

# Positioning paradata:

## A conceptual frame for AI processual documentation in archives and recordkeeping contexts

Scott Cameron\*

School of Information, University of British Columbia; Vancouver, Canada

Pat Franks

San Jose State University; San Jose, California

Babak Hamidzadeh

College of Information Studies, University of Maryland; College Park, Maryland

The emergence of sophisticated artificial intelligence and machine learning tools poses a challenge to archives and records professionals, who are accustomed to understanding and documenting the activities of human agents rather than the often-opaque processes of sophisticated AI functioning. Preliminary work has proposed the term 'paradata' to describe the unique documentation needs which emerge for archivists using AI tools to process records in their collections. For the purposes of archivists working with AI, paradata is here conceptualized as information recorded and preserved about records' processing with AI tools; it is a category of data which is defined both by its relationship with other datasets and by the documentary purpose which it serves. This paper surveys relevant literatures across three contexts to scope the relevant scholarship which archivists may draw upon to develop appropriate AI documentation practices. From the statistical social sciences and the visual heritage fields, the paper discusses existing definitions of paradata and its ambiguous, often contextually dependent relationship with existing metadata categories. Approaching the problem from a sociotechnical perspective, literature on explainable AI insists pointedly that explainability be attuned to specific users' stated needs – needs which archivists may better articulate using the framework of paradata. Most importantly, the paper situates AI as a challenge to accountability, transparency, and impartiality in archives by introducing an unfamiliar nonhuman agency, one which pushes the limits of existing archival practice and demands the development of new concepts and vocabularies to shape future technological and methodological developments in archives.

**CCS CONCEPTS** • Applied computing → Computers in other domains → Digital libraries and archives

**Additional Keywords and Phrases:** paradata, archives, explainable artificial intelligence, XAI, processual documentation, metadata, records management, records, accountability

**ACM Reference Format:** Cameron, S., Franks, P., and Hamidzadeh, B. 2023. Positioning paradata: A conceptual frame for AI processual documentation in archives and recordkeeping contexts. *ACM. J. Comput. Cult. Herit.*

## 1 Introduction

As sophisticated artificial intelligence (AI) tools proliferate, their application in diverse information processing contexts will spread widely. AI researchers have developed increasingly sophisticated and effective machine learning (ML) techniques which, although capable of analyzing vast datasets with minimal oversight, often achieve their sophistication at the expense of our ability to understand the logic of the tool. Working with masses of statistical data, AI tools are often inscrutable even to their designers. If AI is to be used in contexts which demand some degree of accountability and transparency, inscrutability poses a problem. This position paper

Hamidzadeh@umd.edu

\*Author addresses: Mr. Cameron: [scottm.cameron@mail.utoronto.ca](mailto:scottm.cameron@mail.utoronto.ca); Dr. Franks: [patricia.franks@sjsu.edu](mailto:patricia.franks@sjsu.edu); Dr.

discusses paradata as a framework for documenting AI applications in diverse archival processes. It argues that by synthesizing concepts from multiple fields including the empirical social sciences, explainable AI (XAI), and archival studies, paradata provides an approach for supporting the accountability needs of records and recordkeepers in documenting AI processes in specific technical and organizational contexts. As defined by InterPARES for archival contexts, paradata consists of “information about the procedure(s) and tools used to create and process information resources, along with information about the persons carrying out those procedures” [19]. This paper intends to elaborate on this concept in the AI context specifically; while it is implicit from this definition that paradata exists outside of AI contexts, the essay will limit itself to the direct challenge to transparency of archival processes which emerges in AI implementations.

This paper envisions paradata as a key framework for articulating the accountability needs of archivists in the assessment, application, and documentation of AI software for archival purposes. A recent literature review on the topic of AI applications in archives suggests “that more work should concentrate on improving the trust into these AI techniques by developing a stronger ethical framework and a better understanding of their impact on research practices” [15]. This paper intends for paradata to directly address this gap. As such, this paper draws upon paradata as discussed in three distinct contexts. It opens with a discussion of the ‘paradata’ term as it has been used to date, largely drawing on the statistical social sciences, visual heritage, and documentation fields. Across these and other disciplines, paradata has been used to describe information which records the process of creation or curation of other data or datasets. As the term has proliferated across several fields to refer to processual data in discipline-specific contexts, this paper draws on these existing conversations to better articulate the needs of archives in documenting the activity of sophisticated but opaque artificial intelligence tools. The second section discusses existing XAI literature, detailing some approaches to the black-box problem of understanding the logical pathways leading to an AI tool’s decision, when even the tool itself may not be able to articulate its own logical structure. The challenges which the XAI literature identifies have been discussed within existing literature on applying AI tools within archival contexts, which will be addressed here as well. Finally, the paper will contextualize the previous topics of paradata and XAI in the literature on accountability and transparency in archives, suggesting that the term may prove useful to improve archivists’ accountability in their own roles as archivists while also encouraging broader social accountability in the position of AI in society more broadly. To conclude, the paper will point towards directions for further research in AI documentation practices.

## 2 Paradata: origins in statistical sciences and propagation across contexts

The term ‘paradata’ has emerged in the last 25 years to describe information documenting research processes which fall outside of the traditional limits of the data collected. Sociologist Mick Couper first suggested the term, which emerged from his discussions of the potential uses of keystroke data gathered by a computer survey tool [18,39]. In addition to collecting survey responses themselves, Couper realized the potential value of data recording keystroke entries, cursor movement, and page scrolling behaviour of survey respondents, both as a means of adding depth to the primary survey data and as a means of evaluating survey instruments [18]. While Couper’s initial 1998 publication on the topic did not include the term ‘paradata,’ he reportedly introduced the term into conversations within the Survey Research Methods Section of the American Statistical Association around this time [52]. Couper noted that the complexity and volume of paradata collected in computerized survey programs posed challenges for researchers, as did the implications of its collection for research ethics. Ongoing research noted the depth of these challenges, tempered by the impressive use-value of paradata for evaluating and refining interview techniques and questionnaire design [50].

By the mid-2000s, paradata had gained a strong and increasingly institutional foothold in the statistical research methods community. For instance, in June of 2008 the National Centre for Health Statistics published the first-ever National Health Interview Survey Paradata file along with the survey’s main data corpus for 2006 [57]. The National Institute of Statistical Sciences (NISS) convened an expert panel on metadata and paradata to investigate “existing standards and practices for metadata and for paradata that are currently in use for federal

data." While their report noted that the boundary between metadata and paradata was "meaningful, but nebulous," their 2010 report proposed the following definitions for the two terms:

"Metadata: Formalized data about statistical data needed to search for, display and analyze those data.

Paradata: Formalized data on methodologies, processes and quality associated with the production and assembly of statistical data." [36]

From an information science perspective, NISS' definition of metadata is rather limited. It describes perhaps the metadata a typical researcher might encounter, but it excludes the administrative and use metadata which most repositories would collect. While the NISS definition of paradata may be limited to "formalized data," more recent definitions tend not to include this qualifier. For instance, the US Census Bureau as of 2022 defines paradata as "a term used to describe data generated as a by-product of the data collection process." The Bureau does not limit its definition to specific formats or content, indicating instead that these will vary based on "the system that generated the data" [59]. Researchers such as Edwards et al. increasingly consider paradata to include qualitative observations made by researchers, including "the context in which questionnaires are completed [and] interviewer-generated observations about the process of data collection." Its collection, preservation, and distribution should allow for the "interrogation of the place of the researcher in the data-gathering process," allowing greater transparency into research processes for future researchers using the dataset [24]. Comments left by questionnaire administrators in the margins of their pages may even comprise paradata, and have been systematically analyzed as offering commentaries on the research design itself [24]. Paradata's definition as provided in the SAGE Research Methods Foundations series acknowledges its potential breadth:

"para' data are those data which fall outside the intentionally or purposefully collected data. They are data captured beyond the confines of any methodological specification or data that are collected alongside and in addition to the primary research concerns." [46]

The two extremes of formalized, quantitative information collected by design and qualitative observations generated incidentally throughout the research process would be considered paradata under this definition. SAGE's reference source suggests an emergent capacious application of the term to diverse research contexts. The role of paradata in offering explanation of the process of creating or processing information recurs across the domains in which paradata has emerged in context-specific technical and professional lexicons. For the reader's convenience, a summary of the definitions of paradata and metadata across the fields in which the term has emerged may be found in [Table 1](#). This may elucidate further discussion of the terms' meanings across contexts as this paper develops.

As a term referring to data which documents the process of collection of primary data, 'paradata' is well established within the statistical social sciences. With the growth of computerized research methods, the term has cross-fertilized into other fields to refer to digital process-related research documentation. For instance, in education, digital learning platforms' user data such as page view times, navigation pathways, and cursor tracking have been described as paradata, much as the term is used to describe the same phenomena in survey administration in Couper's original definition [42]. Researchers in the medical sciences have noted the utility of similar types of user behaviour data for questionnaire design and data analysis. Paradata collection through computerized systems presents a new consideration for research ethics, given that it can readily be collected without participants' knowledge [55]. The ways in which a person interacts with a given information or technological system can reveal valuable information for research or commercial purposes. Cursor tracking, for instance, can easily reveal pain points in web design, but may easily be used to predict users' age and gender with significant accuracy using very basic techniques [38,37]. While the term 'paradata' may not be in vogue for the firms likely to collect, record, and analyze the micro-data generated in users' interactions with their apps or websites, this user data is fundamentally similar to paradata generated and preserved in a research process and recorded by computerized survey instruments. While researchers may call this paradata and consider the ethical implications of collecting it in their research contexts, private concerns may be less scrupulous, and research ethics for paradata collection in social sciences contexts remains a key concern.

