Chapter 5

Research Data Management at CERN

Vera J. Lipton

This chapter is the first of the two in this book documenting experiences with implementing open data. Specifically, it outlines the practices of research data collection, processing, curation, and release as open data. Some early examples of the use of open data are also provided.

The chapter includes:

1. Organisational approaches to research data management
2. Research data management at CERN

Introduction

The explicit policies mandating open data make it clear that the curation and release of scientific data in electronic formats are no longer an issue. Rather, the discussion has shifted to issues such as what specific data to curate and share, how to do it, in what format, at which time, and according to what conditions. The management of research data is dynamically evolving and presents many challenges to research organisations. While data sharing among peer researchers has been an established practice for many years, the digital curation of data for public release is both very recent and complex. Indeed, making scientific data available and useful to unknown audiences, and for unanticipated purposes, may not be easy to achieve.

This chapter deals with some of the evolving aspects of research data management (RDM). It examines data-driven experiments at the European Organization for Nuclear Research (CERN) and documents some emerging best practice with open data. It is acknowledged that it is not feasible to address, within the purview of a single chapter, all unfolding issues associated with the curation and use of open scientific data. The discussion here starts with a brief overview of the data management approaches taken by research organisations.

This is followed by a detailed discussion of these practices at CERN, including analysis of organisational policies underpinning open data. The chapter concludes by summarising the key lesson learnt from open data practice.

5.1 Organisational approaches to research data management

There are no established definitions of RDM in the context of open scientific data. Rather, data management is defined as a set of organisational practices that lead to specific outcomes. Universities have introduced RDM as a new library service to help researchers to ensure compliance with the mandates recently introduced by funders. The service typically involves assistance with planning, creating, organising, sharing,

and looking after research data, whatever form it may take (Cambridge) [256]. Universities acknowledge the key benefits of data sharing for the conduct of science and the benefits for researchers. Some point to successful case studies, and others state that research data represents a significant investment of money, effort, resources, and time (Princeton) [257]. At this stage, most universities tend to view RDM as a short-term function spanning the duration of research projects.

Taken as a whole, most of the reasons (and incentives) for universities now to implement the RDM function appear to be external. This was the case for such well-known universities as Cambridge,[1] Oxford,[2] Harvard [258], Princeton,[3] Stanford [259], Yale,[4] Cornell [260], and Johns Hopkins,[5] as well as the research-intensive Group of Eight universities in Australia.[6] Some less well-known universities—such as Purdue University in the United States and the University of Edinburgh, home to the Digital Curation Centre—seem to have developed more advanced expertise in RDM and so view data preservation as an integral part of their own research processes. Purdue, Yale, and Cornell have also developed data preservation strategies[7] that set out expectations and limit on data preservation and maintenance—including content migration and software and hardware dependency preservation.

Notwithstanding their operating constraints and technological limitations, some of these universities state that preserving the underpinning publications with the data is a high priority, along with any stand-alone data publications and datasets with high research value.[8] However, these policies do not go further to spell out the processes for internal decisions about what is worth preserving.

The experiences with RDM at universities are at early stages. Yet librarians have already positioned themselves as the key players in the RDM process—they link a project's lifecycle to data management because the techniques used by librarians slot nicely into the different parts of the data lifecycle. The six stages of the matrix (**Figure 3**)[9] make up a cycle, with the expectation that data curated by universities will be reused in future research projects. While this data cycle provides a simplified view of RDM, it appears adequate for the purposes of assisting researchers to manage

---

[1] The University of Cambridge defines the stages of RDM as creating, organising, accessing and looking after data.

[2] RDM at Oxford includes planning how research data will be looked after, how researchers deal with information on a day-to-day basis over the lifetime of a project, and what happens to the data in the longer term (http://researchdata.ox.ac.uk/home/introduction-to-rdm/).

[3] The RDM Team at Stanford offers assistance and training that will help researcher create data management plans for grant applications, identify appropriate repositories for research data, understand repository requirements, and deposit data into DataSpace at Princeton University (http://library.princeton.edu/research-data-management).

[4] Yale University has also developed the Library's Digital Preservation Policy Framework, which outlines the scope of digital preservation services at Yale University (https://web.archive.org/web/20160329191611; http://wiki.opf-labs.org/display/SP/Home).

[5] RDM at Oxford includes planning how research data will be looked after, how researchers deal with information on a day-to-day basis over the lifetime of a project, and what happens to the data in the longer term (http://researchdata.ox.ac.uk/home/introduction-to-rdm/).

[6] The University of Adelaide, the Australian National University, the University of Melbourne, Monash University, the University of New South Wales, the University of Queensland, the University of Sydney, and the University of Western Australia.

[7] A good repository of published data preservation strategies is available at https://web.archive.org/web/20150224021208/http://wiki.opf-labs.org:80/display/SP/Home.

[8] *Ibid.*

[9] *Ibid.*

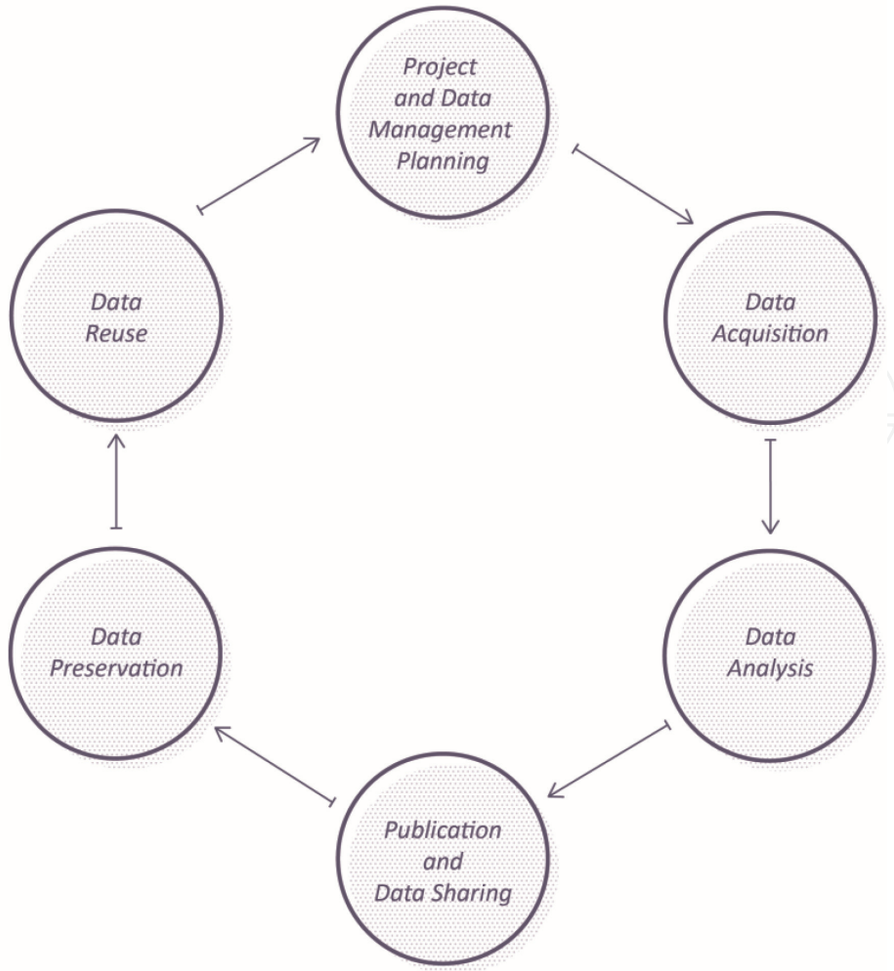

Figure 3.
Research data lifecycle (Source: Briney [261]).

their data, to create a data management plan, and to become aware of the data policies that apply to their work [261]. Although university libraries serve researchers across all scientific disciplines, the curation and preservation of data in the social sciences and humanities are less complex than RDM in other branches of science.

