CARTOGRAPHIC AND GIS PRODUCTS AND THEIR LONG-TERM DIGITAL PRESERVATION

Martin Rechtorik, Otakar Čerba

Mgr. Martin Rechtorik, main author 99%
Department of Geomatics, University of West Bohemia, Faculty of Applied Sciences
Univerzitni 2732/8, 301 00 Plzen
National Archives of the Czech Republic, Archivni 2257/4, 149 00 Prague 4
maarty48@kgm.zcu.cz; martin.rechtorik@nacr.cz

Doc. Ing. Mgr. Otakar Čerba, Ph.D., co-author (1%), supervision
Department of Geomatics, University of West Bohemia, Faculty of Applied Sciences
Univerzitni 2732/8, 301 00 Plzen

Abstract
Digital preservation is about ensuring that data remains accessible and usable. This involves data extractions from various environments and its conversion into an interoperable format to prevent the need of the original software or hardware. Such an approach has to be applied to the geospatial products to ensure the long-term availability of this type of data. However digital preservation is not solely about preserving the data itself, but also about preserving the context and meaning of that data. Metadata provides crucial information about the content, context, and structure of the data, making it a key component of digital preservation. A crucial aspect of the research is a reflection on the anticipated future use of geodata considering its long-term, readability, trustworthiness, integrity and accessibility. Regarding the fragility of geodata in its complex form, preservation poses an increasingly interdisciplinary scientific challenge.

Keywords: *Big Data, Digital preservation, Geospatial Analysis, Maps production, GIS*

INTRODUCTION

Whether we discuss maps or other outputs from GIS systems, we inevitably discuss data and its transformation. ”*Data is a discrete, limitless entity that has an unstructured and unprocessed shape.*”[1] But how to preserve such an entity? We must define its structure and process it in various ways into a product that can be handled as an artefact as proposed by Trevor Owens.[2] In this case, the result is something specific and real that can be preserved, archived and undoubtedly be reused in the future. This paper provides an overview of the ongoing research. On the basis of the project which was carried out from 2020 to 2022 identified gaps that have drawn my attention for future research.

ARCHIVING BELONGS TO DATA MANAGEMENT

Quality data management has gained increasing importance in recent times and it is rightfully receiving more attention. Data has become an important economic resource, facilitating development across all fields. Data fundamentally increases the efficiency of processes, enables the modeling of potential situations in many fields either in short or long periods and helps anticipate developments and risk situations.[3] Data is a matter of concern both from national governments and from the international community.[4] Good quality and accurate data is attracting interest from the private sector, as it enables the creation of new products that generate significant profits. There is currently a meteoric rise in the use of machine learning algorithms, which will undoubtedly lead to increased interest in, and value of, data.[5] A similar trend can be expected for archived data in the future, which will be used by its creators, but also by those interested in other, often unrelated fields.

Recognising this need on the part of the government, the Digital and Information Agency in 2023 to oversee the development of digitalisation in the Czech Republic succeeding the Ministry of the Interior. Among the many tasks is the task of mapping data collections in Czechia. This highlights the importance and value of the data that public institutions create, manage and publish.[6] This activity has to attract the attention of archives for future data preservation. Unlike the aforementioned Agency, archives were previously only interested in the end product, not the raw data itself, although this is no longer true and even raw data is gradually becoming the subject of preservation too with regard to its meaning,

content or value. In any case, archives and agencies that ensure the development of digitisation of the state and its organisations are naturally aligned with a shared interest in the complete life cycle of data including geodata.

OVERVIEW OF THE SITUATION OF THE GEODATA PRODUCERS IN THE CZECH REPUBLIC

Geodata producers are required to develop quality data management to fulfil many requests. That is why they face the critical decision of how to manage their data once it transitions from being actively used to the archival stage. There always exists a milestone that compels a producer to decide what to do with their data. This decision is guided by factors such as data disposal schedules, regulatory requirements, and budget and resource constraints.[7] That issue brings us to the common data lifecycle and its milestones.

The data lifeline framework outlines the journey of data from its initial generation to its ultimate destruction, encompassing six main stages. At the beginning, data emerges in the Data Generation phase, originating from a variety of sources such as user inputs, sensors, logs, or external data feeds. This raw data represents the foundational building blocks of the data lifecycle, capturing the essence of information in its nascent form. Subsequently, in the Data Processing stage, raw data undergoes a metamorphosis, transforming into a more refined and usable format. This transformation contains various operations like cleaning to eliminate errors or inconsistencies, normalisation to standardise formats, and aggregation to condense information. The goal is to enhance the quality and relevance of data, making it structured and meaningful for subsequent analysis and use.

Upon undergoing processing, data moves to the Data Storage stage, where it is stored in designated repositories for future access and retrieval. This phase involves selecting appropriate storage systems such as databases, document management systems, data warehouses, or cloud platforms, taking into account factors like scalability, performance, and security. The objective is to ensure the integrity, accessibility, and security of stored data, thereby supporting the organisation's data management requirements. As data becomes readily available for consumption, it enters the Data Usage stage, where it is used for various purposes such as analysis, reporting, decision-making, or application development. Stakeholders, ranging from applications to analysts, data scientists, and business users, utilise this data to extract insights, drive value, and inform strategic decisions. This stage underscores the significance of extracting actionable intelligence from the available data assets to drive organisational growth and innovation.

