



Scientific
Information
Service

Addressing the challenges of research data management, reuse and collaboration: the case for CERN Analysis Preservation and FAIR data services.

Sunje Dallmeier-Tiessen, Pamfilos Fokianos and Artemis Lavasa @CERN

May 2023

“...and the results of its experimental and theoretical work shall be published or otherwise made generally available”

CERN Founding Convention (1953)

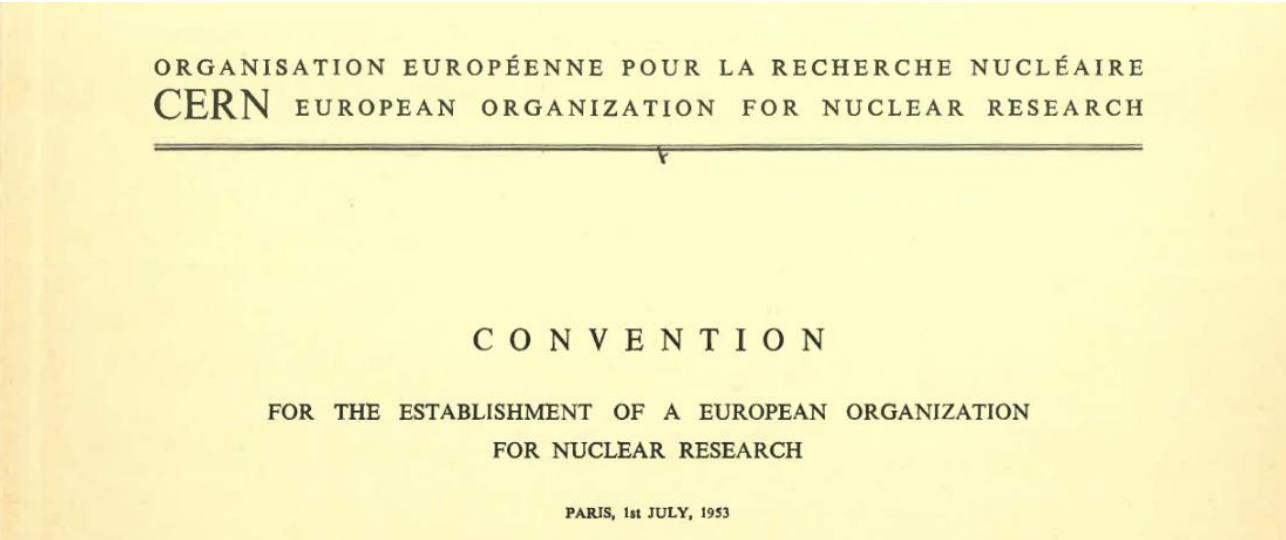
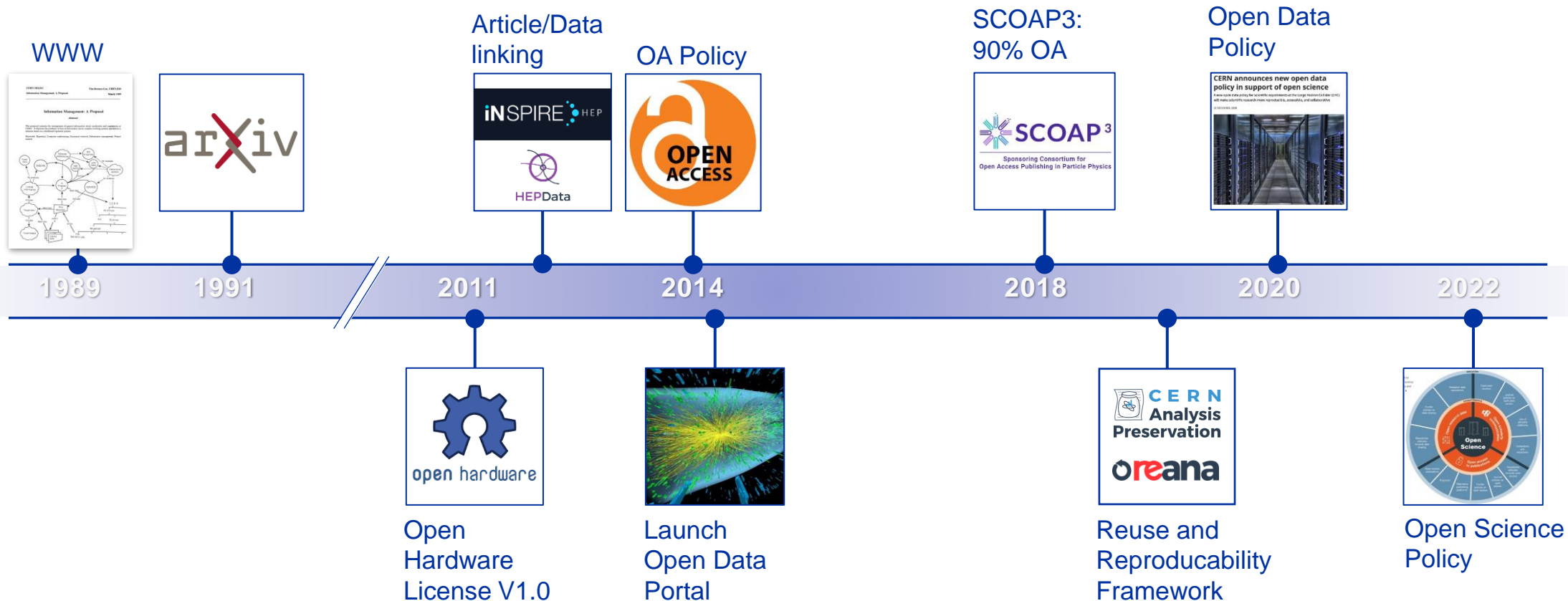
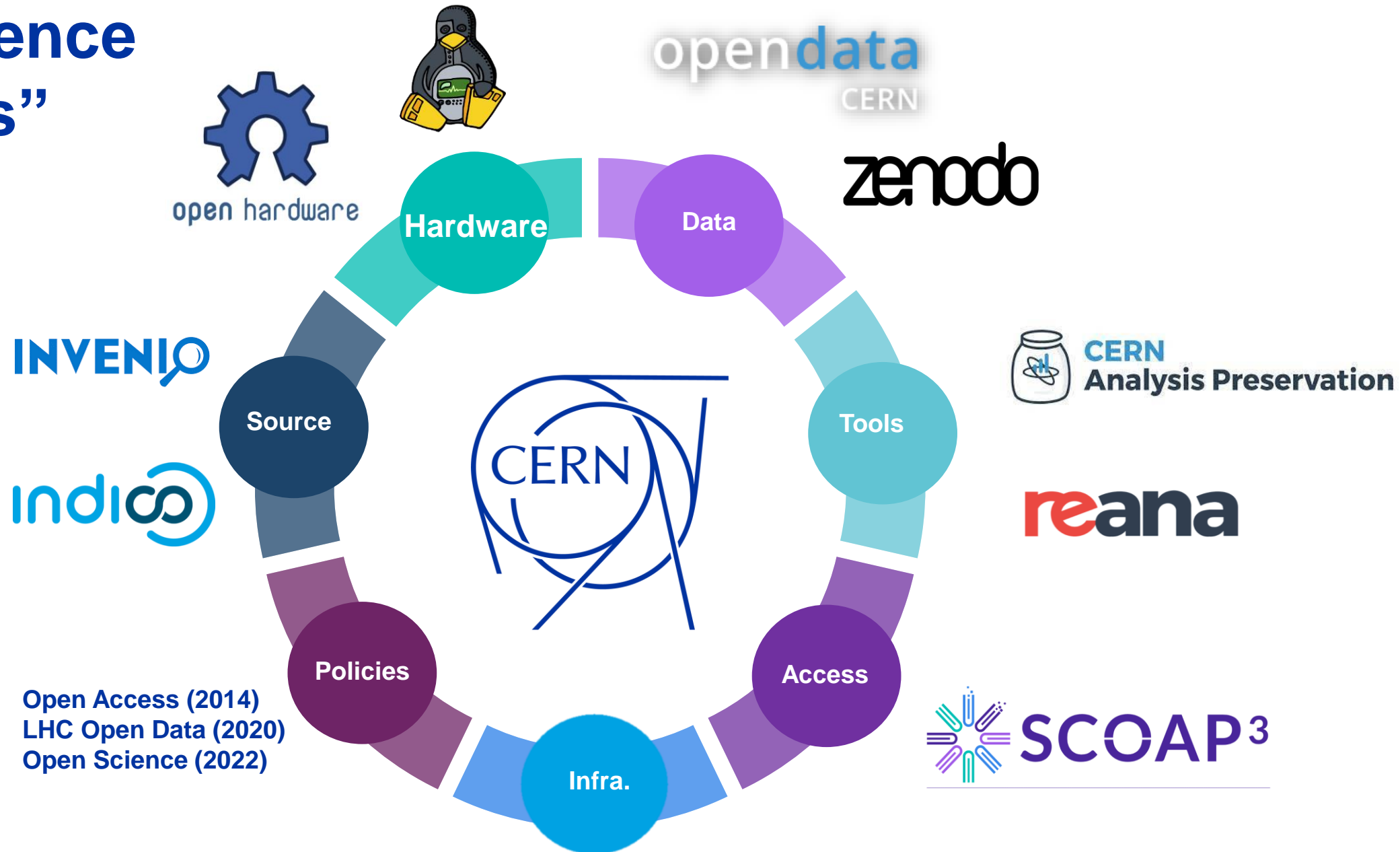