Outside the university sector, RDM in scientific agencies is far better established and forms an integral part of internal research practices. In this context, RDM ensures the long-term preservation of, access to, and the ability to reuse data after research projects have ended. Scientific research organisations and research funders both envisage that preservation should be long term, without defining any specific period.

This flexible approach is also advocated by those institutions that set the standards for data preservation. The leading model in the field, the Open Archival Information System (OAIS), defines 'long-term' as:

> *... long enough to be concerned with the impacts of changing technologies, including support for new media and data formats, or with a changing user community. Long Term may extend indefinitely [93].*

This definition implies that there are two roles in data management—storage (to preserve the data) and curation (to preserve knowledge about the data to facilitate reuse).[10] This definition is user-centric, rather than data producer-centric.

[10] Indeed, this duality is widely discussed by archivists as well as by proponents of open access policies. See, for example, Lee and Stvilia [262]; Digital Curation Centre [263]; and Gladney [264].

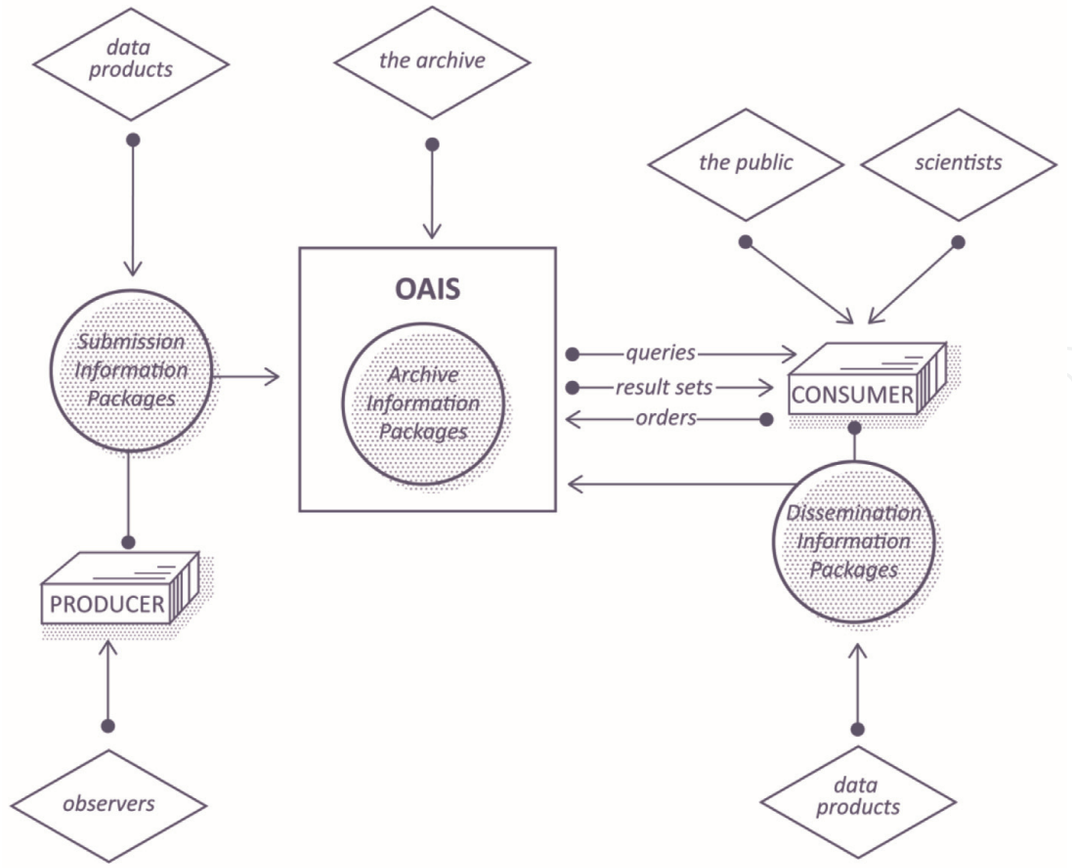

Figure 4.
The highest level structure of an OAIS archive (Source: Reference model for an open archival information system).

Clearly, there is far more to RDM than helping researchers to publish their data so that they comply with the open data mandates. I expect that universities will, over time, both learn from and adopt some of the advanced RDM practices as these evolve and get tested within scientific research organisations. For this reason, the sections below focus on RDM in data-driven scientific research agencies.

So what is required to preserve and maintain access to digital data over the long term? This question is still far from finding a satisfactory answer.

Space agencies have been at the forefront of the debate. The principal model for RDM in large data-driven organisations, including NASA and CERN, is the OAIS reference model. It led to the development of the ISO standard 16363:2012, which has proved useful for research organisations with digital archiving needs. It is the only standard currently endorsed by the Digital Curation Centre[11] for the use in digital preservation planning and management. The structure of the model is illustrated in **Figure 4**, along with the relationships between producers and consumers of data.

In this model,[12] an OAIS archive preserves digital or physical objects for the long term. The archive accepts objects along with metadata and with summaries describing how to interpret the digital objects so as to extract the information within them.

[11] The Digital Curation Centre is an internationally recognised centre of expertise in digital curation with a focus on building capability and skills for research data management (http://www.dcc.ac.uk/).

[12] See *Reference model for an open archival information system (OAIS)*–CCSDS 650.0-B-1. CCSDS Recommendation, 2002. Identical to ISO 14721:2003. Available at: http://public.ccsds.org/publications/archive/650x0b1.pdf [93] and Bicarregui et al. [265].

That information may need further context. The archive receives the bundle of information in the form agreed in a contract between the data producers and the archive. Once the archive receives the information package, it takes over the responsibility for preservation from the producer. The archive distributes its holdings to the data consumers whom the archive is designed to support. It is the responsibility of the archive to determine, either by itself or by way of consultation, which users should become the designated consumers capable of understanding particular data packages. However, the design of the OAIS archive requires the information to be documented in such a way that allows consumers to interpret the data products without any contact with the data producers—an important consideration for future users.

The fundamental OAIS design has become a standard for major digital archives and repositories, including the Library of Congress in the United States, the British Library, the digital library JSTOR, and many others. Some university libraries are already OAIS-compliant. However, the OAIS design is merely a conceptual model that can only be used as a guide for RDM within research organisations. The OAIS model cannot be likened to the 'gold or green' open access standards[13] that were almost uniformly adopted and implemented by research organisations around the globe. There is no 'standard' for RDM, and developing any standards into the future is a far more complex task than was the case with open publications.

There are some major differences between open publications and open data, and these differences underpin the emergence of unique RDM practices that are, and need to be, researcher-centric. At the same time, librarians and research funders tend to approach RDM with a mindset relentlessly focused on creating and applying 'standards' and 'templates', perhaps because they are influenced by their recent experiences with facilitating open access to publications.

RDM is not simply a standardised technical approach to implementing open data mandates. If open scientific data is to be sustainable, then cultural and organisational issues must first be addressed. In particular, a more advanced understanding is needed of the different natures of open data and open publications. The differences between open data and publications and the tools that may be used to improve the availability and reuse of open data are outlined in Chapter 8.

5.2 Research data management at CERN

CERN is one of the earliest and most influential advocates of open science in the world, committed to collaborative research and the dissemination of results in open access publications and, more recently, as open data. Researchers at CERN invented the World Wide Web in 1989, and the organisation is now using it to revolutionise the ways scientists develop, disseminate, and communicate science and to work and to learn collectively in online spaces.

The mandate for openness is embedded in the CERN charter, which states:

[13] Green open access, also referred to self-archiving, refers to the practice of depositing articles in an open access repository, where it can be accessed freely. The self-publication typically occurs after peer review by a journal, the author posts the same content the journal will be publishing to a web site controlled by the author, the research institution that funded or hosted the work, or which has been set up as a central open access repository.
Gold open access 'makes the final version of an article freely and permanently accessible for everyone, immediately after publication. Copyright for the article is retained by the authors and most of the permission barriers are removed' [266].

