Computing and Software for Big Science (2022) 6:6
https://doi.org/10.1007/s41781-022-00081-7

ORIGINAL ARTICLE

Survey of Open Data Concepts Within Fundamental Physics: An Initiative of the PUNCH4NFDI Consortium

Harry Enke[1] · Andreas Haungs[2] · Thomas Schörner-Sadenius[3] · Kilian Schwarz[4] · Markus Demleitner[5] · Achim Geiser[6] · Lukas Heinrich[7] · Michael Kramer[8] · Gernot Maier[9] · Dominik Schwarz[10] · Hendrik Seitz-Moskaliuk[11] · Hubert Simma[9] · Michael Sterzik[12] · Stefan Typel[13]

Received: 29 September 2021 / Accepted: 2 February 2022

Abstract

PUNCH4NFDI (Particles, Universe, NuClei and Hadrons for the NFDI) aims at developing concepts and tools for the efficient management of digital research products in fundamental physics research. At the heart of the research products are scientific data sets that should be made interoperable and available to a broad scientific community and the public for a sustainable usage ("open data"). The first PUNCH4NFDI "Open Data Workshop" gave the opportunity for an initial survey of existing and planned open data initiatives within the PUNCH science field. The paper addresses the conceptual differences and commonalities of the participating communities presented in the workshop. Existing open data collections were presented and discussed. This is an inquiry into the community's requirements for a better use of open data and in this context also of "Open Science".

Keywords Fundamental science · Physics · FAIR data · Open data · Open science

✉ Thomas Schörner-Sadenius
thomas.schoerner@desy.de

Harry Enke
henke@aip.de

Andreas Haungs
andreas.haungs@kit.edu

Kilian Schwarz
k.schwarz@gsi.de

Markus Demleitner
msdemlei@ari.uni-heidelberg.de

Achim Geiser
achim.geiser@desy.de

Lukas Heinrich
lukas.heinrich@cern.ch

Michael Kramer
mkramer@mpifr-bonn.mpg.de

Gernot Maier
gernot.maier@desy.de

Dominik Schwarz
dschwarz@physik.uni-bielefeld.de

Hendrik Seitz-Moskaliuk
Hendrik.Seitz-Moskaliuk@nfdi.de

Hubert Simma
hubert.simma@desy.de

Michael Sterzik
msterzik@eso.org

Stefan Typel
stypel@ikp.tu-darmstadt.de

[1] Leibniz Institut für Astrophysik Potsdam, An der Sternwarte 16, 14482 Potsdam, Germany

[2] IAP, Karlsruhe Institute of Technology, 76021 Karlsruhe, Germany

[3] Deutsches Elektronen-Synchrotron DESY, Notkestr. 85, 22607 Hamburg, Germany

[4] GSI Helmholtzzentrum für Schwerionenforschung GmbH, Planckstr. 1, 64291 Darmstadt, Germany

[5] Astronomisches Rechen-Institut, Zentrum für Astronomie der Universität Heidelberg, Mönchhofstr. 12–14, 69120 Heidelberg, Germany

[6] DESY, Notkestr. 85, 22607 Hamburg, Germany

[7] CERN, 1211 Geneva 23, Switzerland

[8] Max-Planck-Institut für Radioastronomie, Auf dem Hügel 69, 53121 Bonn, Germany

[9] DESY, Platanenallee 6, 15738 Zeuthen, Germany

[10] Universität Bielefeld, Universitätsstr. 25, 33615 Bielefeld, Germany

[11] Nationale Forschungsdateninfrastruktur (NFDI) e.V., NFDI-Direktorat, Albert-Nestler-Straße 13, 76131 Karlsruhe, Germany

[12] ESO, Karl-Schwarzschild-Strasse 2, 85748 Garching, Germany

[13] Fachbereich Physik, Institut für Kernphysik, Technische Universität Darmstadt, Schlossgartenstraße 9, 64289 Darmstadt, Germany

Published online: 09 March 2022

The PUNCH4NFDI Consortium

PUNCH4NFDI [1] is the consortium of particle and astroparticle physics, astrophysics, and hadron and nuclear physics in the German Nationale Forschungsdateninfrastruktur (NFDI, national research data infrastructure) [2]. The consortium—actively supported by more than 40 institutions from German universities, the Max Planck Society, the Helmholtz Association and the Leibniz Association—enjoys a broad community support, not least reflected by the close relations to the elected community representations of the Committee for Astroparticle Physics (KAT), Committee for Elementary Particle Physics (KET), Committee for Hadron and Nuclear Physics (KHuK), and the Rat deutscher Sternwarten (RdS).

In autumn 2020, PUNCH4NFDI submitted a proposal for funding to the DFG which has been accepted [3]. In the proposal, the consortium sketches a broad array of actions towards a management of FAIR data in the "PUNCH sciences" ("Particle, Universe, NuClei & Hadrons"), with key elements being dynamic "digital research products" (RP) and the PUNCH Science Data Platform (PUNCH-SDP)—a complete ecosystem in which RPs can unfold their full life cycle and become a boost to scientific work. With the platform, PUNCH4NFDI will deliver a full set of (modular) services and tools for data management based on a layered data model.

In February 2021, PUNCH4NFDI held a half-day workshop on "Open Data in the PUNCH Sciences", attracting more than 100 participants for an afternoon of presentations and discussions about the status of "open data" in the various fields of PUNCH4NFDI. All the contributions to the workshop can be accessed under https://indico.desy.de/event/28448/. The workshop was intended as a survey of best practices and of immediate needs, and as a starting point for actions to be taken by the consortium. This paper summarises the discussions of the workshop and aims at drawing first conclusions for the future work of PUNCH4NFDI in the field of open—and FAIR—data.

Data Management and Open Data in the PUNCH Sciences

Open data have many different definitions, depending on context and intention. For an initial definition we adopt the Wikipedia one: "Open data is the idea that some data should be freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control." [4].

