Data Management Conceptual Algorithm of Transnational Digital Scientific Infrastructure as an Answer to the Globalization Challenges

Diana Antonova[1] , *Svilen* Kunev[2], *Nataliya* Venelinova[3,*], and *Irina* Kostadinova[4]

[1]University of Ruse "Angel Kanchev", Faculty of Business and Management, Department of Business Development, 8, Studentska Street, 7017 Ruse, Bulgaria
[2]University of Ruse "Angel Kanchev", Faculty of Business and Management, Department of European Studies, International Relations and Security, 8, Studentska Street, 7017 Ruse, Bulgaria

Abstract.

Research background: The research background of the article has been formed by the analyses and expert assessments, made during the development of 10 different, strategic plans for the development of transnational digital distributed scientific infrastructure. This type of scientific network represents the answer of the regional research institutions, scientists, and businesses to the globalization challenges and emerging technologies.
Purpose of the article: The purpose of the paper is to contribute to the enrichment of the limited amount of literature on project management of digital transnational science networks through presenting a conceptual algorithm of a starting-up scientific infrastructure project.
Methods: The approach is based on defining the core inter-related features of a data management, developed for the needs of a newly- establishing virtual transnational scientific infrastructure. It is aiming to justify the defined principles for FAIR data, Data set description, Standards, and metadata, Data sharing, access and preservation, Restricted and Open research data, Data quality, laid down by the data management conceptual algorithm.
Findings & Value added: The algorithm has been presented as an integral part of a won project aiming to establish a virtual trans-national science infrastructure between partners from 5 EU countries. The findings are dealing with the definition of new organizational models based on a common mega-data system. The proper adaptation of the data management algorithm by each partner will improve the information management within its specific infrastructures as well as the visibility of their research outputs.

Keywords: *Data management; Research and Development; Innovation clusters*

JEL Classification: *O31; O32; O36*

* Corresponding author: nvenelinova@uni-ruse.bg

1 Introduction

Traditionally, technological globalization, which is one of the known types of globalization want to be solely offered to the higher social categories that had access to technological improvements. Now, a large scale of completely different as social, cultural and economic status, group of people have access to emerging technologies, networks and popular kind of digital innovations for business, social, educational or research interactions. Technological globalisation makes the individuals connected as well as the organizations, markets, governments and countries. It changes the dimensions of the accessibility to information resources, their organization, processing, structuring and archiving. It would be much more satisfactory if the globalizing trends might be seen only in their positive impact as creativity and innovation booster, but we cannot neglect easier all the challenges faced, together with the advantages of the global digitalization trend. One of them refers to essential changes in the quality, quantity, storage and usability of the information in terms of project management and corporate information management and communication, especially when these processes occur in a multilingual, multi-cultural transnational environment.

The data management (DM) became much more complicated due to the sets of skills needed, technologies introduced, human-software interaction, transparency of decision-making, security and privacy requirements and users-friendly hyper fast access physical and remote to information. All these preconditions evoke the transformation of the traditional data management to one of the most sophisticated, dynamically developing and highly logical and digitalized performing management activity which is based on vital machines, big data, cloud parks [1,2], smart data architecture and permanently developing technologies based on artificial intelligence.

The growing interactions between data, algorithms and big data analytics, connected things and people are opening are giving rise to issues around “data governance” at the national and international levels [1,3]. These include questions around the management of data availability, accessibility, usability, integrity and security, as well as concerns about ownership, impacts on all kind of social, political and business processes. “Data management systems and AI are synergistic. When AI becomes embedded throughout the data management system, it has the potential to improve database query accuracy and performance, as well as optimize system resources, reducing the burden on data base accelerators while improving data access for data scientists and developers “[2].

The use of digital technologies and data underpins digital transformation across all sectors of economies and societies, meaning that any management decisions on data can have wide effects on humans’ life and also, for example, on the characteristics of the value chain starting from the design of basic elements of the industrial process like machine building, and continuing with the digital transformation of the processes of product innovations [4, 5, 6]. The trust in the use of data is a pre-condition for fully realising the global digital transformation [3]. This namely calls to better understand and to take into consideration the heterogeneity of data, to introduce a strategic approach to its governance and to ensure all policy, programs and projects objectives. The last are facing daily the need of permanent upgrade of their own capabilities to harness data for better outcomes.

