11 Big Science and Social Responsibility of the Digital World

Ruediger Wink, Alberto Di Meglio, Marilena Streit-Bianchi, and Shantha Liyanage

11.1 Big Science, Big Data, and Computing

Big Science is often synonymous with ‘big data’. ‘Big data’ generally describes large amounts of data, high data rates, or particularly complicated or unstructured data (Heiss, 2019). The ways data are produced, processed, distributed, made accessible, and analysed are important parts of data management processes in Big Science. Those who are familiar with the Large Hadron Collider (LHC) at CERN know that particles collide in the LHC detectors approximately 1 billion times per second generating about one petabyte (1 million gigabytes) of collision data per second (Gaillard, 2017).

Not all data is useful, and it takes a painstaking effort to determine what data is useful and what is not so useful. Usefulness of data and information is determined by the types of questions asked—not any question, but the right type of question. However, even with advanced filtering techniques, enormous amount of about 330 million petabytes of scientific data from past and present high energy physics (HEP) experiments had been created at CERN by the beginning of 2019 (CERN, 2021b). With the massive amount of data generated by numerous collisions, scientists must be able to go through a process to determine how a rare process differs from a common one. Reliable statistical analysis and probability studies are necessary. An open-source tool set called ROOT[1] developed at CERN and Fermilab computes vast amounts of data very efficiently (See Figure 11.1).

Similarly, the collaboration of eight telescope observatories around the globe to generate the first image of a black hole 55 million light years away from Earth depended on capacities to manage huge amounts of data, as every night of observation generated 1 petabyte of data, which could not be sent via the internet but had to be transported as hard drives from place to place (Castelvecchi, 2019). In

[1] The ROOT system provides a data analysis framework and consists of a set of OO frameworks with all the functionality needed to handle and analyse large amounts of data in a very efficient way (see https://home.cern/news/news/computing/big-data-takes-root).

Ruediger Wink et al., *Big Science and Social Responsibility of the Digital World.* In: *Big Science, Innovation, and Societal Contributions.* Edited by: Shantha Liyanage, Markus Nordberg, and Marilena Streit-Bianchi, Oxford University Press.
© Ruediger Wink et al., (2024). DOI: 10.1093/oso/9780198881193.003.0012

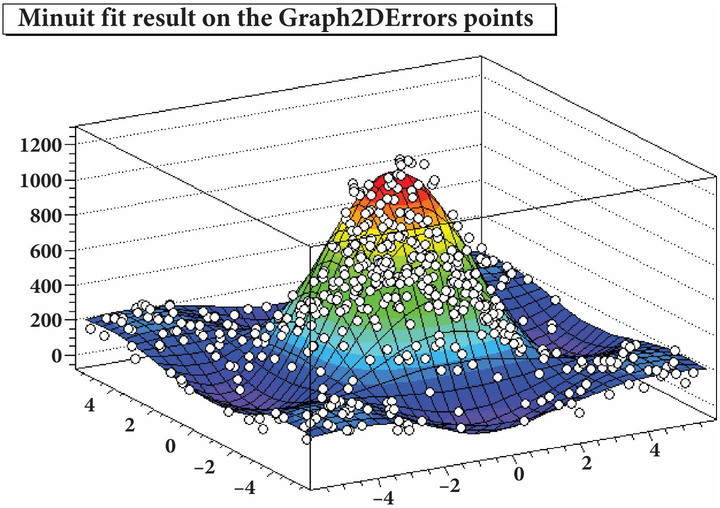

Figure 11.1 An example of a plot created using the ROOT tool

Source: © CERN

fact, Big Science experiments produce extremely high volumes of data. The planned high luminosity LHC (HL-LHC) is expected to produce an annual data volume of approximately one exabyte[2] and the antennas of the Square Kilometre Array (SKA) will produce more than 100 terabytes per second on site and virtually online (Heiss, 2019).

This chapter focuses on three important connections between Big Science, big data, and overall technological developments that impact societal needs in the digital world: the long-lasting experiences in collaborations between the high energy physics community and Information Communication Technology (ICT) companies; the management of big data among scientific communities; and the transfer of knowledge and skills in big data infrastructures in an Open Science and open innovation context to initiate faster and more innovative solutions such as development of vaccines for Covid-19 pandemic. Consequentially, three research questions covered here are:

- Which organisational principles and elements were installed by Big Science organisations such as CERN, ESO, EMBL to maximise the mutual benefits from technological development in computing for the scientific community as well as for commercial partners from the ICT industry?
- Which principles help to organise and manage big data in Big Science projects, and how can these principles meet the expectations of policy and society?

[2] An exabyte is the equivalent of one quintillion bytes, one billion gigabytes, or one million terabytes (TB). In context with other units of digital data and storage: 8 bits equals one byte. 1,024 bytes equal one kilobyte (KB).

- How could the principles of responsive research and innovation in the digital age be transferred into a transformation of Big Science towards open science to maximise and accelerate societal benefits while adhering to data protocols?

Starting with the first research question, Big Science organisations such as CERN were closely connected with the technological edge of computing almost from its earliest days, as the first computer was installed in 1958. It was at the beginning of the 1970s that Lew Kowarski, shared his views on the need for computers in physics and wrote a stimulating paper entitled 'Computers: Why?'. Kowarski reminded us that 'we are only at the beginning to discover and explore the new ways of acquiring scientific knowledge which have been opened by the advent of computers' (Kowarski, 1972: 59). Kowarski listed eight applications for 'the universal black box at CERN', namely: Numerical Mathematics, Data Processing, Symbolic Calculations, Computer Graphics, Simulation, File Management, Pattern Recognition, and Process Control. This work demonstrated the need for data links, forerunners of high-speed networks, to transmit data between small online computers and larger ones in the central computer room.

Already starting in the 1960s, exchanging data between computers and external networking between scientific institutes were core challenges for high energy physics (Hemmer, 2018). The invention of the World Wide Web (WWW) by CERN was the solution to the urgent need of high energy physics community to share data via the Transmission Control Protocol/Internet Protocol (TCP/IP)[3] in a structured way. It is one of the most prominent examples of Big Science and Big computing (Hemmer, 2018). It led to experiences with the transition from big mainframe computers to decentralised networks of small computers at Fermilab (Melchor, 2021). CERN established the Worldwide LHC Computing Grid (WLCG), which allowed scientists to access and analyse data from anywhere in the world.

From an economic point of view, Big Science leads to big gains in knowledge productivity as more researchers with their ideas get access to useful knowledge. Generating, filtering, sharing, preserving, processing, and contextualising data is an important part of these options for scalability. This is one reason, why the close and early collaboration with software and computer firms was so important for the high energy physics community as well as for broader scientific communities in medicine and biology.

