Agris on-line Papers in Economics and Informatics

Volume XV Number 4, 2023

Agriculture Data Platform – Institutional Data Repository – Selected Aspects

Michal Stočes, Jiří Vaněk, Jan Jarolímek, Vojtěch Novák, Jan Masner, Pavel Šimek, Eva Kánská, Martin Havránek, Karel Kubata, Vladimir Voral

Department of Information Technologies, Faculty of Economics and Management, Czech University of Life Sciences Prague, Czech Republic

Abstract

This paper presents selected aspects of a data platform to store agricultural data. It analyses the key user and system requirements for the data platform. The presented aspects were identified through a literature review, interviews and discussions with selected data experts and researchers and future users of the platform. The following issues of the data platform are discussed in the paper: architecture, data types, data source types, metadata, disciplinary interfaces, data sharing, data reusability, Open Science, FAIR data principles, and further data processing options. Part of the knowledge from this article was used in the design and implementation of the Institutional Data Repository called Data Management Platform (DaMP.) CZU (Czech University of Life Sciences Prague) (OPENAI, 2023).

Keywords

Data models, data modelling, metadata, data sharing, open science, FAIR data principles, data platform, agriculture data, life sciences, institutional data repositories, data process, DOI, FAIR data.

Stočes, M., Vaněk, J., Jarolímek, J., Novák, V., Masner, J., Šimek, P., Kánská, E., Havránek, M., Kubata, K. and Voral, V. (2023) "Agriculture Data Platform – Institutional Data Repository – Selected Aspects", *AGRIS on-line Papers in Economics and Informatics*, Vol. 15, No. 4, pp. 127-133. ISSN 1804-1930. DOI 10.7160/aol.2023.150409.

Introduction

With the advent of new technologies, it is possible to share data easily and at a reasonable cost over the Internet around the world. The sharing of data in machine-readable formats was kick-started by the Open Data initiative (Stočes et al., 2018) (Marešová et al., 2019), which primarily aimed at publishing public sector data. The current trend is the concept of Open Science. The idea that scientific research should be free for all was popularised by Robert King Merton in the early 1940s. Data produced by research should be freely shared for the common good. (Merton et al., 1942) Open science has the potential to make the scientific process more transparent, inclusive and democratic and is increasingly recognised as a critical accelerator for achieving the Sustainable Development Goals and a true game changer in bridging gaps in science, technology and innovation and fulfilling the human right to science (Tzanova, 2020), (UNESCO, 2013).

The data is stored and made available through so-called data repositories. The current trend is to link national, thematic and other repositories using metadata catalogues so that data can be retrieved from one place (Thoegersen and Borlund, 2022). The issue of data sharing and repository creation, either at the national or scientific society level, is one of the European Commission's research and innovation strategies. This strategy is presented under the term European Open Science Cloud (EOSC) (Burgelman, 2021). In the Czech Republic, this issue is addressed by the project: the project for the creation of a modernised national large research e-infrastructure e-INFRA CZ, which is formed by a consortium of organisations CESNET, CERIT-SC and IT4Innovations.

In order to increase the reusability of scientific data, it is essential that the data is appropriately described with metadata and standardised. The FAIR Principles initiative addresses this issue. In 2016, "The FAIR Guiding Principles for Scientific Data Management and stewardship" was published in the journal Scientific Data (Wilkinson et al., 2016). The authors intended

to provide guidelines for improving the discoverability, accessibility, interoperability, and reuse of digital assets. The principles emphasise machine action (i.e., the ability of computational systems to find, access, interoperate, and reuse data with little or no human intervention) as people increasingly rely on computational support to work with data due to the increase in the volume complexity, and speed of data creation. This methodological toolkit presents how to publish data based on fifteen principles in four groups: Findability (to be Findable), Accessibility (to be Accessible), Interoperability (to be Interoperable) and Reusability (to be Reusable). The FAIR principles are also reflected by science and research funders such as Horizon Europe, TAČR and others (Prodan et al., 2022).