Illustration by Stephanie van de Sandt

CERN – on the path to universal Open Science



Open Science “products” at CERN



UNESCO Recommendations on Open Science created the momentum

The Assistant Director-General for Natural Sciences

United Nations Educational, Scientific and Cultural Organization

Organization des Nations Unies pour l'éducation, la science et la culture

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura

Организация Объединённых Наций по вопросам образования, науки и культуры

منظمة الأمم المتحدة للتربية والعلم والثقافة

联合国教育、科学及文化组织

Mr Alexander Kohls
Group Leader
Scientific Information Service
European Council for Nuclear Research (CERN)

10 March 2020

Ref.: SC/PCB/SP/1053

Dear Mr Kohls,

During the 40th session of UNESCO's General Conference, Member States tasked the Organization with leading a global dialogue on Open Science with a view to developing a standard-setting instrument in the form of a Recommendation, to be adopted by the UNESCO General Conference in November 2021.

The Recommendation will be prepared through a regionally balanced, multistakeholder, inclusive and transparent consultation process. I would like to invite all UNESCO partners to contribute to this process.

Pushing the Boundaries of Open Science at CERN: Submission to the UNESCO Open Science Consultation
July 2020

Katsumi Niino¹, Tullio Braccaglia², Jelena Branković³, Pentti Hänninen⁴, José Benito González López⁵, María Gracia Páez⁶, Alexander Kohls⁷, Artemis Lavaná⁸, Lars Holm Nielsen⁹, Stephanie van de Sandt¹⁰, Javier Soriano¹¹, Tim Smith¹²

CERN, the European Organization for Nuclear Research, is the world's largest high-energy physics (HEP) laboratory. Since its founding in 1954, the Laboratory has made significant contributions to our understanding of the world and the universe. The mission of the Organization is to provide a unique range of particle-accelerator facilities and related services to the benefit of human knowledge, perform world-class research in fundamental physics, and bring people from all over the world to push the frontiers of science and technology, for the benefit of all. Supported through a global partnership of 23 member states, CERN is home to the world's largest scientific instrument, the Large Hadron Collider (LHC), and hosts over 12,000 scientists and engineers from across the world.

The frontier research conducted at CERN has long exceeded the values that have been recently cited to be defined as the Open Science movement, which describes research and development that is collaborative, transparent and reproducible and whose outputs are publicly available. European Commission (EC), including Horizon 2020, which was endorsed at EC's 10th meeting in November 2019, which states that "the results of its experimental and theoretical work shall be published as extensively as possible and made generally available". CERN Council (1975) providing the Organization with an early Open Science resolution.

On June 19th, 2020, CERN's highest governing authority, the CERN Council, approved its updated strategy for particle physics in the global landscape. This document, the European Strategy for Particle Physics, represents the most important strategic document for CERN, setting the future strategy for the Organization. "Open Science" was given a strong endorsement from the CERN Council, with the strategy stating that: "European science policy is quickly moving towards Open Science, which promotes and sustains the sharing of scientific knowledge with the community at large. Particle physics has been a pioneer in several aspects of Open Science. The particle physics community should work with the relevant authorities to fully share the resulting knowledge on Open Science to the extent a policy of Open Science for the field/European Strategy Group (ESG)".

This paper aims to describe the ecosystem of initiatives, projects and technologies that have been developed at CERN to maximize the impact of our research through building an Open Science infrastructure that is efficient, sustainable, and responsive to the needs of the scientific community. We aim to demonstrate that despite the complexity of the research activities at CERN, Open Science can be achieved through concerted efforts and as such, the CERN example could serve as an inspiration for the global scientific community.

<https://doi.org/10.17181/CERN.1SYT.9RGJ>

United Nations Educational, Scientific and Cultural Organization

Organization des Nations Unies pour l'éducation, la science et la culture

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura

Организация Объединённых Наций по вопросам образования, науки и культуры

منظمة الأمم المتحدة للتربية والعلم والثقافة

联合国教育、科学及文化组织

ESOF2020
EUROSCIENCE OPEN FORUM
TRIESTE

EuroDIG
European Digital Infrastructure for Research Communities

Towards a Global Consensus on Open Science

Online Regional Consultation for Western Europe and North America to the UNESCO Recommendation on Open Science

Online Consultation on Open Science

UNESCO invites scientists, publishers, science policymakers or anyone with experience and interest in Open Science to participate in the online consultation on implications, benefits and challenges of Open Science across the globe.

Open Science

United Nations Educational, Scientific and Cultural Organization

CERN is proud to have joined UNESCO on the journey towards Recommendations on Open Science



Policy framework for Open Science at CERN

CERN Open Access Policy (2014)

- All CERN research articles published OA (CC-BY)
- Central fund available
- Different routes (SCOAP³, Read & Publish, APC payment)

LHC Open Data Policy (2020)