However, not all data remains in its Usage stage indefinitely. When data is less frequently accessed and used still remains important for compliance or historical reasons the Archiving stage begins. Based on a decision made either by producer, data manager or archivist is data split into two groups. The first one is destined for Archiving which means to be transferred to special long-term storage repositories or archives.Such decision making is usually called the Appraisal process. Archiving ensures the preservation of valuable data while optimising primary storage resources for active usage, striking a balance between accessibility and resource efficiency. The second group is destined for destruction. Ultimately, as data loses its usefulness or relevance, it undergoes the final stage of the data lifecycle - Data Destruction. Here, data is securely deleted from storage systems, employing methods that render it irretrievable to prevent unauthorised access and ensure compliance with data privacy regulations. This critical phase safeguards sensitive information, mitigating the risk of data breaches or misuse and culminates the data lifeline journey.

One option available to geodata producers is to partner with digital archives, which are specialised institutions dedicated to the long-term preservation of digital information. By partnering with digital archives, geodata producers can leverage their expertise, infrastructure, and resources to ensure ongoing accessibility and integrity of their data over time.[8] Alternatively, geodata producers may choose to maintain their own data repositories independently taking on full responsibility for long-term preservation. However, this approach requires the implementation of robust data management practices and standards, including regular monitoring, maintenance, and migration efforts which differ somewhat from standard data management practices and standards. It also requires investment in specialised infrastructure and skilled personnel, making it a costly and resource-intensive endeavour. Failure to adequately address the long-term preservation needs of geodata can have significant consequences, including the potential degradation and loss of valuable information. Factors such as bit rot, hardware failures, and technological obsolescence can contribute to the deterioration of geodata over time, ultimately leading to irreversible data loss.[9]

DATA CREATOR ROLE

Geodata producers play a crucial role in the creation, management, and dissemination of geospatial data. Their responsibilities include collecting and acquiring data from diverse sources, processing and analysing it to derive insights, ensuring its quality and integrity through validation and documentation, and adhering to regulatory standards and compliance requirements. Geodata producers facilitate data dissemination and sharing with stakeholders while managing the entire data lifecycle from its acquisition to its archival. They continuously seek to improve their practices through

innovation and the adoption of new technologies, ensuring the availability of accurate, reliable, and accessible geospatial information for a variety of applications and users. The issue of preserving data over time involves several crucial aspects of data management that must be addressed. The first aspect worth mentioning is careful planning which involves establishing a comprehensive plan that outlines the procedures and protocols for data preservation. This plan should include policies for data production, storage, controlled access and a retention period which is usually published as an agency´s disposal schedule. Data production is also tied to the construction of a reliable and secure infrastructure capable of handling variable volumes of data. Redundancy and backup solutions are an essential part of the entire data workflow to mitigate risk of data loss due to hardware failures, cyber attacks or disasters.

In the subsequent step we typically encounter various forms of standardisation. A good strategy is to adhere to industry best practices and standards for data management, such as those outlined by organisations like the International Organization for Standardization[10] (ISO) or the recommendations or best practices published by for example. European Commision.[11] Compliance with these standards ensures compatibility and interoperability with future systems. Once a standard is established and accepted across the entire agency, data integrity must be ensured. The implementation of mechanisms to verify and maintain the integrity of stored data over time like checksums, error detection, and correction algorithms to detect and mitigate data corruption or tampering, requires the usa of a wide range of specialised tools that can ensure such high level features.[12]

If an agency is able to fulfil at least a significant portion of the aforementioned requirements, then the issue of format sustainability arises. Choosing appropriate file formats for data preservation is an important consideration. This aligns with the chapter dedicated to standardisation. Adopting a widely used file format, supported by an open community with published documentation, ensures that the agency's data remains accessible over time. If there is a risk that migrating to an open source file format could potentially damage the original data, it is advisable to preserve data in multiple file formats.[13] This is a common archival strategy, to maintain data in both the original format and in a format suitable for preservation. However, it should be noted that archives are not an omnipotent organisation, and after ten or twenty years, it cannot be guaranteed will still be possible. This is why the initial selection of data format should be made by the producer but in compliance with archive requests. Another option is to use a database if any environment is already in use, it is worth rationalising the infrastructure to a limited number of software applications to mitigate data management risks associated with employee turnover. Preserving a large number of data assets in multiple versions necessitates the production of some metadata alongside the data. There is a range of metadata related to geodata, for example, Dublin Core, ISO 19115, EU INSPIRE standard etc.

The final feature is regular monitoring. Implementing this ensures maintenance to prevent a bit of rot or other forms of degradation and promotes continuous education and training of employees. All these steps must be followed to preserve data over time but when we talk about a real long-time preservation, it introduces similar requirements but at another layer and multiplies the complexity of the issue. Why undertake this if there exist or will soon exist archives capable of doing this for you? Overall, the role of a geodata producer is multifaceted, requiring expertise in data collection, processing, quality assurance, documentation, dissemination, compliance, and innovation. Geodata producers play a vital role in generating and maintaining accurate, reliable, and accessible geographic information for various applications and stakeholders. Their work forms the backbone of various applications and is invaluable to a wide range of stakeholders.