The Data Management Algorithm (DMA) in a research infrastructure unit is specifically designed and applied in the development of a scientific infrastructure project „Digital technological systems for the clean and secure environment – 5D ALLIANCE", elaborated by a research team of University of Ruse „Angel Kanchev", Bulgaria. The overall objective of the project is the establishment of Distributed Digital Scientific Infrastructure (SI) with a Potential for Impact within the Danube Macro region (DMR) for the period 2019-2028, through advanced interdisciplinary research for intelligent, secure, environmental-friendly management of interconnected systems and their business applications to achieve a clean

and secure environment in terms of shared responsibility for macroregional sustainable development.

This is planned to be implemented by a realisation of the following specific goals: (1) Investment and construction works of Distributed Digital Scientific Infrastructure (SI) with a Potential for Impact in the Danube Macro region with headquarter at the University of Ruse (7 laboratories), 1 remote back-up laboratory and 17 "remote access points" with the Scientific R&D Consortium partners; (2) Developing conditions for integrated research solutions in support of sectoral policies management, related to achieving a cleaner and more secure environment, based on digital transformation of conventional technological systems; (3) Approbation, multiplication and internationalisation of scientific achievements because of shared responsibility for macro regional sustainable development.

Those specific goals imply research work in the following main fields: I. Precise technologies for sustainable agriculture and clean and secure environment; II. Low carbon mobility and intelligent transport systems; III. Multi-modal human-machine interface and 3D kinematics in technological systems for a clean and secure environment and transnational projects; IV. Digital energy systems for a clean and secure environment.

2 Methods: The need for a data management algorithm

The here proposed Data Management Algorithm (DMA) describes the procedures used in the project for research infrastructure unit (RIU) for the processing of data during the project and after its end, defines the various kinds of data that need to be collected, stored and used, and synthesized, which approaches and standards will be applied during data collection and usage, suggests the procedures for sharing and open access to the scientific infrastructure's data and for curation and preservation of the data. Furthermore, procedures concerning the General Data Protection Regulation (GDPR) are defined and how the scientific infrastructure ensures the protection of the involved partners' data, information and privacy rights.

As part of the public scientific virtual infrastructure and being distributed among present and future partners who will participate in a range of subsequent scientific projects and pilot or experimental actions, the DMA establishes the framework on open research data that might be produced [1], used, re-used, shared etc. The aim is to provide indications as to what kind of data would be collected, how the data will be preserved and which sharing policies will be adopted towards making these data readily available to the research community [7].

The project's efforts in reference of open research data are deeply explained giving focusing on the following issues: the variety of open and non-open data that will be created or collected by the consortium, via research and development activities, during the project's duration implemented by all, one or more Consortium's partners; the technologies and infrastructure solutions that will be used to securely store the data long-term; the standards used to encode the data; the data exploitation plans; the sharing/access policies applied to each data-set.

The content of this paper builds upon the empirical input of the RIU partners. A short questionnaire, outlining the DMA's objectives and stating the required information in a structured manner, has been edited by the RIU's Lead partner and will be disseminated to the partners. The compiled answers will be integrated into a coherent plan. The present DMA can evolve as the RIU progresses in accord with the project's efforts in this area. At any time, the DMA should reflect the validated status of the consortium's agreements regarding data management, exploitation and protection of rights and results. The specifications of the storage facilities needed should also be outlined and guidance documents should be developed to ensure their proper use.

3 Results

3.1 Key characteristics of the data management algorithm

One of the most important issues of data-intensive science is to facilitate creation and transfer of knowledge by assisting researchers and the technological environment in their creation of, access to, integration and analysis of, specific scientific data and their associated algorithms and workflows. Here, we apply the FAIR DATA - a set of guiding principles to make data Findable, Accessible, Interoperable, and Re-usable [8, 9].
To be able to distinguish and easily identify data sets, each data set will be assigned with a unique name. This name can also be used as the identifier of the data sets.

