

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

July 2020

Kamran Naim*^{id} Tullio Basaglia^{id} Jelena Brankovic^{id} Pamfilos Fokianos^{id}
Jose Benito Gonzalez Lopez^{id} Maria Grazia Pia^{id} Alexander Kohls^{id}
Artemis Lavasa^{id} Lars Holm Nielsen^{id} Stephanie van de Sandt^{id} Javier Serrano^{id}
Tim Smith^{id}

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 that enable research at the forefront of human knowledge, perform world-class research in fundamental physics, and unite 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 embodied the values that have more recently come to be defined at the Open Science movement, which describes "research and development that is collaborative, transparent and reproducible and whose outputs are publicly available" (European Commission (2018)). Indeed these values were enshrined in 1953 in CERN's founding Convention, which states that "...the results of its experimental and theoretical work shall be published or otherwise made generally available," (CERN Council (1953)) providing the Organization with an early Open Science manifesto.

The embrace of the values of Open Science at CERN is not considered an obligation, but rather an expression of the collective moral and financial responsibility to member states and the global scientific community to advance the boundaries of human knowledge. The community at CERN helped build the Open Internet, were early adopters of Open Source, helped usher in the preprint culture, and helped to pioneer initiatives to allow Open Access to scientific publications. Yet, the complexity of the Laboratory makes the comprehensive pursuit of Open Science an ongoing and ever-evolving effort. Supported by long-term human and financial investments by the international community, CERN hosts dynamic research teams consisting of thousands of researchers from

around the world, who collectively publish almost 1,000 research articles per year. These scientists are engaged in multiple experimental collaborations that operate largely independently of each other and have different tools and practices. The high-energy physics research conducted at CERN is one of the most data-intensive branches of science: particle collisions at CERN produce about 90 petabytes of collision data per year. The complexity, scale and value of these data presents unique and unprecedented challenges for the research community at CERN. The responsible stewardship, analysis and preservation of these data requires a supporting organizational research culture, but also a range of tools and services to optimize how data are used, verified, and built upon to advance knowledge.

On June 19th, 2020, CERN's highest governing authority, the CERN Council announced 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, outlining 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 accelerates 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 help shape the emerging consensus on Open Science to be adopted for publicly funded research, and should then implement a policy of Open Science for the field"(European Strategy Group (2020)).

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 effective, collaborative, and responsive to the needs of the scientific community. We aim to demonstrate that despite the complexity of the research undertaken at CERN, Open Science can be advanced through concerted efforts and as such, the CERN example could serve as an inspiration for the global scientific community.

* kamran.naim@cern.ch

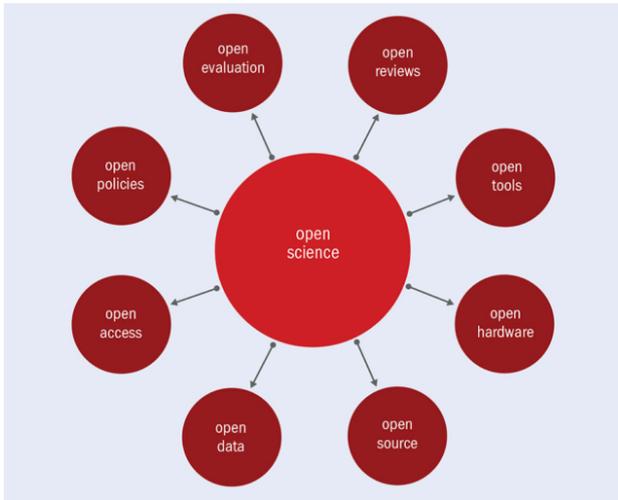


Figure 1: Open Movements. Open Science encompasses all aspects of how scientific research is governed, performed, shared, published and evaluated (Dallmeier-Tiessen & Šimko (2019)).

DEFINITION OF OPEN SCIENCE: THE KEY PILLARS

Open Science. Open Science is an umbrella concept that combines multiple schools of thought based on the aim to solve defects of the traditional research ecosystem. The Open Science movement may aim to make knowledge freely available for everyone, it may be about opening the process of knowledge creation, or about the creation of openly available platforms, tools and services for scientists. But it also refers to public aspects such as making research accessible for citizens or meta-aspects such as the development of alternative metrics for measuring scientific impact (Fecher & Friesike (2014)).

Open Science is not just about opening science in a way that makes knowledge accessible without cost for everyone. It also refers to opening legal barriers that prevent the free usage of research resources. The following terms describe the most common key pillars of Open Science. However, the selection is not complete and the understanding of each term may vary between different stakeholders and disciplines.

Open Access. The European Commission defines Open Access as “as the practice of providing online access to scientific information that is free of charge to the user and that is re-usable” (European Commission (2020)). Thus Open Access refers to principles and practices that foster the openness of research publications in a way that enables others to access and interact freely with them.

The Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities (Max-Planck-Gesellschaft (2003)), signed by CERN in 2004, defines two main re-

quirements for Open Access contributions. All users must be granted “free, irrevocable, worldwide, right of access to, and a license to copy, use, distribute, transmit and display the work publicly and to make and distribute derivative works, in any digital medium for any responsible purpose, subject to proper attribution of authorship.” Additionally, the publication and all supplemental materials, should be deposited “in at least one online repository using suitable technical standards [...] that is supported and maintained by an academic institution, scholarly society, government agency, or other well-established organization that seeks to enable Open Access, unrestricted distribution, interoperability, and long-term archiving.” Ideally, Open Access publishing includes the original scientific research results, raw data and metadata, source materials, digital representations of pictorial and graphical materials and scholarly multimedia material.

