PRAGUE UNIVERSITY OF ECONOMICS AND BUSINESS

Acta Informatica Pragensia
2023, Volume 12, Issue 2, pp. 327–341
https://doi.org/10.18267/j.aip.219

Article Open Access

Digital Archives as Research Infrastructure of the Future

Michal Lorenz, Michal Konečný

Department of Information and Library Studies, Faculty of Arts, Masaryk University, Brno, Czech Republic

Corresponding author: Michal Lorenz (lorenz@mail.muni.cz)

Abstract

While a new paradigm of scientific research based on data centres and research infrastructures is gaining ground in science, and convergence between infrastructures and scientific domains is growing in cyberspace, epistemic cultures, particularly conservative in some fields, play a significant role in the dynamics of knowledge production in general and the adoption of data-intensive scientific practices in particular. In the present study, we focus on the transformations of scholarly communication through the perspective of digital curation of research data in the humanities, which certainly belong to these conservative epistemic cultures. The aim of this paper is to explore perspectives on the evolution of data curation in the context of the transformation of scholarly communication and research infrastructure in the humanities, specifically static archives, into living, continuously enriched data archives supported by artificial intelligence tools. To explore this perspective, we have chosen to compare scholarly communication in the humanities and in high-energy physics, in addition to analysing the practices of data curation itself. We further thematize the identified differences in terms of virtual research environments that can help humanities scholars exploit the potential of data-intensive research infrastructures.

Keywords

Artificial intelligence; Digital curation; Humanities; Knowledge; Scientific research.

Citation: Lorenz, M., & Konečný, M. (2023). Digital Archives as Research Infrastructure of the Future. *Acta Informatica Pragensia*, 12(2), 327–341. https://doi.org/10.18267/j.aip.219

Academic Editor: Adela Jarolimkova, Charles University, Czech Republic

1 Introduction

As the term data displaces the formerly overused concept of information, we can more clearly discern a shift in the perspective of scientific inquiry. Computational processing of information – a pioneering tool of modern scientific knowledge – is incorporated into large database systems that change the perspective of cognition. They integrate different areas of knowledge production and methods of knowledge. Cognition modelling, as an inspiring line of thinking about information technology, is giving way to modelling social and community influences on knowledge production and distribution. Beyond understanding how science works and the way we come to know, modelling enables the design of information space and technology interfaces, which entails new or innovative research methods and their implementation in database systems. Supported by the technical sciences, digital research methods and the rapidly emerging e-science are finding applications in the natural sciences and are penetrating quite rapidly into the social sciences. The slowest pace of digital research has been in the domain of the humanities, although pioneering work dates back to the 1950s (Schreibman et al., 2004). This uneven development of different disciplines and scientific fields is not new. Already canonically mentioned is the gap between the two domains of knowledge, between the culture of literary intellectuals and the culture of the natural sciences, first reported by C. P. Snow in 1959 (reprinted in Snow, 2014). Since its inception, scientometrics has explored interdisciplinary differences in the citation and use of scholarly sources (Murugesan & Moravcsik, 1978; Garfield, 1979). Cetina (1999) analysed disciplinary differences under the label of epistemic culture in the process of knowledge creation and communication in science, while Bates (2018) analysed data practices under the label of data culture). Data culture varies not only between disciplines, but also between different organizational units of disciplines, and can therefore be seen as part of the broader epistemic culture of a discipline.

Especially departments conducting digital research in the domain of the humanities rely on quantitative data processing based on the digitization of cultural heritage and the availability of such data in databases, digital libraries and digital archives. In the humanities, this creates tensions between researchers and departments that subscribe to different traditions of knowledge: traditional hermeneutic approaches and nomothetic approaches typical of the digital humanities. This also widens the gap associated with advanced information and computing technology skills and programming. Disciplines in the natural sciences domain with established research infrastructures adapted to work with big data, codified data flow processing procedures (data pipelines) and best practices for data management create virtual research environments with new types of documents and research objects, and open new frontiers of research and innovation (Hey et al., 2009; Candela et al., 2013).

We anticipate that these developments will also affect the e-infrastructure of the humanities, where these trends have been slower in taking hold. Thus, their custodians and curators need to reconcile the care of digitized resources aiming at the preservation and conservation of cultural heritage data with the practice of a living archive, that is not only expanded and enriched by new research data and connections to other data sources, but is also continually co-creating, negotiating, changing, in short, open in both access and open-endedness (Rudy, 2010; Hall, 2001). What implications will the living archive have for digital curation practices? To answer this question, we will trace a parallel between data practices in high-energy physics and digital curation practices in the digital humanities.

2 Data and Epistemic Cultures

Science, like many other areas of human activity, is undergoing a transformation brought about by information and communication technologies. The most significant is the involvement and exploitation of database technology in information systems applied in scientific practice. This process has implications for the synchronization and reconfiguration of social and natural ordering (Bowker, 2005), for the convergence between disciplines and domains (Barrios et al., 2019; Coccia & Wang, 2016) and between research communities (Pollock, 2022), and for the convergence of memory institutions (Duff et al., 2013) that constitute the documentary infrastructure of science. This contributes, in effect, to a transformation of the scientific method, that is, of how science is done. Increasingly, science is organized as e-science, a standard- and evidence-based collaborative research activity, grounding its practices in a research infrastructure that enables computational operations and sophisticated data processing. Building large research infrastructures is an expensive process. It requires large investments, which is one of the important factors that make science an important sector of national economies. The phenomenon was first noticed by Weinberg (1961), who referred to science using large research infrastructures as big science. Large research infrastructures then generate and process big data for scientific research (Borgman, 2015).