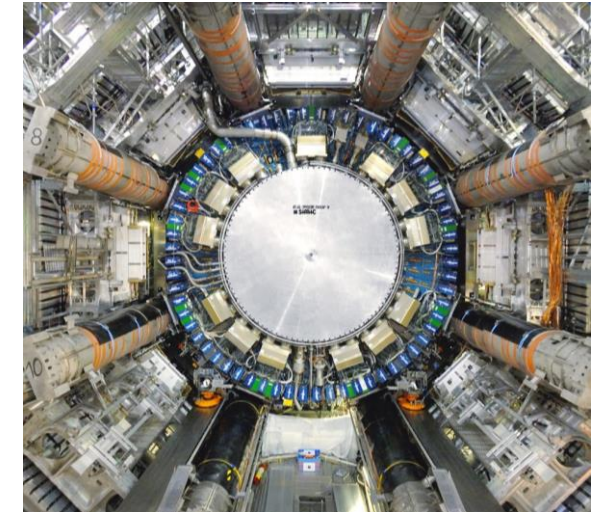
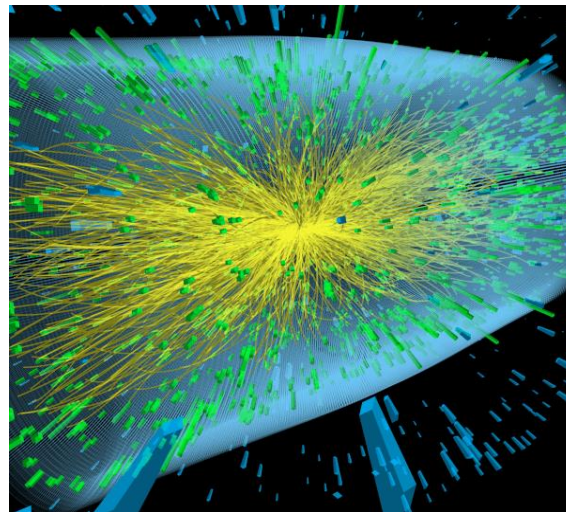
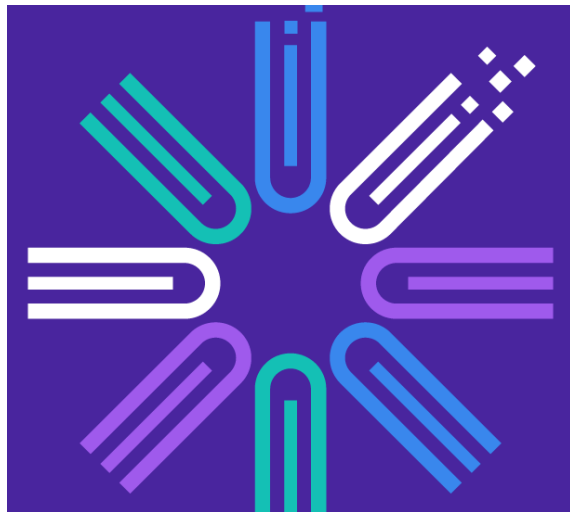
- 4 LHC collaborations will release all level 3 data (+ level 1 and 2)
- Gradual release will start ~5 years after collection
- Other experiments to follow

European Strategy for Particle Physics(2020)

- OS recognized as organizational issue for the discipline
- Should develop and implement an OS policy for the field

Funder Open Science Policies

- Funding agencies supporting experimental collaborations have specific open data requirements
- CERN will establish central support office for compliance



CERN Open Science Policy Published October 1st, 2022

After 12 months of consultations and collaborative drafting, CERN's first institutional Open Science Policy was formally adopted by CERN Council

- August/September: Presentations and discussions with directorate
- 29th September: Policy presented to CERN Council
- 1st October: Policy formally in place

More information: <https://openscience.cern>

CERN Accelerating science

Signed in as: akohls (Drupal) Sign out Directory

CERN Open Science

OPEN SCIENCE POLICIES OPEN SCIENCE ELEMENTS HISTORY NEWS ABOUT SEARCH

Welcome to the CERN Open Science portal

At CERN, we believe that the practice of open science is key to delivering on our organizational mission: to perform world-class research in fundamental physics at the forefront of human knowledge; provide a unique range of particle accelerator facilities that enable this research, educate the next generation of scientists; and unite people from all over the world to push the frontiers of science and technology, for the benefit of all.

Supported by long-term financial investments from its Member and Associate Member States, with significant contributions also from non-Member States, CERN is committed to the advancement of science and wide dissemination of knowledge by adopting practices to make scientific research more open, global, collaborative and responsive to societal changes. By embracing open science, we aim to cultivate a dynamic and evolving ecosystem of initiatives, projects and interoperable technologies to maximise the global impact of our research.

View Edit Delete Manage display Revisions

CERN publishes Open Science Policy

CERN's core values include making research open and accessible for everyone. A new policy now brings together existing open science initiatives to ensure a bright future of transparency and

2022-10-01

SCOAP3 reaches 50'000 articles milestone

The [Sponsoring Consortium for Open Access Publishing in Particle Physics \(SCOAP³\)](#)—the world's largest disciplinary open access initiative—has reached the milestone of over 50'000 research articles

2022-05-19

First CMS Open Data from LHC Run 2 released

As the experiments at the Large Hadron Collider (LHC) brace for the start of Run 3 of the accelerator's programme in 2022, the CMS collaboration has released a new batch of research-quality open data

2021-12-20

more

CERN Open Science Policy

- Captures current practice and states progressive vision across multiple Open Science domains:
 - Open Access to Publications
 - Open Research Data
 - Open Software
 - Open Hardware
 - Research Integrity, Reuse & Reproducibility
 - Infrastructure for Open Science
 - Research Assessment & Evaluation
 - Education, Training & Outreach
 - Citizen Science
- Policy to be regularly updated to reflect changes in landscape, practices, funder requirements & community demands
- Policy and its implementation plan are developed and governed by the community.
- V1.0, Oct 2022: <https://cds.cern.ch/record/2835057>

Research integrity, reuse and reproducibility

5. Research integrity, reuse and reproducibility

CERN is committed to ensuring the integrity of research. In order to facilitate the reuse of its research products, CERN provides infrastructures to accommodate the scale and complexity of its research outputs. Reuse and reproducibility are facilitated by practising comprehensive analysis preservation to capture relevant research objects, such as research data releases with supporting metadata, auxiliary data, linked software, reproducible analysis workflows, documentation, etc.