> *The Organisation shall provide for collaboration among European States in nuclear research of a pure scientific and fundamental character, and in research essentially related thereto. The Organisation shall have no concern with work for military requirements and the results of its experimental and theoretical work shall be published or otherwise made generally available.*[14]

In 1953, when the CERN Convention was signed, the focus for research in pure physics was to understand atomic structure—hence the name the European Organization for Nuclear Research. Over time, the focus of experiments conducted at CERN has shifted towards particle physics, and organisational practices have also moved towards being more open, inclusive, and capable of forming research teams spanning the entire planet.

There are some 2400 permanent staff and 1300 contractors working on the CERN campus at any time, along with 1000 or so visiting researchers. There are 12,500 scientific users off-campus, in 70 countries and of 105 different nationalities. According to CERN, this number represents more than half of the world's particle physicists.[15] The number of member states has also increased to the current 22, since the opening to non-European members in 2010 when the State of Israel became a full member. The scope of membership possibilities has also expanded. Another seven countries hold associate member status or are in a pre-stage to membership.[16] Countries including the United States, the Russian Federation, and Japan hold 'observer status', and it is envisaged that they may join the organisation in the future. Many other countries, including China, Argentina, Australia, Canada, and South Africa, have signed cooperation agreements with CERN.[17]

The global expansion of CERN in recent years can largely be attributed to its workforce, collaborative spirit and the second-to-none research infrastructure the organisation has developed over the years. It continues to modernise as quickly as technologically possible, with continuing funding and resources received from the CERN member states and other participating institutions. Perhaps even more importantly, CERN has put significant emphasis on publicising its research to the outside world, to both lay and expert audiences.

The experiments conducted at CERN are fascinating, if perhaps largely mysterious to the outsider. They are becoming more and more accessible to the general public—whether through Hollywood movies, particle physics masterclasses

---

[14] Convention for the Establishment of a European Organization for Nuclear Research, signed in Paris on 1 July, 1953 as amended on 17 January 1971, Article II.(1) (https://council.web.cern.ch/en/content/convention-establishment-european-organization-nuclear-research).

[15] CERN estimates there are some 20,000 physicists in the world today. See for example, CERN [267].

[16] Serbia, Cyprus, and Slovenia are associate members in the pre-stage to membership, and Turkey, Pakistan, Ukraine, and India are associate members. Source: CERN [268].

[17] Observer states and organisations currently involved in CERN programmes include the European Commission, Japan, the Russian Federation, UNESCO, and the United States.
Non-member states with cooperation agreements with CERN include Albania, Algeria, Argentina, Armenia, Australia, Azerbaijan, Bangladesh, Belarus, Bolivia, Brazil, Canada, Chile, China, Colombia, Costa Rica, Croatia, Ecuador, Egypt, Estonia, Former Yugoslav Republic of Macedonia (FYROM), Georgia, Iceland, Iran, Jordan, Korea, Lithuania, Malta, Mexico, Mongolia, Montenegro, Morocco, New Zealand, Peru, Saudi Arabia, South Africa, the United Arab Emirates, and Vietnam. Source: CERN [268] at point 23.
CERN also has scientific contacts with Cuba, Ghana, Ireland, Latvia, Lebanon, Madagascar, Malaysia, Mozambique, Palestinian Authority, the Philippines, Qatar, Rwanda, Singapore, Sri Lanka, Taiwan, Thailand, Tunisia, and Uzbekistan.

directed at school children, a strong presence on social media, or popular culture seeking to understand the foundations of the universe. People of all ages and professions are increasingly becoming aware of the experiments and discoveries coming out of CERN and are naturally drawn to them.

Open data forms an intrinsic part of these outreach activities.

5.2.1 Data collection and processing

Most experiments conducted at CERN today concentrate on understanding the data collected in the Large Hadron Collider (LHC)—the largest and most powerful particle accelerator in the world. With a 27 km circumference, the LHC accelerates protons in clockwise and anticlockwise directions at almost the speed of light before colliding them at four points on the LHC ring. The temperatures resulting from collisions in the LHC are over 100,000 times higher than in the Sun's centre [269].

This unique research environment presents unique challenges for data collection and processing. The volume of data generated and collected as part of LHC experiments is staggering. In June 2017, the data centre at CERN reached a new peak of 200 petabytes of data in its tape archives. This is about 100 times the combined capacity of academic research libraries in the United States [270]. Data is gathered from the particle collisions, of which there are around 1 billion per second in the LHC that result in approximately one petabyte of data per second [271]. Existing computing systems cannot record such a data flow; hence it is filtered and then aggregated in the CERN Data Centre.

The centre also performs initial data reconstruction and archives a copy of the resulting data on long-term tape storage. However, even allowing for the vast quantity of data that is discarded following each experiment, the CERN Data Centre processes an average of one petabyte of data per day [272]. This volume is growing and is predicted to continue to grow well into the future, mostly due to the ever-increasing complexity of the experiments and the increasing capacity to process and store data at CERN and other participating institutions.

The demand for data transfer and network capacity is increasing, too. The Worldwide LHC Computing Grid (WLCG) was created to provide the computing resources needed to analyse the data gathered in LHC experiments. Work on the design of the grid began in 1999. At that time, the computing power required to process the LHC data was much lower but still exceeded the funding capacity of CERN. A solution was found involving collaboration with laboratories and universities that have access to national or regional computing facilities. The LHC Grid was created on the basis of a memorandum of understanding signed among these institutions in 2001 [273], and their services were integrated in 2002 into a single computing grid. This facilitates storage and provides the computing power to distribute and to analyse the LHC data nearly in real time and all over the world. Some 10,000 researchers can access the LHC data from almost anywhere [274].

The number of institutions participating in the LHC Grid has grown to over 170, with 13 institutions participating as Tier 1 centres [275] and the remaining organisations as Tier 2 centres [276]. The LHC computing grid consists of two principal grids—the European Grid Infrastructure and the Open Science Grid based in the United States. There are many other participating regional and national grids, such as the EU–India Grid.

The distributed infrastructure has proved to be a highly effective solution for the challenge associated with the LHC data analysis. Not only is the Herculean task of data distribution and storage shared among the participating institutions but the technical advantages of the grid offer unprecedented possibilities for data access,

curation, use, and preservation. The advantages are many and are well-summarised on the CERN home page:

> *Multiple copies of data can be kept at different sites, ensuring access for all scientists independent of geographical location. There is no single point of failure; computer centres in multiple time zones ease round-the-clock monitoring and the availability of expert support; and resources are distributed across the world and are co-funded by the participating institutions [277].*

Data processing at the LHC computing grid occurs at four levels, internally known as Tier 0, Tier 1, Tier 2, and Tier 3. Each tier includes several participating institutions with their own computing resources and data storage facilities.

- **Tier 0** is the CERN Data Centre, which is responsible for the collection and initial reconstruction of the raw data collected from the LHC. The centre further distributes the reconstructed data to Tier 1 participating institutions and also reprocesses the data when the LHC is not running. This data centre accounts for less than 20% of the grid's total computing capacity.[18]

- **Tier 1** includes 13 major data storage and processing centres around the world, connected by optical fibre links working at 10 gigabits per second [278]. This high-bandwidth network is generally restricted to data traffic between the CERN Data Centre and Tier 1 sites and among the Tier 1 sites themselves. These institutions provide a round-the-clock support to the grid and take responsibility for storing their share of the raw and reconstructed data, as well as for reprocessing and storing the resulting output. Each Tier 1 site has connections to a number of Tier 2 sites, usually in the same geographical region.