FAIR always refers to the list of requirements for scientific data, stated in FAIR standards [5]. FAIR is an acronym of *F*indability, *A*ccessibility, *I*nteroperability, *R*eusability. These four requirements are extended by more detailed criteria allowing for a detailed assessment of the "FAIRness" of data collections. FAIR data refer to data collections which claim to fulfill these requirements.

Origins of "Open data" are demands of civil organisations like the "Open Knowledge Foundation" with its mission of "an open world, where all non-personal information is open, free for everyone to use, build on and share; and creators and innovators are fairly recognised and rewarded" [6] and fighting for this goal using legislation such as 'Freedom of Information' acts in many countries worldwide. In the EU, in the context of the Digital Single Market (see, e.g. the EU Parliament [7]), one of the main ideas is that public sector information should be made publicly accessible in a machine-readable form and free of charge. This is supported by legislation of many EU members, see, e.g. the German Open Data Act [8]. Since most of the data gathered in the PUNCH field of science are publicly funded, the relation of the current approaches to make scientific data publicly available and demands for Open Data raises many questions to be resolved.

The requirements of Open Data implies that restrictions on their use are only permitted to safeguard the origin and openness of the knowledge, for example, by naming the author or using a share-alike clause. However, there is also criticism that no legal entitlement to data sets exists and far-reaching exceptions to their publication are allowed [9].

The concept of publishing freely accessible data (meaning: accessible to the scientific community) has been pursued in PUNCH sciences for a long time, e.g. by the Sloan Digital Sky Survey (SDSS) [10] pioneering the use of relational database technology in astronomy, and the Virtual Observatory (VO) showcasing a federated information infrastructure.[1] Another example are data (and their associated metadata) collected by astronomy satellite missions, which regularly are made publicly available, e.g. from the Gaia satellite [11].

Open data and Open science were furthered by initiatives from our communities, and the publication of data collections have a long tradition, even the development of FAIR criteria partially build on these. But the realisation of the FAIR and Open Data principles depends strongly on the (sub-)communities, the abstraction-level of their data, the status of data curation, and are therefore realised to vastly varying degrees.

[1] VO is often used as an abbreviation to refer to the ongoing efforts to standardise data, metadata, and access protocols for astronomical data collections, organised by the International Virtual Observatory Alliance (IVOA).

In particle physics, some aspects of the FAIR and open data principles are well realised. The raw data that are gathered by the scientific instruments, the lower level-derived data, and the corresponding simulations are typically only available within the producing collaborations. Within a collaboration, the data and analysis software are freely available for all researchers. Nevertheless, there are efforts ongoing to gradually release more and more (older) data together with the required software to allow re-analyses by non-collaborators. There is also a long-standing tradition to release higher abstractions of the data and relevant simulations (distributions, histograms, likelihoods) digitally in searchable public archives, e.g. hepdata.net [12]. These archives are connected with databases of research results such as inspire.hep.net. Also, high-level (statistical) analysis and simulation software is freely available.

For hadron and nuclear physics, the same holds true as for the particle physics community in many experiments at CERN (Conseil européen pour la recherche nucléaire) and elsewhere. The smaller hadron and nuclear physics communities still require significant development to fulfil the FAIR and open data principles. The future experiments at the FAIR (Facility for Antiproton and Ion Research) accelerator at GSI (Gesellschaft für Schwerionenforschung), finally, have not yet started to collect data.

In astroparticle physics, whose goal is to conduct astronomy with particle-physics methods, the degree of application of the FAIR and open data principles lies in between those from particle physics and astronomy. Lower data levels and the associated software packages are findable and accessible only for researchers inside the corresponding collaborations. Notable exceptions are, e.g. the open portal KCDC (see below) providing public access to raw data and software tools. Final data products in reusable data formats (e.g. FITS) are often connected to the publication through the ADS service [13] and accessible through the collaboration's websites. The gamma-ray community is in the process of moving towards the usage of open community software (e.g. Gammapy [14]) and data format standards on the reconstructed gamma-like event level ("open gamma-ray data format"), which enables one to publish these data and combine multi-messenger data sets from different observatories in a consistent manner. Observatories under construction (e.g. CTA) will follow FAIR and Open data principles in their research data management.

In astronomy, the requirement of sharing and using data collections from archives and sites across the world began long ago and has accelerated in the digital age. For sharing (initially satellite) data, the FITS format with a limited and controlled metadata vocabulary has been established as a ubiquitous data storage format in all astronomy. Its development started in 1979, first standard was released in 1981, and "evolved out of the recognition that a standard format was needed for transferring astronomical data from one installation to another." [15].

The growing size of astronomical data collections, the need for online selection of subsets on the one hand side (using databases and SQL) and the notion of a "multi-wavelength view of the universe", requiring combination of data sets from distributed data collections on the other hand started the Virtual Observatory initiative (2002). Their goal has been to develop standards and protocols for the exchange and re-use of digital data collections [16]. Many astronomy specific standards include or extend common metadata systems (e.g. OAI-PMH [17]). Its focus has been the service paradigm, and one of its main achievements has been the specification of a data access layer with protocols and standards. Especially in optical astronomy, recent data collections use (partially) VO recommendations for metadata or data access protocols even with non-public data. Also, the development of open source tools such as Astropy [18] and PyVO [19] have boosted the use of VO protocols. Often, public access to new data collections is restricted only for a limited embargo period. For survey data the curation of data collections for public access often takes years and results are organised as data releases, along with scientific papers evaluating the scientific benefits of these data releases.

Survey of Existing Open Data Concepts and Initiatives

To get a first overview of the approach to implement the FAIR and open data principles among the different experiments and fields within the PUNCH4NFDI community an Open Data workshop was initiated to gather the needs, requirements, and available solutions of the participating communities and use this as a starting point for future work. The goal of the PUNCH Science Data Platform to provide access to all the different data resources of the involved communities depends on integrating the various existing concepts which were formed by the needs of the individual research fields and their specific users.