The shift of scientific methods to big data and databases embodies the fourth paradigm of scientific research, as Jim Gray explained in his presentation to the US National Research Council reprinted in Hey et al. (2009). Gray described the evolution of approaches to scientific inquiry. In the earliest times, scientists relied on observation and direct interaction with the phenomena under study, often in the form of experiments. Empirical science has an immediate, and thus relatively limited, range of experience. A paradigm shift in scientific inquiry was signalled by Kepler's laws of planetary motion. Kepler discovered them through analysis of archived astronomical observations in a catalogue carefully prepared by Tycho Brahe. Systematization and accumulation of observations allowed data analysis and generalization in the form of models and theories. The newly emerging theoretical science sought to understand the laws of nature (e.g., Newton's laws, Maxwell's equations). The variety and quantity of data, together with the complexity of the models, gradually began to cause complications in the analytical processing. The computer entered scientific practice, opening the way for another form of research. A third paradigm of scientific research emerged. Thanks to computers, complex phenomena could be simulated, thus transforming theoretical science into computational science. The first paradigm led science to the generation of data, the second to its generalized theorization through archiving, and the third paradigm used computational modelling. This gradually created complementary activities that contribute to excellent science: experimentation and analysis, modelling and theorizing, programming and predictions. As science continued to strive to understand reality, it continually moved away from the objects of its interest; the mode of knowing became fragmented into multiple roles. The same is true of the fourth paradigm of scientific inquiry, which moved even further away from the phenomena under investigation, but sought to unify experimentation, theorizing and simulation. The new paradigm is a paradigm of data-intensive science. Data come from various generators: from simulators and supercomputers, from specialized scientific instruments such as telescopes, CCDs, accelerators, sequencers, etc., from sensors and sensor networks or from the Internet and social networks, from digitization technology and OCR applications. The data are processed by software and stored in data centres and research infrastructures, which are responsible for their management and possible enrichment with additional metadata. It is only at this stage that scientists enter the research process by data analysis using quantitative and statistical methods and using the functions of the knowledge organization and data management systems. Scientists are therefore very far from the phenomenon observed. They do not interact with the phenomenon directly in the field, they are often not even in contact with the phenomenon indirectly through instruments. They investigate the phenomenon in a mediated way, through their computer, through data from processors, which is made possible by a complex production data assemblage (Kitchin & Lauriault, 2018). Not only data archiving, data management and knowledge organization gain importance in such conditions. New disciplines that contribute to data capture and data processing are emerging, such as the domain of sensor networks, data science and new kinds of computer science that interdisciplinarily connect and apply information science and computer science to a specific research field (Dalrymple, 2016; Fourman, 2003). Data practices are gradually spreading to fields for which data analysis and data mining methods have not traditionally been a subject of interest, often not even accepted, as in many humanities disciplines. However, as digitization progresses and digital culture take shape, we are seeing an increasing use of digital research in the humanities and arts as well. Here too, research infrastructures are taking shape (e.g., LINDAT/CLARIAH-CZ or the Archaeological Information System) and the research domain of digital humanities is developing. Small data domains are moving towards big science, infrastructures, standardization and new forms of scientific communication and collaboration.

As tools change, so does communication in science. New journal publishing models, new communication channels and new forms of communication are emerging. However, their adaptation and use are strongly influenced by different epistemic cultures. Epistemic cultures are "those amalgams of arrangements and mechanisms […] which, in a given field, make up how we know what we know. Epistemic cultures are cultures that create and warrant knowledge …" (Cetina, 1999, p. 1, emphasis original). As Cronin pointed out, "conventions favoured by different epistemic cultures will have a bearing on how ICTs, specifically electronic publishing technologies, are adopted and co-opted." (Cronin, 2003, p. 8) The process of knowledge making is shaped by differences in various domains, such as local rules and principles collectively applied by groups of experts. Cronin considered the use of media, reviewing and publishing practices, including genres and subgenres of academic writing, outcomes recognized in the career development system, disciplinary discourse and the rhetorical construction of arguments, self-archiving, credibility criteria and various forms of authorship. Data culture also plays a role, the spectrum of which is bounded on the one hand by scholars who do not acknowledge any knowledge that is not based on generalizing quantitative data, and on the other by researchers who distrust data and consider the description of particularities and the understanding of uniqueness as the basis of knowledge.

Physics, for example, emphasizes distributed collaboration, publication of preprints in archives and repositories, co-authorship of online papers, and the fastest possible process for publishing and distributing the results of scientific research. A set of papers to which an expert has only contributed can be submitted as a habilitation paper. It can be said that trustworthiness is built socially in physics. In the humanities, on the other hand, individual authorship prevails, trustworthiness is institutionally based, and monographs are held in high regard, especially if they are published by reputable publishers. They are also an outcome that is evaluated at habilitation, the time span of the sources of a scholarly paper is not limited to the most recent production, but spans decades, which correlates with the slow pace of knowledge production and the acquisition of expertise. With the digital turn, however, a number of new practices are pervading the humanities, widespread in faster-moving, technology- and data-driven domains. Traditionally closed archives converted to digital form are being opened up, enriched, intertwined, and become part of a broader, data-driven journey of knowledge making. This opens questions about not only the adaptation of curatorial practices to new conditions, but also the incorporation of new modes of scholarly communication, including data sharing and distributed cognition of epistemic cultures (Giere, 2002).

3 Digital Curation of Research Data

The idea of the researcher as a co-creator of data has become increasingly important in the last two decades. The fact that researchers are not just passive consumers of data, but active creators and stewards of them, is becoming increasingly apparent. In the past, researchers often relied only on data they collected themselves or on data available in closed institutional repositories. Advances in technology, together with the spread of open data and open science principles in the scientific community, have not only opened the door to new ways of data acquisition, but have also enabled the discovery and use of data acquired by others, including data in the domains of other research disciplines. Increasingly, there is a convergence of research communities without any examination of the role of the data itself in this process, and the ways in which data are used in digital repositories (Pollock, 2022). The activities of repositories fall within the field of long-term preservation of digital information, which is an established discipline, and institutions such as the Digital Curation Centre, the Digital Preservation Coalition, and the National Data Stewardship

Alliance are systematically engaged in the development of digital curation and its practices. The concept of data stewardship has been adopted as a key component of data governance for the practical application of curatorial principles to the protection of research data held in closed as well as open repositories. It is a set of "the people, organizations, and processes needed to ensure that the appropriately designated stewards are responsible for the governed data." (Plotkin, 2021, p. 3) When we talk about research data in living archives, we are referring to data located in open repositories where data are accessible to different types of users and systems across different research communities. Whether these are institutional repositories or repositories within research infrastructures, the ways users interact with them bring new opportunities and also new challenges. Users of these repositories use the data for their own purposes, but they can also participate in adding value to the data in ways that Beagrie (2008, p. 5) described in the context of digital curation of research data: "... added value is derived from annotation, linkage, and the management, validation, and editorial input of domain specialists..."