<https://cds.cern.ch/record/2835057>



CERN Analysis Preservation: what is it?

Flexible and collaborative tool to link and preserve “everything” around an analysis, metadata, data, software...

Version content and metadata. Link persistently all the elements of an analysis needed to understand and rerun an analysis several years later.

Standardize analysis components so that they are reusable (ex. scripts or CI/CD, writing tools, workflow engines, push to other services).

Ensure that users are always in control of when and if their work is shared.

Accommodate the needs of each collaboration or team to integrate into local tools



The screenshot shows the homepage of the CERN Analysis Preservation service. At the top left is the CERN logo and the text 'CERN Analysis Preservation'. At the top right is a navigation menu with links: 'Home', 'What is CAP?', 'Get Started', 'Integrations', 'Documentation', and 'Log in'. The main content area features the title 'CERN Analysis Preservation' in large blue text, followed by the tagline 'capture, preserve and reuse physics analyses' in smaller black text. To the right of the title is a 3D illustration of a virtual meeting room with several avatars and data visualizations. Below the title and tagline are three columns, each with an icon and a brief description: 'Capture' (a list icon) with the text 'Collect and preserve elements needed to understand and rerun your analysis'; 'Collaborate' (a group of people icon) with the text 'Share your analysis and components with other users, your collaboration or group'; and 'Reuse' (a circular arrow icon) with the text 'Run containerized workflows and easily reuse analysis components'.

Starting an analysis in CERN Analysis Preservation

The screenshot displays the CERN Analysis Preservation web interface. At the top, there is a dark navigation bar with the CERN logo, a search bar, and links for '+ Create', '? FAQ', and 'Account'. Below this is a breadcrumb trail: '< Demo July >'. The main content area is divided into a left sidebar, a central form, and a right-hand metadata panel.

Left Sidebar: Contains navigation options: Overview, Edit (selected), Connect, Workflows, and Settings.

Form Fields: The 'Edit' section has two tabs: 'Initial Input' (active) and 'Later Input'. The 'Initial Input' tab contains the following fields:

- Initial Input Status:** A dropdown menu.
- Short Title:** A text input field.
- Publication Title:** A text input field with the placeholder text 'Provide a title for your publication'.
- Description:** A rich text editor with the placeholder text 'Provide a description' and a toolbar with icons for bold, italic, underline, list, link, and other text formatting options.

Right Panel: A metadata table with the following entries:

ID	FASER-2022-9
Collection	FASER Analysis v0.0.11
Status	draft
Creator	sunje.dallmeier-tiessen@cern.ch
Published URL	Not published yet
Created	9 seconds ago
Last Updated	9 seconds ago

Below the table, there are sections for 'Files | Data | Repos' and 'All Files', both showing 'No files uploaded yet' with a printer icon and a menu icon.

FAIR @ CERN Analysis Preservation

Findable

Each analysis with a unique ID, with rich metadata, that facilitates advanced search

Accessible

Metadata are accessible and retrievable (if permitted by the experiment/team)

Interoperable

JSON schemas. Challenge are diverse community terminologies...

Reusable

Rich metadata for each analysis, i.e. automated ingestion from experiment tools, contextual linking

CAP: preserving context

User Analysis

Gitlab repositories of the analysis



No Items added

+ Add Item

Docker images of the analysis



No Items added

+ Add Item

Additional Repositories



No Items added

+ Add Item

Metadata

Reviewers



No Items added

+ Add Item

Review eGroup

Status

Institutes Involved

Keywords

Additional Resources

Please provide information about the additional resources of the analysis

Internal Discussions



No Items added

+ Add Item

Presentations



No Items added

+ Add Item

Publications



No Items added

+ Add Item

Repos, Docker images

Presentations, Discussions

Preserving repositories

The screenshot displays the CERN Analysis Preservation web interface. At the top, there is a dark navigation bar with the CERN logo, a search bar, and links for '+ Create', '? FAQ', and 'Account'. Below this, a breadcrumb trail shows 'Demo July'. A left sidebar contains navigation options: 'Overview', 'Edit', 'Connect' (highlighted), 'Workflows', and 'Settings'. The main content area is divided into three sections: 1. 'Repositories' section with explanatory text about downloading and connecting repositories. 2. 'Create a repository' section featuring a 'Create' button and a message: 'You can create a new repository, and add'. 3. 'Add new repository' section with a text input field for 'Github/Gitlab URL' and a placeholder: 'Please provide a valid Github/Gitlab repository or file URL'. On the right side, there is a metadata table for a repository and a file management section.

