NECESSITY AND OPPORTUNITY IN MANAGEMENT OF NUCLEAR PHYSICS DATA

PÉTER LÉVAI*, DÉNES LAJOS NAGY, AND EDIT SZILÁGYI
Wigner Research Centre for Physics,
29-33 Konkoly-Thege Miklós Street, Budapest, 1121 Hungary
**E-mail: levai.peter@wigner.hu*

Abstract: We describe recent challenges in management of scientific data, especially in nuclear science. During last years, common effort has been started in Europe to optimize the existing data banks and to create data repository for easy access. We display the Hungarian activities and discuss the influence of such repositories in scientific research.

Keywords: *Nuclear physics, experimental data, data bank, data repository, FAIR, EOSC.*

1. INTRODUCTION

The fundamental feature of nuclear physics and related subfields of nuclear science and technology is the wide range and large size data collection, and the application of problem-specific phenomenological models to investigate the created huge amount of data. The researchers analyze these data sets in terms of their own (semi-) theoretical models and draw conclusions influencing multibillion-dollar projects in the energy sector, ranging from energy production to the management of the electric network and the public/industrial consumption. Usual expectation is the reliability and the repeatability of the conclusions witnessed by the trustfulness of the source (to be an expert person or an expert company) in most of the cases.

Nowadays data storage is relatively cheap and widely used data analysis methods (of open source) are available so that the demand of transparency of scientific studies can be fulfilled in a more general manner. However, to maintain this transparency we need trustful data bases with long-term availability. This request meets the general expectation of funding agencies to store the collected scientific data and to make them public if the data collection was supported by public money. Furthermore, the existence of trustful and open databases can successfully decrease the expenses of scientific investigations thereby avoiding the multiple repetition of data collection in equivalent situations.

This problem is known in many fields of science including particle physics, statistical physics, biology, chemistry or medical sciences. The widely applied solution is creating data banks where the available data are stored by the owner (expert) who knows precisely the meaning and the structure of the data sets. This way we may avoid data loss. However, the abundance of the data and the complexity of different research fields may result in a far-from-optimal situation when exabyte of data is stored in the data banks, however the possibility of understanding these data has been lost lacking the missing expert who originally created and stored those data but, meanwhile, is no longer available. Recently, a justified request has appeared from the funding agencies to create data bases of long standing and independent from the original creators clearly displaying the scientific information. This request was manifested in the foundation of data repositories with open access and in-built transparency.

The creation of these data repositories led to recognizing and determining the value of stored data. First of all, the cost of data creation can be estimated after defining the quality and the deepness of the data sets (CapEx cost). Furthermore, the cost of data storage including the technical circumstances and the human contribution to the continuous data management (OpEx) can be clearly determined. If we can estimate the expected scientific and industrial benefit of future applications of the stored data then we can estimate the value of these data and recognize the appearance of a data treasure. This recognition changed the attitude of many funding agencies and modified the structure of these data repositories enhancing user-friendly features and wide-scale availability.

2. Recent projects in Europe and in Hungary on management of scientific data

2. 1. European achievements in data collection

In 2016, an international collective of authors under European guidance put forward the set of requirements for scientific research data FAIR that, meanwhile, has become a worldwide accepted standard. The abbreviation 'FAIR' stands for 'findable', 'accessible', 'interoperable' and 'reusable', respectively. 'Findable' means that metadata and data should be easy to find for both humans and computers. Data and metadata are 'Accessible' once one knows how they can be accessed, possibly including authentication and authorization. The requirement 'Interoperable' requires that, as a rule, data can be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing. Finally, the term 'Reusable' indicates the requirement that metadata and data should be well-described so that they can be replicated and/or combined in different settings.

The European Open Science Cloud (EOSC) is an environment for hosting and processing research data to support EU science. The ambition of EOSC is to provide European researchers, innovators, companies and citizens with a federated and open multi-disciplinary environment where they can publish, find and re-use data, tools and services for research, innovation and educational purposes. Accordingly, EOSC is not and will never be a pan-European research infrastructure akin to projects and landmarks listed and regularly updated in the Roadmap of the European Strategy Forum on Research Infrastructures (ESFRI)

EOSC ultimately aims to develop a Web of FAIR Data and services for science in Europe upon which a wide range of value-added services can be built. These range from visualization and analytics to long-term information preservation or the monitoring of the uptake of open science practices.

EOSC is recognized by the Council of the European Union as the pilot action to deepen the new European Research Area. It is also recognized as the science, research and innovation data space, which will be fully articulated with the other sectoral data spaces defined in the European strategy for data.

In July 2020, an EOSC Association was set up to provide a single voice for advocacy and represent the broader EOSC stakeholder community. This association became operational in 2021 and is rapidly expanding its membership.

