1-1-2022

Tracking the Functions of AI as Paradata & Pursuing Archival Accountability

Jeremy Davet
The University of British Columbia

Babak Hamidzadeh
University of Maryland, College Park

Patricia Franks
San Jose State University, patricia.franks@sjsu.edu

Jenny Bunn
The National Archives, UK

https://doi.org/10.2352/issn.2168-3204.2022.19.1.17

Tracking the Functions of AI as Paradata & Pursuing Archival Accountability

Jeremy Davet; School of Information, University of British Columbia; Vancouver, Canada
Babak Hamidzadeh; College of Information Studies, University of Maryland; College Park, MD
Patricia Franks; San Jose State University; San Jose, CA
Jenny Bunn; The National Archives; London, United Kingdom

Abstract

While a familiar term in fields like social science research and digital cultural heritage, 'paradata' has not yet been introduced conceptually into the archival realm. In response to an increasing number of experiments with machine learning and artificial intelligence, the InterPARES Trust AI research group proposes the definition of paradata as 'information about the procedure(s) and tools used to create and process information resources, along with information about the persons carrying out those procedures.' The utilization of this concept in archives can help to ensure that AI-driven systems are designed from the outset to honor the archival ethic, and to aid in the evaluation of off-the-shelf automation solutions. An evaluation of current AI experiments in archives highlights opportunities for paradata-conscious practice.

Introduction

As machine learning algorithms continue to grow more accurate, adaptable, and affordable, the moment at which artificial intelligence is formally implemented in archives draws correspondingly near. In recent years archives of all stripes have begun experimenting with different AI-powered software, often in the hope of automating the more menial and time-consuming elements of appraisal, selection, description, and arrangement. In reviewing several such experiments, it has become clear that the functioning of AI is as yet too opaque to meet the bar of accountability and transparency set for archives by the publics they serve.

The primary goal of this piece is therefore to explore the value and feasibility of documenting and preserving the procedural functioning of AI as it is applied to data and records. In order to properly conceptualize this sort of documentation, it is necessary to introduce the term 'paradata' into the archival vocabulary, drawing from the fields of cultural heritage, social science research, and archaeology. Defined within the InterPARES Trust AI research group as 'information about the procedure(s) and tools used to create and process information resources, along with information about the persons carrying out those procedures,' this piece will highlight some opportunities to collect and organize 'paradata' in the context of the application of AI to archives.

In this context of collection, paradata can be used to allow auditors, archivists, and/or members of the public to identify functional weaknesses in automated systems. In so doing, it should facilitate experimentation with those systems so that they might be made more efficient or responsive to the specific needs of archives users. More broadly, it may also be used to ensure that the archives remains an accountable institution even when certain archival functions are mediated by automated systems, which themselves cannot be subject to the same standards of scrutiny as human beings.

While these objectives are all aspirational and supposed upon the current, developing InterPARES definition, they nonetheless reflect the spirit in which the concept was developed. The ends to which paradata are applied are ultimately the purview of the archives which have collected them. However, the general character of paradata is use- and user-agnostic.

As current experiments with automation illustrate, there is a growing need to collect and disseminate information about AI augmented processes in language suitable for archivists, administrators, programmers, and members of the public alike. Furthermore, those efforts which address this need must remain mindful of changing emphases in archival ethics, as well as the current capabilities of AI and their designers to articulate their workings.