ID	FASER-2022-9
Collection	FASER Analysis v0.0.11
Status	draft
Creator	sunje.dallmeier-tiessen@cern.ch
Published URL	Not published yet
Created	3 minutes ago
Last Updated	3 minutes ago

Files | Data | Repos

All Files

No files uploaded yet

All Repositories

No repos uploaded yet

Capturing repositories

The screenshot displays the CERN Analysis Preservation web interface. At the top, there is a navigation bar with the CERN logo, a search bar, and links for '+ Create', '? FAQ', and 'Account'. Below the navigation bar, a sidebar on the left contains menu items: 'Overview', 'Edit', 'Connect' (highlighted), 'Workflows', and 'Settings'. The main content area is titled 'Demo July' and shows the 'Connect' configuration page. It features a 'Github/Gitlab URL' input field containing 'https://github.com/cernanalysispreservation/analysispreservation.cern.ch'. Below this, a section titled 'You have selected the following repository:' shows the repository name 'cernanalysispreservation/analysispreservation.cern.ch'. There are three main options for uploading: 1) 'Upload snapshot of repository' with an 'Upload' button; 2) 'Automatically Upload on release' with an 'Upload onRelease' button; and 3) 'Automatically Upload on push event' with an 'Upload onPush' button. Each option includes a brief description of the upload process. On the right side, there is a metadata table and a file/repo list section.

ID	FASER-2022-9
Collection	FASER Analysis v0.0.11
Status	draft
Creator	sunje.dallmeier-tiessen@cern.ch
Published URL	Not published yet
Created	5 minutes ago
Last Updated	5 minutes ago

Files | Data | Repos

All Files
No files uploaded yet

All Repositories
No repos uploaded yet

Full control for the users: permissions and reviews

Publish your Analysis
Publish

Publishing is the way to preserve your work within CAP (and CAP only). It makes a snapshot of everything that your analysis contains - metadata, files, plots, repositories - assigning to it an unique versioned identifier. All members of your collaboration can search and reference published content. Once published analysis cannot be deleted, but can be modified and published again with a new version tag.

Access & Permissions
Add

Email	Type	Permissions	Action
pamfilos.fokianos@cern.ch	user	Write ▾	
sunje.dallmeier-tiessen@cern.ch	owner	Admin ▾	

< 1 >

Delete
Delete permantly your analysis and all metadata
Delete

Give user/egroup permissions ✕

Search For

Users E groups

Name	Email	Department	Permissions	Action
atlas-cms-publication-committees	atlas-cms-publication-committees@cern.ch	egroup	Read ▾	Add
cms-publication-committee	cms-publication-committee@cern.ch	egroup	Read ▾	Add
cms-publication-committee-admins	cms-publication-committee-admins@cern.ch	egroup	Read ▾	Add
cms-publication-committee-chair	cms-publication-committee-chair@cern.ch	egroup	Read ▾	Add
cms-publication-process-review	cms-publication-process-review@cern.ch	egroup	Read ▾	Add

Cancel OK

CAP is already being used by some experiments

The screenshot displays the CAP interface for editing a questionnaire. The main content area is titled "1. Analysis Context" and contains several sections for data entry:

- 1.1 Your name:** A text input field with the instruction "Please insert the name of the person filling up the questionnaire."
- 1.2 Your email address:** A text input field.
- 1.3 Working group:** A text input field with the instruction "PAG/POG identifier".
- 1.4 CADI entry number, if available:** A text input field with the instruction "If a CADI entry is available, please make sure you fill it correctly, as an email address. If no CADI entry is available, please provide below the title and the CADI line. If no CADI entry is available, please provide below the title and the CADI line." (Note: the instruction text is partially obscured).
- 1.5 Title and references (if CADI number not available):** A text input field with the instruction "If you don't have a CADI entry number yet, please fill this field with the title and references (Analysis Note numbers, etc.)".
- 1.6 Next deadline date (typically, preapproval date):** A date selection field with the instruction "Please cite the preapproval date (if it has been set), or the closest next deadline date (assuming you fill it when freezing for preapproval)".
- 1.7 Three-line summary of the analysis:** A text input field with the instruction "Please describe briefly the analysis (what is being measured or sought for, and why it is important)."

On the right side, a metadata table is visible:

ID	7adba2926263400b8d9d8b0fd53b49f0
Collection	CMS Statistics Questionnaire v0.0.2
Status	draft
Creator	info@inveniosoftware.org
Published URL	Not published yet

Below the main form, a secondary window shows the "3. Multivariate Discriminants (ML)" section with the following questions:

- 3.1. Have you read the SC TWiki page on MVA recommendations?** (Yes/No radio buttons)
- 3.2. Have you read the ML Groups page on MVA recommendations?** (Yes/No radio buttons)
- 3.3. Are you making use of any of the centralized CMS ML applications?** (Select the answers that apply. Multiple choices allowed. For any other software, please specify it here.)
- 3.4. Are you using a multi-variate discriminant in your data analysis to discriminate between signal and background?** (If the answer is "No", please proceed with the next page of the questionnaire. Yes/No radio buttons)
- 3.5. What software are you using?** (Select the software from the list. Multiple choices allowed. If the software is not listed above, please specify it here.)