**Table 1: Definitions of paradata and metadata across contexts**

Field	Definition of paradata	Definition of metadata
Statistical sciences	Qualitative or quantitative data about the process of gathering or assembling statistical data [36,46]	Data about a dataset, used for search, display, or analysis; typical uses are discovery and collection management [36]
Virtual heritage visualization (Bentkowska-Kafel)	Illustrates design choices made throughout the process of creation of a heritage visualization	Describes properties of data object; primarily used for discovery and collection management [4]
Research dataset documentation (Huvila)	“data that can help to elucidate past, ongoing and potential processes relating to data”; a category of metadata which may also overlap with contextual and provenance information, depending on the specific case [31]	Information defined in relation to other information objects.
InterPARES’ view (shared by this paper)	information about the procedure(s) and tools used to create and process information resources, along with information about the persons carrying out those procedures [19]	Information about another information resource [66]
Common aspects across all sources	Information about other information resources; recorded as a means of documenting processes of creation, curation, or management of other information resources	Information about other information objects

## 2.1 Paradata in virtual heritage visualization

Moving further afield from its origins in statistical sciences, paradata has become a key term in digital heritage visualization processual documentation. Since the 1990s, 3-dimensional visualizations of artifacts and historical locations have emerged as prominent tools in archaeology, public history, and the humanities generally [30]. Faced with the problems of conveying the decisions which practitioners made in creating visualizations, Drew Baker of the King’s College Visualization Lab in London proposed ‘paradata’ as the descriptor of this process documentation [63]. Seeking to make digital heritage visualization “at least as intellectually and technically rigorous as longer established cultural heritage research and communication methods,” the London Charter proposed paradata as a central concept in establishing heritage visualization’s rigour and scholarly accountability [20]. The London Charter proposes a framework for documentation of the context and content of a heritage visualization, including description of the knowledge claim a visualization makes, whether it depicts an evidence-based or hypothetical restoration, any factual uncertainties undergirding the depiction, and documentation of available research sources [20]. It prescribes the following brief outline for the process documentation, or paradata, of a given visualization. Paradata here is described as follows:

“Documentation of the evaluative, analytical, deductive, interpretative and creative decisions made in the course of computer-based visualisation should be disseminated in such a way that the relationship between research sources, implicit knowledge, explicit reasoning, and visualisation-based outcomes can be understood.” [20]

Here, rather than documenting observations made during the process of data collection, paradata provides an intentional description of the actions taken and agency exercised in the process of a reconstructive activity. Since digital heritage visualizations are a reconstruction of past artefacts and places, rather than an unfiltered view of historic objects themselves, 'paradata' is proposed as a catch-all term to describe deliberate authorial interventions which have shaped the digital object as it is presented. This approaches more closely InterPARES' archival definition of paradata; while the primary goal is to offer a direct view into the remains of the past, practitioners in fact make decisions which influence the objects which researchers encounter, be they visualizations or archival records. These decisions must be transparent to researchers to reveal the roles of their curators, creators, and managers in relation to the information objects they are presented with.

While the London Charter's definition does not so explicitly differentiate paradata from metadata, Drew Baker has drawn clearer distinctions in further publications on the topic in visual heritage. As Baker writes,

"whereas metadata, including contextual metadata, describe the properties of data objects in order to facilitate the management and dissemination of collections, paradata, as the term is used here, describe the processes of interpreting and creating data objects in order to enable understanding and evaluation."

Where metadata comprises static properties of an object recorded objectively, paradata documents the creative and interpretive decisions underlying an object's creation [4]. By Baker's definition, much as with NISS' discussed earlier, properties of an object recorded in its metadata will be evident from the moment of creation onwards, whereas paradata comprises documentation generated about the process of an object's creation, modification, or use. By this definition, metadata and paradata do not overlap conceptually. Further elaboration is provided by London Charter co-authors Bentkowska-Kafel and Denard:

"The London Charter defines 'paradata' as information about human processes of understanding and interpretation of data objects. Paradata is thus constantly being created, irrespective of whether they are systematically recorded or disseminated. Examples of paradata include a note recording method in a laboratory report, descriptions stored within a structured dataset of how evidence was used to interpret an artifact, or a comment on methodological premises within a research publication. Paradata differ in emphasis from 'contextual metadata': whereas the latter tends to focus upon how an object has been interpreted, the central focus of paradata tends to be the processes of interpretation through which understanding of objects [is] communicated [or] sought." [9]

Whereas contextual metadata might offer the results of an interpretation, paradata offers documentation of the process of interpretation which went into the creation of an object. Mark Carnall, writing in the same edited volume, describes 'making of' featurettes embedded in DVD or BluRay issues of documentaries as a common type of paradata, seamlessly incorporating the documentation of a visualization's creation alongside the visualization itself [12]. In the same volume, Martin Turner traces some of the consequences of the expansion of metadata recording and collection from the limits of currently existing metadata, which he describes as a static and comparatively straightforward field, towards paradata as a multi-layered record of interpretative choices made in the course of the creation and curation of an object. Paradata compilers must consider the "question of who should do this annotation and who is responsible for ensuring its accuracy and usefulness," in addition to the audiences the information is presented for. Turner cautions that "the true complexity, as it is human based, of accurately describing metadata or paradata for all purposes is an extremely hard problem" [58]. Engaging with paradata as a record of human deliberation and action remains a key stumbling block in existing approaches to the problem.

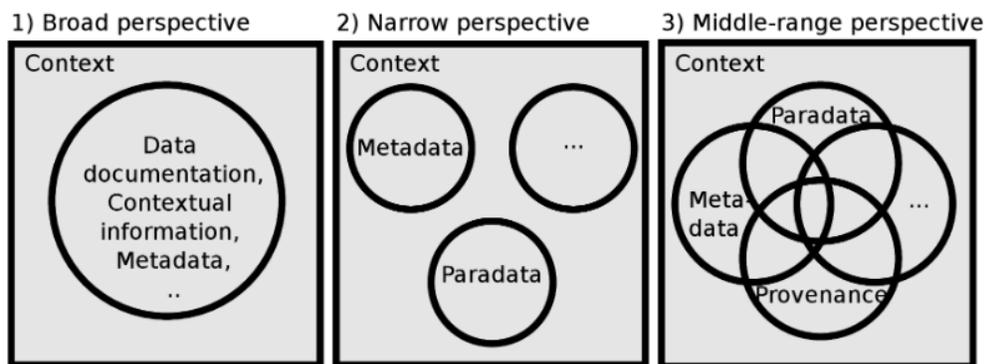
Although the London Charter represented a significant collaborative achievement for paradata in visual heritage, researchers in the field are not unanimous in support of the term's use. Writing in 2012, Bentkowska-Kafel notes that paradata documentation of visualizations remained "the exception" rather than the rule [8]. In the same volume, Mark Mudge targets the Charter's "invention of the term 'paradata'" as needless proliferation of jargon where more familiar terminology might do. Mudge describes the boundaries between paradata and metadata as a "distinction without a difference," though he does note the need for documentation of the visualization's circumstances of creation to foster transparency into a visualization's creators' process. Mudge describes the necessary process documentation in terms of provenance, as both documentation of the real-world

evidentiary basis of a visualization and of the creator’s generative decisions underlying a creation [45]. While Mudge sees the term ‘paradata’ as unnecessary, he remains on board with the case for creating expanded process documentation as a means of ensuring accountability in the visual heritage field. Like with NISS, the distinction of the boundaries between paradata and metadata remain inadequate from an information sciences perspective.

The London Charter thus is perhaps best seen, as Huvila has written, as an impressive conversation opener on the topic of paradata in virtual heritage visualizations, if not the last word [30]. While no major new standards or publications similar in scope to the Charter have emerged since, the phrase has become increasingly common within the field and attracted additional champions. Working in virtual archaeology to recreate Haudenosaunee woodland longhouses, William Carter credits paradata as a key factor in enabling participatory approaches in historical visualization generation. Flagging moments in which researchers have made interpretative choices opens those choices up to worthwhile debate in community-oriented projects [13]. More recently, Rachel Hann has suggested a ‘paradata map’ schema to incorporate paradata into heritage visualizations. Hann’s visualization schemas include toggle controls to allow viewers to quickly surface alternate scenarios, conjecture on the researcher’s part, and information originating from a given historical source. Hann positions paradata as a key “critical framework for reading and designing heritage visualization, which allows the field to “represent the provisional and tacit knowledge latent within the selective and interpretive processes” of creating visualizations [27]. In these contexts, paradata is discussed less as a defined set of questions and more as a problem of documentation for a creative and technical process. What is most important for our purposes perhaps is the emphasis on transparency into the creative and curatorial process, by which paradata becomes a way of offering insight into ambiguities and autonomous decisions undertaken by the creator.