Data set description: each data set that will be collected, processed or generated within the project must be accompanied by a brief description, history of changes table with the precise indication for the version number, the date of the change, the type of change, the person initiating the change, the person executing the change as well as the dissemination of the data set after changes [10]. Participating institutions and teams should provide a very good comprehensive description, including the scientific area and technical methodology, to enable alignment of their data sets with specific research themes.

Standards and metadata: Partners must describe the procedures and standards they apply to structure their data (i.e. fully reference the metadata) so that other researchers to be able to assess and re-use the set of data. Since the partners agree to interact within a complex system like the 5D-ALLIANCE, their decision-making processes could be described under the principles and influencing factors of the dynamics of the complex systems [11]. If available, they should provide a reference to the community data standards with which their data conform and that make them interoperable with other data sets of similar type. The selected versions and types of the software should not include complicating metadata requirements that might be confusing with the produced 5D-ALLIANCE data sets and might create obstacles to produce integrated applications at the commercialization phase.

These file formats that might be produced in Microsoft OS environment must be preferably chosen because they are accepted standards and in widespread use. Files should be converted to open file formats where possible for long-term storage and open access re-use in conformity with the rules for securing the sensitive data as well to protect the authors' ownership. The content management system must be applied to the created contents at the horizontal level of the project in terms of creating an opportunity for further integration, redefinition and archives management. Metadata is recommended to be kept two formats – contextual information about the data in a text-based document and ISO 19115 standards metadata in an xml file. Two formats for metadata should be chosen to provide a full explanation of the data (text format) and should ensure compatibility with international standards (XML format).

Data sharing, access and preservation: The digital data created by the project should be diversely curated depending on the sharing policies attached to it. For both open and non-open data, the aim is to preserve the data and make it easily accessible to different stakeholders for the whole duration of the project and after its end. From the viewpoint of the software systems, and architecture of a distributed knowledge-based system might be appropriate to be developed [12]. A public Application Programming Interface (API) should be provided to registered users allowing them access to the platform.

The database compliance aims to ensure the correct implementation of the security policy on the databases verifying vulnerability and incorrect data. The target is to identify excessive rights granted to users, too simple passwords (or even the lack of password) and finally to perform an analysis of the entire database [13]. At this point, the minimum

requirements for assuring proper management of data could be: (1) registers for controlling access for users and data authentication; (2) monitoring and Log of activity; (3) design and application of an alert system; (4) liability. When developing solid data management plans, researchers have the task to analyse the following topics and to respond to the following questions: what approach will be used to collect or produce new data and/or how will existing data be re-used? What approach will be used to collect or produce data (for example the types, file formats, and volumes) for open data sharing from the scientific activities?

Non-Open research data: The non-open research data should be archived and stored long-term in digital infrastructure servers and at a back-up lab at the territory of a dedicated RIU partner being specifically employed to coordinate the project's activities and to store all the digital material connected to 5D-ALLIANCE. If under any legal and contractual reasons certain datasets cannot be shared (or need restrictions), this should be explained.

Open research data: Access should be acknowledged. This is especially important as they transition to an open, transparent research infrastructure can only take place if as many stakeholders as possible are invited to participate [15,16]. The open research data should be archived on their platform enabling users to share and preserve research data and other research outputs in any size and format: datasets, images, presentations, publications and software. The digital data and the associated meta-data would be preserved through high-quality methods such as mirroring and periodic backups. Each uploaded dataset should be assigned a unique DOI rendering each submission uniquely identifiable and thus traceable and referenceable.

Metadata and data preparation: to make stored data easy to find, with good access, interoperable and reusable (FAIR), it is not enough to store 'raw data'; they need to be properly documented and described using informative metadata. Defining appropriate metadata depends on the discipline and/or the methodology that is used to produce the data, either for research or education that includes some widely used modern software solutions like 3D, for example [13]. Discipline-specific repository facilities usually have comprehensive requirements for data description that are stored in that depository. A one widely applied minimum standard for describing information on the web, suitable also for research data, is Dublin Core [17].

Documentation and data quality: Under this DMA four levels of documentation have been established. All descriptions of qualitative or quantitative data used by the RIU members should be categorized within one of the following categories: Deliverable no; Responsible Partner; WP no. And title; Task no. and title; Version; Version Date; Dissemination level – PU – Public, PP - Restricted to other project participants; RE - Restricted to a group specified by the consortium; CO - Confidential, only for members of the consortium.