Conceptual Origins and Development

While the intellectual origins of the concept of paradata as applied to the aforementioned fields of cultural heritage, social science research, and archaeology are diffuse and date at least as far back as 1989, the term itself is generally attributed to a presentation given by the sociologist Mick P. Couper at the 1998 Joint Statistical Meeting in Dallas, Texas.[1] Originally used to refer to data created as a byproduct of automated systems used during the research process,[2] paradata has since been generalized outside of the archival context to also mean information about human processes of understanding and interpretation, unintentionally created but nonetheless instructive and analytically useful in its own right.[3], [4] The lack of a singular definition of paradata reveals its multifocal application. Depending on the field, paradata has been used to "[communicate] uncertainties and the different phases of the process of interpretation that were often impossible to discern [otherwise]";[5] to improve social science survey design, especially to improve response rate and quality; and to capture information about the researcher-subject relationship, aiding analysis of the conclusion-forming process.[6]

In every context where it has been applied thus far, paradata has been used to pursue intellectual accountability, assure operational transparency, and enable review of important intellectual decisions; the introduction of the concept of paradata into the archival sphere is intended to support the pursuit of similar goals.

Relation to Archival Theory

The development of the archival definition of paradata is a reflection not only of a particular preoccupation with the functioning of automated systems within archives, but also as an extension of evolving ideas around archival transparency, accountabilities, and record integrity. Archivists and archives administrators have increasingly been interested in enumerating their positionalities and redressing their personal biases, in service of improving the function of the archives.[7] Some archives have also taken to including contextual metadata in their descriptive schema, explaining what interpretations an archivist has made of a record and serving to underscore that these interpretations are mediated, subject to revisitation and revision.

Paradata seeks to serve a similar but distinct purpose; to define the steps in and character of the process by which these interpretations were synthesized. The same motivations which have informed the use of positionality statements and contextual metadata have shaped the conceptualization of paradata; namely, refinements to archival ethics and an expansion of the publics which archives are to serve. Moreover, it is intended to fill a perceived void in the critical evaluation of archival arrangement and description, especially in cases where the archival agent cannot be conversationally interrogated. By collecting, preserving, and evaluating the information that said agent used to inform their final interpretation, the functional transparency of the archives is improved, and new avenues by which archival decisions may be redressed are opened.

While this work and the current research of the InterPARES Trust AI group emphasizes the relationship of paradata and AI-enabled automated systems, it should also be recognized that this is not a limitation inherent to the concept. Rather, it can be applied in any case where a person or intelligence makes archival decisions in the normal course of archives' operations or during periods of experimentation.

Application to Artificial Intelligence (AI)

When applied to the AI-powered systems with which archives are currently experimenting, the concept of paradata is intended to describe the often-murky internal operations of these automated solutions. Regularly these programs or systems are in part or in whole 'black box,' meaning that the methods by which they interpret an input and return an output are obscured from the user.[8] The reasons for the development of these black-box solutions are numerous and range from the particular competencies of different programmers to the financial resources of the contracting party. The means by which they come about are also multitudinous and include everything from the cobbling together of many different programs of diverse origins to complete complex tasks, to the software automatically self-iterating as more data is input. Whatever the reasons for the inclusion of black-box elements in a particular AI-utilizing system, their ubiquity is undeniable.

This reality has already begun to inform experimentation on the part of public institutions like DARPA and private concerns pursuing projects in 'explainable AI' (XAI) and leading directly to the development of tools such as model cards and impact assessments.[9] Conscious of this experimentation, paradata can and should be used to frame archivists' understanding of how the processes of AI can be explained, and of what information needs to be collected in order to do so. None of the aforementioned projects have been specifically created with archivists in mind, and as such it is incumbent upon archivists to use a concept like paradata to articulate their needs to AI developers vis-à-vis the archival mission and ethic.

Of course, archives and archivists are not always in a position to dictate the form of the technologies upon which they rely. In the (likely many) cases where archivists will have to use an off-the-shelf AI solution, paradata may also be employed to evaluate deficiencies in the function and output of these tools, identifying areas in which automated work must be hand-corrected and where systems may be improved. As an example, the examination of linked training objects within a corpus of paradata might reveal that the corresponding AI system was not exposed to a particular file type during training, explaining why it fails to categorize that file type correctly in practice.

Ambiguities & Directions for Development

The primary ambiguity embodied in the current definition of paradata is that of the extent of information collected and classified under that heading. The exact amount of information necessary to fully understand the process of archival interpretation will likely differ case-to-case, and so it is difficult to articulate the full extent of what might be included in a corpus of paradata absent specific knowledge of an archive's operations.