The high energy physics (HEP) community became familiar, relatively early on, with the use of common data formats, communication standards, and sharing practices. Due to historical experience in collaboration among the HEP community and ICT companies, formal organisational structures have been established (Zanella, 1990; Williams, 2004). For other scientific disciplines such as oceanography, human genome project, climate change, the transition towards common data management required considerable adjustment to social and organisational practices, making it

[3] CERN named Ben Segal as its first 'TCP/IP Coordinator' and the TCP/IP protocols (as Internet protocols were then called) were introduced at CERN in early 1990s, inside a Berkeley Unix system.

even more difficult to exploit the benefits of common research (Bos et al., 2007; Meyer, 2009). CERN and ESO are at the forefront of using big data infrastructure such as grid computing which facilitates faster dissemination, analysis, and retrieval of the enormous amount of data being produced by the LHC experiments and ESO detectors (see image of CERN computer centre in Figure 11.2). Such sophisticated infrastructure is not readily available for small laboratories and is a major drawback in analysing research data.

CERN's experiences with computing were transferred into biological research communities in 1994, when Paolo Zanella became the Director of the European Bioinformatics Institute (EBI) after having been the leader of the Computing and Data Handling Division at CERN during the period when major outcomes occurred with the start of computing online, computer automatic event recognition, PET development, and the World Wide Web (1976–1989). Zanella reported: 'The discovery of a new gene or the determination of the genomic variations related to a particular illness may have an impact on bio-medical research and on the pharmaceutical industry. The delay between a discovery and its effects on healthcare is shorter. This has to do with the strength of research performed by Industry and with the size of molecules, a billion times larger than that of quarks and leptons, thus resulting in less expensive research and development' (Zanella, 2014: 155),

In February 2001, both Nature and Science published the initial sequencing and analysis of the human genome. Nature published data from the international human genomic mapping consortium, which included several thousand Human

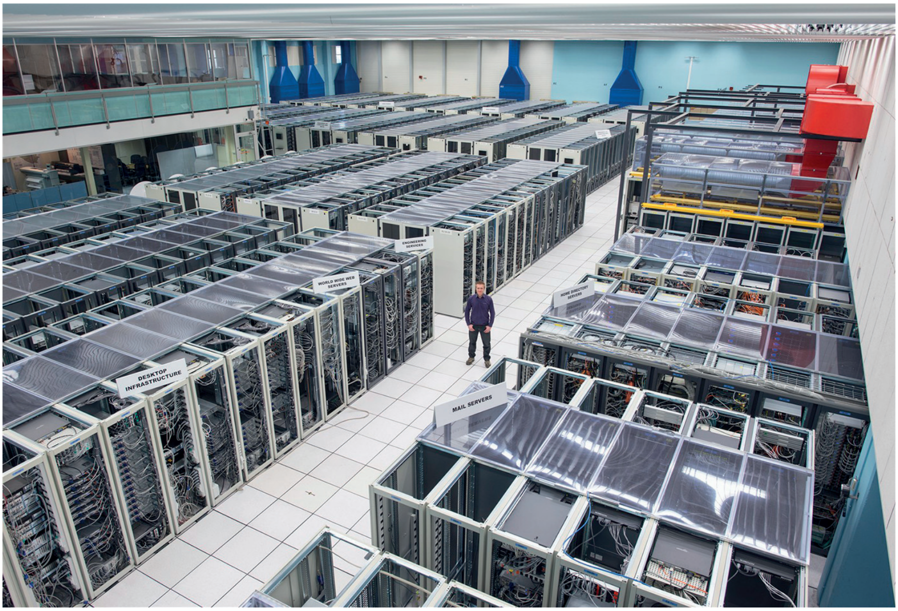

Figure 11.2 CERN's Computing Centre

Source: © CERN

Genome Project (HGP) researchers led by Francis Collins (The International Human Genomic Mapping Consortium, 2001). Science published data from Craig Venter's joint private academic venture (Venter et al., 2001). The use of big data in biology, followed by a significant media launch and the sharing of bioinformatic information, was an important and unavoidable lever in many areas of research.

Similarly, the use of big data has been recognised as a critical prerequisite in neuroscience for better understanding of mouse and human brain functionalities (Koch and Jones, 2016).

An important step towards further collaboration between Big Science and leading global ICT firms was achieved by the foundation of CERN openlab in 2001. CERN openlab, which will be described in detail below, is a public–private partnership with different levels of intensity in the way that companies are integrated as members, ranging from partners (with at least a three-year commitment to a common programme of work), contributors (with a formal collaboration in joint tactical projects for one to three years) to associates with a formal collaboration on a specific joint targeted project (Di Meglio et al., 2017).

Current partners include the companies Google, Oracle, Micron, Siemens, and Intel. When companies partner with CERN openlab, they have access to challenging requests for their technological advancements. CERN Openlab's expertise stimulates the development of new ICT infrastructures and technologies, and it provides ideal testing conditions for new technologies in a demanding, pre-competitive environment (Grey, 2003; Hemmer, 2018).

Accordingly, CERN combines the roles of a lead user who defines avantgarde and specified needs and directions for technological solutions, a new technology testbed with challenging tasks and applications, and co-developers who collaborate with industrial labs to produce new and leading-edge technological solutions.

In CERN's White Paper to describe the priorities for the sixth three-year phase ending at the end of 2020, it has identified four major R&D priorities for CERN openlab (Di Meglio et al., 2017; Albrecht et al., 2017).

- Data centre technologies and infrastructures dealing with storage and processing needs for even extremely large scales of data generated in new scientific experiments;
- Computing performance and software to modernise coding techniques and optimise the exploitation of features offered by modern hardware architectures;
- Machine learning and data analytics to maximise the value to be generated from data while optimising resource usage;
- Applications in other scientific fields with large quantities of data and ICT challenges are comparable to high energy physics, like life sciences, medicine, astrophysics, and urban/environmental planning.

In 2020, CERN management announced the quantum technology initiative (QTI) as a new three-year activity to focus on further investments and research in quantum technologies, in particular quantum computing, with huge potential to be used in high energy physics as well as comparable other scientific fields (Di Meglio et al., 2020; Melchor, 2021). QTI has defined a medium- and long-term roadmap and research programme in collaboration with the HEP and quantum-technology research communities.

Again, the existing structures in CERN openlab with industry partners and other research institutes provide an excellent precondition for developing and exploiting the full potential of these new dimensions of computing (Di Meglio, 2021), and potential fields of applications in high energy physics (Tavernelli and Barkoutsos, 2021, on the industry perspective from IBM). CERN openlab is an important case example of Big Science and big data initiatives.