DIGITAL ARCHIVES ROLE

Digital archives are institutions that specialise in preserving information for a very long time, or even indefinitely. They are so-called memory institutions that are dedicated to preserving historical artefacts for the future, for further research and education. History shapes us, connects us, whether in a positive or negative sense. Traditionally, archives have been concerned with the care of hand written or printed sources, for example materials on which information has been superimposed. However with the explosive growth of digitisation, these traditional and conservative organisations must gradually change into digital archives. Although the term ´digital´ and ´digitising´ archive would be far more accurate, as the transformation of originally analogue information into digital form is also one of the key activities carried out in archives these days.[14]

Digital preservation is based on the International Standard ISO 14721:2012, titled ´Space data and information transfer systems Open archival information system (OAIS) Reference model´. This is a recommended technical practice to establish a common framework of terms and concepts which make up an Open Archival Information System (OAIS) which provides a fuondation for further development and promotes standardisation within the archival sector. It also aids in enhancing vendor awareness of OAIS and supports archival requirements.[15]

GEODATA PRESERVATION PROJECT

The first real encounter with geodata archiving occured in May 2019, thanks to participation in the 1st Geospatial Preservation Conference, which took place in Ljubljana.[16] When an offer of cooperation came from the Research Institute for Soil and Water Conservation[17] in September of the same year, the National Archives accepted it and thus the preparation of a joint project began. Both sides decided to compete together for support from the Czech Technical Agency for a geodata archiving project within the Éta programme. The Éta programme aims to promote interdisciplinary cooperation, particularly between the humanities and the exact sciences and geodata archiving fulfils exactly this.[18] This application was successful and from 2020 to 2022 the project was carried out. The VÚMOP team consisted of experts in spatial data production, including collection of real data in the field, and insiders from imaging and making it available to the public. Among the members of the National Archives' team were digital methodologists. At the time of the project, the situation with geodata production In Czechia was mapped with results important for the development of tools.

As the National Archives is the Central Archives of the state and is responsible for the National Digital Archives too, according to legislation, it brings many challenges related to the wide scale of data producers, including those producing any type of geospatial data and in any format. To solve such a big challenge, strict standardisation on data level structure as well as on metadata level is required. We discovered that the basic metadata requested for digital archiving which is based on Encoded archival Description[19] is not sufficient for complex and sophisticated data which geodata truly is. That is why we decided to implement EARK structure[20] as a content of the Czech submission information package (hereafter referred to as SIP) and to map our own implementation of Dublin Core metadata[21] which harvest the most important metadata values from ISO 19115 standard[22] or must be produced by SIP Packagers before ingesting. Further development of archival cataloguing software used for collections arrangement is expected too.[23] All these are necessary features for an expected future reuse of geodata. Strict standardisation and the high demands associated with it proved to be a real problem, but loosening it would mean a big risk that non-standardised data would be acquired into the National Digital Archives and could not be easily shared in the future.

Within the project, a questionnaire survey among producers was conducted, which told us that only a half of them produces quality metadata describing datasets content, and the majority of them publish data using ESRI Shapefile format.[24] Other half produces metadata either compliant with EU INSPIRE standard and ISO 19115. That information led us to try producing a specialised SIP Creator[25] suitable for geodata within the project. Our partners and leaders of the project from the VÚMOP took up this challenge and developed a pilot software, which turned out to be functional, and we are about to implement it into our appraisal workflows. During the testing phase of it, we suddenly realised one important thing: it is annoying to fulfil the same metadata for different purposes, and automation of it is absolutely inevitable. That led us to think out of the box to solve such complex tasks. The main idea is to use geospatial metadata as much as possible, to parse published XML files compliant to the EU INSPIRE Standard[26] for preservation needs but keep the original metadata too. On the other hand it can be useful to develop tools which can produce not only archival metadata but geospatial metadata too in a simple way, and such tools can be handy to an agency which has still not implemented metadata production yet. Furthermore, for long term preservation it is preferred to preserve well-known text representation of geometry too.That led us to dive deeper in this issue, and in house, we developed two simple Python scripts.

The first tool is being used for scraping published open data datasets. Based on its ID, the dataset is downloaded and SIP is produced in an automated way, including all requested metadata. When data is published as a service and no file can be downloaded, it can be used to produce SIP too, but requires a path to the location where data files for preservation are stored within the file system. Besides the automated metadata production, there are some nice features implemented too, like migration of PDF files to PDF/A format, static image check if they are corrupted or not too. In the future we want to enhance the tool also with format identification based on Pronom registry and with validation of Geopackages[27] and GML[28] files against its schema files.[29] The second script is intended for geodata producers who neither run their own metadata catalogue nor publish metadata elsewhere. Its purpose is to produce various metadata files which are common in Czechia, to work like all-in-one-tool. If a user of this script inserts metadata in an interactive way to a terminal, the script produces a SIP package which contains not only metadata files for preservation purposes but also metadata files for further deployment and expected future re-use.