How does the requirement for systematic, targeted and responsible data stewardship relate to the idea of open access and sharing of data produced by researchers themselves? The answer to this question is not simple, for several reasons. Data governance is applied in the context of research data management (RDM) by institutions and individuals to ensure that the investments associated with the acquisition and reuse of research data are protected. Thus, descriptions of RDM in varying degrees of detail can be found on the websites of a number of institutions, many of which are exemplified by the University of Edinburgh (Why do research data management?, 2022), Monash University (Research Data Management at Monash, 2022), or in the Czech Republic by the Charles University (Výzkumná data, 2022). However, as Birkbeck et al. (2022) pointed out in their review of the literature on this topic, there is significant inconsistency in the conceptualization and implementation of RDM, and its requirements are often implemented by researchers who have neither the knowledge nor the necessary support for the task. For a change, Lefebvre et al. (2018) highlighted the absence of standards for research data management.

Another challenge is the requirement to adhere to the "FAIR" principles of reusability of research data (Wilkinson et al., 2016) that are often associated with RDM. FAIR principles are useful not only for stakeholders in the conventional sense, i.e., researchers, publishers or software developers, but also for applications and software tools themselves, which are referred to as computational stakeholders. However, their implementation requires additional knowledge and skills that researchers would have to acquire in addition to their own expertise. RDM practices very often rely on data management plans (DMPs). Smale et al. (2018) concluded from extensive research conducted among Australian institutions that although data management plans are considered a central element of RDM and their elaboration is often a prerequisite for the awarding of research funding, their benefits to researchers and funding institutions are in fact negligible and often represent instead an additional administrative burden. The reusability of data and their exchange and sharing in practice faces a number of difficulties related to both the technical aspects of management, sharing or retrieval and the issue of proper understanding of context. These difficulties are exacerbated when the data are outside the researchers' own domain of expertise. In such cases, it is necessary to use a unified and simplified terminology and avoid technical jargon to maintain proper understanding and context of the data (Pollock et al., 2022).

What options do researchers have if they want to meet their commitments to data management, sharing and interoperability with a clear conscience, while at the same time dealing with the obstacles that are placed in front of them? One possible solution is Repository as a Service. This concept was described by Lewis et al. (2012). In their paper, they discussed the future of digital repositories, especially community repositories. In doing so, they referred to the already established and nowadays growing trend of providing IT solutions in the form of third-party services: repository as a service is, according to them, a specialized variant of platform as a service. Another opportunity is offered by research data management services provided by libraries. Librarians collaborate with researchers and assist them in areas related to the research data lifecycle. However, they encounter a lack of knowledge and experience in RDM (Mushi, 2021). If, for various reasons, neither of these options is satisfactory, then methodologies and guides such as The Turing Way Guide (The Turing Way Community et al., 2022) or the UK Digital Curation Centre's website on creating data management plans (Data Management Plans, 2022) can be helpful for those who decide to go their own way.

Reproducibility and replicability are key concepts that define the way we think about sharing research data within and across domains. However, these terms are not used consistently across scientific communities, as Plessner (2018) demonstrated. Experimental sciences approach their interpretation differently, and different interpretations are used by computational sciences. In the computational sciences, Claerbout and Karrenbach understand reproducibility as "running the same software on the same input data and obtaining the same results" and, complementary to this, Rougier et al. understand replicability as "writing and then running new software based on the description of a computational model or method provided in the original publication, and obtaining results that are similar enough . . . " (both cited in Plesser, 2018, p. 1). However, these definitions are at odds with the established terminology used by the Association for Computing Machinery in the experimental sciences, which inverts the meaning of the terms (Plesser, 2018). The Turing Way Guide to Reproducible Data Science offers the following demarcation, extended with two other related terms: robustness and generalizability (The Turing Way Community et al., 2022):

- **Reproducible** are results "when the same analysis steps performed on the same dataset consistently produces the same answer".
- **Replicable** are results "when the same analysis performed on different datasets produces qualitatively similar answers".
- **Robust** are results "when the same dataset is subjected to different analysis workflows to answer the same research question […] and a qualitatively similar or identical answer is produced".
- **Generalisable** are results formulated by combining replicability and robustness.

One of the key principles of a curatorial approach to data protection is to preserve the trustworthiness and integrity of the data. The fluid nature of research data in living archives makes adherence to this principle more difficult than it is for more static digital documents. Quite subtle and hardly detectable changes in the content of datasets can have a crucial impact on reproducibility and replicability. Open science repositories dedicated to preserving research data, such as Zenodo (https://about.zenodo.org/policies/) or Dryad (https://datadryad.org/stash/faq), allow versioning of published datasets. Thus, it is essential to determine what changes are major and what changes are only cosmetic (minor) and to number the versions appropriately according to a defined policy. In combination with the checksums that repositories automatically generate for all datasets, researchers and authors of publications can link directly to specific versions of datasets to verify their integrity. Using checksums in UNF format is also a practice recommended by the Digital Curation Centre for citing datasets (Ball & Duke, 2005). The requirements for fulfilling the FAIR principles and implementing DMPs place demands on researchers that they are not always able to meet. Ensuring sharing, interoperability and reusability creates a demand for the role of a "data mediator" headed towards information professionals and data management organizations. They can help research teams share the context of their data before and after the publication of results. They can also help locate data from nearby disciplines that are housed in distributed repositories, provide training in the use of tools, data management and data analysis (Pollock et al., 2022). The importance of information professionals supporting researchers in managing research data was also mentioned by Lefebvre et al. (2018). In 2017, they conducted case study research, which included analyses of 13 selected RDM regulations of Dutch universities and interviews with 22 respondents. Based on the results, two professional profiles were identified to facilitate successful research data management. The first are governance supporters who assist researchers in open science. The second profile is research supporters, which include data managers and data stewards.