2. 2. Hungarian project: ARP

In Hungary, the Eötvös Loránd Research Network (ELKH) also recognized the need of a digital infrastructure required for the construction of the data repository and set up the project ELKH Data Repository Platform (ELKH ARP). Within the framework of this project, three research centers of ELKH, viz. the Institute for Computer Science and Automation (SZTAKI), the Centre for Social Sciences (TK) and the Wigner Research Centre for Physics (Wigner RCP) will implement an internationally cutting-edge data repository by the end of 2024. Such a system is expected to provide high availability, high data security, high bandwidth data connectivity, long-term storage, the ability to accommodate huge data volumes and, last but not least, the ability to efficiently search across data.

The aim of ELKH ARP is to create a Hungarian research data repository that will comply with the FAIR storage and management principles of research data, the long-term storage requirement of research datasets associated with publications and their sharing (open or closed) with the scientific community, between disciplines or even internationally, thus ensuring the possibility of future integration into an international research infrastructure.

The ELKH ARP system will support the introduction and widespread use of the Research Object tool that enables the new, FAIR-sensitive level of research data management. The use of internationally accepted metadata schemes and the development of new ones enable the national application of internationally spread discipline-specific metadata schemes in the repository and also the creation and maintenance of national, institutional and research-field-specific metadata scheme application profiles.

The internationally accepted metadata schemes ensure that the data description will be rich enough and suitable for reusing the data (even by the software used in a given research field) without the expert who originally created these data. Although nowadays only a few internationally accepted metadata schemes exist, their use is expected to spread across all fields in the near future. To develop such suitable internationally accepted metadata schemes for the various research fields, competent researchers will play an essential role. The storage of research data in this sort of repositories will offer also new possibilities for the researchers as discussed in the next section.

2. 3. Opportunities connected to the scientific data repositories

The necessity of creating long standing, user friendly and easily manageable data repositories is justified among the scientists, even if their creation and continuous management will generate extra burden personally and extra cost financially. The future generation of scientists will be requested to follow pre-determined protocols on their experimental research activity to create and store their experimental data. However, these protocols will ensure the later usability of the data without extra cost (decreasing the budget request of the original owner or other potential users). Even more, a well-structured data base may highlight missing data sets, which can ignite extra activities to improve the coverage of the repository – leading to building new experimental instruments if the existence of missing data has objective barrier.

We should not forget about the existing data stored in various data banks and in the hard drives of thousands of personal computers and laptops. Fortunately, in most of the cases the expert persons of these data are active and capable of converting their data into any pre-determined structure. Of course, this conversion and archiving is not only an intellectual challenge but also require extra time and energy. However, the prize is the acknowledged contribution to a long-lasting data treasure of mankind.

Furthermore, the arrangement of the existing data may lead to the demand of re-measurement, which opens the opportunity for bachelor and master students to create an excellent and scientifically recognized thesis. This activity gives a chance for rejuvenation, finding committed followers on the given field. The importance of this outcome cannot be emphasized enough, especially in the field of nuclear science and engineering.

From the beginning of the pan-European activity on data repositories, it was clear that the advertised aim of high-level scientific data archiving is beyond the capability of average scientists as well as of average librarians. A new profession has been born, namely the ‘data shepherd’, ‘data assistant’, or ‘data steward’ who has knowledge on the scientific fields as well as on the archiving activities, and is capable of communicating with those IT-experts who are operating the background computing resources. An interesting question of the future is how much activity can be assigned to a data shepherd to make the repository-related activities easier for scientists.

The challenge of the near-future is the inclusion of artificial intelligence into this field: if AI will appear already in the improvement phase or it will use the opportunity of a well-established, close-to-perfect data base in interdisciplinary research and related innovation activities just later. Very possibly a new generation of scientists will think about this opportunity very soon. The appearance of AI on data management is unavoidable.

3. CONCLUSION

Recent activities in data management have been reported. The establishment of complex, well-determined and long-lasting data repositories has a strong influence on scientific data collection and application in nuclear physics and related fields. The existence of trustable, user-friendly, precise and widely accessible data repositories may improve the R&D activities of scientists and engineers, may help the education and the training activities, and is able to optimize (in most of the cases to decrease) the expenses. European countries, including Hungary, are using cutting-edge information technologies to establish such a system to build a competitive advantage for European science. However, since science is a world-wide endeavor of the mankind, European (and Hungarian) organizations are open for collaboration on this intellectually challenging field.

4. ACKNOWLEDGMENTS:

The authors thank the Eötvös Loránd Research Network (ELKH) for the financial support of the ARP project, which made necessary and available the reported studies. One of the authors (P. Lévai) thanks the organizers of the 15th Vietnam Conference on Nuclear Science and Technology (VINANST-15) for their invitation and their financial support.

5. REFERENCES

1. FAIR Principles: https://www.go-fair.org/fair-principles/
2. EOSC WEB-page: https://eosc-portal.eu/
3. ELKH WEB-page: https://elkh.org/
4. ELKH ARP WEB-page: https://science-research-data.hu/
5. WIGNER RCP WEB-page: https://wigner.hu/en/news