Open Data. The Open Data movement is based on the thought that research data should be made freely available to everyone in a timely manner. Data does not only refer to datasets that are produced within the process of research: administrative or governmental data should be openly available for research. Ideally, Open Data includes the right that others may use, transform or share the data as they wish. Thus, the Open Data Handbook defines Open Data as the following: “Open data is data that can be freely used, reused and redistributed by anyone - subject only, at most, to the requirement to attribute and sharealike” (Open Knowledge Foundation (2020)). In this context, Open Data does not only require that data are published under open licenses such as CC0 and CC-BY that allow the usage and transformation of the data, it also means that data are available in a convenient and modifiable format. The FAIR data principles (Wilkinson et al. (2016)) provide guidance on how Open Data can be turned into a reality by making data human- and machine-readable. The principles declare that data should be Findable, Accessible, Interoperable, and Reusable.

Open Data does not enforce all data to be openly available without restrictions. It is rather the philosophy that data should be available as open as possible. Some data require special protection such as clinical or sensitive data, or data that includes the geolocation of endangered species. In these cases, only the metadata may be accessible in a designated research data repository.

Open Tools. The term Open Tools is often used in a similar context as Open Data. Tools refer in this context to (research) software that should be made publicly available under open licenses as a supplement to the research publication. It may also refer to the way software is being developed, which should be as open and transparent as possible from the beginning using community develop-

ment platforms like GitHub¹ for example.

Open Notebook Science. This includes the opening of the whole research process and insights in every stage. Entire research projects are made openly available from the beginning, granting others access to virtual research workspaces.

Open Source. Open Source or Free Software describes software that is publicly available under an open license that grants others the right to modify, expand, use or share the source code, design, or blueprint, etc. The Free Software Foundation defines free software as software where “the users have the freedom to run, copy, distribute, study, change and improve the software,” (Free Software Foundation (2019)) in contrast to proprietary software. The Open Source Initiative expands this definition, by stating that Open Source provides not only free access to the source code, but includes a license that allows free distribution. The source code must be included in the software release. The chosen license must allow modifications, derived works, and sharing under equal conditions. Additionally, the content may not be discriminating and the technology should be neutral (The Open Source Initiative (2007)). Commonly used open licenses that are approved by the Open Source Initiative are the MIT license, the GNU General Public License, and the Apache License 2.0 among others. Open software can also be openly developed, e.g. on shared platforms such as GitHub.

Open Research Assessment. This describes ways of organized research assessment beyond standard methods such as the double-blind peer review. Open Research Assessment makes all or different parts of the review processes transparent, which is meant to decrease the possibility of arbitrariness. There are three ways in which peer review, a standard for research assessment, could be organized to be Open:

- The identity of reviewers is given and clearly visible, which is meant to increase the quality of the review as it is no longer anonymous.
- Secondly, the way reviews are submitted can be open, which means that review reports are publicly available to everyone as well as/including the history of submission.
- Lastly, there is the concept of open participation, which means that not just assigned reviewers can comment but the whole community.

In praxis, journals like PeerJ Preprints allow early communication or feedback before the actual peer review process starts². After the peer review, the article is still open for comments from the community and the whole

review process is transparently available.

Now, with the proliferation of studies related to the COVID-19 pandemic, rapid peer review of preprints has become even more essential. This crisis has increased the popularity for platforms such as the Outbreak Science Rapid PREreview.³

Citizen Science. The Oxford English Dictionary defines ‘Citizen science’ as “scientific work undertaken by members of the general public, often in collaboration with or under the direction of professional scientists and scientific institutions” (Dictionary (2020)). Citizen Science means that research is conducted fully or in parts by the broader public, which is meant to either raise the public awareness or increase the knowledge of individuals.

But involving amateur scientists can also benefit the research itself as solutions to unsolved problems may be solved by adding a different perspective or tasks that were too big before may now be manageable. In general, there are different approaches on how to include amateur scientists. Ecology studies often require large datasets, which is why there are many Citizen Science projects in this area. There are expert-assisted citizen science projects, where citizens are asked to send material to experts who analyze it (e.g. taxonomic classifications) and these experts then send the structured data to the actual scientists. An example for this approach is the bee sampling and identification project in Switzerland (Le Féon et al. (2016)). Another approach requires that the citizens directly send their observations to the scientist. Zooniverse⁴ is a platform for citizen science projects across multiple disciplines. Another approach is sharing computing resources by participating in distributed computing projects.

OPEN ACCESS AT CERN

The Open Access (OA) movement aims to grant anyone, anywhere and anytime, free access to the results of scientific research. Although OA was formalized as a global movement in 2003 with a number of declarations advancing the free availability of scientific research, the high-energy physics community at CERN had been practicing OA through its culture of sharing preprints for over six decades. First achieved through mass-mailing for maximum circulation, with the advent of the internet the HEP community spearheaded the culture of online research repositories with the establishment of arXiv at Los Alamos Laboratory in 1991. Broader distribution of knowledge was achieved through the World Wide Web, created 1989 by Tim Berners Lee while working at CERN, and its underlying source code being placed in the public domain by CERN in 1993.