## 2.2 Paradata across contexts

In light of emerging conversations about paradata across visual heritage and other fields, Isto Huvila has offered some clarification of the relationships between metadata, paradata, and provenance within an argument for the term’s broad utility across research fields. Taking on Huvila’s conceptual model for the purposes of this paper opens paradata up to archival contexts. He offers an encompassing definition of the term: “paradata is data that can help to elucidate past, ongoing and potential processes relating to data” [31]. Huvila notes the increasing use of the term to refer to research process data across multiple fields, and the emerging problem of a lack of



**Figure 1: Paradata in broad, narrow, and middle range perspective. Diagram by Isto Huvila [Creative Commons Attribution 4.0], via Open Information Science. <https://doi.org/10.1515/opis-2022-0129>**

standardization for how to define, collect, maintain, and make accessible a dataset’s paradata. Rather than offering a prescriptive definition of paradata, Huvila instead points out that as one scales inwards or outwards, the

definition of paradata and its relationship to provenance and metadata will change accordingly. While he does view paradata as a subset of metadata, Huvila advocates for a dynamic middle perspective (as illustrated in section 2 of [Figure 1](#)) towards paradata, noting that when addressing a particular question, a given piece of information may appear as part of one or more of an object's paradata, metadata, or provenance. For Huvila, the term is an "analytical" rather than "substantial" concept, and a restrictive definition is less important than the sufficient documentation of the processes which a given information resource has been subjected to [\[31\]](#).

The definition of paradata will therefore vary across contexts – but with this variability acknowledged, "as a subset of all contextual information, a complete utopian set of paradata would make it possible to reconstruct and follow all doings and decisions related to a particular data in minute detail" [\[31\]](#). Huvila notes that the difficulties of selecting paradata to be documented comprises a challenging task, let alone to begin the challenge of articulating standards for professional documentation. However, without paradata documenting a dataset's origins, that data's reusability and comparability are severely limited:

"Without proper documentation of the human processes of creating, understanding and interpreting data objects, there is a risk of creating and archiving large collections of data that are incapable of supporting research and other types of reuse, and even worse, leading researchers and others to work and conduct research on faulty premises and drawing erroneous conclusions on data that has been created under incompatible bases." [\[31\]](#)

While paradata may demand greater effort be dedicated to the documentation of research processes than researchers are currently accustomed to, it may also reduce duplication of effort and ease the process of comparing datasets and information objects. By illuminating the context of their creation and processing, paradata may allow readers to understand the information objects in greater depth.

To conclude, we may delineate more explicitly the boundaries between metadata and paradata as the terms are used in the visual heritage fields and by records professionals. Existing literature in heritage visualization and statistical sciences tends to view metadata as limited to static descriptive fields: who created an object, when, where, and so on. From this perspective, paradata, in contrast, may change over time, as it describes dynamic processes which have influenced the primary data throughout its lifespan. This approach to metadata is more restrictive than the encompassing definition of metadata as 'data about data,' and more restrictive than the perspective of any archivist or records manager who may deal with administrative, use, integrity, or preservation metadata in the course of their daily workflows [\[66\]](#). To records professionals, metadata consists of both static and dynamic information points related to other information objects and must be maintained and updated as needed to responsibly manage a collection. Paradata is proposed in this project as an analytical category which allows archivists to interrogate both the AI or XAI tools which they may apply in archives and to record the actions and decisions which they make themselves or delegate to AI which concern the records under their care. In this specific context, paradata is a relational category which would fall into the broader category of metadata defined as data about data. However, whereas metadata serves a variety of purposes, paradata records the "procedure(s) and tools used to create and process information resources" [\[19\]](#). As Huvila's middle perspective notes, a piece of information which may be understood as paradata at one moment may serve as metadata, provenance, or contextual information from another perspective. Paradata may be 'data about data,' but its distinct status as paradata emerges from its relationship with other information objects and the role it plays in elucidating the processes which led to the information's present form. Specific schema for metadata existing in current institutions may contain some aspects of AI paradata, but not necessarily all. The specific forms of this documentation will vary based on institutional, social, technical, and legal contexts – a subject to which the third section of this paper will return. For now, let us address some of the technical challenges to process documentation and accountability which have spurred this paradata conversation in archives.

### 3 Explainable AI and machine learning tools in archives

What then is the nature of the technical problem which paradata is proposed to address? This section will elucidate some of the problems which arise from the applications of AI tools in contexts necessitating

accountability and transparency. Existing AI literature largely addresses these problems through the lens of explainability, leading to the emergence of a subfield known as explainable artificial intelligence or XAI. XAI literature operates on the premise that given user groups will need explanations which are designed to meet their domain-specific needs [5,61]. Explainability is in this context an evaluable property of AI systems, one which can be assessed in relation to its suitability to a given application. Paradata is here proposed as a useful term to articulate archivists' need for processual documentation, both for the benefit of archivists' own accountability and for researchers' understanding the context and processing history of the records in front of them. AI tools may be explainable or not and may provide adequate paradata for an application or not; paradata is meanwhile a property of records emerging from related documentation. While literature on AI and XAI is thus closely related to the topic, it does not address explicitly the documentation problem as paradata does. AI tools fundamentally change the nature of archival processing by delegating agency to non-human actors, a phenomenon which, it is argued here, necessitates new approaches and lenses for documentation.

XAI addresses the problem of AI tools' relative inscrutability to human oversight. Algorithms which are capable of learning from vast datasets and their ongoing data processing are typically not capable of articulating the basis of their decisions, leaving those responsible for employing the tools at a loss when called to justify their actions. At present, XAI literature recognizes that explanations satisfactory for computer scientists will not necessarily be satisfactory for the general public, or vice versa. This is a challenging situation for paradata, which can never be all things to all audiences; Huvila has critiqued the authors writing in *Paradata and Transparency for Visual Heritage* for "portray[ing paradata] as a complex but still primarily technical, rather than theoretical, issue" [30]. While solving the technical challenge of understanding what an algorithm is doing and why is certainly part of the puzzle when considering the use of AI in archival contexts, the availability of XAI tools will not instantly render AI tools appropriate for archives and their parent organizations' needs. XAI in archives must produce explanations satisfactory to archivists, their parent organizations, and the specific institutions and individuals to whom they answer.

### 3.1 The black box problem

The most sophisticated AI tools tend towards greater and greater opacity, leading to a 'black box problem' in understanding their decision-making processes. Black box tools may produce the desired outputs from the provided inputs but are incapable of providing an explanation of the reasoning they employ in reaching their conclusion. As Vilone and Longo write in a systematic review, the black box problem emerges when AI systems' "underlying structures are complex, non-linear and extremely difficult to be interpreted and explained to laypeople" [61]. This poses an inherent problem for trustworthiness in applications entailing any degree of liability. The problem is hardly new. An early approach to the black box is provided in Ross Ashby's 1956 *Introduction to Cybernetics*, which poses the problem as a logical and ontological one. As Ashby writes,

"The theory of the Black Box is merely the theory of real objects or systems, when close attention is given to the question, relating object and observer, about what information comes from the object, and how it is obtained."

At a fine enough level of granularity, a task as basic as riding a bicycle becomes a black box problem for Ashby, since very few of us could explain the operation of forces at a subatomic level propelling our motion forwards [3]. The black box problem emerged in the medical field as early as 1975, as computerized diagnosis tools proved effective but incapable of providing explanations for their results [54]. Despite nearly a half-century of development, the same problem remains in the medical field's AI applications. For instance, one medical machine learning tool successfully recognizes the signs of cancerous growth from images of a patient's body, as verified by further examination by physicians; however, the same tool is unable to identify the indicators which suggested the presence of abnormalities within its samples. This leaves physicians unable to refine their own techniques by comparison with the ML tool's process, refine the model's techniques, or understand its limitations [22]. As AI tools have developed, explainability of their capabilities continues to lag.

A further complication to the explainability challenge is that more complex AI tools deliver superior predictive accuracy relative to simpler tools, if the real-world problem and the dataset the tool addresses are of greater complexity than the AI tool itself. For instance, complex machine learning techniques offer the greatest sophistication of analysis and, conversely, the least transparency in the inevitable “trade-off between model interpretability and performance” [5]. Arrieta et al. note that while performance often comes at the expense of transparency, models which are well understood may perform better than those which are opaque, given users’ increased ability to understand, assess, and fine tune a given application for its context. While explainability can be worked on, complex models such as deep learning neural networks will never be as understandable as simpler models such as decision trees, due to the inherent complexity of given tools and the abstracted nature of most explanations [5].

AI researchers addressing the black box problem and adding explainability to existing AI models have emerged from many corners. In 2021, the National Institute of Standards and Technology (NIST) published a summary of four principles which might characterize truly ‘explainable’ AI and offered a list of the strategies which XAI models could take towards cracking open the black box. NIST’s principles are laid out as follows for AI systems intended to be explainable:

“Explanation: A system delivers or contains accompanying evidence or reason(s) for outputs and/or processes.

Meaningful: A system provides explanations that are understandable to the intended consumer(s).

Explanation Accuracy: An explanation correctly reflects the reason for generating the output and/or accurately reflects the system’s process.