The interface includes a sidebar with navigation options (Overview, Edit, Connect, Workflows, Settings) and a top navigation bar with a search bar and user account options.

Control centre for a project or an experiment (aka the “admin panel”)

What is this?

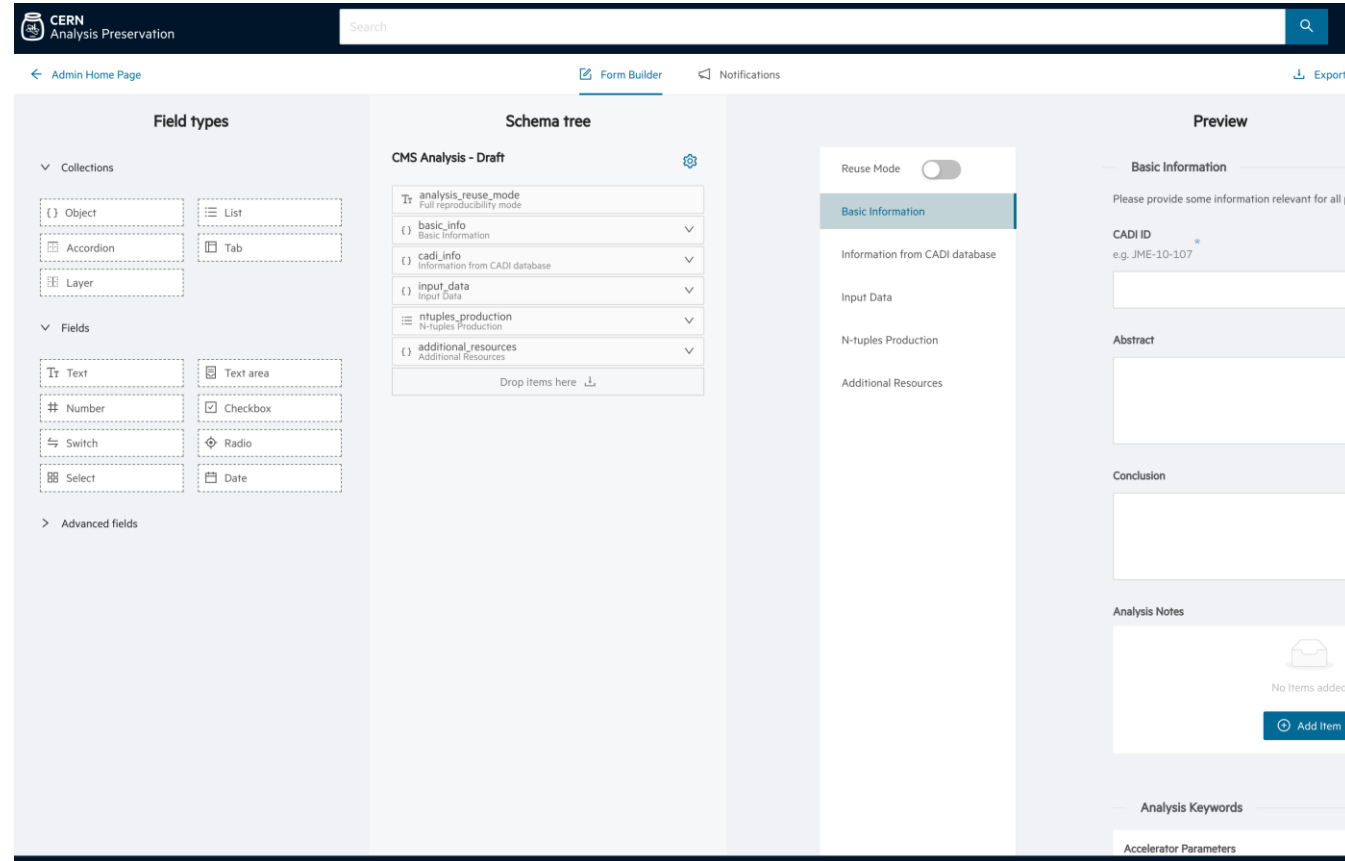
Easier control over communities, schemas, interfaces

Who is this for?

Conveners, managers, boards...

Why?

No need to ask us for smaller changes, you got full control over communities and settings



Example admin panel for a CMS analysis

← Admin Home Page Form Builder Notifications

Field settings

{ } root / { } basic_info / abstract

basic_info [✎](#)

Schema Settings UI Schema Settings

UI Schema

UI Options

Field Width

25% 33% 50% 66% 75% 100%

Rows
The number of the textarea rows

Max Length
Provide a number as the maximum limit of characters, infinity if not provided

Min Length
Provide a number as the minimum limit of charactes, empty if not provided

Placeholder
Provide a placeholder for the field

Schema tree

CMS Analysis - Draft ⚙️

- Tr analysis_reuse_mode
Full reproducibility mode
- { } basic_info
Basic Information ^
- Tr cadi_id
CADI ID
- abstract
Abstract
- conclusion
Conclusion
- ana_notes
Analysis Notes v
- analysis_keywords
Analysis Keywords v
- Drop items here ↓
- cadi_info
Information from CADI database v
- input_data
Input Data v
- ntuples_production
N-tuples Production v
- additional_resources
Additional Resources v
- Drop items here ↓