In the following chapter, we summarize the contributions of the workshop to the survey as well as their interpretation with an outlook on the course of action within the PUNCH-4NFDI consortium.

Open Data in the NFDI

The vision of the NFDI is to provide research data according to the FAIR principles where the data sets are openly available as possible and as closed as necessary. In its final stage, the NFDI aims to cover all research disciplines in up to 30 domain-specific consortia [20]. The potential for open data sharing depends on several aspects: are there information

privacy reasons which prevent the open publication of data? Are there questions of licensing and copyright to consider? Does the researcher gain reputation from the open data publication? The consortia of the NFDI will answer these questions differently, therefore, also a different level of openness will be achievable.

To enable the development of interdisciplinary solutions, sections will be formed within the NFDI association. In these sections, experts from all consortia come together to discuss cross-cutting topics of open and FAIR data sharing, e.g. legal aspects, PIDs, metadata standards and ontologies. The first cross-sectional work will start in 2021.

HEP and CERN Open Data

In contrast to Astrophysics, high-energy physics (HEP) has a tradition of 'unlimited embargo periods' for low level research data of most experiments. However, data for educational purposes have been made openly available for a long time, and recently most journal publications by large experimental groups are open access, with published results available in machine-readable formats [12]. Some of the older low level data are also available to physicists beyond the original members of the respective collaborations. Furthermore, CERN has very recently adopted a formal Open Data policy [21] aiming at making all LHC data available via its Open Data platform [22] with an embargo period between about 5–10 years.

Partial archived research level data sets are available via the DPHEP initiative for most of the CERN LEP e^+e^- experiments [23] which were taking data in the 1990s, and for one of the DESY PETRA experiments, JADE [24], for which data were taken in the late 1980s. With close to 10% of the JADE publications arising from the corresponding archived data, they represent the most successful case of useful data preservation in HEP so far, albeit based on a private initiative by a few individuals [25]. The data from the HERA experiments at DESY, for which data taking ended in 2007, have been preserved in a more formal way by the laboratory. They are accessible by individual agreement [26, 27] and are on the way to rival the JADE achievements, although no resources to make them openly available have been allocated.

The CERN Open Data initiative for low-level research data [22] was pioneered by the CMS experiment since 2014, and meanwhile more than 2 PB of CMS data (more than half of the CMS Run 1 data, plus related simulated data sets) are openly available to outsiders in their original analysis format, accompanied by the original analysis software, without any restriction [28]. Despite the need to solve associated non-trivial challenges, they are successfully used for physics publications by non-collaboration members at a current rate of about five publications per year. Large parts of the recent CERN Open Data policy [21] have been modeled after the successful CMS policy, also integrating the corresponding policies by the other LHC experiments. Related projects of data reuse at higher abstraction levels are being addressed in Sect. 4.6.

The PUNCH4NFDI project will hopefully further accelerate this process, and facilitate linking the available HEP data with those from other physics communities.

ESO's Approach as (Open) Science Data Provider

The European Southern Observatory (ESO) is an intergovernmental organisation that builds and operates most advanced ground-based astronomical facilities. All scientific data gathered by ESO facilities, being a publicly funded observatory, are returned to the public. Based on FAIR principles [29], ESO has established an infrastructure that stores, preserves and grants access to scientific and calibration data through a data archive [30]. ESO also recognises that the scientific usefulness and legacy value of astronomical data products increases with their processing level and only then enables their scientific exploitation beyond the original goals of a specific scientific question.

Establishing and preserving data archives would not be possible without an entire ecosystem of data standards, metadata definitions, interoperability protocols, associated software, and documentation of various forms. Defining, maintaining, and supporting that ecosystem is challenging, as it requires international agreements, e.g. through the International Astronomical Union, Commission 5, or through the Virtual Observatory initiative.

A systematic process that maps the astronomic research cycle into concrete work- and data-flow processes has been implemented at ESO. It ensures that calibrated data products can be generated through data pipelines for the most actively used modes of operational instruments. The data products consist of data with instrument signatures removed and calibrated into physical units, while instrument pipelines are a product of thorough scientific analysis. Data stewardship is spearheaded by ESO, in collaboration with instrument consortia and external community experts.

Consistent data verification and curation is warranted by operating a dedicated process (aka "ESO Phase 3" [31]). The vigorous application of science data product standards are critical to ensure homogeneity and coherence across a wide range of different data stemming from vastly different instruments and modes. They are continuously evolved to meet the ever increasing complexity of new instruments, and their corresponding data formats, and processing histories.

The ESO Archive Facility is the interactive and programmatic access point that serves the large volume and high quality of ESO data. By making data freely and easily available to the community at large, the ESO Archive fosters

the use of data at various processing levels by scientists not involved with the original proposal for their novel and independent science goals."Archive enabled papers" (see the ESO bibliography [32]) contribute to roughly 1/3 of the yearly output of ESO in terms of refereed papers—an impressive demonstration of the scientific value of a high quality archive.

The Virtual Observatory VO

The Virtual Observatory (VO) is a global data infrastructure held together by standards on how to find data collections and services (primarily through our Registry), work with them (using protocols either tailored to specific product types like spectra or images or general-purpose protocols typically transporting SQL-like queries), and ensuring they can be worked with by defining a set of interoperability conventions on several levels. More than 50 data centers adhere to these standards, offering about 30000 individual resources publishing 100s of millions of data sets and dozens of billions of table rows for use from clients like TOPCAT [33] or Aladin [34], libraries like pyVO, or web pages like ESAsky. There are no user-visible central components. Data providers can run their own publication infrastructure (like ESA does) or have their data published by project-independent data centers (like GAVO's, the CDS, or CADC). More information can be found at the IVOA [16] and worked-out examples of VO workflows can be found at GAVO [35].

