An Intelligent Class: The Development Of A Novel Context Capturing Framework Supporting The Functional Auto-Classification Of Records

Nathaniel Payne
School Of Library, Archival, and Information Studies (iSchool)
University Of British Columbia
Vancouver, Canada
nathaniel.joseph.payne@gmail.com

Abstract— **The need to accurately classify records is a core problem in many domains. Current methods for auto-classification focus on a record's content and not its context. As a result, current auto-classification methods are unable to achieve the levels of precision, accuracy, and recall that match or exceed the levels generated by human classifiers. In order to address this challenge, a new methodology is needed that specifies how to extract contextual features from a record in order to improve the auto-classification accuracy, precision, and recall of records at scale. This paper closes this gap, using the diplomatic definition of context to specify a mapping that will operationalize the capturing of context from a record. This mapping, makes it possible to continue developing a formal method for functional auto-classification and contextual feature extraction that will utilize a record's context to improve functional auto-classification accuracy, precision, and recall.**

Keywords—context, auto-classification, classification, machine learning, functions

I. INTRODUCTION

The need to accurately classify records is a core problem in many domains [1]. Historically, the classification of records was a manual process [2, 3, 4, 5]. This classification was usually done by records managers and archivists who manually "read" records as they were received, categorized those records by distinct attributes, and indexed record attributes accordingly [6]. Unfortunately, due to the significant growth in the volume of records, the classification of records has become not only more complex, but manually infeasible [7]. Furthermore, while machine classification tools have been cited as a solution, sparse literature on these tools within the records domain raises questions about their effectiveness when it comes to records classification [8, 9, 10, 11].Today, one open research question relating to classification is whether automated classification tools achieve categorization that is at the same level of accuracy, precision and recall as a human classifier. It is possible to surmise that poor performance in the classification accuracy, precision, and recall of auto-classification arguably exists because current tools are designed to read the content of a record rather than the context. In order to solve this problem, new tools, new algorithms, and new approaches are needed so that we can overcome the limitations of current auto-classification methods and achieve classification accuracy, precision, and recall that are as good or better than human classifiers. Such tools would significantly accelerate records classification.

In order to begin addressing this gap, this paper will present a new methodology to capture context from records that can be used to create a new method to auto-classify records. Today, current methods for auto-classification **focus on a record's content and not its context**. These methods are unable to achieve the levels of precision, accuracy, and recall that match or exceed the levels generated by human classifiers because of this. Unfortunately, no methodology exists today that specifies how to extract contextual features from a record that can be used to improve the auto-classification accuracy, precision, and recall of records at scale. This is the gap that this paper will address.

Overall, this foundational work is part of a larger body of research that will seek to create a formal method for functional auto-classification and contextual feature extraction that will utilize a record's context to improve the functional auto-classification accuracy, precision, and recall. Without the completion of this paper, this would could not progress. This progression will pave the way for a long term research roadmap that will enable the development of new methods that extract contextual elements from digital records of all types for a variety of use cases.

II. LITERATURE REVIEW

A. Documents vs Records

Today, discussions of records often start by mentioning the term document. For example, if we reference the glossary published by the Society of American Archivists in 1992, we see that a record was defined as 'a document created or received and maintained by an agency, organization or individual in pursuance of legal obligations or in the transaction of business' [12]. Similar definitions of a record can be found in works by Luciana Duranti [13]. Importantly, document, derives from the Latin word docere, which means to teach, and documentum: lesson, proof, instance, specimen [14]. In medieval Latin the word was used to mean "instrument," "charter," or "official paper." In the twelfth and thirteenth

centuries it denoted "lesson" or "written evidence" (old French) [15]. Clanchy suggests that the document may also be an oral instruction (Norman, eleventh century) [16]. Duranti, also proposes that the simplest definition of document is *recorded information* [17, 18]. Recorded information, can thus be thought of as a construct that allow us to organize, navigate, and understand information [19, 20, 21, 22].