¹<https://github.com/>

²<https://peerj.com/preprints/>

³<https://outbreaksci.prereview.org/>

⁴<https://www.zooniverse.org/projects>

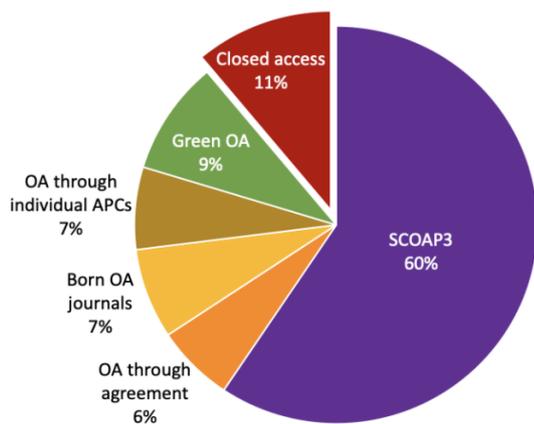


Figure 2: Open Access Status of CERN Publications (2019)

CERN Open Access Policy

CERN issued its Open Access policy in October 2014, requiring OA publishing of all original high-energy physics results, whether experimental or theoretical. In 2017, the policy was expanded further to all publications in related fields such as instrumentation, scientific computing and accelerator physics (CERN (2017)). As a result of this policy, the overwhelming majority of CERN research is published OA. In 2019, CERN authors published a total of 941 scientific articles (and in addition 663 conference proceedings). 80% of scholarly articles from CERN were published immediately Open Access, with an additional 9% of the articles deposited as a preprint in arXiv⁵. A total of 89% of CERN research articles during 2019 were freely accessible and reusable by researchers worldwide (CERN Scientific Information Service (2020)).

Open Access to CERN’s research output is achieved through a number of agreements with individual publishers and CERN’s Scientific Information Service (including new ‘transformational agreements’), with the majority supported through CERN’s participation in the SCOAP³ project. As a consequence of the diversity of CERN research, a number of scientific articles are also published outside such agreements. CERN still aims to also publish these articles OA on a by-article basis. To facilitate the administration of a wide variety of OA arrangements, CERN is a pilot partner of the OA Switchboard initiative⁶ which aims to simplify the exchange of metadata to facilitate OA funding requests at the article level.

⁵<https://arxiv.org/>

⁶<https://www.oaswitchboard.org/>

The Sponsoring Consortium for Open Access Publishing in Particle Physics (SCOAP³)

The Sponsoring Consortium for Open Access Publishing in Particle Physics (SCOAP³)⁷ is an Open Access collaboration hosted at CERN, involving over 3,000 libraries, national funding agencies and research institutions from 43 countries and three intergovernmental organizations. It functions on the basis of a ‘redirection of funds’ business model, as a collaboration between libraries, national funding agencies and publishers of HEP journals. By centrally covering the costs involved in providing Open Access, SCOAP³ pays the publishers directly, thus removing subscription fees for individual journals and any expenses scientists might normally incur to publish their articles openly (i.e. Author Processing Charges or APCs). This way, authors from anywhere in the world publish without any financial burden and retain the copyright of their work. Libraries pay their membership fees to the consortium, reusing funds previously spent on subscription fees for the journals which are now Open Access and financial contributions to the collective fund are calculated based on a function of their GDP as well as their respective national published output in supported journals. Since its launch in 2014, SCOAP³ has made more than 35,000 scientific articles published by authors from over 100 countries freely accessible to everyone. Through the partnership between libraries and the world’s leading outlets in high-energy physics, today, nearly 90% of scientific articles in this field are available to everyone without restriction and authors from anywhere in the world can publish their papers without any financial barriers.

Open Access Services at CERN

CERN maintains a range of services for the more ‘traditional’ outputs, such as papers, proceedings, multimedia, presentations, etc. INSPIRE is the core HEP information system operated and curated by a consortium consisting of DESY, Fermilab, IHEP, IN2P3, SLAC, and CERN. The service focuses on aggregating scholarly works from the global HEP community from various sources ranging from publishers to arXiv. INSPIRE also connects to data providers, such as Zenodo, DataVerse⁸ and Figshare⁹, and indexes the community data repository HEPData. To support the fair assessment of research and researchers, INSPIRE also computes transparent citation metrics for scholarly articles including citations from and to preprints. To further foster this important cause, the INSPIRE collaboration is in the process of signing the Declaration on Research Assessment (DORA)¹⁰.

The CERN Document Server (CDS)¹¹ is CERN’s insti-

⁷<https://scoap3.org/>

⁸<https://dataverse.org/>

⁹<https://figshare.com/>

¹⁰<https://sfedora.org/>

¹¹<https://cds.cern.ch/>

tutional repository which gives access to CERN works and related HEP scholarly literature (preprints, articles, etc.), as well as administrative and multimedia content. Through CDS, CERN also publishes various CERN reports, most notably the CERN Yellow Report series¹², which was started in 1955. It provided a medium for communicating the work done at CERN in open access. The Yellow Report series includes the proceedings of schools and of workshops having a large impact on the future of CERN, as well as reports on detectors and technical papers from individual CERN divisions.

OPEN DATA AT CERN

The four particle detectors of CERN's LHC are home to the experiments ALICE, ATLAS, CMS and LHCb. These detectors, operated by international collaborations of scientists and engineers, generate data at the scale of about 90 petabytes per year, likely making the LHC the world's most prolific source of experimental data. Members of the LHC experiments perform their research using these data, studying the fundamental particles and forces of nature, and testing various theories and models that have been proposed to explain their behavior. CERN experimental data are unique and the result of vast long-term human and financial investment by the international community, and form an important part of the scientific legacy of the LHC.

CERN upholds the principle that Open Access to data and the adoption of FAIR standards in making data available will, in the long term, optimize investments and allow the maximum realization of their scientific potential. As such, it represents a realization of the collective moral and fiduciary responsibility to member states and the global scientific enterprise more broadly. Opening experimental data from CERN presents unique opportunities but also unprecedented challenges, due to the complexity, value, and scale of these data. Making data available responsibly—at different levels of abstraction and at different points in time—ensures that it can be appropriately used, verified and built upon to advance knowledge. Preserving associated software and analyses supports reproducibility and can underpin further research.

The data generated by the CERN experiments take many forms. Starting from either raw experimental or simulated data through to reconstructed data and the datasets of higher abstraction generated by analysis workflows, and finally all the way to data represented in scientific publications. Each of these layers has the potential to afford different opportunities for long-term reuse and poses different challenges for openness and preservation.