- **Tier 2** involves over 150 universities and scientific organisations that originally were intended as centres for performing specific data analyses. As time went on, Tier 2 centres also became involved in data reprocessing and data offloading, particularly during a peak grid load that arose without warning due to higher-than-expected data collection that was beyond the capacity of the Tier 0 and Tier 1 centres. Each tier centre has at least one staff member dedicated to maintaining the LHC Grid.

- **Tier 3** nodes, apart from contributing processing capacity as required, enable individual scientists to access the grid though local computing resources. These may be part of a university department or simply the laptops of researchers.

There is no formal connection between the grid and the final users, as the agreements are with hosting institutions. However, the end users can choose from a broad range of services—including data storage and processing, analysis software, and visualisation tools. The computing grid verifies user identity and credentials and then searches for availability on sites that can provide the resources requested.[19] As required, users can access the grid's computing power and storage. They may not even be aware of the hosts of the resources.

---

[18] *Ibid.*

[19] CERN [277] at point 33.

Essential for the smooth functioning of the grid was the commitment of all participating organisations to use open-source software to power the grid. The CERN legal department played a central role in driving the early discussion among the participating institutions. In line with its commitment to an open Internet, CERN is also committed to open software, open hardware, and open source. As a leading software developer at CERN recently puts it:

> *We are a pure Linux shop from the point of view of real computing and real software development. That enables us to work fast and cut some corners [279].*

Crucially, the use of Linux, FLOSS, and other open platforms allows the grid centres to contain costs by deploying entirely generic components in processing and storage networks.[20]

Accordingly, CERN relinquishes all intellectual property rights to the software code, both in the source and binary forms. Permission is granted for anyone to use, duplicate, modify, and redistribute it. Similarly, all participating institutions warrant and ensure that any software that they contribute to the grid can be integrated, redistributed, modified, and enhanced by other members ([273], Article 10.1). Several participating institutions in the United Kingdom reported that the choice of Linux also made it easy for more centres to offer resources [279]. In using open software to power the grid, CERN is leading the development of open standards for distributed computing. Maarten Litmaath recently suggested that this CERN infrastructure can be used as a model for cost-effective collaborative computing in other fields of scientific research. The model can also be easily implemented in developing countries, which often do not have the resources to invest in data processing and storage [280].

5.2.2 Open data policies governing access to research data

Access to the LHC data stored in the grid centres occurs at various levels and combines multiple phases of data processing, access, use, and control. The LHC experiments generate large datasets, and before these enter the analysis phase, they undergo intricate quality assurance processes. The result is a trail of research outputs with varied stages of refinement and usage [281]. Direct access to the grid and raw data is enabled for some 10,000 physicists engaged in specific projects grouped around one of the four primary LHC data collecting detectors (particle collision points), internally referred to as 'four LHC experiments' (see **Figure 5**).

Each of these detectors has a separate team of researchers accessing and analysing the data collected. The largest are the ATLAS and CMS experiments, with some 6000 researchers working in one of the two collaborations. These are some of the largest scientific teams ever formed, as evidenced in the list of thousands of authors included at the end of their publications.[21]

The LHC data powers these mega collaborations. Access to the initial LHC data is restricted for several years, to the members of a specific experiment, as explained below. In fact, the principal motivation for building and operating the experiments is access to and a shared understanding of that data, along with the right to author publications subsequently.[22]

---

[20] *Ibid.*

[21] A recent physics paper from CERN has listed 5154 authors and has, as far as anyone knows, broken the record for the largest number of contributors to single research article. See Aad et al. [282].

[22] Bicarregui [265] at point 19.

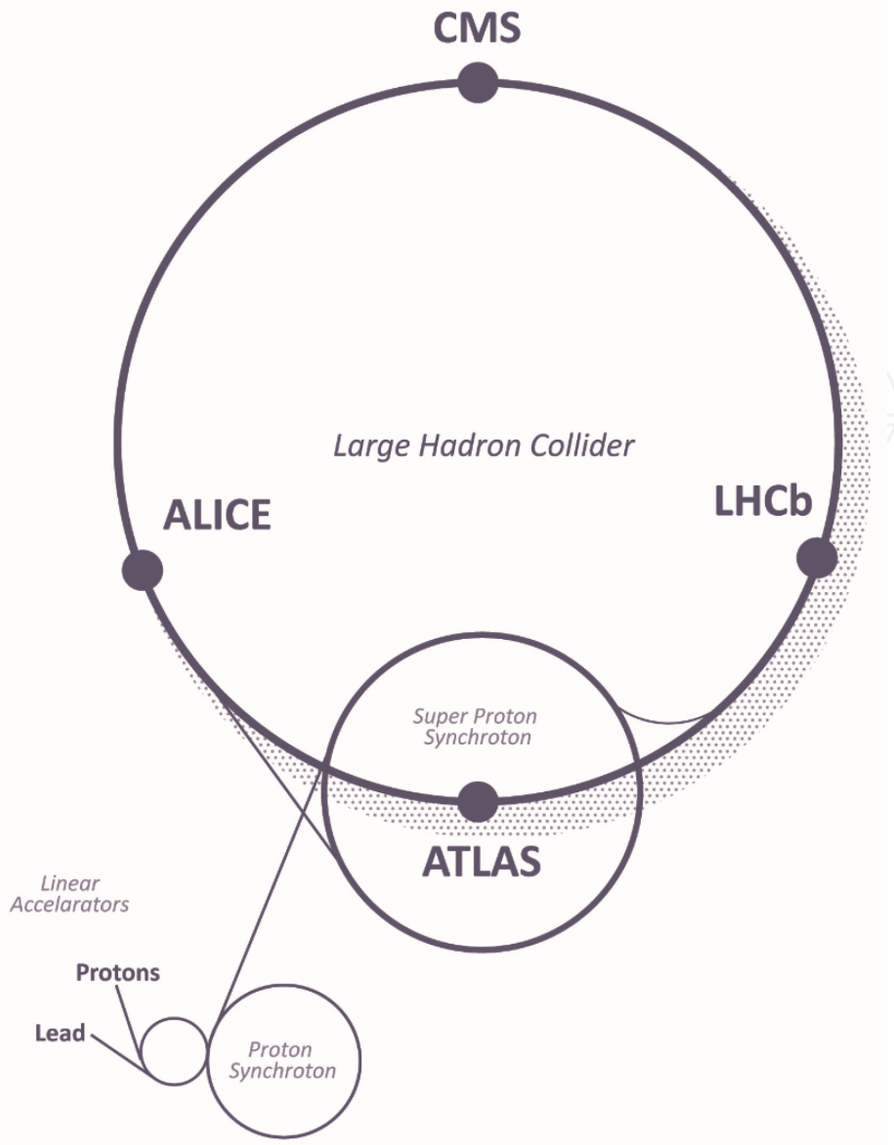

Figure 5.
Data harvesting points at the large hadron collider (Source: Wikimedia Commons).

The ATLAS and CMS experiments use detectors designed for general purposes to investigate the broadest ranges of particles possible. The two teams compete, rather than collaborate, with each other. Such competition is an effective way to cross-validate the outcomes of analyses produced by either of the two teams. As such, members of the ATLAS collaborations do not have access to the CMS data and research methods and vice versa. However, cross-migration of researchers between the two collaborations can occur, and such transfer also facilitates access to the data of the competing experiment. A level of secrecy about data processing and research methods remains essential due to the nature of scientific research performed by the two teams. The fact that the two detectors were independently designed is vital to the cross-validation of any discoveries [283]. For these reasons, it is unlikely that the first analyses of ATLAS and CMS real and raw data—representing the lowest and most guarded level of access—will ever be available as open access.

The two remaining experiments at CERN are known as ALICE and LHCb. They focus on research-specific phenomena. Instead of using an enclosed detector at the collision point, as is the case in ATLAS and CMS, the LHCb experiment uses a series of subdetectors to collect data concerning particles thrown forwards in one

direction by the collision.[23] One subdetector is mounted close to the collision point, with the others lined up over 20 m. The positioning of detectors enables examination of the slight differences between matter and antimatter by monitoring the movements of a particle called the 'beauty quark' [284].