4 Digital Curation: Two Epistemic Cultures

If differences between various domains with distinct epistemic cultures are manifested in publishing technologies and at the same time a convergence of research communities in the electronic environment is taking place, trends in the technological implementation of scientific communication of traditional big sciences should be adopted by the sciences that are gradually establishing themselves as big sciences thanks to the electronic environment. However, given the slower pace of development, it can be assumed that many tools have not yet been adopted and their future implementation will be subject to modification in accordance with the evolving epistemic preferences of the domains. In the following section, we will focus on a comparison of two domains, representing traditional big science and a domain newly emerging as big science. In the first case, we will focus on theoretical high-energy physics and its arXiv platform, recognized for its contribution to the development of scientific communication (e.g., PAM Award, 2017; Einstein Foundation Award, 2021), and on the domain of digital humanities, in which a number of fields based on digital methods and data handling are emerging (Lorenz, 2022).

A high degree of collaboration is a necessity in high-energy physics. Experiments use expensive instruments that are rare. To exploit the full potential of these facilities, collaboratories are being set up. These virtual research environments (VREs) allow control and management of the instruments of a physical laboratory through an interface over the Internet (Finholt, 2002). The collected data are stored in a research infrastructure repository, which often also provides a platform for communication among researchers, including the publication of their results. The arXiv project is a successful example of a repository provided as a service and at the same time an example of an environment that helps converging scientific communities in activities related to sharing research data and information not only in theoretical high-energy physics. It is a community archive dating back to 1991. It also focuses on other areas such as mathematics, computer science and electrical engineering. Its derivations are also penetrating into other technical and natural sciences. During the COVID-19 pandemic, it permeates especially biology and medicine. Since 2001, the project has been operating under the auspices of Cornell University (Ginsparg, 2021).

ArXiv has become a platform for rapid communication of scientific research results in the form of preprints. The quality of papers is checked automatically through machine learning, and potentially problematic papers are further reviewed by humans. Expert comments can be further evaluated before the paper is published in a peer-reviewed journal. Principles of digital curation are applied in the management of research data, with a strong emphasis on enhancing the value of the information held. This can take the form of, for example, creating organized collections, providing metadata to documents, or finding and describing relationships in the stored data. The focus of the arXiv platform on publishing preprints primarily defines the type of data stored within the archive as scholarly papers in the form of documents. It is essentially a homogeneous open repository to which registered users can contribute. Contributions are not subject to peer review but are assessed ("moderated") after submission for completeness, correct classification, form and other criteria. Submissions are accepted in PDF, HTML and TeX/LaTex formats, with the latter being the preferred format. Digital objects in the archive are structured internally in a manner common to scholarly works. Thus, in addition to the actual text that is the subject of the work, they contain the usual content elements such as an abstract, annotations and a bibliography. Although these content elements are not formalized in any way and their specific implementation is up to the author, they enable analytical and data operations used to extract metadata and relationships that are further used by the portal. Documents can be supplemented by other files such as image attachments, datasets, source code or compiled applications. Papers are organized into basic subject categories (e.g., Physics, Mathematics), which are further subdivided into subcategories. Beyond this categorization, arXiv does not work with any additional taxonomy that would, for example, allow searching with a controlled vocabulary or the creation of collections. Instead, it offers, through additional services and integration modules, linking based on content analysis (using machine learning technologies) and the datasets used. As part of the submission, authors provide metadata for the documents; author(s), title and abstract are required. Optionally, other data may be added, including DOI identifiers (if available). Links to related documents in the archive, to scholarly works on the same topic, to datasets and machine learning models, or to source code are created automatically by the archive infrastructure but can also be added directly by users.

Although the main focus of arXiv is the publication of scientific preprints and papers, i.e., the open repository services, this is not its only activity. The arXivLabs framework allows archive staff and community members to create tools that extend the capabilities beyond conventional repositories. One such tool, for example, is linking published papers to the Papers with Code portal, which offers free machine learning resources. With this link, arXiv users have immediate access to the source code and datasets associated with published papers. Another extension of arXiv offers access to related papers through the Connected Papers service, which, by analysing tens of thousands of papers, creates a power-driven graph that visualizes relatedness not only based on direct citations but also on mutual similarity (Eitan et al., 2022). The Litmaps extension is a tool for retrieval and visualization of bibliographic relationships between papers and also enables team collaboration (Discover the World of Scientific Literature, 2022). The aforementioned services and tools (and many others) are integrated directly into the arXiv portal site using the arXivLabs framework, making them immediately usable.

Application of value enhancement principles can also be found in research infrastructures in the digital humanities. They connect users, curators and providers of information content and provide a technical framework for advanced sharing. Analogous to research infrastructures in physics or astronomy, which give their users access to technologies that would otherwise be completely out of their reach, users of the DARIAH infrastructure can use aggregated content from a large number of sources and search it using indexed metadata to describe text or music records (Heinrich & Gradl, 2013). Another European infrastructure, CLARIN, is linked to DARIAH through collaboration and offers access to a large number of language corpora and processing tools. The two infrastructures are merging in a number of countries (this connection then bears the name CLARIAH), and their collaboration has given rise, for example, to the PARTHENOS project, which offers as one of its outputs a virtual research environment, a communication channel we have already encountered in high-energy physics.

