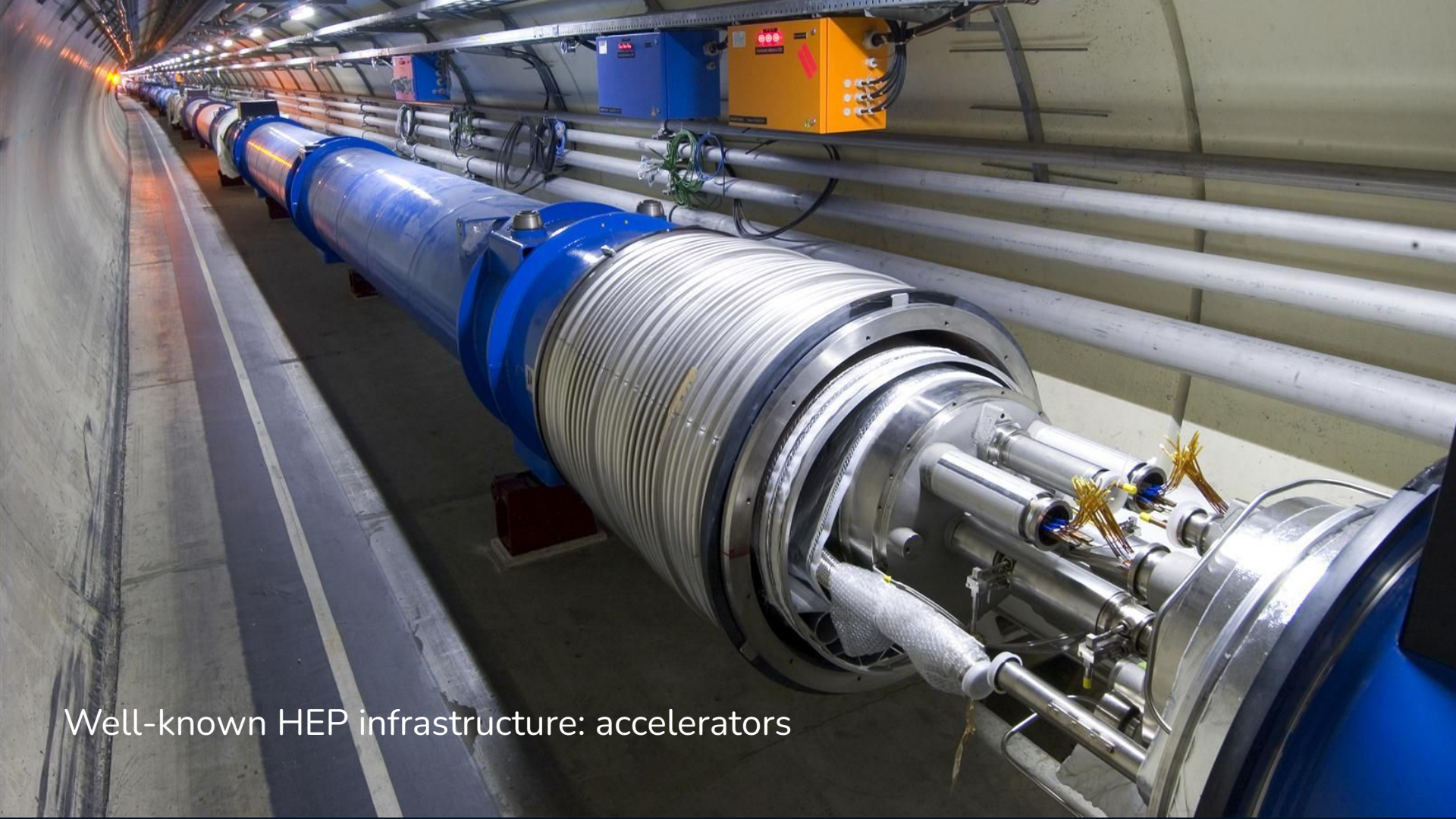


---

# Sustaining the inconspicuous information infrastructure

IUPAP-C11 meeting, July 2024

Micha Moskvic (CERN) on behalf of the INSPIRE collaboration  
With input from arXiv, HEPData, PDG, SCOAP<sup>3</sup>