11.2 CERN openlab: Partnership in Scientific and Technological Innovation

11.2.1 What Is CERN openlab?

CERN openlab is a unique public–private partnership that works to accelerate the development of cutting-edge ICT solutions for the worldwide LHC community and wider scientific research. Through CERN openlab, CERN collaborates with leading ICT companies and research institutes. Within this framework, CERN provides access to its complex ICT infrastructure and its scientific and engineering experiences—in some cases this collaboration even extends to institutes worldwide. Testing in CERN's demanding environment provides the ICT industry collaborators with valuable feedback on their products and outcomes, while enabling CERN to assess the merits of new technologies in their early stages of development for possible future uses. In a similar way, research laboratories and academic institutes worldwide can join forces with CERN scientists and technologists to advance knowledge in computer and data sciences for large-scale scientific applications. This framework also offers a common ground for carrying out advanced research and development activities with more than one company or institute, thus accelerating innovation through collaboration and cooperation.

11.2.2 Brief History

CERN openlab was established in 2001 at the start of the construction of the Large Hadron Collider (LHC) at CERN to provide a framework through which CERN could collaborate with leading ICT companies to accelerate the development of cutting-edge ICT solutions needed by the HEP community. The complexity of the scientific instruments and infrastructures needed by CERN and the LHC experiments, presented extreme challenges and provided the ideal environment to carry out joint R&D projects and evaluation of new technologies in large-scale operations.

The joint collaborations were at the beginning aligned along a sequence of three-year phases, a balanced duration long enough to go beyond simple short-term investigation, but still short enough to ensure impact and deployment of results in production within reasonable life-cycle expectations of the technologies being assessed. The transition between phases would also provide a natural boundary to update the research priorities and collect requirements from the research community. Today the phase mechanism is still in place, CERN openlab has entered its seventh phase in January 2021. However, the projects are not necessarily aligned anymore with the phases in order to follow more closely the rhythms of the LHC experiments, and computing infrastructure upgrades, and the evolution of technologies and products.

The first four phases of CERN openlab between 2001 and 2013 were primarily focused on industrial collaborations in ICT and infrastructure technologies, from networks for distributed computing to computing platforms, from storage and databases to security for control systems.

The typical model for collaborations was based on a small number of large companies able to shape the broad technology landscape, such as Intel, Oracle, or IBM. As worldwide distributed infrastructures based on cloud computers became more established and reliable and new computing paradigms such as artificial intelligence and quantum computing started showing promising potential for scientific research, CERN openlab collaboration gradually extended into more academic research in computer and data science. In 2022, CERN openlab runs more than 30 different collaborative projects with both international companies and academic institutes worldwide.

11.2.3 Collaboration Principles: Win–Win Scenario

The CERN openlab collaborations are based on a few principles, or ‘rules of engagement’, designed to maximise the chances of achieving results and keep the parties engaged and committed for the duration of the projects. Every conversation is based on the rule that collaborations are purely focused on joint R&D and innovation. Explicit commercial interests must be left out of the discussion and taken up with the appropriate procurement services.

The second important rule is that all parties engaged in a project must have something to gain from the collaboration. For CERN and the physics experiment, the interest is to get access as early as possible to innovative technologies, be part of the development process and have direct channels to suggest requirements and improvements, and possibly get access to sponsorships and funding opportunities for students and researchers. For the companies, the interest must be related to the opportunity of assessing their technologies in CERN’s challenging environment or applying them to computationally intensive use cases, so that they can be stress-tested and improved even before becoming products or services on the market. In addition, the association of a company brand with CERN brings tangible marketing value that companies have successfully exploited in many collaborations through case studies or joint participation at events and conferences.

Finally, the rule that ‘there is no such thing as a free meal’ must be respected. In order to ensure the continued commitment of all parties to the achievement of the agreed objectives, they must invest resources in the collaboration. Such resources include monetary contributions from the companies to the public activities organised by CERN openlab, such as conferences or training programmes or communications; contributions in the form of grants for dedicated researchers in the joint projects; and ‘in-kind’ contributions such as time of people, access to infrastructure or services, and hardware samples; and dedicated events at CERN to showcase the collaboration with a company and attract interest in the technologies under development.

The lifecycle of any new project goes through several steps, passing through a transition phase from the first introduction to the successful development of partnership. The typical sequence of the CERN openlab steps of the collaboration process is represented in Figure 11.3. The process starts always from with a technical requirement or, better yet, a challenging scientific or technological problem, that both the scientific community at CERN and the R&D teams in a company can relate to. The definition of the problem is followed by technical due diligence, which usually involves technical experts reviewing and brainstorming the technical importance of the problem and solution. If the discussion leads to positive outcomes, the project is deemed implementable on agreed timelines, objectives, and available resources. Finally, the project proceeded to form a collaboration agreement and contract to formally launch it with the approval of legal and financial services.

The parties can decide that the potential results are not worth the investment and explore other directions or part ways. When a project is finally technically and legally validated, it becomes binding on both sides. However, the process ensures that by then the parties share a clear understanding of the potential impact and ‘return on investment’.

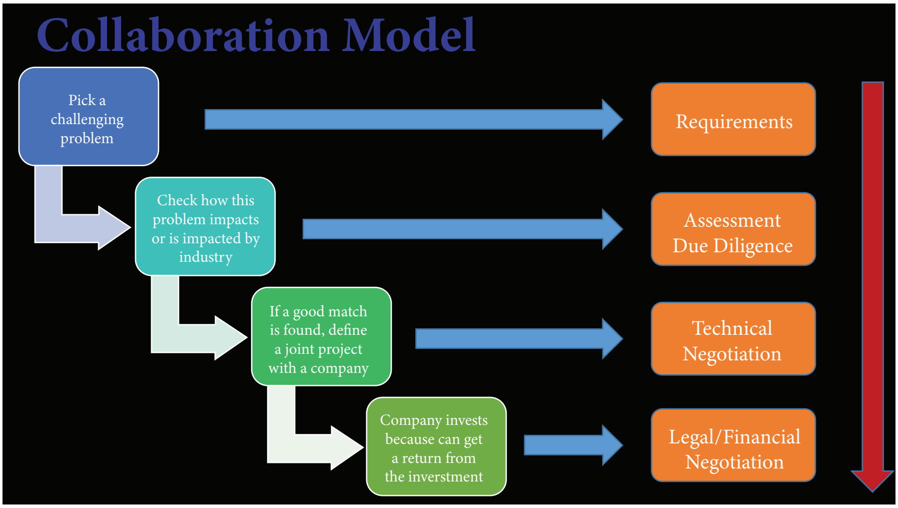
Collaboration Model
Pick a challenging problem
Check how this problem impacts or is impacted by industry
If a good match is found, define a joint project with a company
Company invests because can get a return from the inverstment
Requirements
Assessment Due Diligence
Technical Negotiation
Legal/Financial Negotiation

Figure 11.3 The CERN openlab steps of the collaboration process

Source: Created by Alberto Di Meglio

During the first 20 years of the history of CERN openlab, on only two occasions was a collaboration terminated before the expected end of a project; once because of a lack of agreement on the intellectual property rights (IPR) of the project results, and once because the chosen technology did not live up to expectations very early in the project lifetime and no suitable alternatives could be found. These could have been avoided by a thorough assessment during the preparation phases.