GEOINFOSTRATEGY OF THE CZECH REPUBLIC

The GeoInfoStrategy was developed on the basis of the Government Resolution No. 837 of 14 November 2012, with respect to all international obligations to which the Czech Republic is bound in the field of spatial information (EU, NATO), and in relation to the strategic documents of the Czech Republic. The aim was to define a strategic development framework, set clear rules for the creation, management, and use of spatial information by the whole society, and create conditions for the organic integration of guaranteed spatial information into the decision-making processes of public administration and the life of the whole society. The GeoInfoStrategy was approved by the Resolution of the Government of the Czech Republic of 8 October 2014, No. 815, and elaborates the basic principles of the development of public

administration and eGovernment in the field of spatial information. It focuses on solving specific problems in this area in the Czech Republic and proposes the provision of quality guaranteed spatial information and services over spatial data, not only for the effective performance of public administration, but also for the needs of the whole society.[30]

PROJECT RESULTS

As the project was carried out, it turned out to be interesting for the Ministry of Interior, and the final product, which was methodical guidelines for geodata preservation, was not only approved and certified by the Ministry of Interior, but was also incorporated into the GeoInfoStrategy. This achievement is very much appreciated.[31] Besides the methodical guidelines, there are other findings too. As geodata is a complex GIS output, it requires a unified folder structure to be stored in. A similar way was decided to be used by the Ministry of Regional Development for data exchange within digital spatial planning. Unlike them, we didn't invent our own folder structure. We compared the Swiss model,[32] The newly developed EARK Geospatial Data specification in version 3,[33] and the Danish model too.[34] At last, we decided for the EARK specification as more suitable for our needs. In fact, there are only slight differences between those models, and even if we decided to develop our own structure, it would not differ too much. Moreover, the implementation of the European model can be advantageous in using common tools which are developed within the cooperation of DLM Forum members, or by the EARK projects, or from any other support coming from the European Commission.

Among practical findings are limits connected with usage of the SIARD format[35] However, data managed by GIS can be treated as database data; it is fundamentally different from data managed by common relational databases. Despite the fact that the relational model is used by GIS, it is used only for handling object properties. GIS is a naturally object oriented database system and the relational model is only a useful part of it.[36] Thanks to our colleagues from Estonian National Archives, we can use the SIARD format for Oracle Spatial Database environment preservation, but we can not use it for PostgreSQL plus PostGIS extension.[37] Except for the lack of support for PostGIS extension, we faced issues connected with overloading of that database which is not compliant with SQL standard.[38] Nevertheless the preservation of data managed in PostgreSQL plus PostGIS was successful. Data was exported using QGIS software and SIP was produced manually. But another issue was soon discovered. It was connected with the Geography Markup Language format. One of the files was bigger than 1 Gigabyte and such size is not easy to handle, and it is recommended to be migrated to another file format either to Esri Shape file or Geopackage. This problem is easily solved by slicing a large file into pieces so that every single feature is represented as a single file as proposed to me by one colleague from the Czech State Administration of Land Surveying and Cadastre. It can be solved in that way but only for processing purposes and not for preservation. As digital preservation uses regular fixity and corruption checks over the repository and produces special metadata like PREMIS[39] for each component inside SIP. That approach would inevitably lead to creating unprocessable big metadata files. This is a highly undesirable scenario, and other options to use compressed packages like ZIP file format[40] or TAR.GZ[41] prevent digital archives from making regular checks for each one object inside those packages.While we can say that this is fine now, it is unfortunately insufficient for the coming future when data is going to be shared using various web services[42] and only a few enthusiasts would be satisfied to download data files and deploy it by themselves. As a significant majority of future users of the digital archive is going to expect a rapid access to the preserved geodata, an automated workflow has to be implemented in digital archives processes too. Such automated workflow is only possible when data is preserved in a unified structure, and its deployment can not be technically challenging and has to be software independent.

RESEARCH GOALS

The preceding paragraphs make it quite clear, it is possible to preserve geodata for a long term but at high costs and effort. As the importance and usage of geodata, cartographic and GIS products is rising, another much easier and automated way has to be carried out and suitable tools developed. We are able to preserve geodata only as a deconstructed set which deployment is time consuming and expects some knowledge of the Geomatics and good skills in using GIS systems like QGIS, Grass GIS or ArcGIS too. Granularity of geospatial data plays an important issue and has to be considered. With the support of the European Commission, the Archiving by Design concept will be developed, which will promote the ability to archive data as a standard capability for all systems used in public administration or local government.[43] Its main purpose is not to prevent the problems associated with data migrations, vendor lock-in, but solely to archive data that has a clearly defined purpose and rationale. It is clearly an opportunity for all software developers and vendors to add these capabilities to the products they develop. However, these capabilities should certainly not be one-way only, i.e. for exporting data, but also for importing the DIP package so that the customer can process the data in an automated way to compare changes in the new and historical dataset. Archiving must become an elementary function of these systems.