Finally, the ALICE experiment is a heavy-ion detector designed to study the physics of strongly interacting matter at extreme energy densities. The conditions simulated at ALICE are thought to resemble those that occurred in the universe just after the big bang. The ALICE collaboration studies a phase of matter called 'quark-gluon' plasma, observing as it expands and cools and progressively gives rise to the particles that constitute the matter of the universe today.[24]

Each of the four LHC experiments produces unique data of interest to both the scientific and non-scientific communities around the world. Because of the open and collaborative nature of research at CERN, and the increasing awareness of the LHC experiments involving data, it is often thought that all data collected at the LHC is available as open data. This is incorrect. The data made available in the public domain represents only a tiny fraction of the data collected in the LHC. What becomes available is data requiring a higher level of analysis that directly underpins publications or carefully selected research experiments—the outcomes of which are peer-reviewed and cross-validated by other CERN researchers.

Access to the LHC data is governed by policies for the access and preservation of the data collected and processed by any of the four experiments. Each collaboration team has developed its own data preservation and access policy [285] that share some common characteristics and recognise four different data user groups:

1. Original collaboration members requiring access long after data harvesting is completed

2. The wider high-energy physics community and researchers from relating scientific disciplines

3. Those in education and outreach

4. Members of the public with an interest in science.

Each of these user groups has different data needs and requires the LHC data and supporting analyses at different levels of processing. Therefore, the open data policies of all four experiments have adopted a uniform classification of the LHC data developed by the Study Group for Data Preservation and Long-Term Analysis in High Energy Physics in 2009 [286], as follows (**Table 2**).

While CERN has already shared Level 1 data for a number of years, it needed a central point of access for Level 2 and Level 3 data, noting that Level 3 data can already be accessed through the grid by researchers directly associated with one of the four collaborations.

Level 4 data (raw data) collected at the LHC is not yet available as open data. Given the complexity and costs of data collection and calibration, as well as the technical expertise required, CERN has no intentions to make Level 4 data available in the public domain any time soon. Such data requires a large software, discovery,

[23] *Ibid.*

[24] CERN [269] at point 25.

	Data type	Data access point	Primary users
level 1	Data directly related to publications that provide documentation for the published results.	High-Energy Physics Literature Database: Inspire (http://inspirehep.net/); and HEP Data (https://hepdata.net/)	Interested scientific members or the general public (= any internet user).
level 2	Simplified data formats.	The Open Data Portal: http://opendata.cern.ch/	Outreach and education providers and users.
level 3	Reconstructed data, simulation data, and the analysis software needed to allow a full scientific analysis.	The Open Data Portal (selected datasets are available after an embargo period spanning 3 to 10 years). The WLHC Grid (scientists working in one of the four CERN collaborations and approved scientific members). No embargo period.	High energy physicists.
level 4	Raw data and access to the full potential of the experimental data.	The CERN data centre.	Restricted CERN users (data constructors and data-takers) working in one of the four collaboratons.

Table 2.
Levels of data access at CERN.

processing, and database infrastructure for meaningful use and interpretation of it. Even the members of the four LHC experiments generally cannot access Level 4 data. The data is uncalibrated and meaningless for direct analyses. However, CERN is open to the possibility of making subsets of data available for external use.

Therefore, CERN does not propose to devote resources to providing open access to the full raw dataset, although it might consider providing access to representative smaller samples of the Level 4 data.[25] Furthermore, physicists associated with CERN can access Level 3 data directly through the grid. At the same time, Level 2 data and some subsets of Level 3 data are, after the expiration of the embargo period, increasingly becoming available as open data on a dedicated server [288].

The parameters determining the level of access to the LHC data are based on the credentials of the potential user. CERN strictly differentiates between internal and external users and then between the varying levels of access permitted to individual users within the two main user groups—with Level 3 being the most guarded data and Level 4 being the most restricted data. Level 1 data is available by default—that is, immediately with publications that the data underpins. Level 2 data is carefully selected and tested before the release for educational purposes. The key access decision points are depicted in **Figure 6**.

5.2.3 The Open Data Portal

The CERN data portal is the key enabling platform for Level 2 data and selected subsets of Level 3 data after the expiration of the initial exclusivity period spanning 5–10 years. As shown in **Figure 7**, the home page offers users two profiles—education, consisting principally of visualisation tools and learning resources and research, providing direct access to the working environment along with tools for starting research projects at high school and other outreach institutions.

The portal, launched in November 2014, currently includes public data releases from the CMS, ALICE, ATLAS and LHCb experiments. This data comes with the software and supporting documents required to understand and to analyse it,

[25] CERN [287], ATLAS Data Access Policy released on 21 May 2014, 4.

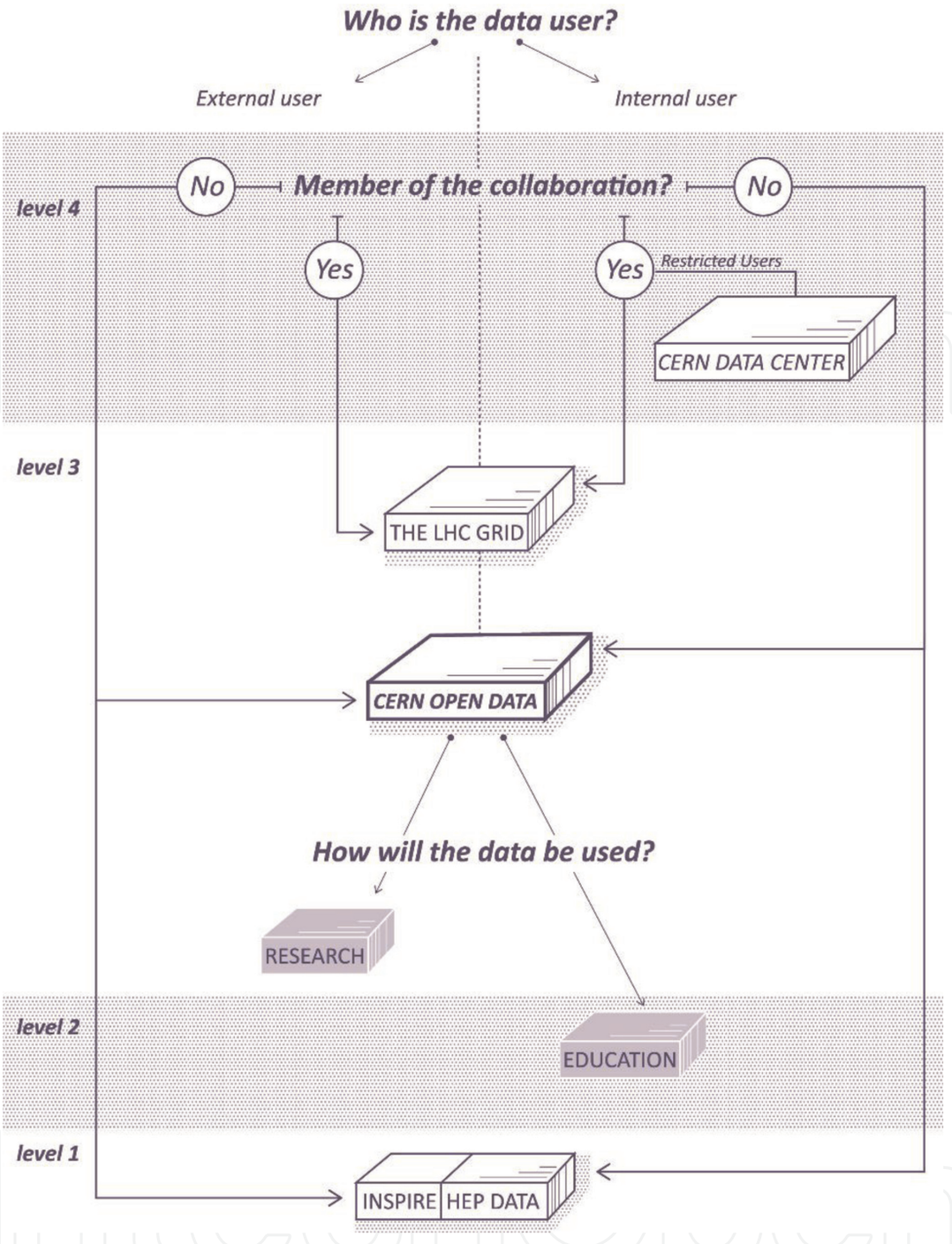

Figure 6.
Key data decision points at CERN.

supplemented with examples illustrating how a user, even from the general public, could write code to analyse the data [289]. There are several high-level tools for working with the data, and it is possible to download virtual machine images to enable external researchers to create tailored work environments.