¹²<https://cds.cern.ch/collection/CERN%20Yellow%20Reports>

Open Data Policies

To ensure that LHC data can be fully exploited for their scientific value by researchers and communities beyond the CERN collaborations, the four LHC experiments have adopted policies for preservation and access.¹³ In alignment with global efforts, and to re-emphasize CERN's leadership in Open Science, CERN is currently pursuing an organizational CERN Open Data policy, which aims to empower the CERN experiments by adopting a consistent approach towards the openness and preservation of experimental data to maximize their long-term value. The policy describes how the LHC experiments believe in making data publicly available—at different levels of complexity—and where practical to enable their reuse by a wide community and a diverse set of use cases including for: collaboration members long after the data are taken, experimental and theoretical HEP scientists who were not members of the collaboration, professional science, educational and outreach initiatives, and citizen scientists in the general public. The LHC experiments will provide open access to their data after a suitable embargo period, allowing experiment collaborators to fully exploit their scientific potential.

Open Data Services at CERN

True Open Science demands more than simply making data available: it needs to concern itself with providing information on how to repeat or verify an analysis performed over given datasets, producing results that can be reused by others for comparison, confirmation or simply to ensure deeper understanding and inspiration. This requires runnable examples of how the research was performed, accompanied by software, documentation, runnable scripts, notebooks, workflows and computational environments. It is often too late to try to document research in such detail once it has been published.

Open Data is identified as a key pillar of Open Science, enabling open research practices through scientific reuse, verification and reproducibility of research results, and accelerating science and discovery through re-exploitation of datasets. At CERN we have learned that the act of simply opening up the data itself is not sufficient to achieve the benefits of open and reproducible research. Indeed, openness alone does not guarantee reuse and reproducibility, and as such should not be pursued as a goal in itself.¹⁴ Openness of data needs to take a more holistic perspective. Rather than opening data after research is completed, considerations regarding the openness of data need to be adopted throughout the research life-cycle: from the collection of data, to its analysis, availability and

¹³Policies for LHC Experiments are available at: <http://opendata.cern.ch/collection/data-policies>

¹⁴For a detailed explanation of best practices established at CERN that demonstrate how research reproducibility requires going beyond openness, see: Chen et al. (2019)

preservation. Further, focusing on data is not sufficient: they must be accompanied by software, workflows and explanations, to allow researchers to derive the maximum meaning and scientific value from them. These associated analysis artifacts need to be captured throughout the regular iterative, closed research cycle, but preserved in such a manner that they complement Open Data releases to ensure their optimal use, reuse and reproducibility. Rather than thinking about Open Data from its inception, it is more useful to consider data to be FAIR throughout the scientific process, from being initially closed to being made meaningfully open later to those outside experimental collaborations. This suggests the adoption of a set of scientific practices and technological solutions that can be readily integrated into existing research processes and procedures, making them a seamless part of a researcher's daily work.

The CERN Analysis Preservation and Reuse framework (Chen et al. (2019)) consists of a set of services and tools developed together with the LHC collaborations, that assist researchers in describing and preserving all components of a physics analysis such as data, software and computing environment. Each of these services, built on free and open source software, aim to provide FAIR compliant data and can be set up for other research communities as they are implemented using flexible data models. Data provenance is a key feature of each of these services, facilitating reproducibility and data sharing.

CERN Analysis Preservation (CAP)¹⁵. CAP is an aggregator of high-level physics information about individual physics analyses that facilitates preservation, discoverability and reproducibility. It is interconnected with collaboration databases and workflows to allow seamless integration into the researchers' workflows. The preservation of physics analyses in CAP ensures long-term accessibility, but does not directly address how they can be reused, which is a core motivation for researchers to preserve their work in the first place.

Reusable Analysis Service (REANA)¹⁶. REANA allows researchers to instantiate preserved research data analyses on remote compute clouds using modern container technologies, as well as run to 'live' analyses that are still ongoing and have not yet been published or preserved.

RECAST¹⁷. RECAST (Request Efficiency Computation for Alternative Signal Theories) is an analysis reinterpretation framework, which enables physicists to test alternative physical theories through simulations of what those theories predict and running the simulated data through analysis workflows previously used for publications. The application provides for restricted queries on

the data that are reliable and high-impact.

The Durham High-Energy Physics Database (HEP-Data)¹⁸. HEPData is a repository for publication-related high-energy physics data that is run in collaboration by the Institute for Particle Physics Phenomenology at Durham University in England, and CERN's Scientific Information Service. It comprises data points from plots and tables from relevant publications and mints persistent identifiers for them.

CERN Open Data (COD)¹⁹. The CERN Open Data portal allows experiments to share their data— currently over two petabytes of HEP experimental collision and simulated datasets, software, configuration files, virtual machines, etc. —with a double focus: for the scientific community, including researchers outside the CERN experimental teams as well as citizen scientists, and for the purposes of training and education through specially curated resources. Data are published on the platform after embargo periods, which are specified by data policies of each individual LHC collaboration. The service captures extensive metadata to describe the complex materials on the portal to ensure preservation and discoverability, and includes usage instructions, related software and other supplementary materials along with the release of data to enable reuse.