With respect to AI, the rapid development of new AI capabilities and the continual creation of composite AI-powered systems of archival automation make it difficult to generalize about what information needs to be collected to understand their 'interpretations.' As will be articulated in the following section, current archival projects using AI and proposed AI-recording technologies broadly suggest the sorts of information that it is currently possible to collect. However, whether this information is suitable to fully explicate the interpretive process or whether it is feasible to collect this information in all cases is still a matter of debate.

It may be that the archival community eventually decides upon and supports a unified AI-based automation system or agrees upon shared structural and reporting standards for their individual systems. The more general conception of paradata should be used to support these sorts of decisions, and in the event that such unified systems or shared standards are developed it may be possible to further iterate upon the concept of paradata in relation to AI.

Paradata Collection

To a large extent, what might be collected as paradata is already generated incidentally or collected as a matter-of-course during the testing and deployment of AI-based automated systems. Data sets used to train AI models represent an exceedingly common form of what can be collected as paradata, although the framing of them as such is novel. As they directly influence the capabilities and competencies of AI, they can be used to explicate the interpretations of that AI as paradata.

However, not all AI-enabled systems may have accompanying training sets; in some cases, the training data may be inaccessible, lost, or be restricted by legal contracts or security measures. Indeed, what can be collected, grouped, interpreted, and disseminated as paradata in any given archival context is reliant on an expansive set of factors which are impossible to explore fully over the length of this work.

In order to begin to understand what might be collected as paradata, some of the practical factors which influence what it is possible to collect and use, and current efforts to explicate the function of AI in archives, consider two relatively recent experiments.