While in the 1990s we saw significant decentralisation in the Czech Republic, we are now facing a reverse process due to the digitisation and automation of numerous administrative procedures. A substantial number of data producers will revert to being directly managed by ministries, posing a real risk of losing high-value data.[44] Many diverse information systems will need to be migrated to new solutions, and an equivalent number will be discontinued without replacement.

Consequently, it becomes crucial to seek solutions for Geographic Information Systems (GIS) that facilitate easy and cost-effective archiving of spatial raw data from a broad spectrum of production systems. This approach will ensure the reliability of these data during appraisal and transfer to digital archives and enable their future re-use.

CONCLUSION

Digital preservation is a multidisciplinary concept that combines policies, strategies, and actions to ensure access to digital content over time with archiving serving as one of its methods. Archiving originating from the humanities, ensures a contemporary context for the preserved data, while other methods like fixity, checksums, etc., are more technical and belong to digital science. Preservation is a fundamental part of data management, although it is often misunderstood or overlooked by data producers. This is because the benefits of digital preservation are usually realised in the long term, while the costs and efforts are typically considered only in a short term. However, as our reliance on digital data grows, so does the importance of digital preservation especially when considering fragility of digital data. This is the main reason why digital preservation is so important.

Digital cartographic and GIS products have become natural part of our life in recent times, making their archiving one of the most important tasks for digital archives.[45] In our reality of digital long-term preservation, the relationship between geodata producers and digital archives is crucial for ensuring the integrity, accessibility, and longevity of geospatial information. Geodata producers are tasked with generating and disseminating accurate, valid, and trustworthy data, often in compliance with regulatory standards and organisational mandates.[46] However, the responsibility for preserving this data over the long term presents unique challenges and considerations. The appraisal and geodata transfer to digital archives also pose significant challenges. When considering that geospatial information contains the majority of systems used by governmental agencies or by municipalities, archivists would expect that almost everything has to be managed as geodata, but that is not the case. A real geodata model like vector, raster, point cloud etc. must be strictly distinguished from common database data because their preservation requires special treatment and validation although databases are often one of its components. This situation generates high demands on both sides, archives and producers especially in the area of technical knowledge, skills, software and hardware, data throughput, data management, and strategies in digital archives. In summary, the relationship between geodata producers and digital archives is pivotal for ensuring the continued accessibility and integrity of geospatial information. By collaborating with digital archives or implementing robust internal preservation strategies, geodata producers can mitigate the risk of data loss and uphold the trustworthiness of their datasets for future generations.

REFERENCES

1) SYED IFTIKHAR HUSSAIN SHAH, VASSILIOS PERISTERAS, IOANNIS MAGNISALIS; DaLiF: a data lifecycle framework for data-driven governments. In Journal of Big Data (2021) 8:89; Available at https://journalofbigdata.springeropen.com/counter/pdf/10.1186/s40537-021-00481-3.pdf; Worth mentioning is that the construct of limitless data is quite close to Owen´s finding about limitless archival information. Please check following reference 2), p. 22-23

2) OWENS, TREVOR. The Theory and Craft of Digital Preservation; Baltimore, MD, Johns Hopkins University Press, 2018, p. 11-25

3) There are many references to Data Management like IBM available at What Is Data Lifecycle Management (DLM)? | IBM; Martin Ofner, Kevin Straub, Boris Otto, Hubert Oesterle; Management of the Master Data Lifecycle: A Framework for Analysis, available at (PDF) Management of the Master Data Lifecycle: A Framework for Analysis (researchgate.net) BHAKTI GALA, MANU T R; Growth and development of Research Data Management in India; Available at (PDF) Research Data Management lifecycle: an overview (researchgate.net)

4) EU Data Act available at Data Act | Shaping Europe's digital future (europa.eu) The EU's data management strategy aims to promote data accessibility and sharing, strengthening EU/EC control over data. It establishes a robust data governance framework, enabling data exchange and setting clear usage rules. This approach fosters innovation, enhancing EU competitiveness, and encourages the publication of government datasets, either freely or with certain limitations.

5) More information about EU Artificial Intelligence Act available at EU AI Act: first regulation on artificial intelligence | Topics | European Parliament (europa.eu)

6) Digital agency was established by legislation, Act n.471/2022 col., available at 471/2022 Sb., 1. 1. 2024, aktuální znění, informativní znění systému e-Sbírka (e-sbirka.cz)

7) Disposal Scheduling, British National Archives, 2004. It is a comprehensive definition of disposal schedule and other concepts helpful to the quality data management. Available at https://cdn.nationalarchives.gov.uk/documents/information-management/sched_disposal.pdf

8) As the importance of geodata preservation grows more attention is being given to that issue. Available at 050523_Programme_GeospatialDataPreservationConference2023_ENG_V30.pdf (dlmforum.eu); Microsoft Word - GeoForum_Prague_2022_Final_programme_01.09.22.docx (dlmforum.eu). In addition to quality data management, there is also a great threat of sophisticated cyber attacks that can irreversibly damage data or completely disable information systems. Available at Hackeři zašifrovali data Ředitelství silnic a dálnic. Profesionální útok, uvedl český kyberúřad | iROZHLAS - spolehlivé zprávy; Podcast s generálním ředitelem ŘSD ČR - Kybernetický útok na ŘSD - ŘSD s. p. (rsd.cz)