Well-known HEP infrastructure: accelerators

iNSPIRE HEP

arXiv



HEPData

SCOAP<sup>3</sup>

PDG  
particle data group

inconspicuous

adjective

UK  / ˌɪn.kənˈspɪk.ju.əs/ US  / ˌɪn.kənˈspɪk.ju.əs/

not easily or quickly noticed or seen, or not attracting attention:



- Open access since its founding in 1991 by HEP-theorist Paul Ginsparg
- Hosts over 2.5 million e-prints and receives an average of over 17,000 submissions every month
- Cost for arXiv to publish can be as high as \$20-per-paper
- Staffing has not kept pace with growing number of submissions (30% increase YOY)
- 60% of funding comes from voluntary members (universities, libraries, research labs, etc.), suggested rate based on usage stats

# The SCOAP<sup>3</sup> Model: How It Works

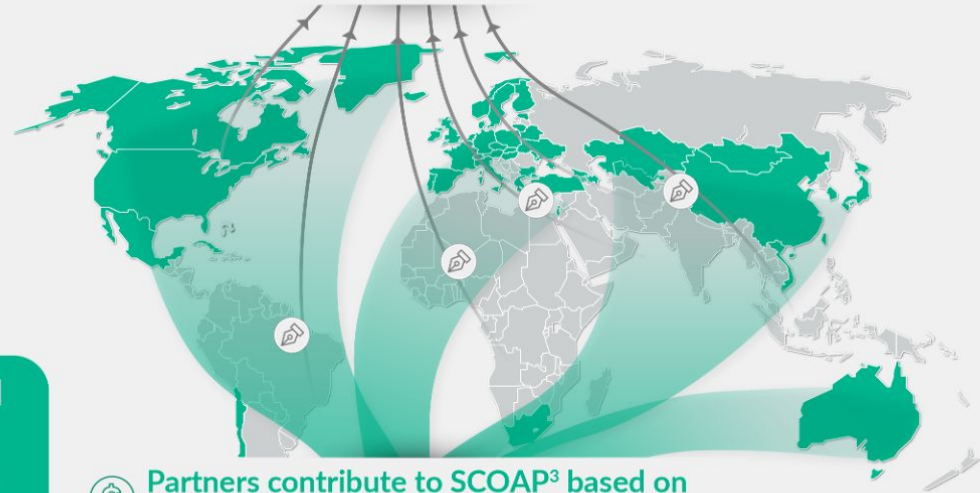
SCOAP<sup>3</sup> centrally underwrites Open Access to research in high-energy physics, enabling free publishing, global access, and re-use



Everyone around the world can access and reuse any SCOAP<sup>3</sup> article for free



No publication costs for authors worldwide



Partners contribute to SCOAP<sup>3</sup> based on their share of the published literature

Research articles are published fully Open Access with CC-BY licenses



SCOAP<sup>3</sup> centrally pays for Open Access publishing services

**SCOAP<sup>3</sup>** Sponsoring Consortium for Open Access Publishing in Particle Physics

Join **SCOAP<sup>3</sup>**  
Open up **high-energy physics** to the world  
[scoap3.org](http://scoap3.org)

- Unique open-access repository for publication-related high-level (predominantly tabular) data
- Almost 140k data tables from more than 10k High-Energy Physics (HEP) publications, linked bi-directionally with INSPIRE records
- Core component of the data management plan of many collaborations
- Since 2016, partnership between IPPP at Durham University and CERN
- UK STFC funds project manager and software engineer, technical infrastructure and operational support provided by CERN
- Funded by STFC as part of IPPP, [need to apply for funding every 3 years](#)