Knowledge Limits: A system only operates under conditions for which it was designed and when it reaches sufficient confidence in its output.” [48]

According to NIST, explanations must consider their intended audience. Jargon-laden or encoded outputs incomprehensible to laypersons will not do if the process must be comprehensible to non-specialists. Furthermore, any explanation must address its own epistemic limits. AI tools can all too readily be used to mystify the actual workings of systems and exaggerate their capabilities; XAI should fight this tendency by delineating the knowledge limits of a given tool. The authors note that XAI is a developing field, but suggest that the four principles enumerated above might be used to evaluate the effectiveness of an XAI tool in producing a comprehensible explanation of its own workings [48]. While the field has not yet developed authoritative standards as to what explainability might mean, the NIST provisions are a strong move in that direction.

NIST goes on to categorize the approaches which XAI tools might take to produce explanations for AI systems’ actions. Self-interpretable models are AI tools in which the tool itself may serve as an explanation of its own workings. For instance, AI models reliant on decision trees or rule lists allow users to walk through the logical paths which lead a given input through the process of it becoming an output. More sophisticated models’ explanations fall into the category of post-hoc explanations. Within this category, local explanations elucidate one or more individual decisions, whereas global explanations recreate an approximate model of the non-interpretable tool. Local explanations vary in their exact techniques; however, many of the current models operate by providing variations of actual data submitted within a given sample to identify which aspects of a dataset or an item provided the decisive factors leading up to an AI tool’s actions. Global models approximate the entirety of a model and tend to illustrate how variables in the inputs lead a model to different outputs. Often this information has been reconstructed from multiple local explanations and may be represented through visualizations [48]. Ashby theorizes similar methods for reconstructing the logic of black box systems through experimentation as early as 1956 [3]. A more thorough catalogue of the myriad approaches for deriving explanations from AI tools is provided by Vilone and Longo’s 2020 systematic review of then-available literature on XAI [61]. However, for our purposes it will be sufficient to note the necessity in most cases, as explained by NIST, of reverse-engineering additional models to trace the paths by which black box tools came to a given conclusion.

In framing their assessment of a model’s successes and shortcomings, NIST advocates for comparison of an AI tool’s explainability to its human counterpart. For instance, NIST explains that forensic scientists’ explanations of

evidence to juries in criminal trials often fail to impart juries with adequate understandings of their findings; such explanations would not be meaningful given that they do not successfully communicate a specialized field of knowledge to laypersons. NIST argues that AI tools should succeed or fail on the same metrics by which humans performing the same tasks succeed or fail [48]. Translating this into the archival realm, it follows that AI tools should be as capable of explaining their own actions as archivists are themselves – and that shortcomings of explainability on the part of either AI or of archivists should be used to develop our capacities to explain both rather than to excuse our failings in one or the other.

The NIST document provides some solutions-in-progress to issues which have emerged in AI explainability over time. While more elaborate schema do exist to break down the categories of explanation available, NIST's four categories suffice for our purposes [61]. The importance of context in any explanation – and the attendant question of 'explanation for whom?' – has emerged as a key theme in ML and XAI literature. For instance, Bhatt et al. caution ML developers to consider who the stakeholders in a given dataset might be, and to ask what questions they may be interested in before devising an explanation model: "Explainability tools cannot be developed without regard to the context in which they will be deployed" [10]. Arrieta et al. propose that the specific audience become part of the definition they propose for XAI: "Given an audience, an explainable Artificial Intelligence is one that produces details or reasons to make its functioning clear or easy to understand." With the endless possible variation of audiences, an explanation must be contextually situated to be meaningful. No explanation will be universal [5]. Tim Miller argues for the situation of AI explanations within social and human sciences literature on cognition and decision-making, given that the problems XAI addresses are fundamentally those of the relationships between people rather than people and technology [43]. Drawing on data visualization methods is an increasingly common method by which XAI practitioners represent their processes to diverse non-specialized audiences, recognizing that explanations need not be in words only [61]. Field-specific discussions of XAI are also very aware of the importance of context. For instance, Wadden defines "the black box problem in healthcare" as occurring "whenever the reasons why an AI decision-maker has arrived at its decision are not currently understandable to the patient or those involved in the patient's care because the system itself is not understandable to either of these agents" [62]. Explanations must satisfy not just the healthcare provider, but the patient as well. Without close attention to the social and ethical context of an AI application, explainability is a meaningless term.

While the NIST framework does offer a current summary of some persistent and fundamental problems in AI explainability, complex AI tools will continue to pose specific and unique challenges to those attempting to understand, let alone document, their activity. For instance, purportedly colour-blind AI tools have in several cases been shown to reproduce racial biases based on apparently unrelated data points rather than on data subjects' race itself. In these cases, Adler et al. have suggested carefully testing input variables to trace the indirect influence of each as a source of unintended racial bias [1]. Challenges of emergent bias are bolstered by an insular approach to AI development, as existing AI ethics across multiple domains often fail to display significant stakeholder engagement. Bélisle-Pipon et al. reviewed 47 major AI ethics guidelines from private and public stakeholders. Their findings show that only 38% reported any kind of stakeholder engagement in their processes; only 9% reported any engagement with citizens at large. However, engagement paid off, as guidelines developed based on stakeholder engagement were widely applicable across different contexts [7]. AI poses novel challenges for use in high-liability-risk environments, and demands responsible management and application, informed by the subjects and contexts of AI processing.

### 3.2 AI and archives

The approach of AI towards archival practice has not come unnoticed in archival and documentation studies, where the challenges of AI's agency to traditional frameworks and practices become apparent. Focussing on social media platforms, Clifford Lynch has identified the chief problem in documenting the behaviour of algorithms as being a "shift from artifacts requiring mediation and curation, to experiences" [40]. Perceptively, Lynch notes that the archivist's job may become an effort to reconstruct the logic of algorithms through reverse engineering their

logic, the exact process which NIST describes as being quite common in reconstructing global models of AI tools [48]. Lynch's focus is on algorithms which exert massive social power but remain inscrutable to the would-be archivist, such as those employed by massive public-facing but privately-owned platforms like Facebook or Twitter. The problem of documenting social life as mediated by these platforms remains, beyond the challenge in documenting AI tools to which we have full access. Lynch writes that basic documentation of algorithmic systems and their actions is not currently produced in society as presently organized. Whereas established and responsibly-run organizations produce paper trails as a matter-of-course, or, more informally, as ethnographers of all stripes produce documentation of societies, no comparable social division of labour exists for the problem of documenting the role of algorithms in society – a gap which Lynch characterizes as a fundamental problem for the archival profession in the “age of algorithms” [40]. At a basic level, archivists face a problem of a fundamentally new character of social actor for which there exists no established social or technical conventions for creating documentation. A recent literature review has divided AI applications in archives according to their role in the creation, capture, organize, and pluralize stages of a record's life under the records continuum model [15,60]. However, the authors do not discuss papers addressing the documentation of AI tools themselves, suggesting that key work remains to be done on this topic.

The records management field furnishes key contributions to the problem of documenting AI activity, offering a vision of what paradata may resemble in practice. For records managers tasked with documenting the actions of an AI tool, limiting potential “liability for harms caused by AI systems” will be a priority, especially in high-risk cases such as those in which introduction of bias may be a risk or areas which may lead to tort cases [44]. Fear of bias in AI systems is widespread: as Mooradian points out, polls suggest that 60% of Americans anticipate AI tools will reproduce the biases of their creators. The risk of ‘emergent bias,’ by which the tool itself may produce biased outputs through its own functioning, is documented in the literature [25]. As such, Mooradian argues that archivists need to develop an appropriate “AI record” to illustrate the decision processes which went into an AI tool's implementation. The AI record he anticipates documenting “actions, transactions, and events that are carried out (fully or in part) by AI algorithms.” This scope will grow along with the expansion of AI tools' domains of action. At a basic level, an AI record would have to offer firstly a preliminary assessment from the firm of the risks and benefits of a proposed AI solution; secondly, technical documentation of the AI's system design and testing; and thirdly, documentation of the specific decision paths taken leading towards a decision. In creating these AI records, situating the loci of decision-making capabilities within an organization and identifying the roles of AI within those sites, will become key tasks for the archivist of algorithmic agency. For each site, Mooradian recommends developing a strategy to document big data-level informational inputs – whether that be documentation in full or part – in order to develop a strategy for AI documentation before AI takes on a greater role within organizations [44]. All of this information would comprise paradata, given that it records AI processes to which given records are individually or collectively subjected. The AI record, or here, paradata, comprises a preservation target in its own right given its documentation of the archivist's AI processing in relation to records at the group- or item-level [23].

AI does not necessarily produce appropriate documentation, and archivists will need to take deliberate approaches to ensuring that such paradata is created. Assessing the extent of necessary AI documentation, Jenny Bunn proposes several questions which recordkeepers might pose when challenged to document the AI or mixed human/AI decision making:

- “What records are created within AI research teams to document their process?
- What records are created of the decisions to procure or deploy systems utilising AI?
- What records are created of the decisions and impact of such systems?
- Are the created records sufficient to meet existing legal provisions?
- Do the created records meet the required standards of quality?” [11]

Records created by XAI systems as designed by programmers and engineers may not always measure up to the demands of the law or of responsible recordkeeping practices; archivists and records managers therefore must interrogate any XAI systems within their own institutional needs. Bunn ultimately suggests reframing the record

as being a document which provides an explanation of the actions taken by an organization. Rather than seeing records as an inherent and trustworthy by-product of actions or processes without respect to their future audiences, archivists must consider present, potential, and future readers within the content of accountability when considering AI documentation practices. Echoing Bunn's conclusions, a report on The National Archives' experiences with AI notes that successful explanations must "consider the individuals, and the environment in which" a given tool will be used [33]. AI tools demand archivists take future archival readers into account at the instant of their application, and devise documentation practices accordingly.