The KASCADE Cosmic-Ray Data Center KCDC

KCDC, the "KASCADE Cosmic-ray Data Centre" [36], is a web-based interface where initially the scientific data from the completed air-shower experiment KASCADE-Grande was made available for the astroparticle community as well as for the interested public. Over the past seven years, the collaboration continuously extended the data shop[2] with various releases and increased both the number of detector components from the KASCADE-Grande experiment and the data sets as well as corresponding simulations. With the latest releases we added a new and independent data shop for a specific KASCADE-Grande event selection and by that created the technology for integrating further data shops and data of other experiments, like the data of the air-shower experiment MAKET-ANI in Armenia. In addition, we made available educational examples how to use the data, more than 100 cosmic ray energy spectra from various experiments, and recently attached a public server with access to Jupyter notebooks.

[2] The KCDC data shops are collections of data with specified formats and properties from various detector arrays.

Within PUNCH4NFDI, KCDC aims for an integration into the PUNCH Science Data Platform. Doing so, KCDC will benefit from the community overarching synergetic developments for a coherent data and metadata description. In addition, KCDC can be the test base for a coherent concept for data storage and access as well as for a widely used Authentication and Authorisation Infrastructure (AAI) in order to enable a PUNCH-wide multi-experimental and multi-messenger analysis platform.

Data and Analysis Reinterpretation at CERN

The data collected by the LHC experiments represent a unique tool to probe theories Beyond the Standard Model (BSM). While open data are useful to probe so-far untested theories outside of the original collaborations, the process is very resource-intensive. An alternative route to openly test new theories is through reinterpretations of existing analyses. Here, existing phase-space definitions and background estimates can be reused and only a new signal contribution must be estimated using an alternative physics model and processing it through the analysis pipeline. Building on recent progress in analysis preservation utilizing software packaging using Linux Containers and the implementation of data analysis pipelines using workflow languages such as yadage [37] or Common Workflow Language, a new reinterpretation portal RECAST [38] is intended to be established at CERN, which allows external researchers to request new theories to be tested. The procedure has been tested within the ATLAS collaboration for a number of initial analyses and is ready to be extended to a public interface. Besides the analysis pipeline the existing background estimates, including systematic and statistical uncertainty estimation, must be archived, ideally in the form of an existing likelihood model. This has been achieved for a large class of models at the LHC through the pyhf [39, 40] package and an associated JSON format, which is used to publish full-fidelity statistical models to the common and open HEP data repository hepdata.net [12, 41].

Smaller Hadron and Nuclear Physics Communities

The smaller communities of hadron and nuclear physics cover a diverse range of topics and produce data of very heterogeneous types. Although the size of research data from experiment and theory is considerably smaller than in bigger communities, their volume is expected to increase significantly. Raw data remain usually at individual institutes and are only partly maintained for long times. The data management relies on available equipment, tools and limited financial/human resources. In particular, there is lack of data specialists. Processed data are stored locally, published in journals or contribute to specific repositories

operated by individuals, universities or special institutions. However, they often do not follow FAIR principles, have very specific modes of public access, if at all, and are often not accompanied by formalized metadata. However, funding agencies increasingly require to establish dedicated data management policies in projects recently. This will boost the process of preparing research data for open access but has to be supported by reasonable and realizable methods and practices for data management. The education of people on all levels and the development of unified tools and standards is required to meet the challenges.

High-Energy Astroparticle Physics

Data providers in high-energy astrophysics range from observatories operated by international organisations like CTAO, ESO, or NASA, international collaborations (e.g., HESS, MAGIC, or VERITAS) operating high-energy experiments, to small-scale instruments operated by institutional groups. The Fermi gamma-ray data and software are publicly accessible [42]. The CTA (Cherenkov Telescope Array) observatories will provide show-cases for FAIR and open data management: high-level data products are available in less then a day (respectively after a proprietary period for CTAO) to the community; software and guest observer support are provided as a service. Access to the large data archives of currently operating gamma-ray observatories is more difficult, as data, data model, and software are mostly behind closed doors. However, an ongoing community effort aims for the definition of common high-level data formats [43] and public software tools, providing a realistic scenario for accessible and open archives for the future. Demonstrator projects [44] show that this allows additionally to perform multi-instruments and multi-wavelength type of analysis in clearly outlined and reproducible workflows. This interoperability is essential for the exchange of transient alerts, which are by definition public. Transient factories provide alert rates, which increasingly require transparent and complex real-time decision making workflows for optimal usage of both the observatories and the archival data. Challenges outlined above are shared in many cases with other members of the PUNCH4NFDI community; the NFDI activities are expected to provide a significant push towards the FAIR treatment of data, workflows and software products from high-energy observatories.

The International Lattice Data Grid

The international lattice data grid (ILDG) represents a grid of inter-operable data-grids for community-wide sharing of so-called gauge field configurations [45–50]. These are the raw data objects in Lattice QCD and are generated by Markov Chain Monte Carlo simulations utilising significant fractions of the most powerful HPC systems. The Lattice QCD workflow is based on these configurations, and includes further measurement and analysis steps which are not yet covered by the ILDG. The ILDG consists of a virtual organisation, working groups and a governance board. They define standards for (meta-)data formats and the services, like a file and a metadata catalogue, operated by each regional grid and publicly accessible. The ILDG architecture and metadata schema already implement FAIR principles. In the context of PUNCH4NFDI the community envisages to include additional data types, to design and converge to a PUNCH-wide authentication service, and to modernise and extend its web services, like the metadata catalog, for usage in other scientific communities.

LOFAR and MeerKAT

The world leading instruments of radio astronomy are interferometers, networks of hundreds of antennas or antenna stations, optimised for different frequency bands covering the radio window of Earth's ionosphere and atmosphere (10 MHz to 1 THz). The science questions addressed by radio astronomy range across the history of the Universe from the very first stars and black holes to the cradle of life, across size scales from exoplanets to the largest cosmic structures, across duration scales from fast radio burst to nHz gravitational waves, and from the the laws of gravity to cosmic magnetism and many more.