Datasets in 5D-ALLIANCE are defined as organised data and exclude un-organised data. Example of un-organised data could be noted from interviews, workshops and exercises that are not directly included in the project deliverables but are only used in the form of secondary, supportive data. Such data might be used for guidance and analysis internally in the project only and might not be structured in a way to make them reusable after the end of the project. It is foreseen that not all data produced from the project will be openly available after the end of the project, and in the cases where a dataset is public, there might still be parts of the dataset that remain non-public. There are five main reasons for this:

1. Data collected from volunteers participating in interviews, workshops and pilot exercises (etc.) contains personal data that is confidential. The project is subject to Ethical Requirements to protect this data and ensure the participants' privacy. Only aggregated, anonymised and analysed data from datasets are included in project

deliverables and/or published in articles and papers [15]. In the cases where datasets are not made public, the main reason is that the data has the potential to be traced back to the individual participants and must remain confidential to protect their privacy.

2. The data collected in this project is context-specific, and the part of it that is publicly available should be very detailed so that it could be interpreted and understood by external users. If more data is added, for example in the form of "raw data", this could cause misinterpretations.
3. The initial collection of data its "raw form" in empirical studies might disclose characteristics of critical infrastructure operations that could be part of the security and organisational sensitive information and concerned organisations might not permit researchers to make this data available.
4. Most data from stakeholders and participants are collected in local languages. This data is then aggregated and analysed, and only the analysis of this data is available in English. To translate all raw material from interviews, workshops etc. to English would require resources beyond the availability of the RIU, and would again potentially lead to the identification of individual participants (or organisations).
5. It is a common rule that data retrieved from research and scientific published resources are in often copyright-protected so that datasets with entries of text taken directly from such resources cannot be reproduced publicly, except for occasional quotes of very limited length.
6. Since all details of datasets have a common structure, the same wording might be repeated between the different descriptions. The name for each data set includes a prefix "DS" for the data set, followed by a case-study identification number, the partner responsible for collecting and processing the data, as well as a short title [19].

3.2 Approach for a preliminary check of the elaborated DMA

Several domains for quality check of the project for DMA could be useful to be applied. Each domain includes several criteria that should be checked. The information needed to answer to the control questions may be already documented somewhere else. Thus, project management is easier by keeping project-related information in one place. Also, if data will be deposit to a data repository it is important to remember that all relevant information about the research project and the data are deposited. The answers of the control questions evaluate their relevance to the project proposal with three options – yes, no, don't know [20]. The suggested domains of the quality check and their internal elements are:

1. **General project description**: (1.1) Project description; (1.2) Responsible researcher (person, institution or organization); (1.3) Participating researcher and/or organizations; (1.4) Project Data Contact; (1.5) Owner of the material; (1.6) Producer; (1.7) Roles; (1.8) Funder; (1.9) Related Policies.
2. **Ethics and Legal Compliance**: (2.1) Ethical review; (2.2) Privacy officer; (2.3) Informed consent; (2.4) Protection of the identity of participants; (2.5) Confidential information; (2.6) Intellectual property rights/Copyright; (2.7) Agreements with other organizations; (2.8) Archiving; (2.9) Restrictions; (2.10) Embargo.
3. **Data Collection**: (3.1) Existing data; (3.2) Type of data; (3.3) Data collection procedure.
4. **Documentation and Metadata:** (4.1) Documentation; (4.2) Metadata; (4.3) Metadata standard; (4.4) Terminologies, ontologies etc.
5. **Data management during the project**: (5.1) Folder structure; (5.2) Organizing your data: (5.3) Data protection or security policy at the organization; (5.4) File naming; (5.5) File format; (5.6) Versioning; (5.7) Storage and backup.
6. **Budget:** (6.1) Staff; (6.2) Hardware and software; (6.3) Storage;

7. **Data sharing:** (7.1) Making data available; (7.2) Contact a data repository; (7.3) Limitations (legal/ethical restrictions); (7.4) Limitations (hardware & software); (7.5) Delays; (7.6) Citation; (7.7) Persistent identifier (PID).
8. **Other issues:** to be specified by the project team according to the project specifications and the needs for assessment of dedicated characteristics. Figure 1 bellow visualize the interconnection among the criteria domains and their internal number of elements.