For the purpose of comparison, the LINDAT/CLARIAH-CZ platform was selected as the representative of archives in the digital humanities. It is a digital research infrastructure for language technology, arts and humanities created as part of the DARIAH research infrastructure, with the mission of providing open access to data in the humanities and providing tools and services for handling these data. The repository accepts language data and/or NLP data and tools: corpora, annotated corpora, dictionaries, but also trained language models, parsers, taggers, machine translation systems, language web services, etc. It is possible to insert only metadata that link to an online resource. This is an open repository to which users can contribute by logging in with institutional or private accounts. As with arXiv, contributions are reviewed by editors who check the quality and completeness of the metadata, the consistency of the files and the intellectual property rights. Unlike arXiv, the data stored in the LINDAT/CLARIAH-CZ repository are fully heterogeneous and include digital objects of different types. Text corpora are most often represented, but audiovisual objects, image objects or datasets can also be found. Both simple objects consisting of a single file and complex objects are represented. This diversity does not allow unambiguous definition of the data structure. A basic classification divides data according to their author, subject and language. The repository allows advanced searching using controlled vocabularies that include, for example, different types of licenses or a selection of collections relevant to the research community. In this respect, the repository shows a greater degree of organization than arXiv. The repository allows much more complex metadata handling than arXiv. This is due to the variety of content, which is reflected in the use of specific metadata; for example, unlike arXiv, there are metadata containing external links or object type specifications. The architecture solution relies more on relationships among objects in the archive identified through their metadata and does not use third-party services or content analysis to create links, but instead activates the scientific community to participate in identifying relationships and links among data.

In many ways, the two repositories are very similar – both allow content to be uploaded by members of the scholarly community, and the content is edited to some extent and basic curatorial principles of selection, evaluation and preservation are applied to it. Both repositories also emphasize the long-term and sustainable availability of their content and the persistence of identifiers. The most significant difference can be seen in the way the repositories approach the representation of data and the relationships among them. LINDAT/CLARIAH-CZ offers a much more structured approach driven by the repository itself, whereas arXiv relies on the use of third-party platforms and services. A fundamental difference is the extent to which research data are used. While the amount of digitized evidence sources in the humanities is growing and there are more and more research teams whose outputs include research datasets, the usage rate is not staggering compared to the arXiv repository. The technological developments are in many ways outpacing the skills of users. For many researchers in the humanities, the digital archive is primarily a ready source of documents for close reading rather than a source of textual data for further computational processing. However, this state of affairs may not be related to a resistance to technology, but to a data culture manifested by a lack of data literacy that creates barriers on the part of humanities-oriented professionals. Overcoming them requires making data operations accessible so that they can be easily integrated into the scholarly practices of humanities scholars. This can be achieved precisely by adapting virtual research environments.

5 Virtual Research Environment in the Role of a Data Professional

Virtual research environments can be understood as virtualized laboratories that mediate access to infrastructure and provide frameworks and user interfaces specifically designed to support activities within particular research teams or communities. Thus, it is not just virtualization that consists of moving computation to the cloud (Allan, 2009). VREs put into practice the idea of collaboratories envisioned by computer and information science visionary W. A. Wulf in the late 1980s: "a centre without walls, in which researchers can perform their research without regard to physical location – interacting with colleagues, accessing instrumentation, sharing data and computing resources, and accessing information in digital libraries" (cited in Finholt, 2002, p. 77). The PARTHENOS project relies on the so-called Backbone Thesaurus, a minimal metadata set used for searching, and the PARTHENOS Entity Model, a data schema that allows easy mapping of data used across different cultural heritage organizations. Registered users have access to the federated datasets with the possibility to search through them. The virtual lab is equipped with natural language processing and text analysis tools and a media management service. The virtual environment is hosted within the D4Science data infrastructure, which provides facilities for more than 175 VREs. Another example of a VRE in the DARIAH infrastructure is the TextGrid (Neuroth et al., 2011), built by ten research institutions in Germany for research communities of philologists, linguists, musicologists, art historians and classical scholars. It includes a lab with various tools and services using the TEI standard for editing, searching, glossing and linking data, repository services for storing and publishing data and critical editions, tutorials for training, and mailing lists for community communication. In addition, NoteEditor allows editing, visualization and processing of notations in the MEI standard. For more detailed computational analysis of music, there is the Digital Music Lab, an infrastructure including databases with audio data, distributed in collaboration with several partners such as universities, the British Library and I Like Music. The infrastructure provides a virtual research environment, integrating audio data with semantic web technologies, which researchers can interact with in a user interface that allows querying, exploring and visualizing datasets (Abdallah et al., 2017).

Tools for handling data that are processed in the cloud are a major benefit of virtual research environments. In this way, the researcher does not have to store the data on their computer before further processing, whether by implementing a tool specific to a particular research environment, or more general tools available as SaaS or open source. An example of such a tool is NameTag (Straka & Straková, 2019), which uses models trained on language corpora to perform named entity recognition (NER) on submitted text. It is available as part of the LINDAT/CLARIAH-CZ digital research infrastructure and as a web service with a REST API, making it available for use in other systems. From these tools, but also from the results of their application, the interface of a virtual research environment can be composed. That makes it easier to use the data, shortens many of the sophisticated data handling steps and makes data navigation easier. Named entity recognition can then be used in practice, for example, for extracting geographic information from text and in combination with other geographic tools for geocoding (marking referred locations and areas on maps).

In addition to dedicated research infrastructures that offer VRE for a specific area of scientific research over relevant datasets, there are also general-purpose online collaborative environments designed for teams in both academia and the commercial sphere. The Nextjournal service (https://nextjournal.com/) allows its users to collaborate in a shared way on scientific publications or machine learning projects. It is possible to embed source code in shared documents and run them directly in a virtualized environment or to perform complex computations in R. The SHARE (Sharing Hosted Autonomous Research Environments) portal allows reproducing research results published in papers, be it numerical computations, diagramming or automatic theorem proving (Van Gorp & Mazanek, 2011). The Observable service (https://observablehq.com/) offers the possibility to create shared "notebooks" to which one can attach one's own datasets, create searchable databases from them and visualize the data. The ResearchSpace platform (https://www.researchspace.com/) is an example of a comprehensive environment for collaborative research teams and offers, among others, the possibility to create workflows, integrate with cloud tools or storage services in compliance with FAIR principles. myExperiment Virtual Research Environment even allows sharing, composing and reusing workflow components (Roure et al., 2009).