In contrast to this application-oriented view of documentation, early archival theorists supposed that documents suitable to illustrate organizational practices could reasonably be expected to be produced in the ordinary course of business. Archival *eminence grise* Hilary Jenkinson, for instance, identified archival records' quality of not having been produced with posterity and future readers in mind as a testimony to their authenticity and integrity. For Jenkinson, this "impartiality" distinguished archival records from documents encountered outside of archival contexts [34,56]. The introduction of useful but entirely inscrutable black box AI tools demands the deliberate creation of documentation strategies which may frequently, especially in the case of XAI, entail additional AI tools used to reconstruct the logic of a primary AI instrument. This is arguably a complication of records continuum perspectives which pose a need to begin at the instantiation of a record, given the obscurity which AI agency introduces into the process. What impartiality might mean in this new context is a question which we will hold discussion of until further investigation, but which will certainly remain at the top of mind for those considering paradata in archival practice.

Archivists working to develop AI documentation practices will also need to stay abreast of legislative developments as new and far-ranging AI legislation emerges in both the EU and USA. Especially in the context of AI-related decision making, the regulatory shift towards explanation rather than institutionally-limited documentation encourages active oversight by records subjects, and lines up with recent calls for explainability in AI from the European Commission High Level [11,28]. Such a move echoes the European General Data Protection Regulation's tenets, which arguably guarantee in law European data subjects' right to an explanation for AI processing of their records [53]. The 'right to an explanation' is bolstered by robust new European data protection regulatory authorities, and is thus likely to encourage practices of 'data protection by design' and wide-scale algorithmic auditing as recommended by the European Commission to prevent the introduction of bias or other forms of unjust dealing in algorithmic data processing involving human subjects [14,65]. Such legislative contexts spur not only the development of XAI techniques, but also of documentation and records management practices for AI tools within firms making use of them. The White House's Blueprint for an AI Bill of Rights and the EU's proposed AI Act both represent pending codifications of the right to an explanation in broad and influential legislative contexts [67]. The White House's Blueprint, for instance, demands users receive notice of when AI processing is impacting their treatment by an organization, and that explanations be available for the AI's actions. The Blueprint draws heavily on XAI literature by emphasizing that explanations be clear and tailored to the comprehension of a given audience [68]. Though these pieces of legislation are currently in process, it is reasonable to anticipate the eventual passage of similar laws and imperative to develop professional practices which will meet or exceed their standards. In dealing with the emerging ethical and legal challenges posed by AI developments, computer scientists have already turned, in one case at least, to archives for an example of ethical conduct in information management. Jo and Gebru write that archives provide an example of a thoughtful approach to fraught issues of "consent, power, inclusivity, transparency, and ethics & privacy" which emerge in the curation of training datasets for machine learning [35]. As machine learning developments may draw on archival approaches, archivists are well positioned to consider how they can push machine learning tools to abide by ethics compatible with archival standards and emerging legislation.

It is thus evident that the increasing social power and technical complexity of AI is followed by emerging legislative definitions of the responsibilities of agencies which make use of AI. For archive and recordkeeping contexts, this will necessitate the development of vocabulary and capacity in a multidisciplinary approach to AI recordkeeping, as Herbjørn Andresen argues in providing his "discussion frame" for AI explainability in records

contexts. In the context of sophisticated AI tools, rather than simpler logic-based tools, Andresen explains that tools which predict the likely reasons for an output given a certain input – much like XAI reconstructive models – might be the most certain forms of explanation available to those seeking one. In more complicated AI contexts, elaborate ML models may intersect with complex organizational structures, merging the traditional recordkeeping problems of identifying responsible agents and documenting their activity with the new technical context. In many cases, Andresen points out that the distinction between general policy records and specific operational records will remain the relevant materials to consult when an explanation is necessary. Whereas general policy documents will presumably still be developed by human actors, operational records pertaining to specific instances of bureaucratic activity may increasingly be produced by AI. An algorithmic ethics approach to AI application might regularly assess the extent to which its AI tools realize the organization's policy goals, and would tweak them as necessary to increase their accuracy. These documents would become part of any explanation necessary to uncover a record's processing context [2]. As such, archivists might approach the explainable AI problem as situated within an institutional documentation context, in contrast with the technology/user focus of the XAI literature.

## **4 Returning to archival principles: accountability and process documentation in archives**

To summarize the paper's content to this point, paradata has emerged as a term for process documentation in several research fields, although too common are simplified distinctions between metadata as a static descriptive field and paradata as a dynamic field. As a descriptor for process documentation, we aim to borrow the term to describe documentation of AI processes in records contexts. AI introduces new problems in archives; it is not necessarily explainable as human or bureaucratic activity usually may be, and to the extent that it is, its explanations are dependent on a bevy of technical documentation unfamiliar to many archivists and to those to whom archivists are answerable. XAI as a field is dedicated to the provision of explanations of AI activity within a given context; paradata, meanwhile, is proposed as a frame to articulate the needs of a record or group of records for processual documentation of AI activity. This is not to say that AI is a field without its own ethical guidelines or notions of accountability, but rather that we should situate these conversations in the archival tradition.

### **4.1 Paradata and current metadata practices**

But why might paradata even be necessary? Would not existing provisions for policy documentation and item-level metadata accomplish these same ends? The challenge in response to this question is that existing documentation practices are premised on the presumption of human agency, a fundamental problem considering the emergence of AI as a potential agent in its own right. AI is not necessarily fully understandable by humans, and within certain parameters its capabilities are far more impressive. As Jenny Bunn puts it, "If business is no longer to be transacted only by human beings, but also by AI agents, or some combination of the two, what will evidence of those transactions look like, what will the record be?" [11]. Following Huvila in response to this question, paradata is proposed as an analytical rather than substantial category, not intended to supplant formalized categories of metadata under a given descriptive standard or schema [31]. Rather, paradata is proposed as a category of related documentation which may uncover the shortcomings in existing documentation practices vis à vis AI applications in archives, and which may be able to help archivists in the development of AI documentation standards in the future. Existing metadata fields and institutional documentation may serve the purpose of recording archival paradata, although new means of capturing this information may also need to be devised. Paradata needs may include the preservation of novel forms of information, such as datasheets explaining training datasets or model cards documenting an AI tool, and of these documents' relation to archival collections processed using them [26]. The intersection of existing archival practices and the needs for paradata or process documentation for AI processing in archives is a topic which demands further attention.

The inadequacy of existing archival categories emerges as the fundamentally different character of AI necessitates novel approaches if archival accountability and transparency are to be maintained. The problem for archivists is the emergence of AI agency. As an illustrative example from records management, although the ISO 23081 Metadata for Records standard includes provisions for documentation of records' processing after creation, it is designed to address contexts in which human agency, rather than AI tools, carry out the processing. ISO 23081 requires that specific metadata for records be maintained at and after the point of record capture. Into the latter category would fall the majority of paradata concerning AI records processing; here could be made the argument that paradata would only be a subset of the category "metadata after record capture" described. The ethic of paradata as proposed is very much in line with the ISO's plain statement that "All records management processes performed upon a record, or on an aggregation of records, should be documented" in "process metadata" [64]. However, AI may complicate the prescriptions of the standard when we turn to its proposed types of metadata. The standard proposes the collection of five types of metadata, to be gathered both at the point of capture and afterwards:

1. "metadata about the record itself;
2. metadata about the business rules or policies and mandates;
3. metadata about agents;
4. metadata about business activities or processes;
5. metadata about records management processes"[64]

Into which category would AI activity fall? Sophisticated AI tools can act as an agent in their own right and may carry out business and records management processes with agency comparable to human actors. ISO 23081's prescriptions for "metadata about agents after record capture" largely serve to refer readers to sections in ISO 15489 outlining recordkeeping principles of authenticity, reliability, integrity, reliability, and security [64]. As such, we may need as archivists to further consider what the implications of AI implementation may be for archival principles.

## 4.2 AI, accountability, archives, and trustworthiness

Ultimately, AI proposes a challenge to recordkeepers' and archivists' goals of ensuring accountability by introducing a less-than-transparent and frequently uninterrogable actor into administrative processes formerly enacted by human agents. In devising approaches to paradata, then, archivists will need to consider the imperatives of accountability and transparency which underlie responsible documentation practices. A brief overview of what exactly accountability entails in archives is perhaps useful here. At present, archivists and recordkeepers of all stripes are reasonably likely to understand their role as enabling accountability in their specific institutional contexts and in society more broadly. The Association of Records Managers and Administrators (ARMA)'s list of Generally Accepted Recordkeeping Principles, for instance, offers accountability and transparency as the first and second priorities for records managers, although their articulated definition of accountability is bureaucratically situated rather than expansive [69]. Representing the expansive social view of accountability, Livia Iacovino has described "Archives as arsenals of democracy" in an encompassing case for the liberal democratic value of archival practices as ensuring that actions and responsibility are traceable to their authors within complex social organizations [32]. Broad, socially oriented views of the value of archives in ensuring accountability have proliferated in archival literature since the 1970s [47,6]. This is not to say that the threads of transparency or accountability are entirely novel; Schellenberg's "evidential value" is certainly a conceptual precursor to later further fleshed ideas about accountability [51]. Similarly, German archivist Fritz Marx saw archives as ensuring transparency, following from the liberal-democratic "insistence that public business be conducted along the lines of public preference and under the eyes of the public" [41]. Paradoxically, the early stage of archival literature was characterized by a focus on the provision of evidence, as Terry Cook has written; however, this concept of evidence was often narrowly defined in terms of bureaucratic needs, rather than situated in broader ethics of social or professional obligations towards accountability [17,47]. Approaches to ethical AI display a similar social vs. institutional division: on the one hand, ethics as a professional and social

obligation, and on the other, accountability as embodied in workflows or reporting relationships and enforced by legislation.