11.2.4 Innovation, Knowledge Sharing, and Impacts

Although the primary goals of CERN openlab is to support the scientific research done at CERN in high energy physics applications, innovation in ICT infrastructure, computers, and data science is of general applicability and potential benefit in many other domains. Innovative multidisciplinary research projects have been developed during the past few years in a variety of disciplines, including earth observation, social and economic sciences, life science and medical research, and earth science.

The main principles remain the same, the starting point of every conversation is a common interest or challenge among the different parties. However, in this type of collaboration there is a strong element of knowledge-sharing across disciplines. The change of perspective that can be provided in multidisciplinary collaborations has proven to be very often the spark that ignites innovation. As an example, a project to 'modernise' the code used in radiation transport problems for high energy physics detectors, that is to optimise it for newer generations of high-performance multi-core architectures, sparked an interest in applying similar optimisation techniques in the simulation of biological cell development. A few initial projects were defined to assess the applicability of cancer cell growth simulation or to validate models of the development of eye retina structures. The activities attracted the interest of new experts, and a formal open-source community called BioDynaMo (Biology Dynamic Modeller) was formed in 2018. Applications to in-silico simulation of immunotherapy applications for the treatment of Alzheimer's disease with industry followed in 2019. In 2020, the agent-based simulation engine used to model cell dynamics was adapted to simulate epidemiologic models of Covid-19 propagation. Today Big Science systems are being adapted to model socio-economic conditions of city areas and populations and predict the possible impact of investments in transport, education, or healthcare systems.

The challenge for Big Science data is to converge different approaches on common problems and facilitate multidisciplinary teams to respond and collaborate on long term and complex questions. Such data convergent designs and models are useful because scholars from different disciplines are accustomed to using different research methodologies, whether qualitative, quantitative, simulation based, or involving interventions and action research (Smart et al., 2012). As discussed in Chapter 8, astrophysics organisations in particular have adopted policies and approaches to sharing data and facilities in an Open Science flat form.

The world has recently witnessed the rise of novel concepts and technologies that have the potential to revolutionise not only scientific inquiry but also industry and society. Artificial intelligence and new computing systems based on quantum technologies have moved from the laboratory to concrete applications, with a strong projected impact on physics research, chemistry, biology, finance, and the social sciences. However, questions of governance, fairness, and ethics cannot be ignored. CERN and its culture of collaboration between research, industry, and society can be a bridge between different disciplines and communities and a catalyst for open, fair innovation.

11.3 Challenges in Managing Big Data in Big Science

With the increased scale of data created in Big Science projects the management of processes to identify useful information within data, make information available along collaborative communities, communicate hypotheses and empirical results and disseminate insights to society became an incredibly complex task (Bicarregui et al., 2013, on practical challenges to implement data management processes in Big Science projects). With the term 'fourth paradigm of science' (Hey et al., 2009), data driven scientific research was even described as a new and different way of scientific exploration added to methods from empirical evidence, scientific theory, and computational science.

Big Science organisations and experiments have become increasingly data driven. These data programs now play a critical role in enabling and accelerating Open Science which open up a more collaborative culture that enables many possibilities for the open sharing and use of data, information and knowledge within and beyond the scientific community that generates such data.

Machine learning as well as artificial intelligence tasks as part of deep learning have been recognised as important tools to accelerate and extend the possibilities of identifying patterns and using big data sets (Hey et al., 2020). Experts in artificial intelligence and machine learning, however, still emphasise the limitations of these tools, when it comes to tasks like reasoning, planning, and acting in complex environments (Kersting and Meyer, 2018).

In general, the collaborative ethos with long traditions of sharing data as well as the strong awareness of the importance of excellent data management processes in Big Science projects are recognised as important preconditions for data management (Bicarregui et al., 2013). In the following, we look at the example of LIGO (Laser Interferometer Gravitational-Wave Observatory) in the US and its data management plan to describe the key ingredients of a systematic organisation of these processes. The Data Management Plan (DMP) of LIGO (2017) is a deliverable as stipulated

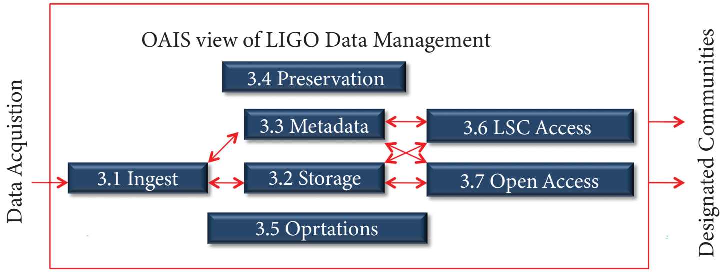

Figure 11.4 LIGO Data Management Plan

Source: Courtesy Caltech/MIT/LIGO Laboratory

in the Cooperative Agreement with the National Science Foundation for the operation of LIGO (see Figure 11.4). The data plan describes how data from the LIGO instruments will be collated, stored, and made available to a range of communities of users of the data and should be preserved for future uses.

LIGO is working closely with the Gravitational Wave International Committee (GWIC) to engage with international partners to coordinate broader access to all gravitational wave data. The LIGO Scientific Collaboration (LSC) provides an avenue for the global community to be involved in the use and analysis of LIGO data.

LSC is a group of scientists focused on the direction of gravitational waves. Gravitational waves are used to explore the fundamental physics of gravity and the emerging field of gravitational wave science as a tool for astronomical discovery and the study of phenomena such as dark matter and dark energy. Such scientific collaborations are significant investments at the scale of Big Science operations. The data collection and analysis consist of multidimensional processes where information from all sources is shared and enriched within the organisation's protocols.