Zenodo²⁰. Zenodo is a multi-disciplinary open research repository hosted by CERN and commissioned by the European Commission (EC) through the OpenAIRE project to support the Open Data and Open Access movements in Europe. The OpenAIRE project²¹, in the vanguard of the Open Access and Open Data movements in Europe was commissioned by the EC to support their nascent Open Data policy by providing a catch-all repository for EC funded research. CERN, an OpenAIRE partner and pioneer in Open Source, Open Access and Open Data, provided this capability and Zenodo was launched in May 2013. Zenodo quickly became a successful product world-wide and got additional project funding from the Alfred P. Sloan Foundation and Arcadia Fund. Zenodo accommodates all kinds of research artifacts that can be grouped into community collections. Zenodo played a key role making software a first class citizen in scholarly publications. Thanks to its integration with GitHub (GitHub (2016)), which allows automatic archival of software releases, Zenodo hosts 80% of the world's software DOIs (Fenner et al. (2018)). Zenodo provides leading features such as versioning (to support reproducibility) and metrics as usage statistics (following Make Data Count and COUNTER guidelines). Zenodo played an active role during the recent scientific efforts to fight COVID-19, by creating a central Open Access sha-

¹⁵<https://analysispreservation.cern.ch/>

¹⁶<http://reana.io/>

¹⁷<https://iris-hep.org/projects/recast.html>

¹⁸<https://www.hepdata.net/>

¹⁹<http://opendata.cern.ch/>

²⁰<https://zenodo.org/>

²¹<https://www.openaire.eu/>

ring space for related research artifacts between disciplines (Ioannidis (2020)).

OPEN TOOLS, OPEN SOURCE & OPEN HARDWARE

The cornerstone of the Open Source philosophy is that the recipients of technology should have access to all its building blocks, such as software code, schematics for electronics and mechanical designs, in order to study it, modify it and redistribute it to others.

To serve the need of CERN's open scientific collaborations to share information between the experiments at the host laboratory and the scientists in universities and institutes around the world, the World Wide Web software was created. CERN and sister institutes instantiated web servers, and put information online, including at SLAC in California whose preprint server SPIRES (predecessor of INSPIRE) became the first web server in the USA. To enable everyone to use the WWW software and more importantly contribute to improving it and building a common web, CERN released the software firstly into the public domain and subsequently under a pioneering Open Source software licence. Within a few years the concept spread from physics institutes to all of science, and on to society, bringing businesses online and transforming life as we know it. Ever since releasing the WWW software under an Open Source model in 1994 (Smith & Flückiger (2020)), CERN has continuously been a pioneer in this field, and makes use of Free and Open Source Software (FOSS) where possible.

One example is **Invenio**²², a free and openly developed integrated digital repository system that was born at CERN around 20 years ago. The technology offered by the software covers all aspects of digital repository management. It complies with the Next Generation Repository (NGR) standard and supports multiple bibliographic standards for metadata. Its flexibility and performance make it a comprehensive solution for the management of document and data repositories of small to very large size. Invenio is a framework and toolbox for building and managing a digital repository server. Together with 18 partner institutions, CERN currently builds a turn-key Open Source research data management platform called InvenioRDM, which main purpose is to enable reproducibility and reuse of research artifacts (Nielsen (2019)).

Indico²³ is another example of an Open Source software, developed by the CERN Document Server Software Consortium to schedule and organize events, from simple lectures to complex meetings, workshops and conferences with sessions and contributions. The tool also includes an advanced user delegation mechanism, allows paper re-

viewing, archival of conference information and electronic proceedings (Indico (2020)). Indico is licensed under a GNU General Public License (GPL).

EOS²⁴ is a multi-protocol software solution for central data recording, user analysis and data processing, developed at CERN since 2010 to store physics analysis and data on physics experiments (including the LHC experiments). The main goal of the project is to provide fast and reliable disk only storage technology for CERN LHC use cases. The system became operational in 2011 for the ATLAS experiment and has grown in scale and capacity since then. (Peters et al. (2015)) Since its release, in version 0.1.1, EOS has been available under GPL v3 license.

The **MAlt Project**²⁵(Ormancey (2019)) is a multi-year effort which aims to investigate the migration from commercial software products (Microsoft and others) to Open Source solutions, so as to minimize CERN's exposure to the risks of unsustainable commercial conditions. The Laboratory aims to play a pioneering role among public research institutions, most of whom have recently been faced with the same dilemma.

Geant4²⁶ is a toolkit for simulating the passage of particles through matter, implemented in the C++ programming language. It provides a wide range of functionality for particle transport including physics models, geometry and materials description, tracking, and detector modeling. Geant4 development was carried out in 1994-1998 by the CERN RD44 project, which gathered an international research team of physicists and software engineers. The Geant4 Collaboration was established in January 1999 to continue the development and refinement of the toolkit, and to provide maintenance and user support. (Agostinelli et al. (2003))

Since its first version in December 1998, Geant4 has been freely released as open source software. It has been applied in particle physics, nuclear physics, astronomy and astrophysics, accelerator design, space and nuclear engineering, medical physics and radiation protection, and in the cultural heritage domain.

RD44 pioneered the adoption of rigorous software engineering and object-oriented technology in the experimental physics environment. While geographically distributed software development and large-scale object-oriented systems are no longer a novelty, the Geant4 toolkit still represents one of the largest and most ambitious projects of this kind in terms of the size and scope of the code, of the broad variety of multidisciplinary applications, and of the widespread use worldwide. Free availability of the

²²<http://inveniosoftware.org>

²³<https://getindico.io/>

²⁴<http://eos.web.cern.ch>

²⁵<https://malt.web.cern.ch/malt/>

²⁶<https://geant4.web.cern.ch/>

code has been a key to its success, along with openness to the evolving requirements of physics research.

ROOT²⁷ is a framework for data processing that is openly developed at CERN for the special requirements of particle physics analysis and the performance of simulations. Several collaborations have developed their own software based on ROOT.

Open Hardware Licenses²⁸. Open Science brings transparency, reproducibility and more efficient dissemination of knowledge. An important part of it is the software and hardware used in scientific facilities. The CERN Open Hardware Licence does for hardware what Free and Open Source Software licences do for software: it provides a solid legal foundation so that designers can easily share their work. It ensures that others can study, modify and republish designs, and hardware based on those designs. It fosters collaboration inside and across institutions, including industry, and maximizes the impact of hardware developments on society (CERN (2015)).