There are several aspects to take into account when talking about open data in the context of radio astronomy: (i) resource allocation; (ii) data policy; (iii) dissemination. Below we give a brief description of the status quo of open data aspects for LOFAR and MeerKAT, two projects that are supported by the BMBF via ErUM-Pro.

At the lowest radio frequencies the International LOFAR Telescope (ILT), carried by 9 national consortia with its base in the Netherlands and a strong German contribution, operates since a decade and has generated over 50 PB of archived data (as of early 2021), which are accessible via the LOFAR Long Term Archive (lta.lofar.eu). The LTA is a federated archive with sites at SurfSARA, FZ Jülich, and Poznan. Most of LOFAR data are kept on tapes and the processing demands of raw LOFAR data require close access to HPC.

LOFAR resources are allocated via cycle proposals (programmes that run within a single semester) and long term proposals (over four cycles), open to everybody in the world. National consortia have a reserved share of access, depending on their financial contribution to LOFAR. Proposals include requests for observing time, computing time, data volume, and user support. The LOFAR Programme Committee (PC) ranks proposals and allocates resources together with the observatory, taking into account the reserved shares of the national consortia. The ILT board endorses the PC

decisions. This procedure is the first element to guarantee open access to LOFAR data.

The ILT knows two categories of data, ILT data and station owner data. All ILT data become public eventually, all ILT observing and standard pipeline processing data remain ILT property. Station owner data (at least 10% of the available observing time) remain at discretion of station owners. Derived data are property of scientists, e.g. LOFAR survey catalogues are released by Key Science Projects. There is a proprietary data period for the proposer, which is 1 year. The PC can change data rights and scope for each ILT project. All ILT data are stored in the LOFAR Long Term Archive.

So far LOFAR gave rise to about 1800 publications that mention LOFAR in the abstract (according to ADS in February 2021). Most of them are on extragalactic astrophysics or pulsars and transients, but solar physics, cosmic ray physics and galactic physics as well as instrumentation and methods are important topics. Besides dissemination via the classic channel of scientific publications, LOFAR surveys are used for citizen science projects like e.g. Radio Galaxy Zoo (LOFAR) [51].

At mid-frequencies, MeerKAT, owned by the South African Radio Observatory and enhanced with contributions from the MPG, is operating in a rather different way. MeerKAT is the mid-frequency precursor of the Square Kilometre Array, which will have its core on the same site in South Africa. Most of MeerKAT access is reserved for Large Scale Projects (LSP). Neither access nor data are open (limited resources). It is the duty of the LSPs to make data products eventually open. For some science programs performed at MeerKAT, i.e. the search for pulsars and transients (TRAPUM and MeerTRAP) the data rates are so high (approx. 2 PB/day) that storing or shipping of the raw data is impossible. Instead, the data are analysed (quasi-) online and only the reduction products (candidates with snippets of data) are stored and transferred for further processing. MeerKAT offers a small fraction of open time, which is open to the world-wide community (so far two calls). Also it does not have an archive like the LOFAR LTA, access to limited archival data can be requested. Data are moved to a compute centre in Cape Town, from where the projects have to retrieve it within a certain time frame. The date volumes are huge, especially if low-level data are required, shipping discs is often still the best option for this purpose.

The recently founded Square Kilometre Array (SKA) Observatory will pose new technical challenges that have their own implications on how we can or have to think about open data. There will be the issue of data irreversibility (online reduction in near-real time) since raw data are of the order of 1 Pb/s and the amount of archived data will reach 300 PB/a. Individual data cubes will be of the order of up to a PB. The access of users to the data is currently under investigation by the SKA Regional Centre Steering Committee. Open Data in the classical sense of raw data will be technically impossible and thus Open Data must be restricted to higher level data products.

To summarise the status quo: Open data are the aim, but currently not the world wide standard in radio astronomy. Openness can be in conflict with resource allocation, granting rules, or missions of funding organisations. Commitment is typically rewarded by privileged access to data and various levels of openness are realised. Open Data does not automatically imply FAIR data since access and reproducibility[3] are challenging due to enormous data rates and volumes, and the problem of data irreversibility.

Discussion and Questions from the Workshop

In this section, we reflect on the main questions and comments received during the discussions at the workshop. In doing so, we do not aim at a full documentation of all answers and the full discussions, but merely try to summarise the main insights.

Structure of Data Publication Methods and Stages of Access

A number of questions and parts of the discussion focused on the different structures of data publication methods (archives, providers, sites) in the various PUNCH fields of research—questions of archival structure, storage sites, etc.

As examples of data providers in high energy physics, CERN and KCDC were discussed and contrasted with the approach in astrophysics. The KCDC "data shop" is an example of an infrastructure offering astroparticle data in the low Terabyte regime to users, with monitored access; the data are free for download after certain selection cuts have been specified by the user, to reduce the necessary download volume. The CERN open data portal offers data from CERN experiments on a much larger scale, in the Petabyte range, including both data and the relevant Monte Carlo samples. Also here, a simple selection is possible (e.g. data set, experiment, ...). In general, at least in high-energy physics, more and more data sets are maintained (and partly made openly available) in what might be called "archive mode"—the collaboration-specific state after operation and funding of the experiment has stopped, but in which data and software are still made available for analysis, either only for the collaboration or also beyond.

[3] The requirement R1.2. of the FAIR principles state: "(Meta)data are associated with detailed provenance" [5]. Detailed provenance enables and implies reproducibility.

It was noted in the discussion that archival and publication of data come with a price tag, and that it would be interesting to learn, for various activities, which fraction of the data taking cost is spent for archiving and publication, and how many results are derived from long-term archived data, especially if these are open data. For example, the ZEUS experiment at the HERA electron–proton collider stopped operation in 2007. Since then, considerable effort has been invested to make data and software usable for the long term, and currently about 10% of the ZEUS collaboration publications are done with these long-term archive data formats after the official funding for the experiment stopped in 2015. More experience needs to be collected.