In the case of LIGO, open data and preservation are done in two phases: Phase 1 involves detection or discovery phases and Phase 2 involves observational phases. During Phase 1, data are released after careful vetting and in batches that detect, for example, the initial discoveries of gravitational waves. These data are continuously corrected and upgraded with better calibration and understanding of instrumental artefacts. As the data management plan explained, the data release to the broader research community makes the most sense for validated gravitational wave data surrounding confirmed discoveries. Detected events in this phase are released with sufficient data so that the broader community is able to comprehend the reported signals. In the observational phase, LSC work closely with the broader research community to establish data requirements, then build and field-test data access methods and tools. These methods and tools are provided to the broader community in Phase 2. The understanding of the data is continuously improved as the field progresses from initial discoveries to the more mature exploration of the astrophysical content of gravitational wave signals.

All Big Science organisations have sophisticated systems for releasing data for scientific research. For example, the LIGO Open Science Centre was created to build data products suitable for public release. LIGO's Open Archival Information System (OAIS) represents the storage, access, and preservation of data management for the scientific community. In this way, Big Science collaborations develop a system of production, dissemination, and storing of data in purposeful and systematic ways. Big Science organisations like CERN, LIGO, and Australian Nuclear Science and Technology Organisation (ANSTO) present real challenges for the creation and dissemination of useful knowledge and information. At the core of high energy physics experiments at LHC, CERN, and astrophysics detectors are the needs to generate data that provide useful information for the scientists. An extensive analysis of publicly available data provides more meaningful insights and knowledge that can be interpreted into new discoveries.

There is always a lag time before data is released to the public. Based on the experience of LIGO, the process of annotating the data that includes identifying artefacts, tagging hardware injections, providing accurate calibrations and performing analyses can take up to 18 months. At this point, an 18-month delay before public release is required for vetted gravitational wave data, at least for the first data from Advanced LIGO (LIGO, 2017: 7). Generally, there is a 12-month delay in data releases, during which time mission scientists have access to the data for analysis, vetting, and clean up before it is made available to the broader community. Similarly, it took two years for the virtual Event Horizon Telescope (EHT) project to start the first Earth-spanning observation campaign until the release of the first image of a black hole (Castelvecchi, 2019).

These data management plans are a reaction to challenges created by the complexity of big data as well as legal and political requirements. Government initiatives on open data, the privacy of personal data, and responsible research have a significant impact on data management issues because most Big Science projects require coordination between multiple governments and are dependent on public funding. A typical example of this is CERN and its wide range of international collaborators (Bicarregui et al., 2013). A closer examination of selected government initiatives is described in the next section.

11.4 Social Responsibility in the Digital Age

During the last three decades, technological progress in data generation, storage, processing, and assessment has stimulated a huge variety of communities in different scientific fields outside high energy physics towards the development of international collaboration projects paving the way towards Big Science (Ulnicane, 2020; Cramer et al., 2020b). Typical fields include molecular biology and neurosciences as well as climate research or spatial environmental planning. These directions towards investments in common basic research infrastructures were accompanied by intensified debates on the expected results from these investments. In the USA, controversies over government spending on basic research culminated in 2010, when the House Majority Leader issued a website called 'you cut' allowing citizens to vote for cutting Federal programmes, including basic research (Fahrenthold, 2013). Political decisions on the allocation of public R&D have been increasingly based on partisan political rationales (see von Schomberg, 2013).

In the European Union, 'responsible research and innovation' became the most influential catchphrase to define political expectations for basic research programmes in the context of Big Science (on the various dimensions of this term Burget et al., 2017; Ulnicane, 2020b). Within its EU R&D Framework Program 'Horizon 2020' the EU defined responsible research and innovation as the main focus. However, the underlying principles and their relevance in the context of digitalisation remained vague at the beginning of the programme. Four specific fields of challenges occurring in the digital age were identified by different authors and public authorities (Stahl, 2013; Elliot and Resnik, 2014 and 2019; Resnik and Elliot, 2016; Filipovic, 2018; Fothergill et al., 2019; Inverardi, 2019; NQIT, 2019; European Commission, 2020):

- Issues of data generation, (re-)use, storage, protection, and privacy;
- Issues of intellectual property;
- Issues of transparency to maintain trust in the verifiability of scientific results based on artificial intelligence and machine learning; and
- Issues of inclusive education and access to scientific data and information.

In the next section, recent developments and challenges in public health surveillance are introduced as a typical example for scientific fields, where 'responsible research and innovation' serves as important ingredients of strategies to cope with new opportunities and (technological as well as regulatory) challenges in the digital world.

11.5 Public Health Surveillance in the Digital World

Public Health Surveillance is 'the continuous, systematic collection, analysis and interpretation of health-related data' (WHO, 2021). Public health surveillance needs led to the definition and establishment of new research directions at the end of the 1990s. Thanks to the development of data management, storage, availability, and the inception of the cloud, an important step from conception to implementation in healthcare has been made possible. The development of online health surveillance platforms was facilitated by the ability to combine and coordinate distributed and heterogeneous computational resources, by gaining access to more advanced information systems and distributed services, by enjoying increased computing and storage capacity, and by being able to use more powerful computers and storage devices.

Since these first emerging steps, big data lakes have been available and used around the world. Numerous epidemiological and scientific queries on knowledge and treatments of specific diseases are now possible due to the implementation of technological findings from high energy physics experiments. One of the outcomes worth mentioning is the scientific grid infrastructure, GÉANT, the broadband and highspeed network for research and education.[4]

Generally speaking, deep-techs such as big data, blockchain, and artificial intelligence are now combined and at the core of research and innovation in many strategic activities of national and private business-oriented institutions and organisations: healthcare services, personalised prevention, the management of healthcare, legal security of digital transactions, or archiving all benefit from disrupting technological approaches. As an example, the recent collaboration at CERN openlab with Beys Research, resulted in a tripartite research team developing a platform based on the blockchain to support European research centres and hospitals in constituting 'Living Labs' where sensitive research data can securely be shared (Frisoni et al. 2011).

Beyond the Living Lab project, disruptive quantum computing has also become a field of collaboration in the public health context. This is the case for Quantumacy,[5] a project launched in January 2021 by Beys Research and CERN openlab.

Quantumacy stands for QUANTUM key distribution for large-scale informational self- determination, privacy awareness, and the distributed processing of personal sensitive medical data.Quantumacy aspires to support the transition to confidential computing and informational self-determination by integrating relevant technologies into core Quantum Key Distribution (QKD)-enabled services requiring key issuance and distribution, resulting in a new approach to making data available and processing under the so-called polymorphic privacy principle.

To that end, Quantumacy is developing a proof-of-concept platform in which existing encryption techniques are combined and extended to use the open QKD infrastructure across the platform's different layers, allowing for testing on concrete healthcare use-cases.