While documents can be thought of as recorded information, records on the other hand, are documents in context. Today, many of the issues that arise when records are seen as evidence recur when we examine the relationship between records and information. For example, knowing who created a record (the persons), the act that created the record, the context of the record's creation, the authenticity of the record, and the bonds that a record shares with other records, are important issues that can impact whether a record can be used to provide evidence and thus be considered informative. In this case, similar questions have to be asked when seeking to understand what a record is. Can records be perceived as a *kind* of information, or is information something that records provide to those who use them? In general, perceptions of records are widely different outside the professional community of archivists and records managers [23]. As Yeo and others note, lawyers, legislators, historians, information technologists, librarians, and members of other professional groups are also likely to have their own, perhaps very different, views of records and recordkeeping, as well as the relationship of a record to information [21, 24, 25, 26, 27]. In fact, as Lemieux argues, from her research in the financial sector, there is no single valid conceptualization of the record. Rather, Lemieux argues that there are many valid conceptualizations arising from particular social contexts, and, further, that meaning in records is engendered over time by all those involved in the processes of inscription, transmission, and contextualization, including record-keepers [25].

In general, writings about records often emphasize their role as a source of memory for organizations and the wider society [28, 29]. Records are often linked with collective memory because they transcend the limits of a single human mind. They are often considered extensions of the human memory [30]. According to Greg O'Shea, "if we revisit the definition of a record we see that the concept of evidence is at its heart" [31]. In fact, as Bearman & Trant argue, the consensus, largely developed since 1990, is that records are evidence of transactions (relationships of acts), means of action and information about acts [32]. In this conception, Bearman stresses the evidence of transaction as an accepted definition of records. That said, other viewpoints have also emerged [33]. Luciana Duranti for example, drew on these ideas when she wrote that in jurisprudence "evidence is not an entity, but a relationship . . . between the fact to be proven and the fact that proves it" [34, 35]. In this conception, Duranti, defines a record, or an archival document as "documents (i.e., recorded information) created (i.e., made or received and set aside for action or reference) by a physical or juridical person (i.e., an

individual or an entity capable of rights and duties) in the course of a practical activity, as a means and instrument for it, and preserved." [34, 36].

From Duranti's perspective, there are seven key components of any record. These components are the 1) persons, 2) acts, 3) archival bond, 4) content, 5) context, 6) form, and the 7) medium [37]. Duranti also stresses the three persons essential for understanding a record: 1) Author 2) Addressee, and 3) Writer [37]. Furthermore, Duranti notes that archival documents must satisfy the five characteristics of archival nature, which include naturalness, uniqueness, authenticity, interrelatedness, and impartiality. Duranti, along with others, also argues that a document cannot be considered an archival document unless it has these characteristics of archival nature [38, 39, 40]. Finally, Duranti stresses the importance of the archival bond when understanding what a record is [41]. As Duranti notes, at the core of archival science is the concept of the archival bond, that is, the network of relationships that each record has with the records belonging in the same aggregation [42, 43]. Duranti notes that the archival bond is originary because it comes into existence when a record is created (i.e. when, after being made or received, it is set aside in the fonds of the physical or juridical person who made or received it for action or reference), necessary, because it exists for every record (i.e. a document can be considered a record only if it acquires an archival bond), and determined, because it is qualified by the function of the record in the documentary aggregation in which it belongs [44].

In addition to the evidential value and role as a source of memory, we also see the characterization of records as by-products or memorialized elements. Angelika Menne-Haritz, Randall Jimerson, and others have all characterized records as *by-products*, *remnants,* or *residues* of activities [45, 46]. Suzanne Briet for example, provides a classical example of an antelope running wild, supporting this abstract notion that a document denotes everything that may be represented or memorialized in order to serve as evidence [47]. Others even suggest that records are *the activities themselves* [48]. While some researchers have argued that documents are memory, still other definitions have sought to position records as being extensions of memory. According to the Association of Records Managers and Administrators (ARMA), records are the evidence of what the organization does. Records capture its business activities and transactions, such as contract negotiations, business correspondence, personnel files, and financial statements [49].