For archivists, the problem of paradata and AI documentation will necessitate both an ethical mandate obliging them towards transparency in AI applications and additionally the development of professional standards, policy, and legislation where necessary to ensure responsible and accountable AI applications. Accountability in archiving has been divided between an institutional/professional approach, which focuses on reporting relationships and documentation mandates, and a broader conception of the social or historical responsibility of recordkeepers [21,32]. The two are not necessarily in conflict with one another: as Terry Cook puts it, "if good records are created for short term, specific accountability requirements, then the longer accountability needs will also be served" [16]. As Iacovino writes, "there is an ethical obligation on the part of record creators and archivists to ensure that records adequately document what happened, beyond mere compliance with the law" [32]. Beyond the archivist's ethical imperative, accountability does not create itself, and must instead be situated within specific institutional, technical, and social contexts. Approaching the problem within the framework of a records continuum, Chris Hurley writes that accountability must be clearly defined before designing a records system: "who is accountable, what are they accountable for, who they are accountable to, and the criteria by which performance is to be judged must be clearly documented in advance" [29]. Hurley provides three main principles to ensuring accountability in archives: firstly, ensuring the preservation of evidence which preserves records' subjects' rights; secondly, ensuring consistency in that similar functional processes are documented in similar ways; and thirdly, ensuring transparency in that retention schedules are consistent and determined in advance [29]. The development of a documentation strategy for AI use in archives will need to take into consideration accountability principles such as those outlined by Hurley. Rather than understanding accountability as a narrowly defined bureaucratic goal, the archival tradition holds space for both a formal, institutionally situated accountability and a broader conception of documentation's role in democratic society. The emergent challenge of AI calls for an approach to accountability encompassing both.

A clear assessment of the context-specific accountability relations must be at the basis of an archivist's task in considering the paradata necessary for a given context. This would vary based on the risk context of the specific application. For instance, the EU's AI Act has recently proposed a multilayered risk assessment approach to AI applications, which proscribes AI applications in unacceptably risky contexts, and which outlines minimum requirements for recordkeeping, transparency, and oversight of implementations in high-risk applications. Unacceptable risks include applications which may directly violate individuals' rights and which individuals are not able to grant informed consent to, while high risk contexts may still risk violating individuals' rights but may be regulated to an acceptable level of risk within the specific context of the application. Non-high-risk applications are regulated in the proposal by voluntary codes of conduct, which prescribe practices in line with those governing high-risk applications. Additionally, Title IV of the proposal also calls for transparency on the part of "certain AI systems to take account of the specific risks of manipulation they pose" by, among other things, offering imitations of their real-world equivalents [67]. Similar guidance on accountable AI implementation can be found in the recent NIST AI and Risk Management Framework or the OECD Framework for the Classification of AI Systems, each of which may aid in the implementation of AI tools in responsible manners according to the context and stakes of their application. These proposed regulations and frameworks leave responsible institutions to devise their own implementation-specific recordkeeping practices, since to prescribe detailed requirements would be a fruitless task in such a dynamic technological context. Archivists must instead draw upon emerging trends in the technical and policy literature to keep archival practice in step with new technological developments, without abandoning traditional archival principles. In other words, we cannot expect regulators to do the work of devising specific approaches to AI documentation for us. Paradata is proposed as a category to describe the necessary documentation of AI processes which may have influenced the form in which readers encounter the archival record or records group. Archivists must work from this imperative to devise adequate practices to describe the AI tools used, their effects, the mandate and purpose of their application, and the

individuals selecting and applying these tools. Determining the methods in which this information is recorded and made accessible is the problem ahead for students of paradata.

## 5 Conclusion

This paper has outlined three literatures and perspectives which contribute to the proposed concept of paradata to document AI activity in archives. Every field of research captures and communicates some information about the process of research data collection. In some fields, this is referred to as 'paradata' and entails the collection and dissemination of quantitative or qualitative data revealing the context of data creation to provide transparency into the research process, reveal the choices made and methods used by researchers, and enable informed usage of the dataset by future researchers. The paper intends that the same term, as applied in the archival context, might be used to ensure accountability in the application of AI tools to archival records, informed by the adjacent developments in the XAI fields.

Further research topics related to paradata and AI abound from practical and theoretical standpoints. The implementation of paradata as an integral part of the record will demand a deliberate approach to prevent dissociation of records from their paradata. It may take the form of, as Mooradian has suggested, an "AI record" which is separate from the primary record or records group but associated with it [44]. Alternatively, a basic set of paradata elements may emerge as a component within a record's metadata. Paradata's overlap with existing categories of contextual and administrative metadata, along with its conceptual similarity to the broader category of 'data about data' may lend itself to this approach. Where paradata may fit into existing metadata schemas would be another fruitful direction for further research [49]. Implementations of paradata design in practice may encounter further difficulties in relation to the commercialized basis of many AI tools. If an off-the-shelf AI tool is used in an archive, how much of its process-data will be available for the archivist? Will this information remain a guarded trade secret, or will archivists be able to understand the logical pathways of the tool without challenging its designers' intellectual property? If a need for processual documentation in service of accountability is built into archivists' approach to AI tools from the beginning, then we may be able to avoid situations in which the chief barrier to AI explainability is proprietary rather than technical.

On a theoretical level, AI in archives may create a need to revisit some fundamental archival principles, or devise ways of using paradata to address the challenges AI poses. For instance, AI processing may subtly change a record or record group's provenance; this the current research group hopes might be documented by the collection of thorough paradata, but still must be minimized. Further theoretical and practical research on this matter would clarify the matter, though recent trends towards multifarious conceptions of provenance remain relevant here. From the perspective of archival diplomatics, how might AI processing change the fundamental character of a record – and more importantly from the paradata perspective, how might this be documented? Beyond its own limits, the archival field must stay abreast of developments in XAI and AI governance in order to track changing technical, social, and legal developments in these adjacent fields. As increasing levels of human agency are delegated to AI tools, archivists must find new solutions to document an archivist's activity carried out at arms-length from human agency – and it is in this context that paradata is proposed as a new category of analysis.

## Acknowledgements

This work has been supported by International Research on Permanent Authentic Records in Electronic Systems (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). Scott Cameron is employed directly as a paid member of InterPARES Trust AI at the University of British Columbia, while Babak Hamidzadeh and Patricia Franks are not employed by InterPARES Trust AI. The authors would like to acknowledge the contributions of Jeremy Davet for his guidance during the early phases of this paper's composition.