Combining privacy-preserving and enabling technologies with blockchain to secure data registration and exchanges across the entire data lifecycle, along with

[4] GÉANT develops and operates a range of connectivity, cloud, and identity services that ensure a safe and secure environment for researchers, educators, and students. See https://geant.org/.

[5] Quantumacy is a privacy-preserving data analytics platform combining the security of Quantum Key Distribution (QKD) protocols QKD protocols and links with state-of-the art homomorphic capabilities to execute machine learning and deep-learning workloads across a distributed federated-learning infrastructure (CERN, 2022, https://quantumacy.cern.ch/home).

full audit trails, while relying on a QKD-enabled network infrastructure, will result in unprecedentedly secure, reliable, and unfalsifiable information systems in the future.

Cancer screening, distant analysis of medical dossiers, and personalised medicine are areas where recent improvements in data analysis and sharing have acted as catalyst for innovative solutions in the health sector. The dawn of all those developments including the assessment of the genome of individuals, the functional genomics developments, and following data collection, storage, and sharing for personalised medicine brings with it many extremely important ethical questions in terms of security and privacy. Diseases with high societal and economic impact, such as cancer, neurodegenerative conditions like Alzheimer diseases, Parkinson diseases, or Soluble Liver Antigen (SLA) pathologies require innovative research methodologies using digitalisation for treatment, follow-up, data analysis, and epidemiological assessment (Fogel and Kvedar, 2018).

In Europe, software developments that handle personal data must adhere to the General Data Protection Regulation (GDPR). These regulations provide data protection principles and new individual rights in business processes. Different approaches to privacy and confidentiality preservation are being explored today. An emerging concept in the field is the Private or Confidential Computing, whereby data is protected (e.g. by means of encryption or anonymisation) across all stages of the data life cycle.

Several commercial cloud providers such as Google or Microsoft have recently proposed pilots using encrypted virtual machines. The future of large-scale data analysis systems, especially in areas where sensitive data is processed, requires new secure ways of protecting data. Typically, a Confidential Computing platform protects four areas of activities:

(1) *Data at rest*: protect data stored in repositories—typically using suitable encryption systems;
(2) *Data in transit*: protecting data as it is transferred across the network, typically by encrypting the contents, the channel, or both;
(3) *Data processing*: protecting data and models as they are analysed/trained/inferred, using techniques such as Secure Multi Party Computation protocols (SMPC) or ML/DL-compatible encryption schemes (homomorphic encryption, functional encryption); and
(4) *Operations*: recording and auditing transactions, managing AuthN/AuthZ,[6] provenance, etc., typically using technologies like blockchain ledgers, x509 security frameworks,[7] etc.

Against this backdrop of growing needs and opportunities to collect and connect big data for public health surveillance while ensuring the protection of individual personal data in the digital world, CERN, and other Big Science organisations' efforts

[6] AuthN is short for authentication and AuthZ is short for authorisation.

[7] *X.509* 'certificates are used in many Internet protocols, including TLS/SSL, which is the basis for HTTPS, the *secure* protocol for browsing the web'.

to build Open Science platforms which serve as critical building blocks for innovative and rapid public health strategies, are explained in the following section.

11.6 Open Science and Covid-19 (CERN and EMBL)

Ideally, all high energy physicists wish to share information openly about their research as enshrined in the values of CERN's original charter. CERN values recognise the universal importance of the fundamental scientific research knowledge produced at CERN to make this knowledge available to everybody for the benefit of the society.

The core foundation of Open Science[8] is making publicly funded research data results available for the public, education and development, and the swift transfer of scientific knowledge. Initiatives have been growing including Open Science access, Open Data, Open Sources and Open Science projects. All of these initiatives aim to disseminate knowledge for the public good. The dissemination of knowledge has been facilitated by the availability of tools and means that facilitate the dissemination, sharing and deposition of big data and information. CERN *Open Science Policy*, adopted in October 2022, provides provisions for all research publications and experimental data, and research software and hardware are to be made publicly available under the new policy, which also incorporates the previous rules for open access, open data, and open source software and hardware. It is important to note that the term 'Open Science' is not just about opening science in a way that makes knowledge accessible without cost for everyone (Naim et al., 2020).

CERN promoted volunteer initiatives such as Rosetta@home and Folding@home, as well as LHC experiment technologies such as the File Transfer System (FTS) and Rucio, a modular, scalable solution for searching high energy physics data files in distributed data centres for monitoring and analysis. Furthermore Zenodo, the open-data repository built and operated by CERN and OpenAIRE to ensure that everyone can join in Open Science, has been extended to research communities collecting research output and information for Covid-19 and SARS-CoV-2, as we will describe at the end of this section.

CERN and EMBL, as members of the European Intergovernmental Research Organisation forum (EIROForum, https://www.eiroforum.org/), have extensive experience in operating data-services for the international community over several decades and are key actors in data intensive science, curating some of Europe's most important and popular datasets. All members of the EIROforum are committed to ensuring their openly available scientific datasets are curated over the long term and maintained through certified digital repositories. Thus, CERN and EMBL, together with the other EIROforum members have stimulated the Open Science movement across Europe and have shown a strong involvement in the European Open Science

[8] Open Science is a term often coined to Steve Mann in 1998 (Wright, 2020), but also already mentioned by Chubin (Chubin, 1985).

Cloud. Open access to research data proved to be not only a selected benefit for the relevant scientific communities dispersed around the globe but also an important stimulus for interdisciplinary science and transversal dynamic research and education. The outcome of such Open Science initiatives has had and will continue to have implications and outcomes for society worldwide. We will now look closer at the variety of initiatives at CERN and EMBL.

11.6.1 CERN

CERN's convention states: '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.'

Open Access has a long tradition at CERN. Since its inception in 1954, particle physicists at CERN have shared their results with fellow scientists across the world by distributing pre-prints. With the introduction of *arXiv* in the 1990s electronic distribution of pre-prints became the norm and since 2014, the Sponsoring Consortium for Open Access Publishing in Particle Physics *(SCOAP)*, hosted at CERN, has made around 90% of all high energy physics peer reviewed journal articles available on open access.

CERN pioneered the largest open access initiative of research data from high energy physics in the world. It promoted the dissemination of research results by hosting SCOAP^3 with over 3000 libraries and research institutes from 46 funding agencies and research institutions from 46 countries and intergovernmental organisations like the European Commission and UNESCO. The aims of these activities go far beyond particle physics. From December 2020, the establishment of a new open data policy to further support Open Science on scientific level 3 data (the type required for scientific studies) from LHC experiments will be released alongside the software and documentation needed to use the data. The data is made accessible via the CERN Open Data Portal.