- Compilation & evaluation of 50k particle properties measurements based on 13k articles + 120 review articles
- Dedicated coordination team at LBNL funded mostly through US DOE
- Coordinates activities of more than 200 authors globally
- Lean operation: 96% of budget in personnel costs, little redundancy
- In 2013, US NSF cut support for PDG after contributing ~10% to PDG funding for decades
- Reduction in scope could ultimately be avoided, but *Review* suffered delays & multiple projects put on hold
- ⇒ Modest cuts can have high impacts to crucial services
- Any further cuts will necessarily affect scope

- Main information platform for HEP and related fields (e.g. nuclear physics)
- 1.6 M papers
- 100k author profiles (excluding automatically generated)
- 25k daily visits (~42% from Europe)
- Fully transitioned to new platform in 2021
- Collaboration of various HEP labs





# Main strength: focus & high quality of information

---

- Focus on HEP (& related subjects) unlike other generalist platforms
- Accurate citation tracking
  - Single record merging info from arXiv & publisher version
  - Extraction of reference lists
- Additional metadata (collaboration, experiment, ...)
- Additional services
  - Jobs
  - Conferences
  - Seminars
- Initially planned for end of year (likely to change due to recent news)
  - Data & software (following Open Science developments in community)

# Behind the scenes: human/machine cooperation

---

- General approach: automation + manual intervention
- High quality of service requires significant resources
- Machine power: ~1k cores in CERN data center
- Human power: ~15 FTE

# Collaboration model

---

- All contributing institutions on equal footing (decision by consensus)
- Each takes care of dedicated subset of tasks
- Technical team (CERN):
  - Development
  - IT operations
- Curation team (distributed):
  - Content selection
  - Metadata quality
  - User support
- Governance: 1 representative from each active lab + Advisory Board

# Dealing with staffing losses

---

- In 2018, SLAC stopped curation activities, not involved at all since 2020
- Very significant loss of 2 FTE
- Mitigation strategy
  - Prioritization of activities
  - Redistribution of essential work across other labs (DESY, IHEP)
  - Dropped non-essential work (e.g arXiv reference correction)
  - Increasing automation
- In June, DESY announced staffing reduction from 2 FTE to 0.3-0.5 FTE by end of 2024 (was 3 FTE until December 2023)
- Very short notice given crucial DESY role
- Currently evaluating further mitigation strategies, will be extremely difficult
- Might require less extensive coverage


























# Main user countries 2023

- Number of visits per country

Trying to expand collaboration:

- Pilot collaboration with UK (STFC)
- In talks with Italy (INFN)
- Finding new potential partners in Germany

## Country

COUNTRY	VISITS
 United States	16.9% 1,460,212
 China	10.5% 909,981
 India	8.4% 722,256
 Germany	7.5% 649,946
 Italy	6.7% 580,948
 Japan	5.6% 487,215
 United Kingdom	5.6% 480,598
 Spain	3.6% 306,287
 France	3.5% 303,108
 South Korea	2.9% 249,117
 Switzerland	2.8% 241,247
 Canada	1.7% 149,756
 Brazil	1.7% 143,847
 Russia	1.6% 141,301
 Netherlands	1.3% 113,355
 Greece	1.3% 108,093
 Hong Kong SAR China	1.2% 100,145
 Poland	1.2% 100,101
 Taiwan	1% 88,747
 Iran	1% 83,775
 Belgium	0.9% 79,789
 Chile	0.9% 77,528
 Mexico	0.9% 73,511
 Israel	0.8% 66,480
 Türkiye	0.8% 65,761

# Summary & outlook

---

- HEP research makes crucial use of information infrastructure
- Largely relies on limited funding that should not be taken for granted
- Sustainable funding is needed to ensure reliable operation & quality of service
- INSPIRE in particular is facing uncertain future after budget cuts at SLAC & DESY
- Would be beneficial for future funding requests if HEP community manifested need for the information infrastructure
- Suggestion: write a paper as input for the ESPP update?