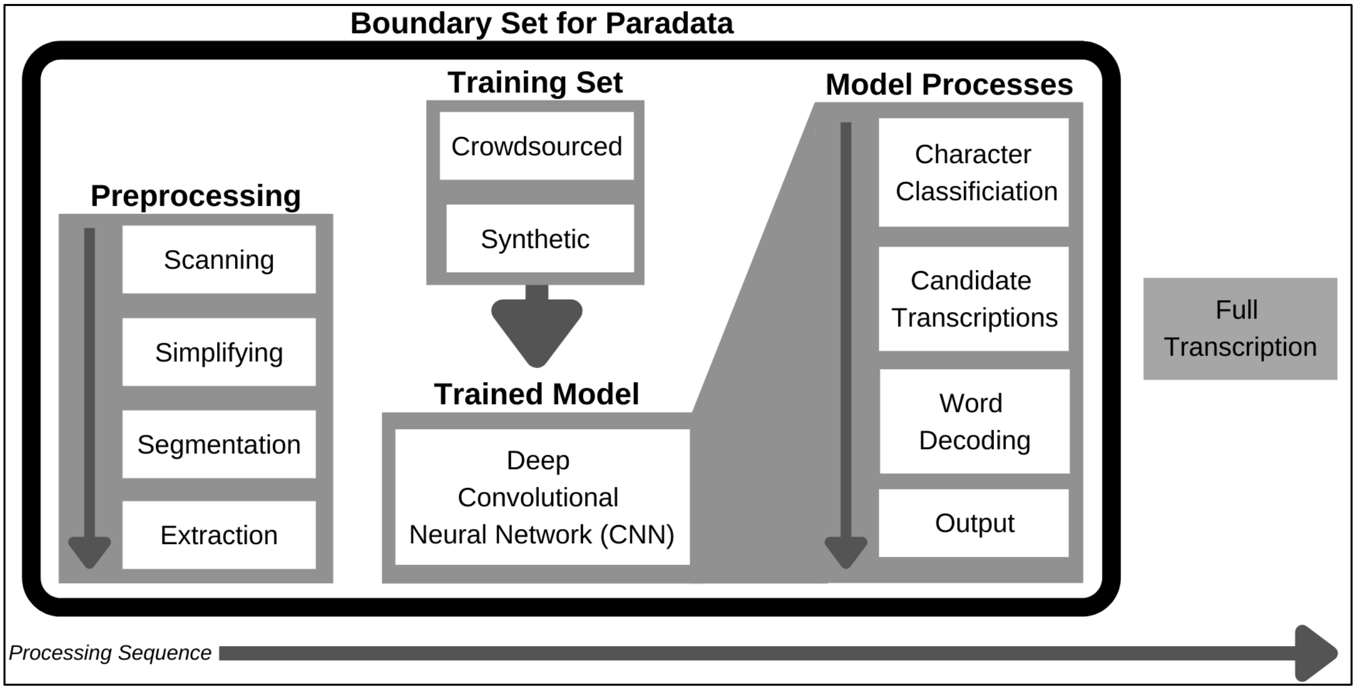

***Figure 1**: A simplification of the flow of records through the In Codice Ratio automated system.*

In Codice Ratio

Based at the Roma Tre University and in collaboration with the Italian National Research Council (CNR), Vatican Apostolic Archive, and the State Archive of Rome, *In Codice Ratio* (ICR) was an experiment launched in 2016 to attempt to improve OCR processing of medieval manuscripts held by the Vatican.[10] Beginning with pages scanned from manuscripts in the Vatican Apostolic Archives, team members from ICR initially relied on dozens of high-school age volunteers to identify and group penstrokes as letters during the development of the training data. Eventually, computer scientists working on the project were able to substitute real images of letters with procedurally-generated facsimiles for the purposes of training their AI, and develop systems to segment handwritten words automatically during the preprocessing phase. Once these words were separated into their constituent letters, they were processed through a deep convolutional neural network (CNN) that identified letter type, the extent of a word, and the likeliest translation based on statistical linguistic models.[11] Publishing the bulk of their findings in 2019, the ICR automated system was eventually able to achieve a 96% success rate in parsing the contents of medieval manuscripts.

The final documentation of the ICR project includes many elements which could be classified as paradata and arrives at conclusions which were unknowingly informed by said paradata. Most obvious is their training data set, which they have published in its entirety on their project website. In evaluating the training set as paradata, one can understand the bounds of the training strategy undergirding the ICR project, as well as the functional limitations that those bounds inspired in the mature ICR system. Specifically, ICR relied on a training set using only one style of medieval handwriting—a derivation of the Caroline style.[12] Were the system to be applied to manuscripts from the same period wrought in a different writing style, it is likely that its successes would have been much fewer in number. Additionally, the training data includes 1,000 'non-character' marks, presumably used to train the CNN to avoid the erroneous classification of letter pairs and flourishes as discrete letters. One familiar with medieval manuscripts might use this information in combination with a knowledge of scribes' fondness for annotation and abbreviation marks to explain the bulk of the failures that the ICR solution manifested. Due to the sheer variety of these sorts of non-character marks, a training set of 1,000 was inadequate to train the system to handle all such cases accurately. In addition to the training data, other elements might have been preserved from the system as paradata, (Fig. 1) including but not limited to the source code of the different iterations of the CNN and design documentation.

The explanations for and conclusions about the function of the CNN as presented in the numerous journal articles published by the ICR team were also informed by what can be classified as paradata. Explanations of how the system segmented each word into letters would have been informed by design documentation and diagnostic information generated by said system. Conclusions about the classification strengths and weaknesses inherent to the system would have relied on performance data and analysis of the training sets. These can all be considered paradata under the Inter-PARES definition, insofar as they speak to the automated procedures and AI tools used to create and process information resources.