11.6.2 EMBL and EMBL-EBI

The European Molecular Biology Laboratory (EMBL) was founded in 1973 with the key mission of carrying out basic research in the life sciences. EMBL endorses Open Science as an integral part of its research activities, as underlined already in its Establishing Agreement of 1973, Article 2§1:

> The Laboratory shall promote co-operation among the European States in fundamental research, in the development of advanced instrumentation and in advanced teaching in molecular biology as well as in other areas of research essentially related thereto, and to this end shall concentrate its activities on work not

> normally or easily carried out in national institutions. The results of the experimental and theoretical work of the Laboratory shall be published or otherwise made generally available.

EMBL-EBI is the European Bioinformatics Institute, based at the Welcome Genome Campus, Hinxton, Cambridgeshire, UK. It is one of the sites of the European Molecular Biology Laboratory (EMBL) and a world-renowned computational biology institute. Since its inception the newly created EMBL has needed leading-edge computing resources to deal with data collection, analysis, and sharing from different projects carried out by the many Member States within their projects. EMBL's biological data resources on nucleotide (EMBLBank) and protein databanks were initially hosted in Heidelberg until the migration of data and services to the site in the UK, thus establishing the EMBL-EBI. Biocomputing activities and computers are integral part of the research activities throughout EMBL. EMBL-EBI, leading computational biological research since 1993, is '*The home for big data in Biology*'. Their data portal allows the user to browse data, perform analysis, and share research among the molecular biology community. Another important mission for EMBL-EBI is the delivery of training in data-driven life sciences.

Open Targets is a public–private partnership that uses genomic data for drug target identification with the aim of helping scientists develop new, safe, and effective drugs. The Open Targets Platform and the Open Targets Genetic Portal are the two sectors containing repositories. In the Platform, information is available for some types of illnesses spanning from RNA expression, genetic, drugs, text mining to animal studies, whereas the Genetic Portal enables the user to explore variant, gene, and trait associations. This fundamental knowledge will allow cohort-based, personalised medicine, and personalised treatment to become part of our daily lives, for which, all research groups need and use computational approaches to drive new experiments. EMBL-EBI's Industry Program has been a key mission since its founding, allowing regular contacts between research groups and pharmaceutical and biotech companies giving outcomes of paramount importance. The results obtained from the joint projects carried out in such contexts are also publicly available.

In addition, EMBL has commitments to Open Science in its Open Science Policy and as a signatory to DORA: the San Francisco Declaration on Research Assessment (https://sfdora.org) upholding the principles and processes defined in the declaration in all its sites.

11.6.3 EOSC

Alongside its commitment to Open Science, EMBL together with CERN and other members of EIROforum is involved in the continuous development of the *European Open Science Cloud* (EOSC). EOSC is a joint initiative between the European Commission, the EU member states and associated countries, and other stakeholders to provide open access to publicly funded research data across scientific domains and without geographical boundaries, including services addressing the whole research data life cycle. Developing iteratively, it aims at finding, accessing, combining, analysing, processing, and storing data in line with Open Science and FAIR principles (findable, accessible, interoperable, reusable). By bringing all available information and data together, it will maximise the impact of publicly funded European research and boost applications.

EOSC can be considered an important catalyst for Open Science and a major player in implementing FAIR data. Integrating EMBL's and CERN's rich open data resources with EOSC aims not only to further underline the organisations' commitment to Open Science but can also act as a catalyst to widen the use of their data resources and make them more accessible beyond the respective physics and life sciences domains. EOSC acts as the missing link between EIROforum's own data resources and those data sets provided by other disciplines, that are not readily accessible, thus fostering transdisciplinary research. Beyond the federation of data, EOSC enables more effective coordination of existing and future research infrastructures and e-infrastructures at the European level with the opportunity to connect with major centres and national initiatives in the European Member States and beyond. Consequently, EOSC aims to enable the federation of the organisations' existing and future computing services to support the distributed analysis of large-scale data. Additionally, EMBL, CERN, and others will be able to provide and also benefit from coordinated training, information and dissemination opportunities, particularly related to data science and FAIR data management.

Even before the EOSC initiative had started, CERN and EMBL were two founding members involved in the Helix Nebula Science Cloud initiative. A forerunner for EOSC, the initiative established a successful pilot cloud platform linking together commercial cloud service providers with the IT resources of ten leading European research centres in the areas of astronomy, high energy physics, life sciences, and photon/neutron sciences.

Throughout its different phases, CERN and EMBL have significantly contributed to the development of EOSC, e.g. through coordination of and participation in key EOSC projects of the European Framework Programmes Horizon 2020 and Horizon Europe (EOSC-Pilot, EOSCHub, EOSCEnhance, EOSC-Life, Covid-19 Data Portal, and others). Additionally, both organisations have been heavily involved in shaping the implementation of EOSC, playing major roles at the EOSC governance level. EMBL was represented on the EOSC Executive Board in 2019–2020 and co-led the Board's Working Group on EOSC Sustainability, which significantly contributed to implementing the EOSC Association. CERN is a member of the EOSC Association directorate.

The EOSC Association AISBL was established to sign a co-programmed European Partnership with the European Commission under the Horizon Europe Framework Programme. The EOSC Association is expected to play a crucial role in EOSC's development by enabling funding by the Commission as well as contributions by

EU Member States and Associated Countries throughout the duration of Horizon Europe through a Strategic Research and Innovation Agenda (SRIA). Another important aspect is that the EOSC Association brings together a rapidly growing number of stakeholders as Association members and both EMBL and CERN became full members of the EOSC Association at the first General Assembly in 2020, joined by other EIROforum organisations. Both organisations continue to be involved in different strategic efforts to shape EOSC, e.g. through the SRIA and by participation in the Association's Advisory Groups.

This engagement has not only brought additional visibility in areas where e.g. EMBL was less prominently known before, but has significantly shaped the direction in which EOSC is moving forward. Setting up EOSC is an activity involving a large variety of communities with different maturity levels in their data and services. EMBL, CERN, and other Big Science centres are forerunners with regards to cloud-based, large-scale, and globally distributed research activities. Aligning EMBL and CERN with EOSC is therefore both a challenge and an opportunity for both organisations to act as role models for others.