The advent of innovative techniques and technologies to acquire data in the field of biology has brought the problem of how to effectively handle the acquired data. Communities have started to emerge in different scientific disciplines to create standards, regulations and ontologies for data handling. An example is the MIAPPE community (Papoutsoglou et al., 2020), which addresses the issue of data standards designed to harmonise data from plant phenotyping experiments. Other initiatives are addressing the issue of how to share data efficiently further. Examples include The Breeding API (BrAPI) project (Selby et al., 2019), an ontology-driven information system designed for plant phenotyping PHIS or the open-source ISA framework.

ISA's open-source framework and tools help to manage an increasingly diverse set of biological, environmental and biomedical experiments that use a single technology or combination of technologies. Built on the "Investigation" (project context), "Study" (unit of research), and "Assay" (analytical measurements) data models and serialisations (tabular, JSON, and RDF), the ISA framework helps you provide detailed descriptions of experimental metadata (i.e., sample characteristics, technologies and measurement types, relationships between samples and data) so that the resulting data and discoveries are reproducible and reusable (Sansone et al., 2012).

In Europe, the non-governmental ELIXIR (David et al., 2020) is the umbrella for activities dealing with bioscience data. ELIXIR brings together biological data resources, which include databases, software tools, training materials, cloud storage and supercomputers. The ELIXIR structure aims to coordinate these resources to form a single infrastructure. This infrastructure makes it easier for scientists to find and share data, exchange expertise, and agree on best practices. Ultimately, it will help them to gain new insights into how living organisms work.

Materials and methods

Building data repositories in organisations working with data reflects the current state of development in data management. Correct data management without tools to store and catalogue data is almost impossible. This paper presents selected aspects of data platforms. The article's authors are actively involved in developing the CZU institutional repository called the Data Management Platform CZU (Czech University of Life Sciences) (DaMP.) In the following sections, selected aspects and requirements for a data platform used to store scientific and research data from the field of agriculture in general from the Life Sciences will be presented.

The procedure for the development of this paper was as follows. The selected aspects presented were developed based on a thorough literature search of current knowledge in the field of data storage and sharing. In the second step, extensive interviews were conducted with scientific lawyers in the field of data sciences. The results were then complemented by interviews with future platform users - life scientists.

The scientific community perceives the need to manage data in a way that enables its reuse. The problem is the theoretical requirements arising from FAIR principles on the one hand and theoretical models and standards based on disciplinary needs on the other. According to a survey (Godem et al., 2022), 46% of biologists do not know how to organise data to reuse it. The object of the project application is to fill the identified research and technology gap, using data from the field of agricultural research and rural development as an example, e.g. crop production, hydrology, economics, etc. Based on the findings, general conclusions valid for other scientific fields will be formulated.

Results and discussion

The current state of data management

In discussions with members of the different scientific teams, it became apparent that there is a notable diversity in the approaches each group employs for data storage. The variations in data storage methods are quite substantial, ranging from the conventional use of USB flash

drives to the utilisation of cloud drives and even the implementation of advanced and intricate storage systems. Each team has adopted a unique strategy tailored to their specific needs and preferences, showcasing the diverse landscape of data storage practices within the scientific community.

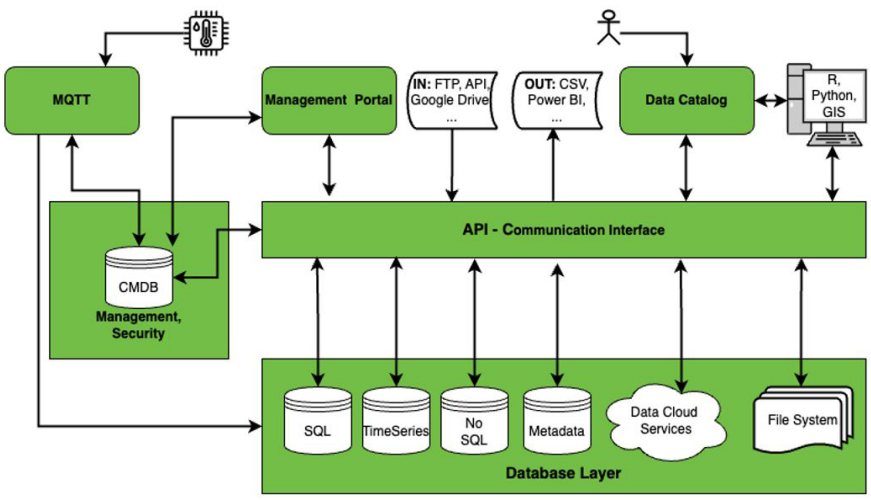
The following main issues were identified:

- Insufficient or non-existent backup procedures were identified as a prevalent issue during discussions with scientific team members. This vulnerability could potentially lead to data loss, highlighting the critical need for robust backup mechanisms.
- Another challenge raised was the inadequate description of data and the time-consuming nature of data retrieval processes. The lack of comprehensive data documentation could hinder efficient data management and retrieval, posing a significant obstacle to research progress.
- Regarding data protection, concerns were expressed about suboptimal measures in place, exposing valuable data to potential risks. Strengthening data protection protocols emerged as a priority to safeguard against unauthorised access or loss of sensitive information.
- Additionally, the difficulty encountered in sharing data with other scientific teams emerged as a noteworthy concern. The existing barriers to seamless data exchange underscored the importance of fostering improved collaboration and interoperability between research groups.

System architecture design

In the analysis of the system architecture, two key requirements were identified and defined: scalability and modularity. Scalability became a priority in view of future growth and increasing demands on the system. This requirement requires the system to scale efficiently and handle increased loads without loss of performance. The second important requirement is modularity, which emphasises a structured and hierarchical approach to system design. A modular architecture allows the separation of functionalities into separate and independent components, making it easier to maintain, extend and update the system. This feature also contributes to the system's overall flexibility, allowing adaptation to changes in requirements and technological environment. Meeting these two requirements together in architectural design is key to creating a robust and sustainable system capable of responding to current and future needs.

The main functional modules of the data platform were defined as data catalogues used for access to data. Database layer used to store heterogeneous data. Which are available through a single API (application programming interface) module. The management portal with the CMDB (Configuration management database) field controls data access through the catalogue and the interface, describing and configuring data sources and describing data collections with metadata. The MQQT broker is an entry point for messages sent by sensors and IoT devices. The modules and data flows of the proposed solution are presented in Figure 1.



Source: author

Figure 1: Data platform - modules and data flow.

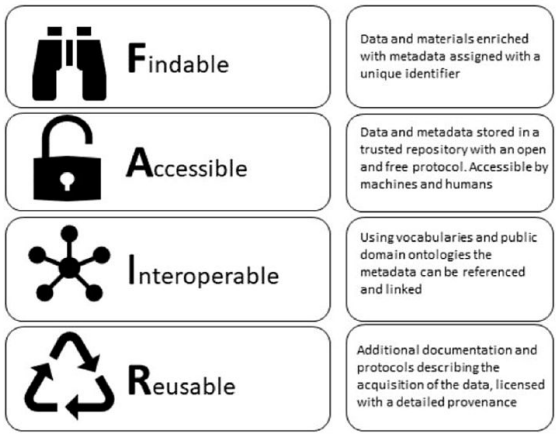
Data sharing

Based on a survey of users (data owners), a key need emerged to enable effective information sharing between different scientific teams. This requirement reflects the urgent interest in creating a mechanism to facilitate seamless data exchange between teams, thereby promoting mutual collaboration and synergy in research.

Given the diversity of scientific projects and the specialisms of the different teams, it proved crucial to ensure that the system allows flexible and secure sharing of data resources. In this way, scientific groups could enrich each other's knowledge and results, ultimately contributing to synergetic progress in scientific research.

Moreover, the emphasis on this need also reflects the importance of communication interfaces and standards that would allow interoperability between the different data systems used by the different teams. This would ensure that data sharing takes place efficiently and without loss of information value, although individual teams may work with various data formats and structures.