The National Archives of the UK

Spurred by 2010 changes to the Public Records Act of 1958 newly mandating that the transfer of public records for permanent preservation should happen no later than 20 years from the date of creation (formerly 30 years), the National Archives of the United Kingdom (TNA) redoubled its efforts to evaluate the treatment of electronic records and to conduct experiments into their automatic

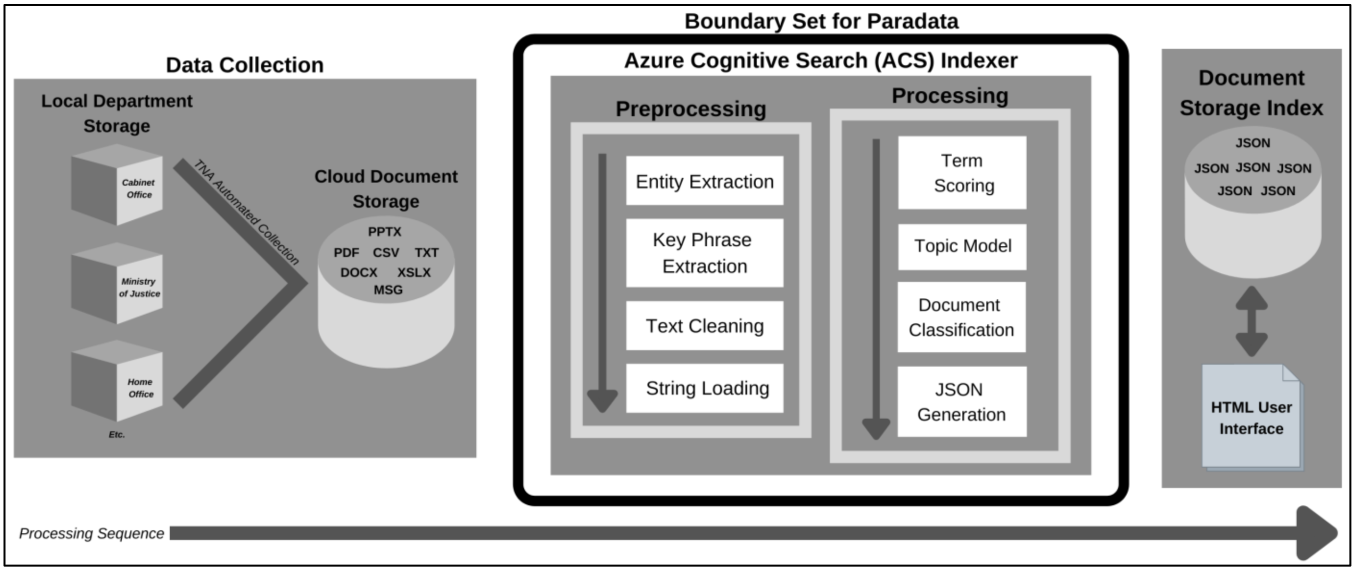

***Figure 2**: A simplification of the flow of records through The National Archives automated processing solution utilizing the Azure Cognitive Search subsystem.*

processing c. 2013. 2016 saw the publishing a report detailing the opportunities for automation offered by existing eDiscovery systems,[13] and this work informed the most recent series of experiments which concluded with the publishing of the 'Using AI for Digital Records Selection in Government' report in October 2021.[14] This report summarized the efforts of five AI vendors to classify a dataset provided by TNA into retention categories related to filetype, subject, department, and the like. While paradata collected by these vendors ultimately informed the conclusions of the report, it is important to note that in this case the paradata were collected as artifacts of experimentation and not as a function of normal operations or operational policy. Such policies have yet to be developed and would likely change the character and extent of captured paradata.

Due to the complexity of the dataset and of the requirements set by TNA, the systems designed and deployed by these five vendors were composed of many different interlinking parts whose individual tasks were distinct. To closely scrutinize each of these systems is outside of the scope of this work, but briefly touching upon the Azure system used by the firm Adatis reveals something of the bounds of paradata and what opportunities exist for its collection within complex, multi-part, AI-based automation solutions.