These datasets are released in batches managed by one of the four CERN experiments. The releases are widely publicised in the media, and early experiences confirm that the publicity has assisted in attracting a large number of first-time visitors to the open data website.

CMS data forms the core of the current open data holdings. The CMS collaboration was the first committed to open data and has, to date, released more than 300 terabytes (TB) of high-quality open data. Included in that figure is over 100 TB collected by the CMS detector in 2011 and around 27 TB collected in 2010 [290].

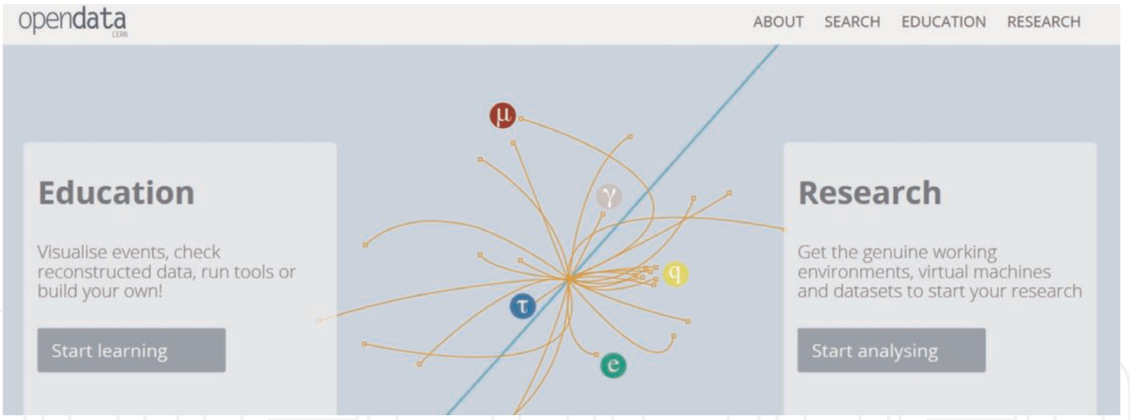

Figure 7.
The landing page of the Open Data Portal at CERN.

With rich metadata and comprehensive documentation, the data and the tools are released under the Creative Commons CC0 waiver, further discussed in Section 7.3 of this book. The data and software are presented in the MARC 21 format for bibliographic data [291], adjusted to accommodate fields for technical metadata or contextualisation. For consistency and to permit easier referencing, each record in the portal is created with a Digital Object Identifier (DOI) 'used for the identification of an object of any material form (digital or physical) or an abstraction (such as a textual work)' [292]. There is the expectation that users will cite the open data and software by the way of these identifiers, permitting tracking of reuse and thus contributing to assessment of the impact of the LHC programme [293]. CERN has adopted the FORCE 11 Joint Declaration of Data Citation Principles [294] and intends to include links to a published result of the (re)use cases in the future [295].

The two entry points on the CERN Open Data Portal, research and education, were adopted with a view to making it easier for users to identify relevant materials. After extensive testing and refinement of both entry points, students from the Lapland University of Applied Sciences in Finland and groups of researchers at CERN reviewed the portal's content, tested the tools, and confirmed that examples were reproducible.[26]

In the education portal, users can find simplified data formats for analysis as training exercises. Each has a comprehensive set of supporting material providing easy use by, for example, high-school students and their teachers in CERN's masterclasses. Students can use datasets, reconstructed data, processing tools, and learning resources to further explore and to improve their knowledge of particle physics.

The research portal presents datasets for research. It also offers reconstructed data, essential software, and guides for virtual machines. The available datasets are explained in detail, including the methodologies for validation and examples for how they could be used.[27]

One of the most popular datasets frequently accessed by users is the data produced as part of the experiments that led to confirmation of the existence of the Higgs boson elementary particle[28] at CERN in 2012. That discovery, made jointly by the ATLAS and CMS collaborations, was acknowledged by the Royal Swedish

[26] *Ibid.*

[27] CERN Open Data Portal at point 52.

[28] The Higgs boson is an elementary particle in the Standard Model of particle physics. First suspected to exist in the 1960s, confirmation of the Higgs boson was formally announced by CERN at the end of 2012.

Academy of Sciences in its announcement of the awarding of the 2013 Nobel Prize in Physics to François Englert and Peter Higgs for their theoretical work on the same subject half a century earlier [296].

CERN has promoted the use of the open dataset through the 'Higgs boson machine learning challenge'. This competition was created with a view of encouraging machine learning techniques using the Higgs boson data. The challenge ran over 6 months in 2014 on the Kaggle platform [297] and was highly successful, with 1785 teams participating and over 35,000 submissions posted on the web. Several of the machine learning methods proposed by the participants have been applied to real data at CERN, and the winners of the competition were invited to CERN to discuss the results with the CERN physicists. This outstanding example of joint work between expert and non-expert teams illustrates in a powerful way the potential that access to open data has to motivate both collaboration and new research.

5.2.4 Data and analysis preservation

CERN is a self-funded organisation, and the open data mandates recently introduced by research funders have not directly impacted the CERN researchers. The mandates have, however, raised the profile of open data and have given a fresh impetus to thinking about data preservation and sharing within the organisation.

When I first visited CERN in 2009, the general view was that data could mean anything and that there were many risks associated with sharing of the LHC data. At that time, the CMS collaboration was experimenting with open data, and the ATLAS collaboration was opposing it. The other two collaborations were closely watching the experiences at CMS. Over time, all four collaborations embraced the sharing of selected subsets of their data, supporting metadata, analyses, and software.

The key incentive for harnessing support for open data across the organisation was the long-established need for data preservation within the high-energy physics community. This discipline is known for its well-developed preprint and data-sharing culture—a practice that also assisted the organisation in rolling out gold open access[29] to all its publications as early as in 2002 [298]. CERN is recognised as a leader in the open access movement and has developed the Invenio digital library software[30] covering articles, books, journals, photos, videos, and other publishing outputs.

The LHC data is unique and forms an important element of the scientific legacy of the organisation. The end of any CERN experiment or scientific project does not usually mean shelving the data. On the contrary, physicists often continue to use the data or they refer to it when cross-validating later results. This can lead to new findings long after the initial experiments are published—for example, when the earlier data is analysed by means of improved methods or software. An outstanding example of this practice is research undertaken by the joint 2004 Nobel Prize in Physics laureates (Davis J. Gross, H. Davis Politzer, and Frank Wilczek), who researched asymptotic freedom in the theory of the strong interaction between nuclear forces. Their work incorporated retrospectively evaluated data from the JADE experiment completed back in 1986 [300].

The need to make Level 3 data openly available to wider audiences presented new challenges in data preservation with a view to achieving reusability, reproducibility, and discoverability of the data. In particular, there was the need to

[29] See definitions of gold and green open access at point 20.