VREs simplify data access and use. They can be a good entry point for humanities scholars who do not have computationally suitable data for their research and rely more on idiographic methods. However, digital data processing methods offer them the opportunity to explore their topics from new perspectives, but here researchers often face the barrier of poor data literacy. Data literacy in this context is understood as "a set of skills and abilities related to accessing research data, understanding, interpreting, managing, critically evaluating, and ethically using it" (Koltay, 2019, p. 14). VRE is particularly helpful in the competencies to acquire, process and manage data, which precede the use of data. It offers access to already prepared data and at the same time tools for data analysis, helping overcome this barrier. As the researcher's point of contact with the data, the virtual research environment in the data pipeline is situated at a higher level of data processing. However, this increases the requirement for higher skills in critical data evaluation, which in some cases may be impossible without thorough knowledge of the data and structure of their processing. Researchers should be able to recognize this situation. Some experts express concern that scientists who do not know their data are unable to do high quality research. However, scientists often do not know the origin of their data, only the formalized context captured in metadata. Scientists commonly use data from data repositories and research infrastructures; working with non-native data is a characteristic of statistics. Moreover, what are one person's research data are secondary data to another, and metadata to the third. Researchers are involved in the preparation and use of data at various stages of their processing, but this does not diminish the weight of their research. We believe that the same is true for humanities scholars working with non-original and preprocessed data.

The trend of deep neural networks and machine learning in general is also essential for the future use of VRE as an access gateway for humanities domain experts. The use of artificial intelligence can help create VREs tailored to the needs of individual researchers as well as to perform various operations on data. With the increasing use of machine learning-related techniques, shared pre-trained neural network models are gaining importance. These serve as a starting point for further work, while significantly saving the time and resources that would be required to create one's own models. A variety of models are currently available for image, audio and natural language processing. One of the problems that researchers may encounter when trying to apply machine learning techniques is the insufficient size of the datasets under investigation. Musicologists researching an area of their field that is very narrowly defined, for example, in terms of time or region, may struggle to obtain the volume necessary to build a deep neural network. Pre-trained models can be applied to other types of data than those for which they were originally developed. Hung et al. (2022) describe techniques for so-called neural model reprogramming, where existing speech processing models are used to classify genres in low-resource music datasets that are not large enough on their own to be used for building own models.

But what about cases where users do not understand the structure and limitations of datasets or cannot select and use the appropriate tool – situations where the role of a data professional or mediator comes into play? It is very likely that this role will be taken over in the future by virtual assistants communicating with users in the form of natural language. Already now, the ChatGPT demonstration language model created by OpenAI is able to search for local names in a given text based on textual instructions, summarize the content of complex articles, simulate virtual computers or provide functional programming code demonstrations. The ChatGPT virtual assistant cannot evaluate the veracity or relevance of information on its own and often presents information that is incorrect or misleading in a very convincing manner. However, it can be used in combination with other systems. Stephen Wolfram (2023) presents a possible way of such a combination and demonstrates how ChatGPT can be "forced" to use data obtained using the Wolfram Language in its responses. ChatGPT thus forms a "user interface" between human and computing machine, allowing the correct answers to questions posed in natural language to be obtained and further manipulated. Artificial intelligence will also be used to test the credibility of findings, for example in the automatic replication of data analysis results. It cannot be assumed that data professionals will be replaced by GPT technology. The intellectual work that a researcher does when processing text or musical scores in TEI and MEI standards will not be done by AI instead, nor will the generation of ideas for new features or methods of data analysis. However, the application of GPT in VRE as a point of contact for many humanities scholars' work with data will significantly affect the spread of digital methods and data practices of that epistemic community. It is only a matter of time before virtual assistants are a common component of complex systems such as virtual research environments, in which they will make advanced features and tools available to users. Indeed, the demonstration assistant itself, built on the ChatGPT model, answers the question "What are the opportunities for virtual assistants like you in VRE?" (Output from free research preview of ChatGPT Dec 15 Version, https://help.openai.com/en/articles/6825453-chatgpt-release-notes):

> *Virtual assistants like me can be used in VRE for a variety of tasks. Here are some examples of possible applications of virtual assistants in VRE:*
>
> 1. *Helping with project organization and management: virtual assistants can help research teams with the organization and management of their projects, for example by helping with planning and assigning tasks, monitoring project progress, and providing regular updates.*
> 2. *Data processing and analysis. Virtual assistants can help automate routine tasks related to data processing and analysis, which can save time and increase the efficiency of research teams.*

3. *Collaboration and communication: virtual assistants can help with organization and coordination of collaboration among members of the research team and with provision of information and updates about the project to other team members.*
4. *Providing expert advice: virtual assistants can provide expert advice and help with finding relevant resources and information for research projects.*

These are just a few examples of possible uses of virtual assistants in VRE. Virtual assistants can be used for a variety of other tasks depending on the specific needs and requirements of individual research teams or organizations.

6 Conclusion

So, what are the perspectives for the development of data curation in living humanities archives in the dawning age of artificial intelligence? Data curation in humanities is evolving in several ways. One major trend is the development and use of machine learning algorithms to automate curation processes over data, which are much more prevalent in high-energy physics. Here we can see inspirations from curation practices widespread in the natural sciences. These algorithms can help identify and extract meaningful information from large datasets such as text, images and audio files, complementing the manual verification and cataloguing of data typically involved in curating humanities data. At the same time, it is evident that data literacy is growing in importance. Other curatorial practices are already characteristic of the humanities domain, and we find little inspiration from the epistemic community of high-energy physics represented by curatorial practices in arXiv.