Particularly, this system example highlights how extensive the paradata may be, as well as the tendency of organizations to already provide for the collection of some types of paradata in experimental contexts. In the case of the Azure component of this system, there exists already fairly comprehensive public documentation, including elements of the source code, information about how the system handles the trouble-shooting process, design notes, et cetera.[15] While not every function is explicated, nor is there much discussion of the nature of changes made between versions, this documentation evidences the implicit understanding that AI designers already have regarding what information is necessary to understand the processes of the systems they have built; even in cases where the final system includes extensive black-box elements.

Additionally, this example suggests how the construction of a system might inform paradata collection, especially in cases where many specialized subsystems are involved. For instance, archivists may determine that paradata related to the separate subsystems responsible for data collection are not relevant to a functional understanding of the ACS subsystem specifically. (Fig. 2) They may then opt to treat these different parts of the overall system as having distinct corresponding paradata. That being the case, strategies for describing the function of these multi-part systems using paradata may still have to rely upon a suite of technologies and procedures, tooled to parse the complex interrelations of component subsystems.

Suggested Paradata Elements

While it is essential to emphasize that the extent of what may be collected as paradata is still a matter under consideration, the two aforementioned examples begin to suggest some elements which should be collected whenever possible for the evaluation of the interpretive processes of AI.

- Training Data: This may include full or partial copies of the training datasets, ideally presented in such a way as to allow easy examination and interpretation. In cases where training data uses electronic records already integrated into digital archival systems, this may be facilitated by the tagging, marking, and/or linking of these records across fonds.
- Performance Information: In addition to quantitative information about the embodied strengths and weaknesses of an AI system, information about performance should also include that which speaks to the reasons for the existence of particular confounding variables. This might also include a representative set of interpretations made by an AI, or information about the function of diagnostic subsystems.
- Versioning Information: Includes proposal and initial design documentation, analyses of changes made to the structure of systems or the training strategies upon which they have been presupposed, and documentation regarding competencies that develop within a system over time. When possible, this might

also include all or part of the source code, and/or maps of the internal architecture as generated by TensorFlow or similar platforms for machine learning.

The extension of a particular paradata schema across archives using the same system or system elements is a potentiality, but it is important to keep in mind that each archive may differ in their capacity to collect, store, interpret, and display paradata. Moreover, guidelines set for a particular system in isolation may not be applicable when that system is part of a more complex, networked AI solution, or in cases where a system can be configured in highly differentiated ways.

Data Structure for Representation and Display

After determining what and how paradata is to be collected, archivists will have to determine the means by which paradata may be organized, interpreted, and displayed. Two current models provide an opportunity to interrogate how paradata may ultimately appear to different stakeholder groups. Experimentation with these structures in the archival environment is likely to generate valuable data about their limitations, leading to improvements in relation to the needs of archivists and the publics they serve.

Google Model Cards

Originally described in a 2018 journal article by Google researchers,[16] model cards represent a compelling opportunity for the articulation of collected paradata and conclusions which they have informed. Model cards are interactive, adaptive modules which briefly explain the use of an AI-enabled program, provide users with basic information about its function, and allow users to interact directly with that program; a working snapshot. Properly configured model cards can provide valuable general information about AI capabilities, allow for rudimentary experimentation, and enable technologically unsophisticated archives users to interact with otherwise inaccessible automated systems.

Current limitations on the usefulness of model cards relate primarily to their reliance on technical experts to encode some of the modular elements that compose each card. While archives may outline the varieties of paradata they wish to describe and the manner in which to describe them, reconfiguration of model cards is still reliant on a degree of specialized knowledge. Additionally, model cards are primarily useful in supporting basic, reflexive understandings of the systems they describe. More complex models, especially those to be used by engineers and savvy archivists, may have to rely on different data structures such as IBM AI Fact-Sheets.