## REFERENCES

- < bib id="bib1">< number>[1]</ number>Philip Adler, Casey Falk, Sorelle A. Friedler, Gabriel Rybeck, Carlos Scheidegger, Brandon Smith, and Suresh Venkatasubramanian. 2016. Auditing Black-box Models for Indirect Influence. DOI:https://doi.org/10.48550/arXiv.1602.07043</ bib>
- < bib id="bib2">< number>[2]</ number>Herbjørn Andresen. 2019. A discussion frame for explaining records that are based on algorithmic output. *Records Management Journal* 30, 2 (November 2019), 129–141. DOI:https://doi.org/10.1108/RMJ-04-2019-0019</ bib>
- < bib id="bib3">< number>[3]</ number>W. Ross Ashby. 1956. *An introduction to cybernetics*. Chapman & Hall Ltd., London. Retrieved from <http://pcp.vub.ac.be/books/IntroCyb.pdf></ bib>
- < bib id="bib4">< number>[4]</ number>Drew Baker. 2012. Defining Paradata in Heritage Visualization. In *Paradata and Transparency in Virtual Heritage* (1st ed.), Anna Bentkowska-Kafel and Hugh Denard (eds.). Routledge, London, 163–175. Retrieved from <https://doi.org/10.4324/9781315599366></ bib>
- < bib id="bib5">< number>[5]</ number>Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bénéttot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58, (June 2020), 82–115. DOI:https://doi.org/10.1016/j.inffus.2019.12.012</ bib>
- < bib id="bib6">< number>[6]</ number>David Bearman. 1993. The Implications of “Armstrong v. Executive of the President” for the Archival Management of Electronic Records. *The American Archivist* 56, 4 (1993), 674–689.</ bib>
- < bib id="bib7">< number>[7]</ number>Jean-Christophe Bélisle-Pipon, Erica Monteferrante, Marie-Christine Roy, and Vincent Couture. 2022. Artificial Intelligence Ethics has a Black Box Problem. *AI and Society* (2022). DOI:https://doi.org/10.1007/s00146-021-01380-0</ bib>
- < bib id="bib8">< number>[8]</ number>Anna Bentkowska-Kafel. 2012. Processual Scholia: The Importance of Paradata in Heritage Visualization. In *Paradata and Transparency in Virtual Heritage* (1st ed.), Anna Bentkowska-Kafel and Hugh Denard (eds.). Routledge, London, 245–249. Retrieved from <https://doi.org/10.4324/9781315599366></ bib>
- < bib id="bib9">< number>[9]</ number>Anna Bentkowska-Kafel and Hugh Denard (Eds.). 2012. *Paradata and Transparency in Virtual Heritage* (1st Edition ed.). Routledge, London, UK. DOI:https://doi.org/10.4324/9781315599366</ bib>
- < bib id="bib10">< number>[10]</ number>Umang Bhatt, McKane Andrus, Adrian Weller, and Alice Xiang. 2020. Machine Learning Explainability for External Stakeholders. Retrieved September 12, 2022 from <http://arxiv.org/abs/2007.05408></ bib>
- < bib id="bib11">< number>[11]</ number>Jenny Bunn. 2020. Working in contexts for which transparency is important: A recordkeeping view of explainable artificial intelligence (XAI). *RMJ* 30, 2 (April 2020), 143–153. DOI:https://doi.org/10.1108/RMJ-08-2019-0038</ bib>
- < bib id="bib12">< number>[12]</ number>Mark Carnall. 2012. Walking with Dragons: CGIs in Wildlife ‘Documentaries.’ In *Paradata and Transparency in Virtual Heritage* (1st ed.), Anna Bentkowska-Kafel and Hugh Denard (eds.). Routledge, London, 81–94. Retrieved from <https://doi.org/10.4324/9781315599366></ bib>
- < bib id="bib13">< number>[13]</ number>William M. Carter. 2017. Virtual Archaeology, Virtual Longhouses and “Envisioning the Unseen” Within the Archaeological Record. Ph.D. The University of Western Ontario (Canada), London, Ontario. Retrieved October 4, 2022 from <https://www.proquest.com/docview/2714866407/abstract/7C5DB4B639D64D95PQ/1></ bib>
- < bib id="bib14">< number>[14]</ number>Bryan Casey, Ashkon Farhangi, and Roland Vogl. 2019. Rethinking Explainable Machines: The GDPR’s “Right to Explanation” Debate and the Rise of Algorithmic Audits in Enterprise. *Berkeley Technology Law Journal* 34, (2019), 145–189.</ bib>
- < bib id="bib15">< number>[15]</ number>Giovanni Colavizza, Tobias Blanke, Charles Jeurgens, and Julia Noordgraaf. 2021. Archives and AI: An Overview of Current Debates and Future Perspectives. *Journal on Computing and Cultural Heritage* 15, 1 (December 2021), 1–15. DOI:https://doi.org/10.1145/3479010</ bib>
- < bib id="bib16">< number>[16]</ number>Terry Cook. 1994. Electronic records, paper minds: the revolution in information management and archives in the post-custodial and post-modernist era. *Archives & Manuscripts* 22, 2 (November 1994), 300–328.</ bib>
- < bib id="bib17">< number>[17]</ number>Terry Cook. 2013. Evidence, memory, identity, and community: four shifting archival paradigms. *Arch Sci* 13, 2 (June 2013), 95–120. DOI:https://doi.org/10.1007/s10502-012-9180-7</ bib>
- < bib id="bib18">< number>[18]</ number>Mick P. Couper. 1998. Measuring survey quality in a CASIC environment. In *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, American Statistical Association, Dallas, TX, USA, 41–49. Retrieved from [http://www.asasrms.org/Proceedings/papers/1998\\_006.pdf](http://www.asasrms.org/Proceedings/papers/1998_006.pdf)</ bib>
- < bib id="bib19">< number>[19]</ number>Jeremy Davet, Babak Hamidzadeh, Patricia Franks, and Jenny Bunn. 2022. Tracking the Functions of AI as Paradata & Pursuing Archival Accountability. In *Archiving 2022: Final Programs and Proceedings*, Society for Imaging Science and Technology, Springfield, VA, USA, 83–88. DOI:https://doi.org/10.2352/journal.110.2022.19.1.17</ bib>
- < bib id="bib20">< number>[20]</ number>Hugh Denard, Richard Beacham, Franco Niccolucci, Sorin Hermon, and Anna Bentkowska-Kafel. 2009. The London Charter for the Computer-Based Visualization of Cultural Heritage. *London Charter*. Retrieved November 8, 2021 from <http://www.londoncharter.org/downloads.html></ bib>
- < bib id="bib21">< number>[21]</ number>John M. Dirks. 2004. Accountability, History, and Archives: Conflicting Priorities or Synthesized Strands? *Archivaria* (May 2004), 29–49.</ bib>
- < bib id="bib22">< number>[22]</ number>Juan Manuel Durán and Karin Rolanda Jongmsa. 2021. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics* 47, 5 (May 2021), 329–335. DOI:https://doi.org/10.1136/medethics-2020-106820</ bib>
- < bib id="bib23">< number>[23]</ number>Luciana Duranti, Adam Jansen, Giovanni Michetti, Mumma Courtney, Daryll Prescott, Corinne Rogers, and Thibodeau Kenneth. 2016. Preservation as a Service for Trust. In *Security in the Private Cloud*, John R. Vacca (ed.). CRC Press, Boca Raton, FL, 47–72. DOI:https://doi.org/10.1201/9781315372211-5</ bib>
- < bib id="bib24">< number>[24]</ number>Rosalind Edwards, John Goodwin, Henrietta O’Connor, and Ann Phoenix. 2017. Introduction: working with paradata, marginalia and fieldnotes. In *Working with Paradata, Marginalia and Fieldnotes*. Edward Elgar Publishing, 1–19. DOI:https://doi.org/10.4337/9781784715250.00007</ bib>
- < bib id="bib25">< number>[25]</ number>Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Trans. Inf. Syst.* 14, 3 (July 1996), 330–347. DOI:https://doi.org/10.1145/230538.230561</ bib>
- < bib id="bib26">< number>[26]</ number>Timmit Geburu, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for Datasets. Retrieved September 19, 2022 from <http://arxiv.org/abs/1803.09010></ bib>