As a concrete example, the CERN Against COVID-19 initiative recently developed a washable 3D printed mask with replaceable filters, which was released via the CERN Open Hardware Repository²⁹ in June 2020. This mask, designed by a team of researchers and engineers at CERN, is designed to offer improved face contact over standard tissue masks, offering wearers a higher level of protection from airborne pathogens if the correct filter is applied. The washable mask is made available as an Open Source solution, allowing anyone around the world to 3D print it using standard 3D printing filaments (Jakobsson (2020)).

CITIZEN SCIENCE AT CERN

CERN and the LHC experiments have been involved in a number of efforts to engage the general public in the scientific efforts undertaken at the Laboratory. Under the umbrella of CERN against COVID-19, CERN contributed computers for the combat against COVID using the Folding@home platform³⁰. In addition, CERN has its own volunteer-computing platform called LHC@home³¹ where users can download software and run HEP simulations in a way that does not interfere with a computer's normal operation. Citizen Cyberlab³² is a multidisciplinary platform that focuses on developing methods and studying motivations for new forms of public participation in research. Together with the UN Institute for Training and Research, UNITAR, and the University of Geneva, CERN organizes events from online crowdsourcing to in-person hackathon

to develop Open Source research tools and training to increase the public involvement in scientific and social research. The Higgshunters project³³ was a successful effort by ATLAS to motivate citizens to find interesting features in LHC data. In total over 37,000 citizen scientists participated in the project and classified more than a million LHC events.

INTERNATIONAL ENGAGEMENT

Over the years CERN has actively participated in many EC projects and is currently a partner in multiple Horizon 2020 projects which operate in the area of Open Science, and more specifically focusing on European research infrastructures and e-infrastructures. Some recent examples include OpenAIRE, EOSC-hub³⁴, THOR³⁵, FREYA³⁶, ESCAPE³⁷. Actively participating in EC projects has allowed CERN and the HEP community at large to have a voice in the Open Science discourse or even influence decision-making. Moreover, it is a way of preventing operating in a vacuum through remaining informed about new developments and initiatives, and by gathering expertise from external partners in different domains.

CERN has been at the forefront of supporting and adopting Open Science-related guidelines (to the degree possible); notable examples include the FORCE11 Joint Declaration of Data Citation Principles (Data Citation Synthesis Group (2014)), the FORCE11 Software Citation Principles (Smith et al. (2016)), the FAIR Data Principles (Wilkinson et al. (2016)), the Plan S Principles (European Science Foundation (2020)) (especially the Requirements for Open Access Repositories in the case of the services summarized above), the Code of Practice for Research Data Usage Metrics (COUNTER (2018)) (in Zenodo), and the recommendations of the Declaration on Research Assessment (DORA (2012)) (in INSPIRE). Examples of projects in this area that include a variety of external partners are the Asclepias project (Ioannidis et al. (2018)) and the Biodiversity Literature Repository (BLR), a joint project of Plazi, Pensoft, and Zenodo that opens data from biodiversity literature. Information like taxonomic treatments, images, and the associated metadata is being extracted from closed access publications, and made publicly available. The data are then immediately distributed to global biodiversity platforms, such as the Global Biodiversity Information Facility, which makes original taxonomic information accessible that would otherwise be hidden behind paywalls (Agosti & Pensoft Publishers (2019)).

CERN has also taken a proactive approach to Persistent Identifiers (PIDs); DOI versioning has been supported

²⁷<https://root.cern.ch>

²⁸<https://kt.cern/ohlv2>

²⁹<https://ohwr.org/project/3dmask/wikis/home>

³⁰<https://againstcovid19.cern/actions/cern-contributes-computers-combatting-covid-19>

³¹<https://lhcatome.web.cern.ch/>

³²<https://www.citizencyberlab.org/>

³³<https://www.higgshunters.org/>

³⁴<https://www.eosc-hub.eu/>

³⁵<https://project-thor.eu/>

³⁶<https://www.project-freya.eu/>

³⁷<https://projectescape.eu/>

on Zenodo for years now and a lot of progress has been achieved in the past few years in terms of increasing adoption of mature PID types within community services (e.g. ORCID iDs for people, and DOIs for data, publications, software, etc.). CERN has also been piloting new and emerging PIDs, such as ROR IDs for organizations, and IDs for funders and grants.

RECOMMENDATIONS

Given CERN's extensive engagement in the domain of Open Science, we hope to extend our experiences, knowledge and tools to further accelerate the global adoption of Open Science policies and practices, and to support scientific communities beyond the discipline of high-energy physics to embrace this vision.

Accordingly, we advance the following general recommendations for UNESCO to adopt as it advocates for broader global adoption of Open Science.

1. Guide research practice with open policies. Transitioning research from traditional closed approaches necessitates a policy environment that both enables and incentivizes researchers to adopt open research practices.
2. Open Access Policies should be aligned with global approaches to support the systemic, large-scale transition to open.
 - (a) As a matter of policy, ensure all funded research outputs are published Open Access with CC-BY licenses.
 - (b) Reduce administrative burden/workflow barriers for authors during the submission process to facilitate seamless OA publishing as the default.
 - (c) Ensure that all subscription arrangements with publishers moving forward are 'transformational agreements', transitioning from payment for reading to payment for publishing.
 - (d) Explore cooperative models for Open Access which redirect subscription spending to support OA publishing. Enable opportunities for researchers to publish in top outlets without having to pay APCs.
3. Open Data is not enough. To ensure maximum scientific value is extracted from research data, releases should be accompanied with all necessary artifacts used to perform research, as tools and methods can inspire and enable other fields as much as the results themselves.
 - (a) Analysis preservation mechanisms are critical to ensure that all analysis artifacts are duly recorded and preserved during all stages of the life of a study. These tools can be closed at first if necessary.