IBM AI FactSheets

Launched in 2020, the IBM AI FactSheets project has used the model of suppliers' declarations of conformity to experiment with the collection and arrangement of information relevant to the creation and deployment of AI models—information which might otherwise be termed paradata.[17] FactSheets utilize the information provided by different roles in the AI development cycle (e.g., clients providing use cases, data scientists providing data gathering strategies) to populate a variety of display templates, each of which are intended to serve different communications purposes and audiences. These templates vary in format and complexity from full-scale technical reports to simplified slideshows, and would theoretically be generated automatically at different stages in a model's operational life cycle.

In comparison to model cards, FactSheets more readily lend themselves to formal standardization and more plainly communicate the underlying paradata to the end user. However, FactSheets are less attractive as a model for displaying paradata to archives users insofar as they are designed and best-suited to support professional functions, rather than casual inquiry. Realtime reporting via Fact Sheets present an intriguing method to analyze iterative AI design and could serve to enhance archives governance. However, the costs of licensing and deploying FactSheets may curtail their use in more financially-limited contexts.

Directions for Further Research

Using the proposed definition of paradata, future research may be directed along any number of routes of inquiry. Extensive theoretical and practical experimentation must yet be undertaken to determine the effects that the collection of paradata will have on the form of AI-based archival automation, and to develop standards.

More information is required about how differing financial, human, and technological resources will affect the ability of archives to collect and utilize paradata. It is not yet clear what the greatest impediments to collecting paradata about AI-based systems of automation may be; nor is it clear how those obstacles might be overcome.

Extensive research will need to be conducted into the ability of archivists to affect the bottom-up development of the AI tools on which they may soon rely. Further codifying a suite of desirable features and their complementary paradata would assist in informing design discussions and ensure that archivists do not find their ethics have been compromised for the sake of operational expediency.

Research may also be directed towards determining effective strategies and appropriate forums for the development of paradata standards and guidelines. Whether the AI-for-archives landscape is eventually dominated by a single system or remains a site of active competition, there may be the need for cooperative bodies to create standards which can be deployed across the archives ecosystem.

Conclusion

In conjunction with increasing interest in the automation of essential archival functions using artificial intelligence, the time for the introduction of paradata to the archival sphere has arrived. An examination of current experiments with artificial intelligence in the archives suggests that considering and encoding paradata at the design stage of any AI-driven automation will be essential to satisfying the archival ethic. However modest their influence may be, archives have an opportunity to encourage the development of tools which are in line with existing archival praxis in pursuit of transparency, accountability, integrity, and usability.

Recent experiments with AI in archives illustrate implicit understandings of the types of information required to parse the interpretive products of AI systems and suggest how future project might iterate with paradata in mind. Existing technologies developed out of and adjacent to XAI projects represent a compelling jumping-off point for further experimentation, and extensive opportunities exist for future research based on the current conception of paradata and its relation to AI, especially about paradata schema and standards.

While the current InterPARES definition of paradata will likely be subject to changes, the problems of automation to which it was created in response will not. Regardless of whether paradata is ultimately the solution, archivists owe it to the publics they serve to consider fully their strategies for maintaining archival accountability in the AI age.

References

[1] M. P. Couper, "Measuring survey quality in a CASIC environment," in Proceedings of the Section on Survey Research Methods of the American Statistical Association, Dallas, TX, USA, 1998, pp. 41–49.

[2] H. O'Connor and J. Goodwin, "Paradata," in Sage Research Methods: Mixed Methods, Thousand Oaks, CA, USA: SAGE Publications Ltd, 2020.

[3] H. Denard, "The London Charter for the Computer-Based Visualization of Cultural Heritage," londoncharter.org, 2009.

[4] A. Bentkowska-Kafel, H. Denard, and D. Baker, Eds., Paradata and Transparency in Virtual Heritage. Farnham, Surrey, England ; Burlington, VT: Ashgate, 2016.

[5] I. Huvila, "The Unbearable Complexity of Documenting Intellectual Processes: Paradata and Virtual Cultural Heritage Visualisation," Human IT: Journal for Information Technology Studies as a Human Science, vol. 12, no. 1, pp. 97–110, 2013.