9) ISO 8000 is an international standard that focuses on data quality management. It provides guidelines and requirements for ensuring the accuracy, completeness, consistency, and reliability of data used within organisations. This ISO standard aims to improve data quality by establishing principles, processes, and metrics for data management practices. It covers various aspects of data quality, including data definition, validation, and maintenance, helping organisations enhance the value and reliability of their data assets. Available at https://www.iso.org/obp/ui/#iso:std:iso:8000:-1:ed-1:v1:en

10) Available at: https://www.iso.org/home.html,

11) European commission published several guidelines aiming on data quality management like: Data management - H2020 Online Manual (europa.eu); General - guidance/guidelines-recommendations-best-practices When aiming only at geodata a comprehensive information is offered by for example czech government; available at GeoInfoStrategie2020+ - Digitální a informační agentura (gov.cz) or by swiss Digitale Verwaltung Schweiz | Landingpage egovernment (digitale-verwaltung-schweiz.ch)

12) OWENS, Trevor. The Theory and Craft of Digital Preservation; Baltimore, MD, Johns Hopkins University Press, 2018

13) Format politics is a popular and sustainable strategy among Digital Archives. More available at File formats for transfer - The National Archives; Alphabetical List of Format Descriptions - Sustainability of Digital Formats | Library of Congress (loc.gov); dokumenty:narodni_standard_formatu_pro_archivaci.pdf (gov.cz); International Comparison of Recommended File Formats - Open Preservation Foundation

14) Basic information about National Archives of the Czech Republic available at National Archives (nacr.cz). Please see The Annual report of the National Archives for 2023, p. 18, too. Available at https://www.nacr.cz/wp-content/uploads/2024/02/VZ-NA-za-rok-2023_def.pdf

15) There exist two standards for digital preservation. ISO 14721:2012 Space data and information transfer systems Open archival information system (OAIS) Reference model available at https://www.iso.org/standard/57284.html and Magenta book, Recommendation for Space Data System Practices - Reference Model for an Open Archival Information System (OAIS) Reference Model for an Open Archival Information System (OAIS) (ccsds.org). Both standards have essentially almost the same content a they define a reference model for an Open Archival Information system but each was published by a different organisation.

16) Research Institute for Soil and Water Conservation is an important geodata producer in the Czech Republic. Institute runs own robust GIS Portal Geoportál SOWAC-GIS (vumop.cz) and its application Anti-Erosion calculator won United Nations award Protierozní kalkulačka získala cenu odborné poroty v Cenách SDGs | VÚMOP, v.v.i. (vumop.cz) Research Institute for Soil and Water Conservation | VÚMOP, v.v.i. (vumop.cz)

17) The Technology Agency of the Czech Republic is an organisational unit of the state and supports applied research and experimental development. Support is divided into many programmes. The Éta programme aimed on multidisciplinary cooperation, available at Program ÉTA - Technologická agentura ČR (tacr.cz)

18) More detailed information about the EARK Geospatial data as a content available at CITS Geospatial data (dilcis.eu)

19) Encoded Archival Description (available at https://www.loc.gov/ead /) is a modern standard which is now implemented by national guidelines. Czech EAD Profile is based on EAD version 3. In the Czech Republic is EAD implemented by guidelines Basic rules for describing archival records, available at https://www.mvcr.cz/clanek/metodiky.aspx?q=Y2hudW09Mw%3D%3D

20) Swiss specification was published in 2018 by the Federal Archives, more information available at Archiving of Geodata (admin.ch) Geo-SIP and Geo-Dossier specification available at https://www.bar.admin.ch/dam/bar/en/dokumente/kundeninformation/Spezifikation%20Geo-SIP%20und%20Geo-Dossier.pdf.download.pdf/Spezifikation_Geo-SIP_und_Geo-Dossier_V1.0_2016-12-05.pdf

21) A comprehensive information about The Dublin Core™ Metadata Initiative available at DCMI: About DCMI (dublincore.org)

22) ISO 19115-1:2014 defines the schema required for describing geographic information and services by means of metadata, available at https://www.iso.org/standard/53798.html

23) Archiving of either digital or analogue records is based on managing (used in Great Britain) or arranging (used in the United States of America) archival collections and making finding aids. For such purposes the Archive Administration of the Ministry of Interior initiated the development of specialised software with support from the Technical Agency of the Czech Republic. Available at https://www.mvcr.cz/clanek/software-elza.aspx and https://www.tacr.cz/elza-elektronicke-zpracovani-archivalii/