In response to the identified need for efficient data sharing between scientific teams, a solution was proposed that uses the principles of Findable, Accessible, Interoperable and Reusable (FAIR). This approach provides a framework for ensuring data findability, accessibility, interoperability, and reusability. Implementing FAIR principles ensures that data are systematically structured and described to make them easily locatable, accessible, and compatible with different systems. (Figure 2)



Source: Kalendralis at al.

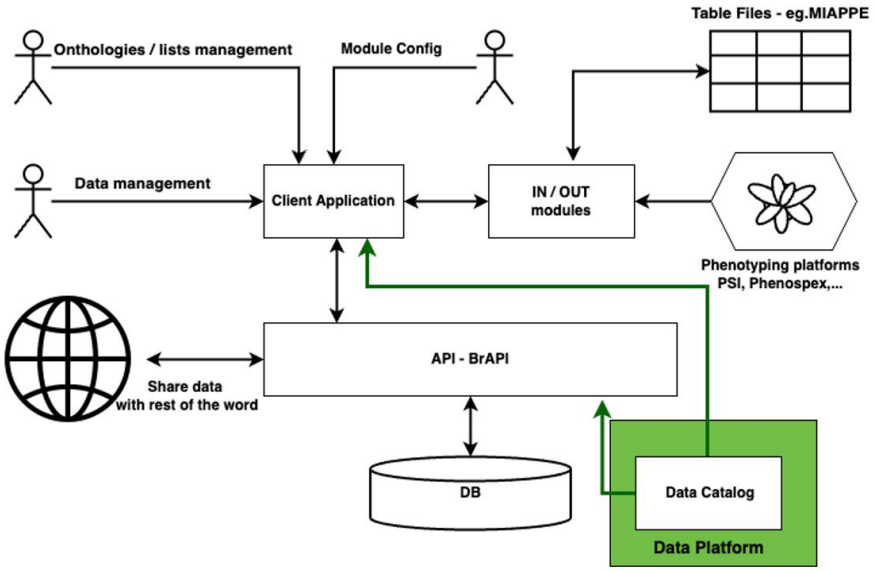
Figure 2: Schematic representation and description of the FAIR data principles..

Another key element of this solution is the use of a Digital Object Identifier (DOI), which is a standardised identifier that uniquely identifies a digital object, in this case, data. The DOI provides a unique and persistent reference to the data, increasing its citability and continued availability. This combination of FAIR principles and the DOI identifier thus strengthens the integrity, reliability and sustainability of data-sharing processes between scientific teams.

Standards for data harmonisation

Scientific information data harmonisation tools and frameworks are key in addressing several important data management challenges. One of the main challenges is the diversity of data formats and structures, which can complicate interoperability and information exchange between different scientific projects. These frameworks facilitate standardisation, providing a unified approach to defining schemas and metadata, eliminating differences and facilitating compatibility between different datasets. Another major problem that these tools address is the lack of a unified approach to data description and metadata. Metadata standards help to create a consistent and structured description of data, which simplifies the processes of finding, interpreting and managing information. This increases the transparency of scientific data and facilitates knowledge sharing among scientific communities. These frameworks and tools support a systematic and consistent approach to scientific data management, contributing to more efficient information exchange, reliable data analysis and overall higher-quality scientific research. The area of plant phenotyping was selected for further analyses in the area of data harmonization. Based on the analysis of the issue and interviews with plant scientists to further evaluate the MIAPPE and BrAPI data harmonisation guidelines and standards.

The following figure shows the connection scheme of the BrAPI module to the data platform modules (Figure 3).



Source: author

Figure 3: Diagram of a possible connection between the data platform and BrAPI.

Conclusion

The following requirements for the Data Platform can be defined from the results. As the data is acquired, it will be stored. This will create the so-called Row data, and then the processed data will be stored. This implies that the Data Platform "does not need to understand the data".

The platform development should be based on pilots' different data sources with data of different natures. The core data types from the agricultural domain are:

- Location data - GIS data
- Sensor data - time series
- Tabular data - e.g. csv format
- Data stored in the field interface
- Photographs, video recordings

Key types of data sources include

- IoT - data from IoT sensor networks
- REST-based interfaces
- Cloud-based file systems - primarily for initialising data retrieval.