An important trend brought about by living archives is the increased emphasis on community curation, where scholars, curators, archivists and other stakeholders collaborate to curate and annotate data. This approach enables a more diverse and inclusive set of perspectives to be included in the curation process, which can help ensure that the archive is comprehensive and representative of different cultural and historical contexts. With the penetration of artificial intelligence into archives and curatorial practice, the importance of ethical aspects of data curation, particularly when working with sensitive or controversial materials, is increasingly recognized. This includes issues around privacy, consent and bias, which need to be carefully addressed to ensure that the archive is trustworthy while respecting the rights and dignity of the individuals and communities represented in the data. It remains an open question whether the curator or the researcher should have the final say in the inclusion of data sources and the description of conceptual relationships. Overall, data curation in AI-powered living humanities archives is evolving towards more automated, collaborative and ethical approaches that aim to maximize the accessibility, diversity and quality of archived materials.

Additional Information and Declarations

Acknowledgement: The work utilized the LINDAT/CLARIAH-CZ research infrastructure (funded by the Ministry of Education, Youth and Sports of the Czech Republic in programme The Large Infrastructures for Research, Development and Innovation, project No. LM2023062).

Funding: This work was funded by the Ministry of Culture of the CR, Programme NAKI II: Applied research and development of national and cultural identity (2016–2020), Grant No. DG18P02OVV03 New Phonograph: Listening to the sound of history.

Conflict of Interests: The authors declare no conflict of interest.

Author Contributions: M.L.: Conceptualization, Writing – original draft, Supervision, Writing – review & editing. M.K.: Writing – original draft, Validation, Resources.

References

Abdallah, S., Benetos, E., Gold, N., Hargreaves, S., Weyde, T. & Wolff, W. (2017). The digital music lab: A big data infrastructure for digital musicology. *Journal on Computing and Cultural Heritage*, 10(1), 2. https://doi.org/10.1145/2983918

Allan, R. (2009). *Virtual Research Environments: From portals to science gateways*. Chandos Publishing.

Ball, A. & Duke, M. (2015). How to Cite Datasets and Link to Publications. *DCC How-to Guides*. https://www.dcc.ac.uk/guidance/how-guides/cite-datasets

Barrios, C., Flores, E., Martínez, M. A., & Ruiz-Martínez, M. (2019). Is there convergence in international research collaboration? An exploration at the country level in the basic and applied science fields. *Scientrometrics*, 120, 631–659. https://doi.org/10.1007/s11192-019-03133-9

Bates, J. (2018). Data cultures, power and the city. In Kitchin, R., Lauriault, T. P. & McArdle, G. (Eds.) *Data and the City*, (pp. 189–200). Routledge.

Beagrie, N. (2008). Digital Curation for Science, Digital Libraries, and Individuals. *International Journal of Digital Curation*, 1, 3–16. https://doi.org/10.2218/ijdc.v1i1.2

Birkbeck, G., Nagle, T., & Sammon, D. (2022). Challenges in research data management practices: a literature analysis. *Journal of Decision Systems*, 31(sup1), 153–167. https://doi.org/10.1080/12460125.2022.2074653

Bowker, G. C. (2005). *Memory Practices in the Sciences*. MIT Press.

Borgman, Ch. L. (2015). *Big Data, Little Data, No Data: Scholarship in the Networked World.* The MIT Press.

Candela, L., Castelli, D., & Pagano, P. (2013). Virtual Research Environments: An Overview and Research Agenda. *Data Science Journal*, 12, GRDI75-G RDI81. http://doi.org/10.2481/dsj.GRDI-013

Cetina, K. K. (1999). *Epistemic Cultures: How the Sciences Make Knowledge*. Harvard University Press.

Coccia, M. & Wang, L. (2016). Evolution and convergence of the patterns of international scientific collaboration. *Proceedings of the National Academy of Sciences*,113 (8) 2057–2061. https://doi.org/10.1073/pnas.1510820113

Cronin, B. (2003). Scholarly communication and epistemic cultures. *New Review of Academic Librarianship*, 9(1), 1–24. https://doi.org/10.1080/13614530410001692004

Dalrymple, P. (2016). What's in a Name? A Brief History of Informatics Education. In Seadle, M., Chu, C. M., Stöckel, U. & Crumpton, B., *Educating the Profession: 40 years of the IFLA Section on Education and Training* (pp.149–164). De Gruyter Saur. https://doi.org/10.1515/9783110375398-015

Data Management Plans. (2022). Digital Curation Centre. https://www.dcc.ac.uk/dmps

De Roure, R., Goble, C. & Stevens, R. (2009). The design and realisation of the myExperiment Virtual Research Environment for social sharing of workflows. *Future Generation Computer Systems*, 25(5), 561–567. https://doi.org/10.1016/j.future.2008.06.010

Discover the world of Scientific Literature. (2022, Dec 21). Litmaps. https://www.litmaps.com/

Duff, W., Carter, J., Cherry, J. M., MacNeil, H. & Howarth, L. C. (2013). From coexistence to convergence: studying partnerships and collaboration among libraries, archives and museums. *Information Research*, 18(3), paper 585. http://informationr.net/ir/18-3/paper585.html

Eitan, A. T., Smolyansky, E., Harpaz, I. K. & Perets, S. (2022, October 20). Find and explore academic papers. *Connected Papers.* https://www.connectedpapers.com/

Finholt, T. A. (2002). Collaboratories. *Annual Review of Information Science and Technology*, 36(1), 73–107. https://doi.org/10.1002/aris.1440360103

Fourman, M. P. (2003). Informatics. In Feather,J., & Sturges, P., *International Encyclopedia of Information and Library Science*. Routledge.

Garfield, E. (1979). *Citation Indexing – Its Theory and Application in Science, Technology, and Humanities*. ISI Press.

Giere, R. N. (2002). Distributed Cognition in Epistemic Cultures. *Philosophy of Science*, 69(4), 637–644. https://doi.org/10.1086/344627

Ginsparg, P. (2021). Lessons from arXiv's 30 years of information sharing. *Nature Reviews Physics*, 3, 602–603. https://doi.org/10.1038/s42254-021-00360-z

Hall, S. (2001). Constituting an Archive. *Third text*, 15(54), 89–92. https://doi.org/10.1080/09528820108576903

Henrich, A., & Gradl, T. (2013). DARIAH(-DE): Digital Research Infrastructure for the Arts and Humanities — Concepts and Perspectives. *International Journal of Humanities and Arts Computing*, 7 (supplement), 47–58. https://doi.org/10.3366/ijhac.2013.0059

Hey, T., Tansley, S., & Tolle, K. (Eds.). (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Corporation.