A part of the discussion centered on the availability of astrophysics/astroparticle physics data like those from Fermi-LAT or CTA. The PUNCH science data platform aims at providing means to work on combined data from these data collections. This cannot be achieved by import of the data into the platform or its storage elements. Alternative paths are either implementing the access to the data via virtual organisations[4] or via Virtual Observatory protocols.

In all the open data initiatives, the long-term focus is clearly on providing openly accessible (and interoperable) data for the general public. However, it is also noted that for the time being, the main beneficiaries of open data efforts might be collaborators or other scientists from the same community. It is also clear that changing data access policies across the various PUNCH disciplines towards more open and publicly usable data collections is to a large extent a task for years to come. There is often a justified embargo period involved, and implementing transitions between closed and public access is a technical challenge as well (e.g. the SKA embargo policy for new data and its technical implementation [52]).

Curation Procedures, Provenance, Raw and Pipelined Data

An important point is to define which data are still in the stage of being pre-processed and what data can be released, e.g. in repositories or in the context of publications. In astronomy, data can be released after the instrument signatures have been removed—the pictures need to be clean of instrument artefacts. These data is often referred to as "level-1" data. A question is also how observation proposals and data results should or can be connected, since FAIR data put also a strong demand on provenance information. ESO facilities, for example, run surveys as well as PI mode observations, where observation time is awarded in a highly formalised way based on proposals[5].

In high-energy physics, data are in general confronted with theoretical predictions, e.g. in the form of Monte Carlo simulations. To facilitate meaningful work with openly released data, the corresponding Monte Carlo samples also need to be released. Various abstraction levels of data offer themselves for release, depending on the specific physics use case[6].

Publication in astronomy journals typically require the relevant data to be uploaded to archives, preferably with Virtual Observatory (VO) standards. It has been realised that data are used more productively if they are easily accessible via the VO developed protocols. The publishing of data collections is only achievable through distributed resource providers, the collaborations or participating institutes. This cannot be done by the participants of the IVOA [16] working on the standards. Implementation of the VO standards makes the data queryable across different data collections. Currently, about 50 data centres are connected to the VO, but still many data sets and data sources are not connected.

For the quality control of published data, the current refereeing process of the respective journals is not sufficient. Although most public data collections used by astronomers are already attached to published papers, the curation of the data sets is not yet handled sufficiently strict.

To a large extent, quality control and data curation are also the responsibility of the observatory, for example, in the case of ESO, the rationale being that the use of expensive infrastructures with low-quality data is a waste of resources. The processes for quality control and data curation is still evolving.

Combining Data Sets and Integrating Experiments—The Interoperability Question

An important aspect of research in the PUNCH disciplines is the combination of data from different experiments and observatories. Such combinations facilitate entirely new questions or even research directions. They are, however, technically challenging, and difficult questions of data access and compatibility (data formats etc.) need to be addressed.

In high-energy physics, platforms like HEPData offer easy access to a multitude of experimental data at high abstraction level, with links to the relevant publication. But

[4] Virtual organisations are the organisational form of collaborative access to the data used chiefly in High-Energy Physics.

[5] Survey mode: a predefined catalog of targets is observed by the operators of the facility. PI mode: a Principal Investigator (PI) is leading the observation

[6] Levels of data here refer to the fact that processing raw data to data for publication has different intermediate stages, often referred to as levels, that are relevant for re-analysis and combination with data from other experiments or communities.

these high abstraction levels are not conducive to combining data sets (or selections) from experiments.

However, the combination of different data sets has recently been advanced a lot. Typically, re-analysis is performed by a team different from the original one, by combining with data sets from other collaborations, or by appending theory interpretations later. Combining systematic uncertainties from various measurements is also a crucial technical problem and the REANA project [53] is an effort to address this. The theory interpretations and the necessary theory input are made available in a unified format in which theorists make their predictions available for the experimentalists. REANA uses a couple of workflow languages that were partly developed by the HEP community, and also includes an effort to look for more common workflow languages. REANA is trying to support multiple languages, e.g. from the life sciences.

In astroparticle physics, the multi-messenger approach to correlate gravitational wave events with other observations is an example of data combination. Reconstruction is, however, still done individually by the participating collaborations based on real-time alert systems. However, KCDC is in its beta-version to provide data from different experiments which can be selected and combined on the same open platform.

During the workshop aspects of the interoperability of metadata semantics in lattice QCD (quantum chromo dynamics) were discussed, the usability of lattice QCD workflows and procedures outside of of ILDG (International Lattice Data Grid), the semantic enhancement of the data of high-energy physics, and the necessity of machine-readable/actionable metadata in general.

A question that is of concern for all PUNCH disciplines is how to integrate experiments that have not yet started their own open data initiative or are simply too small to make resources available for this into developments of PUNCH-4NFDI. No simple solution has so far been put forward.

Licensing, DOIs, and Data Management Plans

A final block of the discussion centered around the topics of licensing, DOIs (digital object identifiers), and data management plans, and in particular around the question whether a coherent approach in using DOIs and licences for open data will be possible, i.e. whether the way that high-energy physics typically deals with the subject will be accepted also by, e.g. the providers of the Virtual Observatories in astronomy?

As an example, KCDC was mentioned again: in KCDC, the user license is provided by a custom-made EULA. If KCDC will be integrated in a larger frame (e.g. by PUNCH-4NFDI), one has to consider the varying community usances (culture) and legal aspects, while continuing to ensure good scientific practice.

It was also pointed out that data management tools need to enter the thinking of all activities in the PUNCH disciplines, also for smaller collaborations. In fact, funding agencies more and more require a data management plan, also providing the necessary funds to implement the plans.

Summary—Open Data in PUNCH Sciences

The workshop, its presentations and the corresponding discussions are a good starting point and will serve as input to the activities and measures of the PUNCH4NFDI consortium towards a coherent PUNCH open data policy and the PUNCH Science Data Platform. We have seen very different approaches and now need to discuss and agree on levels of preparation of data we will make accessible as open data in the PUNCH sciences. The data must be prepared, appropriate metadata need to be provided, and a coherent data curation procedure must be established to be scientifically beneficial.