24) Although the ESRI Shapefile format, which is commonly used for storing geospatial vector data, there are several issues with this format, such as its lack of a standard coordinate reference system, the need for multiple files to store a single dataset, and limited support for attribute fields. These limitations make it difficult for users to effectively manage and share geospatial data. As alternatives to the Shapefile format members of the geospatial IT industry propose GeoPackage and FlatGeobuf. These formats are praised for their modern features and wide support in GIS software. They are presented as promising solutions that can address the shortcomings of the Shapefile format and better serve the needs of today's geospatial data users. Besides these two formats there are other potential alternatives like GeoJSON, GML and SpatiaLite. Each of these formats has its own strengths and is suitable for different scenarios in geospatial data exchange. As GeoJSON and GML are text based formats, they appear to be more suitable for long term preservation purposes. More information available at https://www.esri.com/content/dam/esrisites/sitecore-archive/Files/Pdfs/library/whitepapers/pdfs/shapefile.pdf; https://www.loc.gov/preservation/digital/formats/fdd/fdd000280.shtml; https://training.gismentors.eu/open-source-gis/formaty/vektor.html#shapefile https://www.nationalarchives.gov.uk/PRONOM/Format/proFormatSearch.aspx?status=detailReport&id=328https://www.nationalarchives.gov.uk/PRONOM/Format/proFormatSearch.aspx?status=detailReport&id=1017; https://www.cagi.cz/upload/documents/givs2015/cepicky_shapefile_must_die.pdf

25) The ArchiGIS tool is PHP web client Transformace digitálních prostorových dat pro účely trvalého uložení v digitálním archivu – NARP (nacr.cz)

26) The INSPIRE Directive is a European initiative aimed at establishing a Spatial Data Infrastructure (SDI) to support environmental policies and activities with potential environmental impact. It enables cross-border data sharing, public access to spatial information, and assists in policy-making. The directive involves 34 spatial data themes and is implemented by Member States' infrastructures. Key entities include the INSPIRE Maintenance and Implementation Group (MIG) and National Contact Points (NCPs), which coordinate the maintenance and evolution of the directive. The directive also adheres to the European Commission Privacy Policy. More detailed information available at Overview - European Commission (europa.eu)

27) Open, platform-independent and standards-based data format for geodata exchange, implemented as a SQLite database container. It was defined by the Open Geospatial Consortium (OGC), published in 2014. The rules and requirements for the SQLite container follow the GeoPackage Encoding Standard, which defines the schema, including table definitions, integrity, format restrictions, and content restrictions. Validation is possible. References available at https://www.ogc.org/standards/geopackage ; https://www.geopackage.org/ ; https://www.loc.gov/preservation/digital/formats/fdd/fdd000520.shtml; https://kost-ceco.ch/cms/geopackage; html https://www.cagi.cz/upload/documents/givs2018/30-givs-2018-cepicky.pdf; http://switchfromshapefile.org/; https://portal.nacr.cz/cro/transformace-digitalnich-prostorovych-dat-pro-ucely-trvaleho-ulozeni-v-digitalnim-archivu/katalog-formatu/#geopackage

28) An open, platform-independent format based on the XML specification defined by the Open Geospatial Consortium (OGC) for expressing geographic features. It is a modelling language for geographic systems and also an open format for exchanging geographic transactions and spatial data. It is based on the ISO 19136:2007 standard, created in 1998. It allows the description of schemas and datasets, and the selection of profiles for specific communities (for example application schema for weather, aviation, etc.). Its ability to integrate all forms of geographic information, including conventional vector or continuous objects, coverage and sensor data, is important. References available at https://www.ogc.org/standards/gml; https://en.wikipedia.org/wiki/Geography_Markup_Language; https://www.loc.gov/preservation/digital/formats/fdd/fdd000296.shtml ; https://is.muni.cz/th/p42lc/text_DP.pdf https://www.nationalarchives.gov.uk/PRONOM/Format/proFormatSearch.aspx?status=detailReport&id=1852; https://portal.nacr.cz/cro/transformace-digitalnich-prostorovych-dat-pro-ucely-trvaleho-ulozeni-v-digitalnim-archivu/katalog-formatu/#gml

29) List of Python packages used for automation: Pillow, pillow · PyPI Reliability of provided checks were validated against JHOVE GitHub - openpreserve/jhove: File validation and characterisation with satisfying results. PDF to PDF/A conversion is done using Open Source tool Ghostscript ghostscript · PyPI; EARK package structure is produced using Commons IP jar file from the

KEEP Team GitHub - keeps/commons-ip: Commons IP is project that provide a command-line tool and Java Library to validate and manipulate E-ARK Information Packages, so to create or process E-ARK SIP and AIP and also validate them against official specifications.

30) See reference 11)

31) Available at https://www.mvcr.cz/clanek/metodiky.aspx?q=Y2hudW09Ng%3d%3d and https://www.mvcr.cz/soubor/metodicky-navod-c-3-2022-narodniho-archivu-transformace-digitalnich-prostorovych-dat-pro-ucely-trvaleho-ulozeni-v-digitalnim-archivu.aspx

32) See reference 18)

33) See reference 17)

34) SIARDDK is a subvariant of the SIARD 1.0 format developed by the Danish National Archives. It is intended as a full SIP (Submission Information Package) that packages the associated documentation with a database object in SIARD format 1. SIARDDK is used for the long-term archiving of digitally created data, including relevant documents where appropriate. The format was inspired by SIARD 1.0 and it is used in Denmark, Greenland and the Faroe Islands to preserve valuable data from governmental agencies and Ministries IT systems in versions intended for archiving. If used for geodata preservation then vector data has to be exported using GML format. More detailed information available at DBPTK introduction (rigsarkivet.dk); SIARD (Software Independent Archiving of Relational Databases) Version 1.0 (loc.gov)