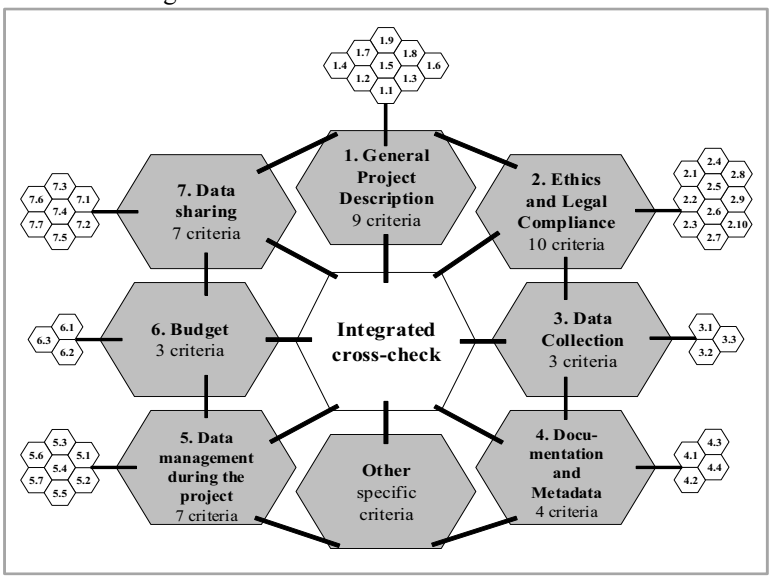

Fig. 1. Integrated set of criteria for a preliminary check of the elaborated DMA
Source: authors' elaboration

The evaluation of the proposal for DMA is based on experts' opinions and could be more precise if each question, examining the relevance of a specific criterion is answered by the experts with the help of a gradient Likert scale with five options, for example, 1 – not relevant, 2 – slightly relevant, 3 – I cannot decide, 4 – relevant, 5 – fully relevant. Also, a second scale for each question could be introduced for evaluation of the current status of the criterion for the specific proposal. In this scale, the five options could be 1 – very weak, 2 – weak, 3 – satisfactory, 4- good, 5 – very good. The usage of the two scales together will bring a very strong benefit because will allow researchers, evaluators or the authors of the project for DMA to identify the most relevant criteria for the project by ranking according to the average score, and on the next step to measure the quality or the current status of most relevant characteristic by their average score for the status.

4 Conclusion

1. The concept of data management algorithm (DMA) will be realized at the successful completion of the transnational scientific infrastructure project "Digital Technology Systems for a Clean and Secure Environment - 5D ALLIANCE" in the period 2020-2027 between scientific structural units (7 multi-disciplinary laboratories, 1 back-up laboratory for minimizing risk) and 17 access points linked in a common metadata system) in Bulgaria, Germany, Slovakia, Romania and Poland.

2. The contribution to the project management literature of our conceptual algorithm for data management in transnational digital science infrastructure network can concentrate on the formation of the main interrelated characteristics of FAIR data, the description of the

data set, standards and metadata, data sharing, accessing and retaining them, documenting unopened research, open-source research data, metadata and data preparation, as well as documentary records and data quality control. Their detailed presentation is the subject of other publications.

3. The prerequisites for the success of the 5D ALLIANCE Project are correlated with the proper adaptation of the DMA, which will have a significant impact on the quality of the planned results, impacts and prospects for the development of the Danube Macroregion through the developed Digital Science Infrastructure Network.

4. The 5D ALLIANCE project proposal was evaluated by an independent European Commission expert panel at the end of 2019 using a Preliminary Check Approach based on several core domains with specific internal criteria that allow integrated cross-checking with statistical processing of quantitative indicators.