11.6.4 Covid-19

The Covid-19 outbreak and continuing pandemic called for rapid and synchronised interactions by sharing data and research results across borders. Open Science and the tools and infrastructures already developed to share scientific information and data rapidly responded to this critical and important need facilitating the sharing, analysis, monitoring, and assessment of breakthrough advancements to explore and find possible solutions. In April 2020, EMBL-EBI launched the Covid-19 Data Portal to include a wide range of data types including genomics, protein and microscope data, as well as a repository of scientific literature. One can be impressed by the rapidity acquired in analysing the virus variants, but this is the result of the strong involvement of EMBL in European Open Science in the development and sharing of many years of previous genomic research and ad hoc development of technologies, as well as the availability and freely shared international scientific information in a dedicated and rapidly designed secure portal. CERN also got involved in ventilator and computer technology to deal with the Covid-19 pandemic. Following the request for actions from the European Commission, Zenodo and OpenAIRE joined their competencies and collaborated to make scientific information available via the OpenAIRE Covid-19 Research Gateway. The OpenAIRE Covid-19 Research Gateway provides a single point of access where publications, data, software, and other research outcomes are made available due to the collaboration with pan European research infrastructures and national and international alliances. Trusted sources and scientific material with bibliographical references are made freely available. All these initiatives to accelerate and facilitate the exchange of leading-edge scientific knowledge formed the basis for the relatively fast responses to the challenges caused by the global pandemic.

11.7 Big Science's Contributions to Societal Challenges

At the beginning of this chapter, we introduced three research questions to guide us through our observations:

- which organisational principles and elements were installed by CERN to maximise the mutual benefits from technological development in computing for the scientific community as well as the commercial partners from the IT industry?;
- which principles help to organise and manage big data in Big Science projects, and how can these principles meet the expectations of policy and society?; and
- how could the principles of responsive research and innovation in the digital age be transferred into a transformation of Big Science towards Open Science to maximise and accelerate societal benefits while protecting personal data?

CERN openlab provides key messages to answer the first question. Transparent processes and joint agreements with commercial partners on exclusively pre-commercial activities and mutual benefits served as foundations for joint and collaborative research on leading-edge technologies for the digital age. The example of the data management plan at LIGO took us to the answer to the second question.

Despite massive and continuously growing volumes of new data, it is called for new data management protocols to facilitate scientific progress. These data protocols also stimulate data management and linkage systems in other scientific fields such as medicine and other social research environments by improving tools to filter and combine meaningful data and forming open data infrastructures.

Our examples from public health surveillance, particularly their applications to research during the current Covid-19 pandemic, helped to answer the third research question. The emergence of open data infrastructures to be used in transdisciplinary environments and the guarantee of protection for personal data by the FAIR principles formed the preconditions for fast reactions to unprecedented research questions during the pandemic.

Especially in times of sovereign budget cuts and austerity policies, funding Big Science projects comes under high scrutiny. The digital age cannot be imagined without the important technological and social developments initiated by collaborations between scientific communities, industrial partners from the IT sector as well as governments and political organisations. From the first steps of computing to collaborations using the Internet and the future of quantum computing, Big Science organisations such as CERN and ESO provide stimuli for new technological developments, create challenging innovation environments to test leading-edge appliances and stimulate creative ideas for new forms of use and applications. The case of CERN openlab reveals the potential of clearly focused organisational frameworks for Big Science–big industry collaborations in the context of pre-market developments and testing. Programmes, processes, and learning environments were developed to maximise the positive impact on technological infrastructures for scientific research as well as on innovative output in the IT sector.

Simultaneously, the transition towards big data led to new challenges for responsive scientific research. These examples from public health emphasise the importance of Open Science infrastructures using experiences from scientific communities in high energy physics as well as bioinformatics. These infrastructures do not only help to follow the pathways of well-ordered science but also to initiate fast common scientific efforts in urgent crises like the Covid-19 pandemic. They also serve as a model for how the FAIR principles of dealing with big data (findable, accessible, interoperable, reusable) can be implemented in a transparent way. These experiences form the basis for broad applications of novel regulatory approaches in commercial uses of big data. Again, the experiences of Big Science communities can help to understand and imagine how the protection of privacy in everyday digital applications based on big data can be achieved in a meaningful way. A challenge, however, remains for any societal use of open data infrastructures, as the collaborative ethos and mindsets of Big Science projects are missing in commercial and competitive contexts.

11.8 Conclusions

We live in a digital society that requires fast and rapid changes due to the way we access, share, and use data. The LHC, LIGO, and other large telescopes produce enormous amounts of data. The present and next-generation Big Science instruments are even going to be laden with massive data generators. The quest for collecting, storing, and analysing data is a challenging task.

Volumes of data, data analysis, and data management in Big Science from HEP to astronomy and molecular biology became ever more important and highly sophisticated. As detection, computing, and digital technology advances, large volumes of data can be captured and stored and need to be disseminated to collaborations located worldwide. Therefore, in Big Science organisations and experiments, well designed research, data handling methodologies and coordinated approaches to carry out and execute the experiments as well as transfer the collected information to the stakeholders are vital and necessary to establish from the inception.

It is important to outline the joint partnerships that Big Science has established with leading IT companies and other research institutes, with openlab at CERN and the industry programmes at EMBL. It is impossible to predict how newly developed algorithms, rare signal discrimination, and methodologies to reconstruct images used in high energy physics and astrophysics will affect our daily lives in the near future. Furthermore, the more intriguing question is how the massive volume of data produced in cosmology or by the LHC experiments can be used and contribute to the broader physics and astrophysics communities and/or educational purposes by companies for practical use and public benefit. The LIGO and ESO for example, have data protocols for more public access to all their data within accepted data management protocols.

Another challenge for IT is how to design memories that are fast, large, and responsive. The ATLAS at CERN has already generated 140 petabytes of data, distributed between 100 different computing centres, with most of it concentrated in 10 large computing centres like CERN and Brookhaven. New large-scale astrophysics projects are producing data and information at unprecedented scales and facing computing challenges at the 'exascale' level and beyond. Astrophysics now uses large data surveys like the Dark Energy Survey, Sloan Digital Sky Survey, and will generate petabytes of data from LIGO, telescopes such as the Square Kilometre Array and the Large Synoptic Survey Telescope. The high energy physics community and astrophysicists must define data collection and storage and use strategies to deal with the massive amounts of data that these infrastructures will generate over the next few years.

Furthermore, this calls for the harmonisation of data usage and ongoing efforts are being made by national governments and the European community. In EMBL's Open Science initiatives, for example in the development of the Covid-19 Data Portal or in the development of the European Open Science Cloud (EOSC), the EU member states and associated countries have collaborated on open data usage.

Big data and Big Science are closely coupled and have the potential to revolutionise the way we organise and conduct research. Open access to publicly funded research data across scientific domains and without geographical boundaries, including services addressing the whole research data life cycle, is increasingly becoming valuable for science-making inroads to social, educational, and economic development. Rapid progress in artificial intelligence with neural networks will also assist societies in making use of data from Big Science projects.