35) A software-independent, open, standards-based format for archiving relational databases that is used to permanently store relational databases and allows migrations between SQL Standard-based relational environments. The format was developed in Switzerland for the needs of the Swiss Federal Archives. It is a ZIP format container that contains tabular data, the necessary metadata in XML format and optionally extracts BLOB objects. References available at https://www.loc.gov/preservation/digital/formats/fdd/fdd000426.shtml; http://www.nationalarchives.gov.uk/PRONOM/Format/proFormatSearch.aspx?status=detailReport&id=2006; https://kost-ceco.ch/cms/siard_de.html; https://github.com/DILCISBoard/SIARD; https://github.com/keeps/dbptk-developer/releases; https://github.com/sfa-siard/SiardGui/releases;

36) RIGAUX, Philippe; SCHOLL, Michel a VOISARD, Agnès. *Spatial databases: with application to GIS*. 2nd ed. San Francisco: Morgan Kaufmann, c2002. ISBN 1-55860-588-6.; and SHEKHAR, Shashi a CHAWLA, Sanjay. *Spatial databases: a tour*. Upper Saddle River: Prentice Hall, c2003. ISBN 0-13-017480-7

37) PostGIS is a powerful extension for PostgreSQL database which enables its usage for spatial purposes. Available at PostGIS

38) Although PostgreSQL RDBMS is highly compliant to SQL Standard (available at ISO/IEC 9075-1:2023 - Information technology — Database languages SQL — Part 1: Framework (SQL/Framework)) popular function Overloading which is based on arguments extends PostgreSQL (available at PostgreSQL: Documentation: 16: 38.6. Function Overloading) is not compliant with SIARD Standard (see chapter 5.15, available at https://siard.dilcis.eu/SIARD%202.2/SIARD%202.2.pdf

39) The de facto digital preservation metadata standard is called PREMIS which means PREservation Metadata: Implementation Strategies (PREMIS). It is intended as requirements to ensure the long-term usability of digital data and to offer access to it in the future Available at https://www.loc.gov/standards/premis/

40) ZIP is a file format that supports lossless data compression. PRONOM | Search by format (nationalarchives.gov.uk); Alphabetical List of Format Descriptions - Sustainability of Digital Formats | Library of Congress (loc.gov)

41) TAR, Tape Archive format. It contains multiple files that have been packaged together for easier storage and sharing. TAR.GZ variant is Tape Archive format compressed using the gzip algorithm. Available at Tape Archive (tar) File Format Family (loc.gov)

42) For example Web Map Service available at Web Map Service - Open Geospatial Consortium (ogc.org) is a simple HTTP interface for requesting geo-registered map images from one or more distributed geospatial databases. OGC defined other web services like Web Coverage Processing Service, Web Coverage Service, Web Feature Service, Web Map Context (ogc.org), OpenGIS Web Map Tile Service Implementation Standard (ogc.org), Web Processing Service (ogc.org), Web Service Common (ogc.org) too.

43) It is one of the new principles of European information governance Available at e3cf4d38-ee41-42f9-8994-f6589ffad458_en (europa.eu) . It was defined by a working group of archivists from European archives. Its core comes from Netherlands approach called DUTO Available at (e-ark-foundation.eu);https://www.nationaalarchief.nl/en/archive/knowledge-base/archiving-by-design

44) Available at Zlepšení fungování, nebo boj o peníze? Vědci se bouří proti novele zákona o vědeckých institucích | iROZHLAS - spolehlivé zprávy; Zveřejňujeme znění návrhu novely zákona č. 341/2005 Sb. o veřejných výzkumných institucích | Věda žije! (vedazije.cz)

45) Available at eArchiving Geospatial Digital Records Preservation Conference (europa.eu), see reference 8) and most recently SNOW, Meagan A.. Preserving Geospatial Data, the Digital Preservation Coalition, 2nd Edition, ISSN: 2048-7916 http://doi.org/10.7207/twr23-01

46) The Czech Republic still faces problems related to wide availability or accuracy and the limitations that come with it. More information available at Projekty MV za 200 mil. Kč k modernizaci veřejné správy nevedly: úřady práci nezjednodušily, občané jim stále musí hlásit i to, co stát už dávno ví | NKÚ (nku.cz);

BIOGRAPHY

Martin Rechtorik, has a Master´s degree in History from Palacky University Olomouc and has been employed as an archivist in the National Archives since 2018. He is a member of the methodological team where he is responsible for databases and geo data preservation. Internationally, he is a member of the DLM Forum's Relational Database Archiving Group, NARA's Database Archiving Working Group, and the DLM Forum's Geoforum Geodata Archiving Group. In 2023 he started his PhD studies at the University of West Bohemia in Pilsen, Faculty of Applied Sciences, Department of Geomatics. His postgraduate thesis deals with long term geodata preservation issues and related challenges.