[30] Invenio was originally developed to run the CERN document server, administering over 1000,000 bibliographic records in high-energy physics [299].

thoroughly document and preserve metadata, along with the need for data format and software version control that had already been well-identified before the development of the Open Data Portal. These processes are internally known as the CERN Analysis Preservation Framework, and they have involved prototyping a central platform for all four LHC collaborations to preserve the supporting information about their analyses and about the tools used for them.

The library team, supported by the four collaborations and the IT team, conducted a number of pilot studies and collected information about how researchers record their research workflows [301]. This was followed by an extensive consultation process and testing that eventually resulted in the new CERN analysis preservation (CAP) library service, hosted by the Invenio digital library platform. The service was designed with a unique disciplinary research workflow, which captures each step and the resulting digital objects [302].

To facilitate the future reuse of multiple research objects, researchers need to plan, from an early stage of their experiments, how they will preserve data. They also need to provide sufficient contextual information around the analysis. A standard analysis (i.e. a record) stored in the CAP server contains detailed information about the processing steps, the datasets that are used, and the software (and version) used. In addition, detailed information about the physics involved is included, along with detailed notes on the scientific measurements.

The ATLAS collaboration made an important contribution to the process. Some of the researchers felt extremely uneasy about the possibility of someone else independently reproducing their Level 3 data experiments without the same knowledge as the members of the collaboration of the intricate internal processes. Members of the collaboration have studied the concept of data reproducibility intensely and, in order to facilitate (in their own words) 'preservation of the recipe, not the pizza' [303], they have developed a useful internal distinction and vocabulary for describing the subtle differences between what they framed as 'data reproducibility' and 'data replicability'.

Reproducibility, analogous with the 'pizza', describes the concept of archiving existing software, tools, and documentation used in the analysis procedures. The proof that an analysis is reproducible is the ability to redo the steps, in close detail, as they were undertaken by the original analysis team. To succeed, all the 'ingredients' that produced the original outcomes need to be preserved as they were at the time of publication. Those ingredients include computer configuration (e.g. operating system and architecture), the software releases used at the time, and the datasets as then reconstructed. These requirements are mostly useful for short- and medium-term preservation. Reproducibility, they concluded, has the most application in the confirmation and clarification of the published result.[31]

Replicability, analogous to the 'recipe', refers to the process of ensuring that the original analyses are repeatable using the most recent version of software tools and data formats. Since the amounts of data involved are enormous, storing indefinitely those datasets reconstructed with old software releases will not be possible. There is an imperative, therefore, to ensure the old data remains readable by newer versions of the software.[32] Those working on the ATLAS experiment are investigating options to ensure replicability in this sense. The ATLAS collaborators believe replicability might be achievable via code migration and regression testing as well as detailed human-readable information about how the analyses were performed. This information will be invaluable for newer members of the collaboration who would

[31] *Ibid.*

[32] *Ibid.*

not be familiar with the older software and analysis procedures. For this reason, the ATLAS team argues, relying solely on reproducibility is not sufficient for preserving data for future access.

In the meantime, the other collaborations continue to 'preserve the pizza' wherever this is deemed necessary and achievable within the resources available. For example, the current ALICE data access policy states that:

> *... while formats can change with time, the collaboration provides software releases suitable to read and process any format, or alternatively to migrate data from one format to another. Since processed data can exist in several versions, only the version used for the final publication of the results is considered as a candidate for data preservation [304].*

Like ATLAS, the CMS experiment is committed to preserving Level 3 data by 'forward-porting'—that is, by keeping a copy of the data reconstructed with the best available knowledge of the detector performance and conditions. This data includes simulations and is capable of analysis by the central CMS analysis software. While at this time it is not possible to reconstruct the CMS data [293], the analysis procedures, workflows, and code are preserved in the CMS code repository.

The pilot CAP testing revealed that, while there were many similarities in the data workflows and processes among the four collaborations, these practices do not allow for the later reproduction of the analysis in a uniform way across the four experiments. The key challenge, therefore, was to establish, firstly, interoperability with a variety of data and information sources and, secondly, connectors between the various tools used by each collaboration. The CAP is not an effort to enforce a standard across experiments, which is the push in other scientific disciplines. Rather, the CAP aims to flexibly accommodate the requirements of the four data collaborations.[33]

The data preservation processes included in the four open data policies have been embedded in the internal research workflows and have become part of daily practice. This is an unintended, yet probably the most tangible, benefit accrued from the internal work on open data at CERN so far. The CERN Library reported that the new CAP service helps researchers to better manage their research workflows by making internal work practices and data more accessible and discoverable.[34] It is believed that the CAP practices will, eventually, also save researchers time and effort as they will be able to utilise the work of others more readily.

Following the CAP pilot, the ATLAS collaboration reported that the key learning outcome was the planning for data preservation from an early stage of any experiment. The ATLAS event data model took this into consideration, among other matters [305]. Also resulting from the CAP implementation is improved access to past corporate knowledge for new members of the collaboration.

Finally, and perhaps most importantly, the improved data curation at CERN has confirmed the organisation's potential to conduct open science and has provided physicists with new means for looking at ongoing and past data analyses. As well, the data enables physicists at CERN to look at novel ways for engaging colleagues outside their individual collaborations. For example, the ATLAS collaboration is exploring the potential of the recasting of analyses. This might result in providing a robust mechanism for the testing, by those outside the collaboration, of new physics models against well-validated analysis chains [306].

---

[33] Chen et al. [302], at point 70, 354.

[34] *Ibid*, 349.

Despite the tangible outcomes achieved through experiments with open data at CERN, there remain researchers at CERN who are yet to be convinced about the utility and value of making lower-level data available to external users as open data. Their concern is a possible lack of interest from non-experts outside physics to meaningfully interrogate the datasets ([89], p. 111). As mentioned earlier, processing CERN lower-level data requires access to high computing power, and it is unlikely that many external users would have such access. Knowledge of physics and data practice in the field is also required to understand the data and the experiments—even in cases where data is meticulously described and when all necessary software and algorithms are made available to the users. The sceptics have a point here, and only future developments in technology and the uses of CERN open data will tell whether their concerns can be overcome.

5.2.5 The use of open data

5.2.5.1 Research

Research activity on the open data website seems to respond to new data releases. Following the release in 2014 of the CMS data compiled in 2010, some 82,000 users visited the site. Of these, 21,000 viewed the data in more detail. The portal had almost 20,000 visitors who used at least one of the tools (event display or histogramming). On average, the web page was used by 1000 people a day. Of these, 40% looked at the detailed data records and 1% downloaded a Level 3 dataset.[35] Just over a year later, in April 2016, the CMS data compiled in 2011 was released, totalling some 300 TB of data. This release saw 210,000 users visiting the site, of whom 37,000 viewed the data in more detail and 66,000 used the event display facility.

When a new batch of open data is released, it is accompanied by extensive press and social media coverage, followed by a peak of interest from the public. In these periods, CERN sees some 70,000 distinct users visiting the site a day. After several weeks, the interest drops to a normal level, which is around 2000 distinct users per day. CERN also sees smaller peaks in the non-release periods due to social media events, such as a recent Reddit 'Ask Me Anything' session that attracted some 10,000 users to the site.[36]

In October 2017, the CMS open data team was excited to see the publication of the first independent study produced reusing CMS open data. The CMS team had put extensive effort into describing the datasets, supporting tools, configuration parameters, workflows, other auxiliary information, and all the 'insider' knowledge that went into constructing the dataset. It was therefore rewarding for the team to see Jesse Thaler's group from MIT succeed in understanding and studying the data independently from the CMS team. The MIT study revealed a universal feature within jets of subatomic particles, which are produced when high-energy protons collide [307]. This research would not have been possible without access to the CMS data.

5.2.5.2 Education

To identify the technical tools and instructions necessary to bring the CMS open data to a wider audience, CERN ran a number of pilot projects in Finnish high

[35] Cowton et al. [295] at point 59, 4.