To view the current situation in a more general way, we need to discuss the developments also under the angle of Open Science. Here, everyone tries to fit their achievements and ideas, partly developed with a long history, into the new concept. This does not make the task any easier.

There are two different development paths: the first originates from the societal demand of free and unhindered access to data from publicly funded institutions and the data and knowledge sharing stance of the scientific communities. Open Data have started with a very strong push towards opening up all the data of governmental institutions accessible to the citizens by digital, modern means. Processes of government institutions should be brought into the digital ages, and allow for efficient interactions with citizens and society. This development is driven by the Open Knowledge Foundation and other similar civil organisation demanding the opening up of governmental data collections for public access in suitable digital formats. One of the most recent outcomes—in Germany—is the 'Guidance for Open Data' [54] and the 'Musterdatenkatalog fuer Kommunen' [55].

The other development path, the FAIR data requirements, is in fact even older, and originates in the scientific communities' demand of having data exchangeable and with sufficient metadata descriptions (geophysics in 1957 and NASA in 1970, with shared data policies for satellite data). This scientific demand clearly has contributed to the various approaches to Open Data as presented above. The efforts unfolded within the various disciplines, but only the advent of the global digital network (aka internet) made the idea of cross-disciplinary use of digital data in a new,

multidimensional space feasible. At this development stage of the new realm different tasks have to be addressed: Making data FAIR is chiefly a science driven intent to maximise the benefit of digital data for the progress of scientific knowledge. Its requirements open up and reorganise the scientific processes, align them with the advanced digital methods and possibilities.

Going from FAIR data to Open Data is addressing societal processes as well, since the current scientific systems are intertwined with their respective societal structures. As the different examples above show, publicly accessible data are one aspect, which also requires many changes to transform the paradigm of the "single scientist and his data", originating from the end of the 19th century, into a current one, where teamwork, division of scientific labour and public funded research on a digital platform play the most significant role, particularly in PUNCH sciences. For Open Data, we need to develop improved policies and agreements for data sharing across the disciplines. We need to give more scientific credit to the collaborative character of the process. This has started with, e.g. creating platforms like arxiv, but needs complementary development of review/evaluation procedures of scientific results, since the current ones were formed more than a century ago, on a completely different platform.

Open Science also emphasises the importance of the software. One could even think the Open Source software and their collaborative development approach as the first and incredibly fruitful example for freeing the results of scientific work from the confinements of narrow corporate interests. The combination of Open Data and Open Source are the building blocks of Open Science.

As a further step, the role of science as an essential factor for the economic and societal development finally is addressed within the frame of Open Science. Since the power of digital science is very unevenly distributed across the world, there are many further tasks to solve. This is only pointed out here, since many more actors have to be included in this discussion in the future.

For the purpose and scope of starting a thorough assessment of the situation for building up FAIR data infrastructures for the physics disciplines and with NFDI, the workshop gave a first overview and material for further discussion and research. Some topics need immediate follow-up: a more detailed review of data publication policies, a fitting metadata scheme, a collection and refinement of data curation procedures from our disciplines, and also the implementation of data provenance. These are, therefore, the main tasks and direct items for the activities within the PUNCH4NFDI consortium.

Acknowledgements The authors would like to thank Hannah Elfner (GSI and Frankfurt) and Jan Mayer (Cologne) for their input to section 3.7 and Carsten Urbach (Bonn) for his input to section 3.9. The authors also thank all the participants of the workshop for their contributions to the discussion. This work was performed in the framework of the PUNCH4NFDI consortium which will be supported by DFG fund "NFDI 39/1", Germany.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

1. PUNCH4NFDI (2019) http://www.punch4nfdi.de
2. NFDI (2020) http://www.nfdi.de
3. PUNCH4NFDI (2020) PUNCH4NFDI proposal, https://www.punch4nfdi.de
4. Wikipedia (2022) Open data. https://en.wikipedia.org/wiki/Open_data
5. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone SA, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop A, Waagmeester Janand, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3
6. Open-Knowledge-Foundation. Open Knowledge Foundation Mission Statement, https://okfn.org/about/ (2021)
7. European-Parliament (2019) Directive of the European Parliament and of the Council on open data and the re-use of public sector information , https://op.europa.eu/en/publication-detail/-/publication/a75e8d18-77dd-11e9-9f05-01aa75ed71a1
8. Bundestag (2017) Federal Open Data Act, https://data.europa.eu/en/news/new-open-data-act-germany
9. European-Commission (2016) Legal study on ownership and access to data , https://op.europa.eu/s/slOy
10. SDSS collaboration (2022) The sloan digital sky survey: mapping the universe. https://www.sdss.org/
11. The Gaia Mission (2022) The Gaia mission. https://www.cosmos.esa.int/web/gaia/home
12. HEPData (2021) http://www.hepdata.net
13. NASA/ADS (2021) Astrophysics data system, https://ui.adsabs.harvard.edu/
14. Gammapy Project (2017) Gammapy—a Python package for for gamma-ray astronomy. https://gammapy.org/about.html
15. IAU FITS Working Group (2014) A brief introduction to FITS. https://fits.gsfc.nasa.gov/fits_overview.html