The system for describing data with metadata needs to be made generic so that it is extensible and mappable to different data consumers in the future. The form of the metadata itself must allow the use of different data formats such as XML (Extensible Markup Language), JSON (JavaScript Object Notation) or YAML (Ain't Markup Language).

Other aspects, such as platform security, data backup and archiving, and API access to data, will be addressed in follow-up studies.

The aspects discussed in this article are mainly focused on scientific data from the field of agriculture - generally life sciences. However, similar principles can be implemented when analysing the requirements for a data platform for agricultural enterprises.

Acknowledgements

This work was supported by the EC's Horizon Europe funding in the project CODECS, grant no. 101060179.

The results and knowledge included herein have been obtained owing to support from the following institutional grant. Internal grant agency of the Faculty of Economics and Management, Czech University of Life Sciences Prague, grant no. IGA 2023B0005.

This research was carried out under the project: "Precizní zemědělství a digitalizace v ČR" (Precision Agriculture and Digitalization in Czech Republic), reg. no. QK23020058.

Corresponding author:
Ing. Michal Stočes, Ph.D.
Department of Information Technologies, Faculty of Economics and Management
Czech University of Life Sciences Prague, Kamýcká 129, 165 00 Prague - Suchdol, Czech Republic
E-mail: stoces@pef.czu.cz

References

[1] Burgelman, J. C. (2021) "Politics and Open Science: How the European Open Science Cloud Became Reality (the Untold Story)", *Data Intelligence*, Vol. 3, No. 1, No. 5-19. E-ISSN 2641-435X. DOI 10.1162/DINT_A_00069.

[2] David, A., Barbié, V., Attimonelli, M., Preste, R., Makkonen, E., Marjonen, H., Lindstedt, M., Kristiansson, K., Hunt, S. E., Cunningham, F., Lappalainen, I., and Sternberg, M. J. E. (2020) "Annotation and curation of human genomic variations: an ELIXIR Implementation Study", F1000Research 2020 9:1207, Vol. 9, No. 1207. ISSN 2046-1402. DOI 10.12688/f1000research.24427.1.

[3] Gomes, D. G. E., Pottier, P., Crystal-Ornelas, R., Hudgins, E. J., Foroughirad, V., Sánchez-Reyes, L. L., Turba, R., Martinez, A. P., Moreau, D., Bertram, M. G., Cooper A. Smout, C. A. and Gaynor, K. M. (2022) "Why don't we share data and code? Perceived barriers and benefits to public archiving practices", *Proceedings of the Royal Society B: Biological Sciences 2022*, Vol. 289, No. 1987. ISSN 0962-8452. DOI 10.1098/rspb.2022.1113.

[4] Kalendralis, P., Sloep, M., van Soest, J., Dekker, A. and Fijten, R. (2021) "Making radiotherapy more efficient with FAIR data", *Physica Medica*, Vol. 82, pp. 158-162. E-ISSN 1724-191X, ISSN 1120-1797. DOI 10.1016/j.ejmp.2021.01.083.

[5] Marešová, P., Jedlička, P., Soukal, I. and Novotný, J. (2019) "Open Science, Open Research Data and some Open Questions", *Hradec Economic Days 2019*, pp. 174-181. ISBN 978-80-7435-736-7. DOI 10.36689/uhk/hed/2019-02-017.

[6] Merton, R. K. (1973) "The Normative Structure of Science", In: Storer, N. W. (ed.) "*The Sociology of Science: Theoretical and Empirical Investigations*, p. 267-278. Chicago: University of Chicago Press. ISBN 0-226-52091-9.

[7] Novotný, J. (2019) "Open science, open research data and some open questions", *Hradec Economic Days*, Vol. 2, pp. 174-181. ISBN 978-80-7435-736-7.

[8] OPENAI (2023) "*ChatGPT-4. AI program*". [Online]. Available: https://openai.com/blog/chatgpt. [Accessed: Nov.16, 2023].