[6] R. Edwards, J. Goodwin, H. O'Connor, and A. Phoenix, Eds., Working with paradata, marginalia and fieldnotes: the centrality of by-products of social research. Cheltenham, UK ; Northhampton, MA, USA: Edward Elgar Publishing, 2017.

[7] D. A. Wallace, W. M. Duff, R. Saucier, and A. Flinn, Eds., Archives, Record-Keeping and Social Justice. New York: Routledge, 2020.

[8] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A Survey of Methods for Explaining Black Box Models," ACM Comput. Surv., vol. 51, no. 5, pp. 1–42, 2019.

[9] Information Innovation Office (I2O), "Broad Agency Announcement - Explainable Aritificial Intelligence," Defense Advanced Research Projects Agency, Arlington, VA, USA, Announcement DARPA-BAA-16-53, 2016.

[10] S. Ammirati, D. Firmani, M. Maiorino, P. Merialdo, E. Nieddu, and A. Rossi, "In Codice Ratio: Scalable Transcription of Historical Handwritten Documents," in Proceedings of the 25th Italian Symposium on Advanced Database Systems, Squillace Lido, Italy, 2017, pp. 65–72.

[11] E. Nieddu, D. Firmani, P. Merialdo, and M. Maiorino, "In Codice Ratio: A crowd-enabled solution for low resource machine transcription of the Vatican Registers," Information Processing & Management, vol. 58, no. 5, 2021.

[12] S. Ammirati, D. Firmani, M. Maiorino, P. Merialdo, and E. Nieddu, "In Codice Ratio: Machine Transcription of Medieval Manuscripts," in Digital Libraries: Supporting Open Science, Cham, 2019, pp. 185–192.

[13] The National Archives, "The application of technology- assisted review to born-digital records transfer, Inquiries and beyond," The National Archives, London, UK, Feb. 2016.

[14] The National Archives, "Using AI for digital selection in government," The National Archives, London, UK, Oct. 2021.

[15] H. Steen, "Azure Cognitive Search Documentation," microsoft.com, Jan. 03, 2022. https://docs.microsoft.com/en-us/azure/search/.

[16] M. Mitchell et al., "Model Cards for Model Reporting," in Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta, GA, USA, 2019, pp. 220–229.

[17] M. Arnold et al., "FactSheets: Increasing trust in AI services through supplier's declarations of conformity," IBM J. Res. & Dev., vol. 63, no. 4/5, p. 6:1-6:13, 2019.

Acknowledgements

This paper is an outcome of InterPARES Trust AI, an international research partnership led by Drs. Luciana Duranti and Muhammad Abdul-Mageed, University of British Columbia. InterPARES Trust AI is supported in part by funding from the Social Sciences and Humanities Research Council of Canada (SSHRC).

Author Biographies

Jeremy Davet is a MLIS/MAS student at the School of Information at the University of British Columbia. Currently he supports the research efforts of the InterPARES Trust AI as a Graduate Academic Assistant.

Babak Hamidzadeh has served as the Interim and Associate Dean of the University of Maryland Libraries; Director of the RDC at the Library of Congress; in senior management at Boeing; and in faculty positions at the University of British Columbia and at the University of Science and Technology in Hong Kong. He received his Ph.D. in computer science from the University of Minnesota and is a faculty member at the University of Maryland's iSchool.

Dr. Patricia Franks, CA, CRM, IGP, and CIGO is the co-editor of the Encyclopedia of Archival Science, the Encyclopedia of Archival Writers, 1515-2015, and the International Directory of National Archives. She is also author of Records and Information Management and editor of The Handbook of Archival Practice.

Jenny Bunn is Head of Archives Research at The National Archives. She has over 25 years' experience as an archival practitioner, educator and researcher at institutions including University College London and the Royal Bank of Scotland. Her research interests have always lain at the intersection of archives and technology.