- < bib id="bib27">< number>[27]</ number> Rachel Hann. 2021. Modelling Kiesler's Endless Theatre: approaches to paradata for heritage visualization. *Theatre and Performance Design* 7, 1–2 (April 2021), 96–115. DOI:https://doi.org/10.1080/23322551.2021.1940455</ bib>
- < bib id="bib28">< number>[28]</ number> High-Level Expert Group on Artificial Intelligence. 2018. *Ethics Guidelines for Trustworthy AI*. European Commission, Brussels. Retrieved September 11, 2022 from https://ec.europa.eu/newsroom/dae/document.cfm?doc\_id=60419</ bib>
- < bib id="bib29">< number>[29]</ number> Chris Hurley. 2005. Recordkeeping and accountability. In *Archives: Recordkeeping in Society*. Chandos Publishing, Wagga Wagga, New South Wales, 223–252. DOI:https://doi.org/10.1016/B978-1-876938-84-0.50009-3</ bib>
- < bib id="bib30">< number>[30]</ number> Isto Huvila. 2012. The Unbearable Complexity of Documenting Intellectual Processes: Paradata and Virtual Cultural Heritage Visualisation. *HUMAN IT* 12, 1 (2012), 97–110.</ bib>
- < bib id="bib31">< number>[31]</ number> Isto Huvila. 2022. Improving the usefulness of research data with better paradata. *Open Information Science* 6, 1 (January 2022), 28–48. DOI:https://doi.org/10.1515/opis-2022-0129</ bib>
- < bib id="bib32">< number>[32]</ number> Livia Iacovino. 2012. Archives as Arsenals of Accountability. In *Currents of Archival Thinking*, Terry Eastwood and Heather MacNeil (eds.). Libraries Unlimited, Santa Barbara, 181–212.</ bib>
- < bib id="bib33">< number>[33]</ number> Lise Jaillant, Katherine Aske, and Annalina Caputo. 2021. *The National Archives (UK): Case study*. Aeolian: Artificial Intelligence for Cultural Institutions. Retrieved December 19, 2022 from https://www.aeolian-network.net/case-study-1-the-national-archives-uk/</ bib>
- < bib id="bib34">< number>[34]</ number> Hilary Jenkinson. 1937. *A manual of archive administration*. P. Lund, Humphries & co., Ltd., London. Retrieved April 3, 2022 from http://archive.org/details/manualofarchivea00iljenk/</ bib>
- < bib id="bib35">< number>[35]</ number> Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, ACM, Barcelona Spain, 306–316. DOI:https://doi.org/10.1145/3351095.3372829</ bib>
- < bib id="bib36">< number>[36]</ number> Alan Karr. 2010. *Metadata and Paradata: Information Collection and Potential Initiatives*. National Institute of Statistical Sciences, Washington, D.C. Retrieved from https://www.niss.org/sites/default/files/research\_attachments/Metadata%20vs%20Paradata-FT.pdf</ bib>
- < bib id="bib37">< number>[37]</ number> Jacob Leon Kröger, Otto Hans-Martin Lutz, and Florian Müller. 2020. What Does Your Gaze Reveal About You? On the Privacy Implications of Eye Tracking. In *Privacy and Identity Management: Data for Better Living: AI and Privacy*, Michael Friedewald, Melek Önen, Eva Lievens, Stephan Krenn and Samuel Fricker (eds.). Springer International Publishing, Cham, 226–241. DOI:https://doi.org/10.1007/978-3-030-42504-3\_15</ bib>
- < bib id="bib38">< number>[38]</ number> Luis A. Leiva, Ioannis Arapakis, and Costas Iordanou. 2021. My Mouse, My Rules: Privacy Issues of Behavioral User Profiling via Mouse Tracking. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, 51–61. DOI:https://doi.org/10.1145/3406522.3446011</ bib>
- < bib id="bib39">< number>[39]</ number> Lars Lyberg, Frauke Kreuter, and Mick Couper. 2010. The use of paradata to monitor and manage survey data collection. *American Statistical Association*, 282–296. Retrieved October 16, 2022 from http://sampeuchair.ec.unipi.it/wp-content/uploads/2018/10/Couper-et-al.pdf</ bib>
- < bib id="bib40">< number>[40]</ number> Clifford Lynch. 2017. Stewardship in the "Age of Algorithms." *First Monday* 22, 12 (December 2017). DOI:https://doi.org/10.5210/fm.v22i12.8097</ bib>
- < bib id="bib41">< number>[41]</ number> Fritz Marx. 1947. The Role of Records in Administration. *The American Archivist* 10, 3 (July 1947), 241–248. DOI:https://doi.org/10.17723/aarc.10.3.f354v0x358486416</ bib>
- < bib id="bib42">< number>[42]</ number> Eileen McIlvain. 2013. Paradata. *NSDL Documentation Wiki*. Retrieved September 9, 2022 from https://wiki.ucar.edu/display/nsdl/docs/Paradata</ bib>
- < bib id="bib43">< number>[43]</ number> Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267, (February 2019), 1–38. DOI:https://doi.org/10.1016/j.artint.2018.07.007</ bib>
- < bib id="bib44">< number>[44]</ number> Norman Mooradian. 2019. AI, Records, and Accountability. *ARMA Magazine*, 9–13.</ bib>
- < bib id="bib45">< number>[45]</ number> Mark Mudge. 2012. Transparency for Empirical Data. In *Paradata and Transparency in Virtual Heritage*, Hugh Denard and Anna Bentkowska-Kafel (eds.). Routledge, London, 177–188.</ bib>
- < bib id="bib46">< number>[46]</ number> Henrietta O'Connor and John Goodwin. 2020. Paradata. In *SAGE Research Methods Foundations*. SAGE Publications Ltd, London. DOI:https://doi.org/10.4135/9781526421036948039</ bib>
- < bib id="bib47">< number>[47]</ number> Jane Parkinson. 1993. Accountability in archival science. University of British Columbia. DOI:https://doi.org/10.14288/1.0086151</ bib>
- < bib id="bib48">< number>[48]</ number> P. Jonathon Phillips, Carina A. Hahn, Peter C. Fontana, Amy N. Yates, Kristen Greene, David A. Broniatowski, and Mark A. Przybocki. 2021. *Four Principles of Explainable Artificial Intelligence*. National Institute of Standards and Technology, Gaithersburg, MD, USA. DOI:https://doi.org/10.6028/NIST.IR.8312</ bib>
- < bib id="bib49">< number>[49]</ number> Gregory Rolan. 2017. Towards interoperable recordkeeping systems: A meta-model for recordkeeping metadata. *RMJ* 27, 2 (July 2017), 125–148. DOI:https://doi.org/10.1108/RMJ-09-2016-0027</ bib>
- < bib id="bib50">< number>[50]</ number> Adam Safir, Tamara Black, and Rebecca Steinback. 2001. Using paradata to examine the effects of interviewer characteristics on survey response and data quality. In *Proceedings of the Annual Meeting of the American Statistical Association*, August 5–9, 2001, 1–6. Retrieved from http://www.asasrms.org/Proceedings/y2001/Proceed/00620.pdf</ bib>
- < bib id="bib51">< number>[51]</ number> Theodore R. Schellenberg. 1984. The Appraisal of Modern Public Records. In *A Modern Archives Reader*. Washington, 57–70. Retrieved April 3, 2022 from http://www.archivists.org/prof-education/pre-readings/FAA/Schellenberg\_Article.pdf</ bib>
- < bib id="bib52">< number>[52]</ number> Fritz Scheuren. 2000. Macro and Micro Paradata for Survey Assessment. In *Satellite meeting to the UN/ECE Work Session on Statistical Metadata*, Washington, D.C.</ bib>
- < bib id="bib53">< number>[53]</ number> Andrew D Selbst and Julia Powles. 2017. Meaningful information and the right to explanation. *International Data Privacy Law* 7, 4 (November 2017), 233–242. DOI:https://doi.org/10.1093/idpl/ixp022</ bib>
- < bib id="bib54">< number>[54]</ number> Edward H. Shortliffe, Randall Davis, Stanton G. Axline, Bruce G. Buchanan, C. Cordell Green, and Stanley N. Cohen. 1975. Computer-based consultations in clinical therapeutics: Explanation and rule acquisition capabilities of the MYCIN system. *Computers and Biomedical Research* 8, 4 (August 1975), 303–320. DOI:https://doi.org/10.1016/0010-4809(75)90009-9</ bib>
- < bib id="bib55">< number>[55]</ number> Azizeh K. Sowan and Louise S. Jenkins. 2010. Paradata: A New Data Source From Web-Administered Measures. *CIN: Computers, Informatics, Nursing* 28, 6 (November 2010), 333–342. DOI:https://doi.org/10.1097/NCN.0b013e3181f698fd</ bib>

- < bib id="bib56">< number>[56]</ number>Richard Stapleton. 1983. Jenkinson and Schellenberg: A Comparison. *Archivaria* (January 1983), 75–85.</ bib>
- < bib id="bib57">< number>[57]</ number>Beth L. Taylor. 2008. The 2006 National Health Interview Survey (NHIS) Paradata File: Overview and Applications. In *Proceedings of the Survey Research Methods Section, American Statistical Association (2008)*. Retrieved from <http://www.asasrms.org/Proceedings/y2008/Files/301266.pdf></ bib>
- < bib id="bib58">< number>[58]</ number>Martin J. Turner. 2012. Lies, damned lies, and visualizations: Will metadata and paradata be a solution or a curse? In *Paradata and Transparency in Virtual Heritage* (1st ed.), Anna Bentkowska-Kafel and Hugh Denard (eds.). Routledge, London, 135–144. Retrieved from <https://doi.org/10.4324/978131559366></ bib>
- < bib id="bib59">< number>[59]</ number>United States Census Bureau. 2021. About Paradata. *census.gov*. Retrieved November 8, 2021 from <https://www.census.gov/topics/research/paradata/about.html></ bib>
- < bib id="bib60">< number>[60]</ number>Frank Upward. 2005. The records continuum. In *Archives: Recordkeeping in society*, Sue McKemmish, Michael Piggott, Barbara Reed and Frank Upward (eds.). Chandos Publishing, Wagga Wagga, New South Wales, 197–222. DOI:<https://doi.org/10.1016/B978-1-876938-84-0.50008-1></ bib>
- < bib id="bib61">< number>[61]</ number>Giulia Vilone and Luca Longo. 2020. Explainable Artificial Intelligence: a Systematic Review. DOI:<https://doi.org/10.48550/arXiv.2006.00093></ bib>
- < bib id="bib62">< number>[62]</ number>Jordan Joseph Wadden. 2021. Defining the undefinable: the black box problem in healthcare artificial intelligence. *Journal of Medical Ethics* (July 2021). DOI:<https://doi.org/10.1136/medethics-2021-107529></ bib>
- < bib id="bib63">< number>[63]</ number>2007. Making Space. *King's Visualization Lab*. Retrieved September 9, 2022 from [https://www.kvl.cch.kcl.ac.uk/making\\_space.html](https://www.kvl.cch.kcl.ac.uk/making_space.html)</ bib>
- < bib id="bib64">< number>[64]</ number>2017. *Information and documentation - Records management processes - Metadata for records - Part 1: Principles*. International Organization for Standardization, Switzerland.</ bib>
- < bib id="bib65">< number>[65]</ number>2017. *Guidelines on Data Protection Impact Assessment (DPIA)*. European Commission, Brussels. Retrieved September 24, 2022 from <https://ec.europa.eu/newsroom/article29/items/611236></ bib>
- < bib id="bib66">< number>[66]</ number>2018. Terminology Database: "metadata." *InterPARES Trust AI*. Retrieved November 12, 2022 from <https://interparestrustai.org/terminology/term/metadata/></ bib>
- < bib id="bib67">< number>[67]</ number>2021. Laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. European Commission, Brussels. Retrieved from <https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence></ bib>
- < bib id="bib68">< number>[68]</ number>2022. Blueprint for an AI Bill of Rights. *The White House*. Retrieved October 10, 2022 from <https://www.whitehouse.gov/ostp/ai-bill-of-rights/></ bib>
- < bib id="bib69">< number>[69]</ number>2022. The principles: Generally accepted recordkeeping principles. *ARMA International*. Retrieved October 11, 2022 from <https://www.arma.org/page/principles></ bib>

Received November 2022; revised March 2023; accepted April 2023