[36] I am very grateful to Tibor Simko, Sunje Dallmeir-Tiessen, and Achim Geiser for collating the statistics.

schools [308]. The International Particle Physics Outreach Group began in 2005 and runs masterclasses in high schools in over 40 countries. Currently, these masterclasses utilise Level 2 data from all four data detection centres at CERN. For instance, 10% of the ATLAS data is available for students to search for a Higgs boson. This masterclass is extremely popular and has reached locations other than schools, such as science centres and museums.

The largest national masterclass programme is offered in Germany. Every year more than 100 young facilitators, mostly masters and PhD students, take CERN data to German high schools. Around 4000 students are invited to further their qualifications as part of the masterclass network, often choosing for themselves the topics of their research theses.

Elsewhere, masterclasses offered in Greek schools are combined with virtual LHC visits in which students link with a CERN physicist working on the ATLAS or CMS experiments.[37]

Due to the rising demand for LHC masterclasses, CERN is investing more resources into developing this resource further. In fact, all four collaborations concur that the benefits arising from Level 2 data are clear and represent a good return on the organisation's investment of resources and staff time.

5.2.6 Data embargo period

CERN researchers are of the view that data exclusivity is required before their data can be shared with external parties. Generally, the data embargo period spans from 3 to 10 years from when the data was taken.

There are several reasons for this long embargo period. Firstly, the lead times for the LHC experiments are substantial. For example, the ATLAS collaboration formally commenced operations in 1994, following 10 years of planning. The first ATLAS data was taken in 2009, and its expected lifetime is more than 20 years. What is more, data curation and processing require a huge and ongoing commitment of research effort. The ATLAS collaboration estimated that each ATLAS member spends, on average, 100 days a year on 'data authorship'. The clear majority of researchers regards this as time spent on 'non-publishable work' and describes it as unrewarded effort. For this reason, most physicists view the incentives associated with data curation largely in terms of exclusive access [309].

The data release period varies across the four collaborations. The CMS experiment is most prone to data sharing and has the shortest embargo period of 3 years. On the other hand, the ATLAS experiment requires the longest embargo period—this is defined as a 'reasonable embargo period' ([287], p. 2). In general, data will be retained for the sole use of the collaboration for a period argued to be commensurate with the large investment in effort needed to record, reconstruct, and analyse it. After this period some portion of the data will be made available externally, with the proportion rising over time. The LHCb collaboration will normally publish 50% of its research as open data after 5 years, rising to 100% after 10 years [310]. The ALICE experiment has committed to make 10% of its data available in 5 years, rising to 100% after 10 years [304].

The CMS collaboration has opted to release Level 3 data publicly on an annual basis. Additionally, releases will be made during long LHC machine shutdowns and on the basis of best efforts during running periods. During the lifetime of CMS, the upper limit on the amount of publicly available data, compared with that available only to the collaboration, will correspond to 50% of the integrated luminosity

[37] *Ibid.*

collected by CMS. Usually, data will be released 3 years after collection, even though the collaboration can decide to release particular datasets either earlier or later [293].

5.2.7 The value of CERN open data

The four open data preservation and sharing policies at CERN result from delicate negotiations and robust internal discussions about the needs and principles for data sharing and preservation. The policies have created a shared understanding of open data at CERN, forging a consensus between many divergent attitudes and views. By codifying the key principles, defining the various components of data at various stages of their processing, developing the criteria (incorporating the levels of processing, preservation, and access), and developing supporting documentation for data preservation, the organisation has greatly improved its internal data management flow. The resulting policies are high level, yet the discussion driving the development of these policies has transformed the way data and supporting analyses are preserved, documented, and used.

The processes underlying these developments were thoroughly workshopped and tested as pilots and only then were they embedded in the internal data flow and research conduct. These processes were driven bottom-up, by the CERN physicists who see the benefits to their work—including further analysis and discovery and validation of their efforts. The library and IT teams have provided hands-on support and facilitated the development of data documentation, citation, linking, and discoverability tools, as well as suitable platforms. Open data at CERN facilitated the emergence of a collaborative and open-ended conversation across the entire organisation. At CERN, open data is not the result of mandates imposed on researchers by external funders, even though external mandates initially prompted the CERN researchers to think about open data.

The greatest value of open data at CERN stems from the benefits that the robust experimentation with, and the sustained thinking about, open data has engendered within the organisation. The value lies in the continuous learning and constant improvement of data quality as a result of improved preservation, curation, accessibility and increased potential for reuse. These processes are transforming not only the minds of researchers and their research conduct, but they are also likely to lead to improved research outcomes.

The value accruing from the use and reuse of CERN open data by external parties is yet to be seen. However, the initial experiences with the outreach programmes have been immensely encouraging.

Conclusion

The World Wide Web was invented at CERN, and the organisation is now using it to conduct big data experiments to extend our understanding of data-driven science.

CERN does not have the computing and financial resources to crunch all the data it collects as part of the Large Hadron Collider experiments in Geneva. Instead, it relies on grid computing powered by computer centres in many parts of the world. The Worldwide LHC Computing Grid gives a community of over 10,000 physicists near real-time access to LHC data. In using open hardware, open software, and open standards to power the grid, CERN is leading the way in developing cost-effective solutions for ‘big data’ tasks. And portions of that data are increasingly becoming available in the public domain as open data.

The four open data preservation and sharing policies at CERN are the result of a mix of delicate negotiations and robust internal discussions about the needs and principles for data sharing and preservation—both for internal organisational purposes and for sharing the LHC data with external parties. The policies have created a shared understanding of open data at CERN, forging a consensus between many divergent attitudes and views. The organisation has greatly improved its internal data management flow by codifying the key principles, defining the various components of data at various stages of their processing, developing criteria, and providing the supporting documentation for data and analysis preservation.

At CERN, open data is not the result of mandates imposed on researchers by external funders, even though external mandates initially prompted the CERN researchers to think about open data. The resulting policies are high-level, yet the discussion driving the development of these policies has primarily occurred among researchers and research teams. In this process, the library and IT teams have provided hands-on support and facilitated the development of data documentation, citation, linking, discoverability tools, as well as suitable platforms. Open data at CERN facilitated the emergence of a collaborative and open-ended conversation across the entire organisation. This combined effort and cumulative thinking has transformed the way the LHC data and supporting analyses are preserved, documented, and used.

Not all data produced at CERN is available as open data at this stage. CERN recognises four distinct groups of prospective data users—from collaboration members, to the wider high-energy physics community, to those in education and outreach, and members of the public with interest in science. Corresponding with the needs of these users is the classification of the LHC data into four distinct levels with different access rights. While open data can serve all of these users, lower-level LHC data—that is, Level 3 and Level 4 data—are only available to expert users. Restricting access to lower-level data to expert users is necessary at this stage, as significant computing power and knowledge of particle physics are required to understand and reuse the data. However, portions of low-level data are increasingly becoming available as open data for research.

CERN has become a leader in the open data field because its management and senior researchers appreciated early that a good data management practice will not only satisfy research requirements in the short term but also serve as an organisational blueprint driving continuous improvement in scientific research and scholarly communications for many years to come.

The greatest value of open data at CERN stems from the benefits that the robust experimentation with, and the sustained thinking about, open data has engendered within the organisation. The value lies in the continuous learning and constant improvement of data quality as a result of improved preservation, curation, accessibility and increased potential for reuse.

The value accruing from the use and reuse of CERN open data by external parties is yet to be seen. However, the initial experiences with the outreach programmes have been encouraging, as evidenced by the high demand for the open datasets that committed and enthusiastic outside users are busily downloading and reusing.

Author details

Vera J. Lipton
Zvi Meitar Institute for Legal Implications of Emerging Technologies,
Harry Radzyner Law School, IDC Herzliya, Israel

*Address all correspondence to: vera.lipton@bigpond.com

IntechOpen