[9] Papoutsoglou, E. A., Faria, D., Arend, D., Arnaud, E., Athanasiadis, I. N., Chaves, I., Coppens, F., Cornut, G., Costa, B. V., Ćwiek-Kupczyńska, H., Droesbeke, B., Finkers, R., Gruden, K., Junker, A., King, G. J., Krajewski, P., Lange, M., Laporte, M. A., Michotey, C., Oppermann, M., Ostler, R., Poorter, H., Ramirez-Gonzalez, R., Ramšak, Ž., Reif, J. C., Rocca-Serra, P., Sansone, S.-A., Scholz, U., Tardieu, F., Uauy, C., Usadel, B., Visser, R. G. F., Weise, S., Kersey, P. J., Miguel, C. M., Adam-Blondon, A.-F. and Pommier, C. (2020) "Enabling reusability of plant phenomic datasets with MIAPPE 1.1", *New Phytologist*, Vol. 227, No. 1, pp. 260-273. ISSN 0028-646X. DOI 10.1111/NPH.16544.

[10] Prodan, R., Kimovski, D., Bartolini, A., Cochez, M., Iosup, A., Kharlamov, E., Rozanec, J., Vasiliu, L. and Varbanescu, A. L. (2022) "Towards Extreme and Sustainable Graph Processing for Urgent Societal Challenges in Europe", *Proceedings - 2022 IEEE Cloud Summit*, 20-21 Dec., 2022, pp. 23-30. ISSN 2330-2186. DOI 10.1109/cloudsummit54781.2022.00010.

[11] Sansone, S. A., Rocca-Serra, P., Field, D., Maguire, E., Taylor, C., Hofmann, O., Fang, H., Neumann, S., Tong, W., Amaral-Zettler, L., Begley, K., Booth, T., Bougueleret, L., Burns, G., Chapman, B., Clark, T., Coleman, L. A., Copeland, J., Das, S., … Hide, W. (2012) "Toward interoperable bioscience data", *Nature Genetics*, Vol. 44, No. 2, pp. 121-126. ISSN 1061-4036. DOI 10.1038/ng.1054.

[12] Selby, P., Abbeloos, R., Backlund, J. E., Basterrechea Salido, M., Bauchet, G., Benites-Alfaro, O. E., Birkett, C., Calaminos, V. C., Carceller, P., Cornut, G., Vasques Costa, B., Edwards, J. D., Finkers, R., Yanxin Gao, S., Ghaffar, M., Glaser, P., Guignon, V., Hok, P., Kilian, A., … Wren, J. (2019) "BrAPI - An application programming interface for plant breeding applications", *Bioinformatics*, Vol. 35, No. 20, pp. 4147–4155. ISSN 1367-4811. DOI 10.1093/bioinformatics/btz190.

[13] Stočes, M., Šilerová, E., Vaněk, J., Jarolímek, J. and Šimek, P. (2018) "Možnosti využití otevřených dat v sektoru cukr – cukrová řepa", *Listy cukrovarnické a řepařské*, Vol. 134, No. 3, pp. 117-121. ISSN 1210-3306. (In Czech).

[14] Thoegersen, J. L. and Borlund, P. (2021) "Researcher attitudes toward data sharing in public data repositories: A Meta-evaluation of studies on researcher data sharing", *Journal of Documentation*, Vol. 78, No. 7, pp. 1-17. ISSN 0022-0418. DOI 10.1108/JD-01-2021-0015.

[15] Tzanova, S. (2020) "Changes in academic libraries in the era of Open Science", *Education for Information*, Vol. 36, No. 03, pp. 281-299. ISSN 1465-3400. DOI 10.3233/EFI-190259.

[16] UNESCO (n.d.) "*Implementation of the UNESCO Recommendation on Open Science"* [Online], Available: https://www.unesco.org/en/open-science/implementation [Accessed: 20 Apr. 20, 2023].

[17] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R. Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., C 't Hoen, P. A., Hooft, R., Kuhn, T., Kok, R., Kok, van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Jun Zhao, J. and Mons, B. (2016) "The FAIR Guiding Principles for scientific data management and stewardship", *Scientific Data*, Vol. 3, No. 1, pp. 1-9. ISSN 2045-2322. DOI 10.1038/sdata.2016.18.