References

1. Anderson, H. J., Stejskal, J. (2019). Evaluating the Impact of Marketing, Organisational and Process Innovation On Innovation Output of Information Technology Firms: Czech Republic and Estonia. In O. Dvouletý, M. Lukeš & J. Mísař, (Eds.), *Proceedings of the 7th International Conference Innovation Management, Entrepreneurship and Sustainability*, (pp. 31-41). Prague.
2. IBM Hybrid Data Management (2020, September 17). *Accelerating AI with Data Management; Accelerating Data, Management with AI*. Retrieved from : https://www.ibm.com/downloads/cas/YD5R1XLB
3. OECD (2019, June 20). *"Data in the digital age", OECD Going Digital Policy Note.* Retrieved from : https://www.oecd.org/going-digital/data-in-the-digitalage.pdf
4. Cai W., Liu, F., Dinolov, O, Xie, J., Liu, P., Tuo, J. (2018). Energy benchmarking rules in machining systems. *Energy*, 142, 258-263.
5. Liu C., Cai, W., Dinolov, O., Zhang, C., Rao, W., Jia, S., Li, L., Chan, F. (2018). Emergy based sustainability evaluation of remanufacturing machining systems. *Energy*, 150, 670-680.
6. Stoycheva, B., Antonova, D. (2018). Investigating Factor Interactions in Formalising the Process of Developing New Products. *Serbian Journal of Management*, 13(1), 173-184.
7. Harvard Biomedical Data Management. (2019, August 5). *HMS Data Management Working Group*. Retrieved from : https://datamanagement.hms.harvard.edu/hms-data-manage¬ment-working-group
8. FORCE11. (2020, January 20). *The FAIR Data Principles*. Retrieved from : https://www.force11.org/group/fairgroup/fairprinciples
9. Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 160018.
10. Hitz, Ch., Vojvodic, M., Wicki, Gr. (2019). Principles of Data Definition for the Use of Measuring Governance. In O. Dvouletý. M. Lukeš & J. Mísař (Eds.), *Proceedings of the 7th International Conference Innovation Management, Entrepreneurship and Sustainability* (pp. 265-278), Prague.
11. Ghinea, V., Mihaylova, L., Papazov, E. (2015). Organizational Culture Dynamics. Complex Systems Dynamics. *Quality – Access to Success*, 147(16), 99-105.

12. Marinov, M., I. Valova. (2019). Component Interaction in Distributed Knowledge-Based Systems. *TEM Journal – Technology education management informatics,* 8(3), 721-727.
13. Harvard Catalyst (2019, August 12). *An Investigator's Guide to Research Data Management Practices*. Retrieved from : https://catalyst.harvard.edu/pdf/regulatory/-Investigators%20Guide%20to%20RDM%20practice.pdf
14. Aliev, Y., Kozov, V., Ivanova, G., Ivanov, A. (2017). 3D Augmented Reality Software Solution for Mechanical Engineering Education. In B. Rachev & A. Smrikarov (Eds.), *18th International Conference on Computer Systems and Technologies* (pp. 318-325). Ruse: Association for Computing Machinery.
15. LIBER (2019, February 04). *Open Access Working Group: Statement on Plan S guidelines*. Retrieved from : https://libereurope.eu/blog/2019/02/04/open-access-working-group-statement-on-plan-s-guidelines/
16. Science Europe (2018, November). *Practical Guide to the International Alignment of Research Data Management.* Retrieved from : https://www.scienceeurope.org/media/jezkhnoo/se_rdm_practical_guide_final.pdf
17. DCMI (2020, January 15). *The Dublin Core™ Metadata Initiative (official website)*. Retrieved from : https://www.dublincore.org/
18. Mikhaylov, A., Peker, I. (2019). Spatial Patterns of Innovation Geography: Knowledge Generation Domain in Russia. In O. Dvouletý, M. Lukeš & J. Mísař (Eds.), *Proceedings of the 7th International Conference Innovation Management, Entrepreneurship and Sustainability*, (pp. 624-637). Prague.
19. Dvouletý, O., A. Pilková, J. Mikuš, M. Rimská. (2019). Entrepreneurial Activity in Slovakia: Selected Regional Aspects and The Role of Governmental Environment. In O. Dvouletý, M. Lukeš & J. Mísař (Eds.), *Proceedings of the 7th International Conference Innovation Management, Entrepreneurship and Sustainability*, (pp. 161-171). Prague.
20. DCC (2013). *Checklist for a Data Management Plan. v.4.0. Edinburgh: Digital Curation Centre*. Retrieved from : http://www.dcc.ac.uk/resources/data-management-plans