Hung, Y. N., Yang, C. H. H., Chen, P. Y., & Lerch, A. (2022). Low-Resource Music Genre Classification with Advanced Neural Model Reprogramming. arXiv preprint arXiv:2211.01317. https://doi.org/10.48550/arXiv.2211.01317

Kitchin, R., & Lauriault, T. P. (2018). Towards critical data studies: Charting and unpacking data assemblages and their work. In Thatcher, J., Eckert, J. & Shears, A. (Eds.). *Thinking big data in geography: new regimes, new research*. University of Nebraska Press.

Koltay, T. (2019). Research Data Management and Data Literacy as We See Them Today. In Lichnerová, L. & Steinerová, J. (Eds.), *Library and Information Science XXVIII*, (pp. 7–16). Comenius University.

Lefebvre, A. E. J., Schermerhorn, E. & Spruit, M. R. (2018). How Research Data Management Can Contribute to Efficient and Reliable Science. In *26th European Conference on Information Systems* (paper no. 35). AIS. https://aisel.aisnet.org/ecis2018_rp/35/

Lewis, S., Shepherd, K., Latt, Y. Y., Schweer, A. & Field, A. (2012). Repository as a service (RaaS). *Journal of Digital Information*, 13(1). http://journals.tdl.org/jodi/article/view/5872

Lorenz, M. (2022). Digitální muzikologie a otázka smysluplného zpřístupnění zvukových digitalizátů. In Lorenz, M. et al. *Záchrana zvukového kulturního dědictví: aktuální situace, problémy, možnosti*. Littera.

Murugesan, P. & Moravcsik, M. J. (1978). Variation of the nature of citation measures with journal and scientific specialties. *Journal of the American Society for Information Science and Technology*, 29(3), 141-147. https://doi.org/10.1002/asi.4630290307

Mushi, G. E. (2021). Research data management and services: Resources for different data practitioners. *IASSIST Quarterly*, 45(3-4). https://doi.org/10.29173/iq995

Neuroth, H., Lohmeier, F. & Smith, K. M. (2011). TextGrid – Virtual Research Environment for the Humanities. *The International Journal of Digital Curation*, 2(6), 222-231. https://doi.org/10.2218/ijdc.v6i2.198

Plesser, H. E. (2018). Reproducibility vs. Replicability: A Brief History of a Confused Terminology. *Frontiers in Neuroinformatics*, 11, 76. https://doi.org/10.3389/fninf.2017.00076

Plotkin, D. (2021). *Data stewardship: an actionable guide to effective data management and data governance*. Academic Press.

Pollock, D., Yan, A., Parker, M. & Allard, S. (2022). The Role of Data in an Emerging Research Community: Environmental Health Research as an Exemplar. *International Journal of Digital Curation*, 16(1), 1-15. https://doi.org/10.2218/ijdc.v16i1.653

Poole, A. H. (2015). How has your science data grown? Digital curation and the human factor: a critical literature review. *Archival Science*, 15(2), 101–139. https://doi.org/10.1007/s10502-014-9236-y

Research Data Management at Monash. (2022, July 14). Library. *Monash University*. https://www.monash.edu/library/researchers/data-collection-management/about

Rudy, S. (2010, October 24). The State of Knowledge about “Living Archives, New Media Archives”. *Sustaining Digital Scholarship for Sustainable Culture*. http://sustainableknowledgeproject.blogspot.com/2010/10/state-of-knowledge-about-living.html

Schreibman, S., Siemens, R. & Unsworth, J. (2004). *A New Companion to Digital Humanities*. Blackwell Publishing.

Smale, N., Unsworth, K., Denyer, G. & Barr, D. (2018). The History, Advocacy and Efficacy of Data Management Plans. *bioRxiv*, 443499. https://doi.org/10.1101/443499

Snow, C. P., (2012). *The Two Cultures*. Cambridge University Press.

Straka, M. & Straková, J. (2019). NameTag (version 2.0). [Web Application]. http://hdl.handle.net/11858/00-097C-0000-0023-43CE-E

The Turing Way Community, Arnold, B., Bowler, L., Gibson, S., Herterich, P., Higman, R. … Whitaker, K. (2022, July 27). The Turing Way: A Handbook for Reproducible Data Science (Version v1.0.2). *Zenodo*. https://doi.org/10.5281/zenodo.3233986

Van Gorp, P. & Mazanek, S. (2011). SHARE: a web portal for creating and sharing executable research papers. *Procedia Computer Science*, 4, 589–597. https://doi.org/10.1016/j.procs.2011.04.062

Výzkumná data. (2022, July 20). Centrum pro podporu open science. Univerzita Karlova. https://openscience.cuni.cz/OSCI-61.html

Why do research data management? (2022, July 29). Research Data Service. The University of Edinburgh. https://www.ed.ac.uk/information-services/research-support/research-data-service/research-data-management

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J., Da Silva Santos, L. O. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T. W., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1), 160018. https://doi.org/10.1038/sdata.2016.18

Wolfram, S. (2023, Jan 9). Wolfram|Alpha as the Way to Bring Computational Knowledge Superpowers to ChatGPT. https://writings.stephenwolfram.com/2023/01/wolframalpha-as-the-way-to-bring-computational-knowledge-superpowers-to-chatgpt/

Editorial record: The article has been peer-reviewed. First submission received on 29 January 2023. Revisions received on 28 March 2023 and 2 June 2023. Accepted for publication on 8 June 2023. The editor in charge of coordinating the peer-review of this manuscript and approving it for publication was Adela Jarolimkova .

Acta Informatica Pragensia is published by Prague University of Economics and Business, Czech Republic.

ISSN: 1805-4951