---

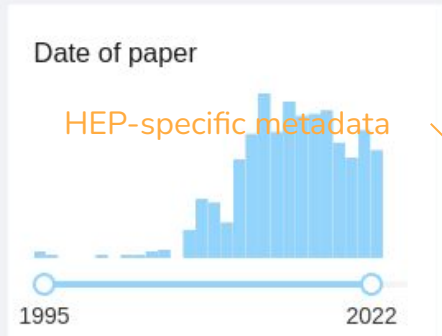
**Thank you!**





---

# Backup



963 results |  Citation Summary

Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC

• Georges Aad (Freiburg U.) et al. (Jul, 2012)

Published in:  • e-Print:

- Date indifférente
- Depuis 2022
- Depuis 2021
- Depuis 2018
- Période spécifique...

[HTML] **Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC** [HTML] sciencedirect.com

G Aad, T Abajyan, B Abbott, J Abdallah, SA Khalek... -

... A **search** for the **Standard Model Higgs boson** in proton–proton collisions with the **ATLAS detector** at the **LHC** is presented. The datasets used correspond to integrated luminosities of ...

gross overestimation:  
double counting arXiv + journal citations












# Key task: Paper curation

---

- Most time-consuming task in INSPIRE
- Matching of arXiv & publisher version
- Reconciliation of conflicting information
- Systematic check & correction of metadata of core HEP papers
  - Author affiliations
  - References
  - Conferences
  - Collaborations
  - Experiments

# Collaboration model

- 5 institutes (formally 6) on equal footing (decision by consensus)
- Each takes care of dedicated subset of tasks
- Governance: 1 representative from each active lab + Advisory Board

	 	 	 	 	 
5 FTE	2.1 FTE	3.25 FTE	2.5 FTE	2 FTE	0 FTE
<ul style="list-style-type: none"> <li>· development</li> <li>· technical operation</li> <li>· paper curation (CERN)</li> <li>· user support</li> </ul>	<ul style="list-style-type: none"> <li>· paper curation</li> <li>· content selection</li> <li>· harvesting</li> <li>· user support</li> <li>· conferences</li> </ul>	<ul style="list-style-type: none"> <li>· author profiles</li> <li>· paper curation (QIS)</li> <li>· jobs</li> <li>· user support</li> </ul> <p>Unique &amp; critical tasks</p>	<ul style="list-style-type: none"> <li>· paper curation</li> <li>· jobs</li> </ul>	<ul style="list-style-type: none"> <li>· paper curation (FR)</li> </ul>	<p>Previously:</p> <ul style="list-style-type: none"> <li>· technical operation</li> <li>· paper curation</li> <li>· user support</li> </ul>

