, the reconstruction and calibrations, as well as to allow time for the scientific exploitation of the data by the collaboration. The size of the released datasets will be commensurate

¹ European Strategy Group (2020), '2020 Update of the European Strategy for Particle Physics'.

² FAIR Guiding Principles for scientific data management and stewardship. Available at: <https://www.go-fair.org/fair-principles/>.

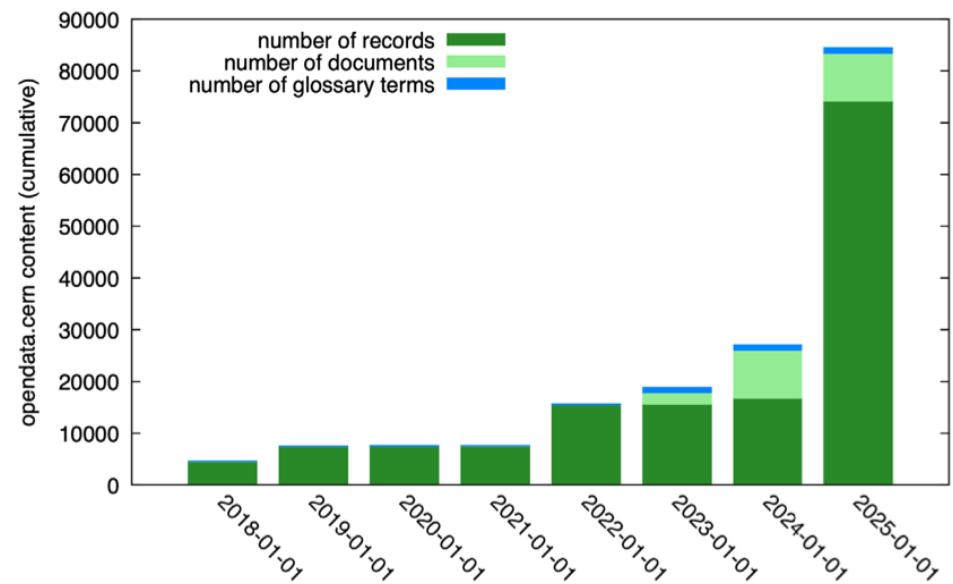
³ Data management plans are defined by the LHC experiments to address the long-term preservation of internal data products. See: Akopov et al., Status report of the DPHEP Study Group: Towards a global effort for sustainable data preservation in high energy physics. arXiv preprint arXiv:1205.4667 (2012).

⁴ Repository for publication-related High-Energy Physics data: <http://www.hepdata.net>.

Explore more than **five petabytes**
of open data from particle physics!

Search

search examples: [collision datasets](#), [keywords:education](#), [energy:7TeV](#)



Focus on

- [ALICE](#)
- [ATLAS](#)
- [CMS](#)
- [DELPHI](#)
- [LHCb](#)
- [OPERA](#)
- [PHENIX](#)
- [TOTEM](#)
- [Data Science](#)

CERN

Analysis Preservation

Capture, preserve and reuse physics analyses



Capture

Preserve elements needed to
understand and rerun your analysis



Collaborate

Share your analysis with other users,
your collaboration or group



Reuse

Run containerized workflows and
easily reuse analysis components

 analysis-preservation-support@cern.ch

 [@analysispreserv](#)

 [@cernanalysispreservation](#)

Supported by:



Yachting Club CERN Match Race, circa 2002