16. International Virtual Observatory Alliance (2022) About IVOA. https://ivoa.net/about
17. Open Archives Initiative Organization (2001) The open archives initiative protocol for metadata harvesting. https://openarchives.org/OAI/openarchivesprotocol.html
18. Astropy-Project (2021) The Astropy Project , https://www.astropy.org/
19. IVOA-Authors (2021) PyVO, https://pyvo.readthedocs.io/en/latest/
20. Kraft S et al (2021) Aufbau und Ziele von Nationale Forschungsdateninfrastruktur (NFDI) e.V., Bausteine Forschungsdatenmanagement Nr. 2, https://doi.org/10.17192/bfdm.2021.2.8332
21. CERNpolicy (2020) CERN Open Data policy, http://opendata.cern.ch/docs/cern-open-data-policy-for-lhc-experiments
22. CERNOD (2014) CERN Open Data platform, http://opendata.cern.ch
23. LEP (2021) Archived LEP data, https://dphep.web.cern.ch/experiment/alephhttps://dphep.web.cern.ch/experiment/delphi, https://dphep.web.cern.ch/experiment/opal
24. JADE (2021) Archived JADE data, https://wwwjade.mpp.mpg.de/
25. Bethke S (2010) Data preservation in high energy physics—why, how and when? Nucl Phys B Proc Suppl 207–208:156. https://doi.org/10.1016/j.nuclphysbps.2010.10.040
26. Bacchetta A et al (2016) Future physics with HERA data for current and planned experiments
27. Geiser A (2020) Preserved HERA data and EIC. https://indico.bnl.gov/event/9287/contributions/4145/attachments/30600/48033/EIC_2020.pdf
28. CMS-Collaboration (2018) CMS data preservation, re-use and open access policy. CERN Open Data Portal. https://doi.org/10.7483/OPENDATA.CMS.7347.JDWH,http://opendata.cern.ch/record/414
29. Retzlaff J, Arnaboldi M, Delmotte N, Farina E, Mascetti L, Micol A (2021) Implementing FAIR guiding principles in the ESO science data archive. https://doi.org/10.17192/bfdm.2021.2. https://doi.org/10.5281/zenodo.5060238
30. ESO (2021) ESO Science Data Archive. http://archive.eso.org
31. ESO (2021) ESO preparation, validation and ingestion of science data products (SDPs). http://www.eso.org/sci/observing/phase3.html
32. ESO-library (2021) ESO telescope bibliography. http://telbib.eso.org
33. Taylor MB (2005) Astronomical Data Analysis Software and Systems XIV, Astronomical Society of the Pacific Conference Series, vol 347, ed. by P. Shopbell, M. Britton, R. Ebert (2005), Astronomical Society of the Pacific Conference Series, vol. 347, p 29
34. Observatoire-Strasbourg (2021) Centre de Données astronomiques de Strasbourg, https://cds.u-strasbg.fr/
35. GAVO. VOTT: Virtual Observatory Text Treasures. http://dc.g-vo.org/VOTT
36. Haungs A et al (2018) The KASCADE cosmic-ray data centre KCDC: granting open access to astroparticle physics research data. Eur Phys J C 78(9):741. https://doi.org/10.1140/epjc/s10052-018-6221-2
37. Heinrich L (2021) Yadage—declarative workflow spec and engine. https://yadage.readthedocs.io/en/latest/
38. Cranmer K (2021) RECAST & CERN analysis preservation. https://indico.cern.ch/event/525142/contributions/2190137/attachments/1291681/1924055/RECAST-CAP-Reinterpretation-Workshop.pdf
39. Lukas H, Feickert M, Stark G. pyhf: v0.6.3. https://doi.org/10.5281/zenodo.1169739. https://github.com/scikit-hep/pyhf/releases/tag/v0.6.3
40. Heinrich L, Feickert M, Stark G, Cranmer K (2021) pyhf: pure-python implementation of histfactory statistical models. J. Open Sour Softw 6(58):2823. https://doi.org/10.21105/joss.02823
41. Maguire E, Heinrich L, Watt G (2017) HEPData: a repository for high energy physics data. J Phys Conf Ser 898(10):102006. https://doi.org/10.1088/1742-6596/898/10/102006
42. NASA. Fermi data policy. https://fermi.gsfc.nasa.gov/ssc/data/policy/summary.html
43. Nigro C, Hassan T (2021) Standardisation of data formats in gamma-ray astronomy. arXiv e-prints arXiv:2101.06018
44. Mohrmann L, Specovius A, Tiziani D, Funk S, Malyshev D, Nakashima K, van Eldik C (2019) Validation of open-source science tools and background model construction in γ-ray astronomy. Astron Astrophys 632:A72. https://doi.org/10.1051/0004-6361/201936452
45. Beckett MG, Joo B, Maynard CM, Pleiter D, Tatebe O, Yoshie T (2011) Building the international lattice data grid. Comput Phys Commun 182:1208. https://doi.org/10.1016/j.cpc.2011.01.027
46. Maynard CM (2009) International lattice data grid: turn on, plug in,and download. PoS LAT2009, 020. https://doi.org/10.22323/1.091.0020
47. Yoshie T (2008) Making use of the International Lattice Data Grid, PoS LATTICE2008, 019. https://doi.org/10.22323/1.066.0019
48. Maynard CM, Pleiter D (2005) QCDml: first milestone for building an international lattice data grid. Nucl Phys B Proc Suppl 140:213. https://doi.org/10.1016/j.nuclphysbps.2004.11.116
49. Andronico G, Barbera R, Falzone A (2004) Grid portal-based data management for lattice QCD data. Nucl Instrum Methods A 534:76. https://doi.org/10.1016/j.nima.2004.07.062
50. Davies CTH, Irving AC, Kenway RD, Maynard CM (2003) International lattice data grid. Nucl Phys B Proc Suppl 119:225. https://doi.org/10.1016/S0920-5632(03)01509-3
51. LOFAR Collaboration (2022) Radio galaxy zoo (LOFAR). http://lofargalaxyzoo.nl/
52. SKA-Consortium (2021) 16231-Factsheets-operational-model-v4.pdf. https://www.skatelescope.org/wp-content/uploads/2018/08/16231-Factsheets-operational-model-v4.pdf
53. REANA team (2022) Reproducible research data analysis platform. https://reana.io/
54. Bertelsmann-Stiftung (2020) Leitfaden offene Daten. https://www.bertelsmann-stiftung.de/de/publikationen/publikation/did/ein-leitfaden-fuer-offene-daten
55. Bertelsmann-Stiftung (2020) Musterdatenkatalog. https://www.bertelsmann-stiftung.de/de/unsere-projekte/smart-country/musterdatenkatalog