- (b) Providing for the revalidation, reinterpretation and reuse of data requires systems that enable access to all components of research data analyses (i.e. the original computing environment, experimental datasets, analysis software, and computational workflow steps).

4. The development and maintenance of robust Open Source Hardware and Software should become a cultural norm within major research organizations, requiring strategic investment and enabling policies that support and incentivize such norms.
5. The assessment of the value of research as well as the performance of scientists needs to be fair and transparent, with a focus on the specific impact of the research on the discipline and for wider society, requiring forms/metrics for evaluation. In particular, the performance of researchers needs to be decoupled from the perceived impact of the outlet where research results are published. New transparent mechanisms for research assessment should include not only direct relevance on other scholarly communication (e.g. through citations), but also future reuse of generated data, processes or workflows and the actual application of results where relevant.

The COVID-19 global health pandemic has brought into sharp focus both the inadequacies of traditional science, but also the embrace of Open Science practices as the most effective mechanism for addressing global crises. It is our firm belief that we are at an inflection point, where Open Science should go beyond an aspiration, to become the default standard for global research. This will not only equip the scientific community to better respond to current and future challenges, but will further accelerate the advancement of the research enterprise itself, making it more efficient, transparent, responsive and dynamic.

We commend UNESCO for the timeliness of their effort to develop an Open Science Recommendation, and offer to share the knowledge and experience of CERN in this domain to advance global efforts.

References

- Agosti, D. & Pensoft Publishers (2019), 'Data mining applied to scholarly publications to finally reveal Earth's biodiversity'. Library Catalog: www.eurekalert.org.
URL: https://www.eurekalert.org/pub_releases/2019-10/pp-dma101819.php
- Agostinelli, S., Allison, J., Amako, K., Apostolakis, J., Araujo, H., Arce, P., Asai, M., Axen, D., Banerjee, S., Barrand, G., Behner, F., Bellagamba, L., Boudreau, J., Broglia, L., Brunengo, A., Burkhardt, H., Chauvie, S., Chuma, J., Chytrcek, R., Cooperman, G., Cosmo, G., Degtyarenko, P., Dell'Acqua, A., Depaola, G., Dietrich, D., Enami, R., Feliciello, A., Ferguson, C., Fesefeldt,

- H., Folger, G., Foppiano, F., Forti, A., Garelli, S., Giani, S., Giannitrapani, R., Gibin, D., Gómez Cadenas, J. J., González, I., Gracia Abril, G., Greeniaus, G., Greiner, W., Grichine, V., Grossheim, A., Guatelli, S., Gumplinger, P., Hamatsu, R., Hashimoto, K., Hasui, H., Heikkinen, A., Howard, A., Ivanchenko, V., Johnson, A., Jones, F. W., Kallenbach, J., Kanaya, N., Kawabata, M., Kawabata, Y., Kawaguti, M., Kelner, S., Kent, P., Kimura, A., Kodama, T., Kokoulin, R., Kossov, M., Kurashige, H., Lamanna, E., Lampén, T., Lara, V., Lefebvre, V., Lei, F., Liendl, M., Lockman, W., Longo, F., Magni, S., Maire, M., Medernach, E., Minamimoto, K., Mora de Freitas, P., Morita, Y., Murakami, K., Nagamatu, M., Nartallo, R., Nieminen, P., Nishimura, T., Ohtsubo, K., Okamura, M., O’Neale, S., Oohata, Y., Paech, K., Perl, J., Pfeiffer, A., Pia, M. G., Ranjard, F., Rybin, A., Sadilov, S., Di Salvo, E., Santin, G., Sasaki, T., Savvas, N., Sawada, Y., Scherer, S., Sei, S., Sirotenko, V., Smith, D., Starkov, N., Stoecker, H., Sulkimo, J., Takahata, M., Tanaka, S., Tcherniaev, E., Safai Tehrani, E., Tropeano, M., Truscott, P., Uno, H., Urban, L., Urban, P., Verderi, M., Walkden, A., Wander, W., Weber, H., Wellisch, J. P., Wenaus, T., Williams, D. C., Wright, D., Yamada, T., Yoshida, H. & Zschesche, D. (2003), ‘Geant4—a simulation toolkit’, *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **506**(3), 250–303.
URL: [https://doi.org/10.1016/S0168-9002\(03\)01368-8](https://doi.org/10.1016/S0168-9002(03)01368-8)
- CERN (2015), ‘Open Hardware at CERN’.
URL: <https://cds.cern.ch/record/2109248/files/CERN-Brochure-2015-002-Eng.pdf>
- CERN (2017), ‘Open Access Policy for CERN Publications’.
URL: <https://cds.cern.ch/record/1955574>
- CERN Council (1953), ‘CERN Convention for the Establishment of a European Organization for Nuclear Research’.
URL: <https://council.web.cern.ch/en/content/convention-establishment-european-organization-nuclear-research>
- CERN Scientific Information Service (2020), ‘CERN Scientific Information Service : Activity Report 2019’. ISBN: 9789290835684 Library Catalog: cds.cern.ch Publisher: CERN.
URL: <https://cds.cern.ch/record/2715694>
- Chen, X., Dallmeier-Tiessen, S., Dasler, R., Feger, S., Fokianos, P., Gonzalez, J. B., Hirvonsalo, H., Kousidis, D., Lavasa, A., Mele, S., Rodriguez, D. R., Šimko, T., Smith, T., Trisovic, A., Trzcinska, A., Tsanaktsidis, I., Zimmermann, M., Cranmer, K., Heinrich, L., Watts, G., Hildreth, M., Lloret Iglesias, L., Lassila-Perini, K. & Neubert, S. (2019), ‘Open is not enough’, *Nature Physics* **15**(2), 113–119. Number: 2 Publisher: Nature Publishing Group.
URL: <https://doi.org/10.1038/s41567-018-0342-2>
- COUNTER (2018), ‘COUNTER Code of Practice for Research Data Usage Metrics release 1’. Library Catalog: www.projectcounter.org Section: News.
URL: <https://www.projectcounter.org/counter-code-practice-research-data-usage-metrics-release-1/>
- Dallmeier-Tiessen, S. & Šimko, T. (2019), ‘Open science: A vision for collaborative, reproducible and reusable research’, *CERN Courier* **59**(2), 25–26.
URL: <https://cds.cern.ch/record/2666160>
- Data Citation Synthesis Group (2014), ‘Joint Declaration of Data Citation Principles’.
URL: <https://doi.org/10.25490/a97f-egykh>
- Dictionary, O. E. (2020), ‘Citizen Science’.
URL: <https://www.oed.com/view/Entry/33513>
- DORA (2012), ‘San Francisco Declaration on Research Assessment’.
URL: <https://sfdora.org/read/>
- European Commission (2018), ‘Open Science Policy Platform Recommendations’.
URL: https://ec.europa.eu/research/openscience/pdf/integrated_advice_opssp_recommendations.pdf
- European Commission (2020), ‘Open Access’.
URL: <https://ec.europa.eu/research/openscience/index.cfm?pg=openaccess>
- European Science Foundation (2020), ‘Plan S. Making full and immediate Open Access a reality. Principles and Implementation’.
URL: <https://www.coalition-s.org/addendum-to-the-coalition-s-guidance-on-the-implementation-of-plan-s/principles-and-implementation/>
- European Strategy Group (2020), ‘2020 Update of the European Strategy for Particle Physics’.
URL: <https://home.cern/sites/home.web.cern.ch/files/2020-06/2020%20Update%20European%20Strategy.pdf>
- Fecher, B. & Friesike, S. (2014), Open Science: One Term, Five Schools of Thought, in S. Bartling & S. Friesike, eds, ‘Opening Science: The Evolving Guide on How the Internet is Changing Research, Collaboration and Scholarly Publishing’, Springer International Publishing, Cham, pp. 17–47.
URL: https://doi.org/10.1007/978-3-319-00026-8_2