Reuse Mode

Basic Information

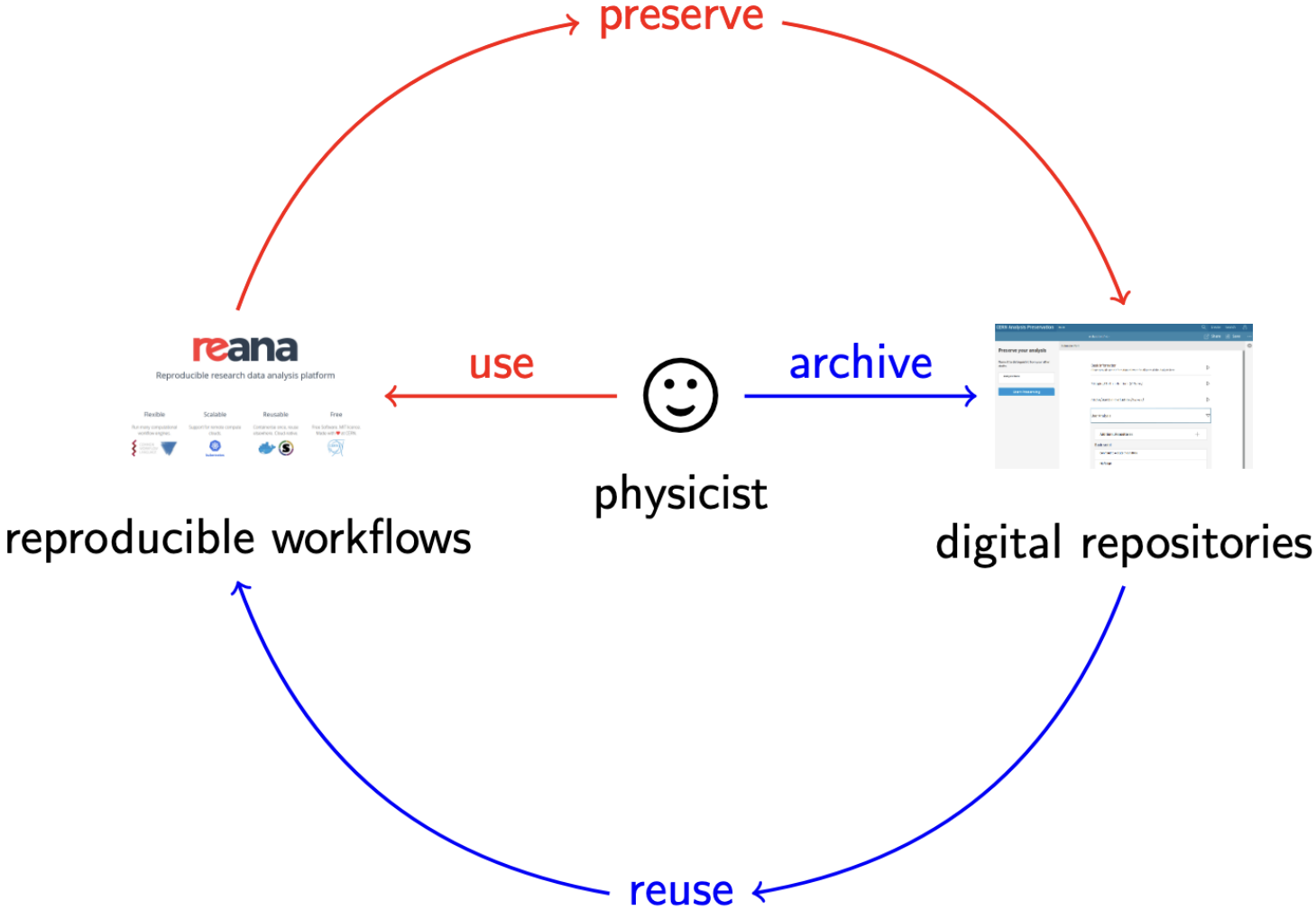
Information from CADI database

Input Data

N-tuples Production

Additional Resources

Reproducibility \Rightarrow Preservation



@tiborsimko

Conclusions

- CERN Analysis Preservation a tool to manage, preserve and share internally research “products” around an analysis
- Preserving integrity of research, enable FAIR research, and prepare for Open Science
- Already being used by some experiments

Future:

- Monitor and enhance FAIR developments
- Integrate more with the experiments, i.e. exploit potential of “admin” panel
- Integrate with REANA

Research integrity: access and reusability

Foundation of research integrity: accessibility of all components, quality assurance, reusability

Enabling research integrity and openness through preservation the “entire” analysis workflow. Ensuring the value of our research beyond contract durations, experiment lifetime etc.

Many funders, political bodies and research organisations stressed the importance of research integrity, open science and FAIR

Thank You!

Questions?

analysis-preservation-support@cern.ch



Scientific
Information
Service

scientific-info.cern