# Future INSPIRE sustainability approaches

---

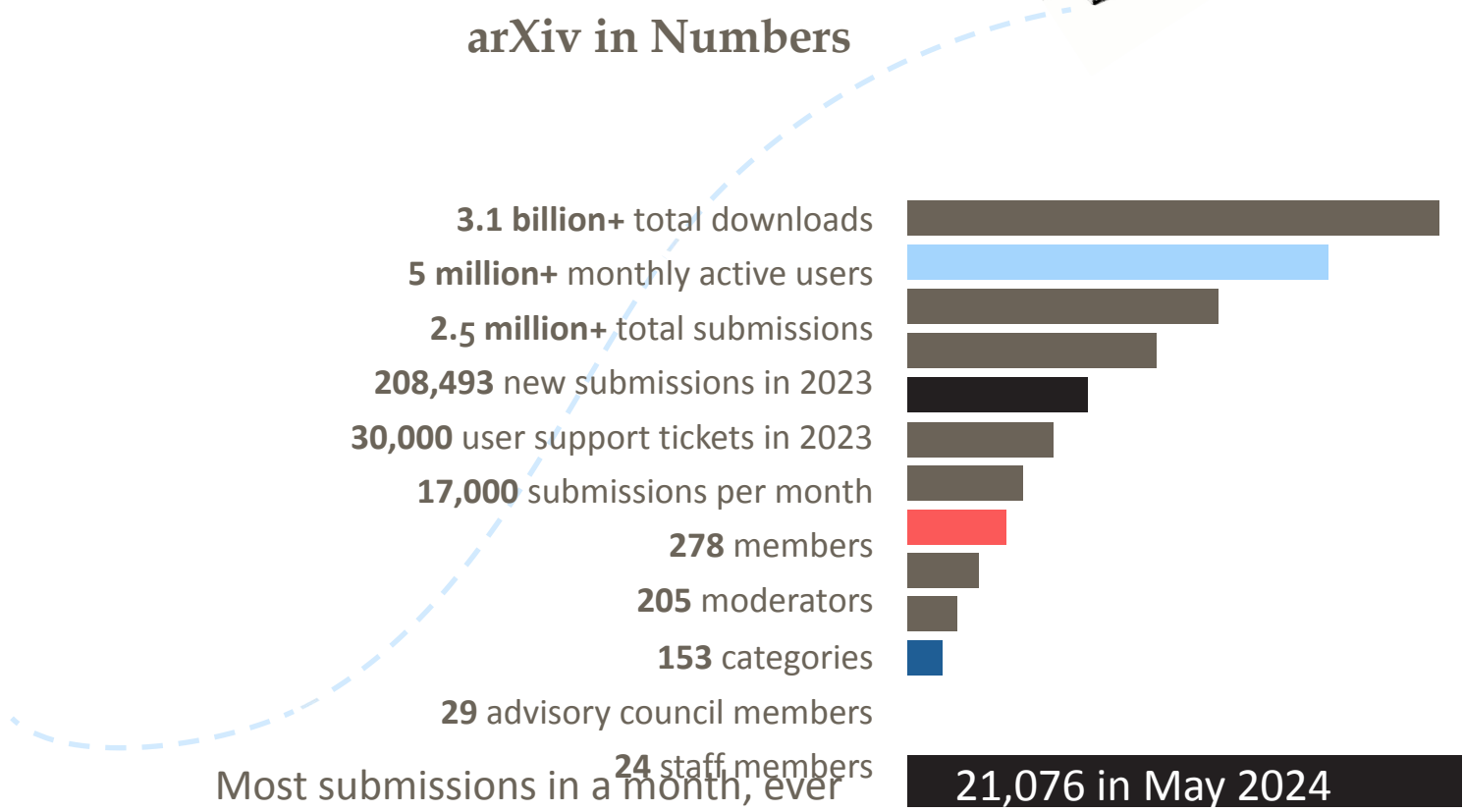
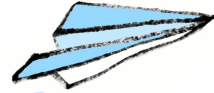
- Strategically important to reduce risks related to staffing losses
- Pursuing several approaches
  - Expansion to new partners
  - Increased automation
  - Crowd-sourcing
  - Identifying less essential tasks (will impact quality)
- All of the options will require additional support, either to set it up (automation, crowd-sourcing) or continuously (new partnerships)

# DESY current responsibilities

---

- Responsible for “harvesting”:
  - Ingestion of most journal articles, all theses and conference proceedings (except APS, Elsevier)
  - Requires specific computing + library expertise
- Fully responsible for content selection:
  - Decide which incoming papers are relevant for INSPIRE and get added to database
  - Requires domain expertise (in multiple fields, theory, experiment, math, quantum, etc. !)
- Contributing to curation effort: responsible for journal articles
- ⇒ Without DESY (& no changes), ingestion pipeline is blocked  
+ important loss in curation capacity

# arXiv in Numbers



We must ensure sustained development, maintenance, and user support for key cyberinfrastructure components, including widely used software packages, simulation tools, and information resources, such as the Particle Data Group and INSPIRE. Although most of these shared cyberinfrastructure components are not specifically tied to projects, nearly all scientists in the field rely on them. A significant investment—at the level of

**Area Recommendation 18: Increase targeted investments that ensure sustained support for key cyberinfrastructure components by \$8M per year in 2023 dollars. This includes widely-used software packages, simulation tools, information resources such as the Particle Data Group and INSPIRE, as well as the shared infrastructure for preservation, dissemination, and analysis of the unique data collected by various experiments and surveys in order to realize their full scientific impact.**