- Fenner, M., Katz, D. S., Nielsen, L. H. & Smith, A. (2018), 'DOI Registrations for Software'. Library Catalog: blog.datacite.org.
URL: <https://doi.org/10.5438/1nmy-9902>
- Free Software Foundation (2019), 'What is free software?'.
URL: <https://www.gnu.org/philosophy/free-sw.html>
- GitHub (2016), 'Making Your Code Citable'.
URL: <https://guides.github.com/activities/citable-code/>
- Indico (2020), 'Learning Indico'.
URL: <https://learn.getindico.io/>
- Ioannidis, A. (2020), 'Fighting the Coronavirus Disease (COVID-19)'.
URL: <https://blog.zenodo.org/2020/04/02/fighting-the-coronavirus-disease-covid-19/>
- Ioannidis, A., Bigarella, C., Nowak, K. & Robitaille, T. (2018), 'Asclepias Broker. Scholarly link broker for the Asclepias project'.
URL: <https://asclepias-broker.readthedocs.io/en/latest/>
- Jakobsson, E. (2020), 'CERN develops washable 3D-printed face mask'. Library Catalog: kt.cern.
URL: <https://kt.cern/article/cern-develops-washable-3d-printed-face-mask>
- Le Féon, V., Henry, M., Guilbaud, L., Coiffait-Gombault, C., Dufrière, E., Kolodziejczyk, E., Kuhlmann, M., Requier, F. & Vaissière, B. E. (2016), 'An expert-assisted citizen science program involving agricultural high schools provides national patterns on bee species assemblages', *Journal of Insect Conservation* **20**(5), 905–918.
URL: <https://doi.org/10.1007/s10841-016-9927-1>
- Max-Planck-Gesellschaft (2003), 'Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities'.
URL: <https://openaccess.mpg.de/Berlin-Declaration>
- Nielsen, L. H. (2019), 'InvenioRDM: a turn-key open source research data management platform'.
URL: <https://inveniosoftware.org/blog/2019-04-29-rdm/>
- Open Knowledge Foundation (2020), 'The Open Data Handbook. Guides, case studies and resources for government & civil society on the "what, why & how" of open data'.
URL: <https://opendatahandbook.org/>
- Ormaney, E. (2019), 'Migrating to open-source technologies. Taking back control using open software'. Library Catalog: home.cern.
URL: <https://home.cern/news/news/computing/migrating-open-source-technologies>
- Peters, A. J., Sindrilaru, E. A. & Adde, G. (2015), 'EOS as the present and future solution for data storage at CERN', *Journal of Physics: Conference Series* **664**(4), 042042. Publisher: IOP Publishing.
URL: <https://doi.org/10.1088/1742-6596/664/4/042042>
- Smith, A. M., Katz, D. S. & Niemeyer, K. E. (2016), 'Software citation principles', *PeerJ Computer Science* **2**, e86. Publisher: PeerJ Inc.
URL: <https://doi.org/10.7717/peerj-cs.86>
- Smith, T. & Flückiger, F. (2020), 'Licensing the Web'.
URL: <https://home.cern/science/computing/birth-web/licensing-web>
- The Open Source Initiative (2007), 'The Open Source Definition'.
URL: <https://opensource.org/osd>
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A. C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J. & Mons, B. (2016), 'The FAIR Guiding Principles for scientific data management and stewardship', *Scientific Data* **3**(1), 160018. Number: 1 Publisher: Nature Publishing Group.
URL: <https://doi.org/10.1038/sdata.2016.18>