Our final historical analysis is a recognition of newer views of the record. Many of these newer views, faced with the pressures of convergence and interdisciplinarity, have started reflecting and acknowledging the role of state in the definition of a record [50]. They have also sought to disassociate the concept of preservation – including the purposeful preservation of data – from the understanding of what a record is [51, 52, 52]. Yeo, for example, defines records as persistent

representations of activities, created by participants or observers or their authorized proxies [54]. While this definition says nothing of transaction, evidence, or authenticity, this representational definition of records has been useful because it is flexible enough to include those criteria without explicitly making them necessary to the definition [55]. At the same time, unlike Bearman, Hunter, and Duranti, Yeo's definition of records also reflects the convergence of disciplines upon which records management theory is built. Thus, his attempt to avoid emphasizing one criteria over another provides flexibility, while seeking to ensure that the concept of a record aligns with changing digital forms [56]. While Lemieux accepts this premise, she also calls out limitations, noting that Yeo's definition is limited and mentions nothing about human perception. She also points out that the aim for any definition should be to ensure that it is not restricted to a certain philosophical stance (e.g. positivist versus post-modernist) [55].

B. What Is Records Classification?

Understanding records classification is not only fundamental to this research, but central to the constantly evolving records and information management domain. Today, discussions surrounding the goals and objectives of records management are varied, with conflicting views as to the best practices for records classification [57, 58]. What's more, the timing of records classification (i.e., throughout the active life of a file, or when the file is closed), as well as the responsibility for it (i.e., records creators or staff dedicated to recordkeeping activities) remain evolving and contentious [59, 60, 61]. In general, the classification of records is primarily function based [62]. This approach comes from the need to make explicit the 'archival bond' that exists among all the records participating in the same activity since the moment of their creation. Functional classification is used because activities are aggregations of functions. Moreover, this approach, is related to the broader documentary, procedural, and provenancial contexts characterizing and uniquely identifying each record [63]. Through utilizing a functional approach to records classification, it is argued that the network of relationships inherent in the nature of any record not only is brought to light, but it is also established and perpetuated. In this way, the meaning of each record in relation to all the others as well as the structure of the whole of records (i.e., the archival fonds) can be understood and transmitted over time [64].

In order to delve a little further into function-based records classification, we must start with understanding what is meant by the term function. Function, is defined in archival diplomatics as the whole of the activities aimed to one purpose [1, 65]. Function, understandably, is an abstraction, and as such it needs a structure made of rules and resources in order to be implemented [66, 67, 49]. While the concept of function may seem logical, its manifestation within the records classification process and the daily activities of archivists can come with challenges [68]. To be precise, determining the function classification of a record, whether manual or automated - is not an easy task [69]. This is because within any organization,

where each function is carried out, as well as the boundaries of either concept (i.e., function and structure), may be so blurred that making a distinction for the purpose of describing only the function, 'abstractly,' is impossible [71]. Arguably, it may be difficult to determine a single function by which a record should be classified, and uncertainty as to whether an archivist has the skills or time to accurately do so. What's more, there is subjectivity inherent in any human classification endeavour. Furthermore, today, with so many records coming in, it may be impossible without automation to accurately classify the function of a record without the aid of auto-categorization methods [72].

To understand the root of this functional approach, we now turn our focus to history. Historically, during medieval times, records were either spontaneously accumulated as they were sent or received, or were deliberately selected, always for practical and operational reasons, to make up series consisting principally of legal titles [59, 72]. One of the first institutions to seek to understand and define best practices when it came to records classification was the Prussian State [73]. During the 17th and 18th century, the Prussian state, developed a new method to organize the records made and received by the government that paved the way to a larger movement towards functional classification [74, 59]. In their approach, all records related to a given business transaction, activity, or procedure - independently of their status of transmission (i.e., degree of perfection, that is, draft, original, or copy), form, or value - were incrementally placed together in discrete physical and logical units, called dossiers or files [75]. These dossiers or files, would then, in turn, be aggregated organically according to various homogeneous criteria (e.g., names of persons or corporate bodies, geographic units, subject-matters, dates) [76]. Within the Prussian approach, competence was deemed to be of key importance, where competence utilized a function based approach with organizational structure corresponding to the hierarchy of functions and activities attributed to any given entity [77, 59].

With this foundation laid in the early 17th and 18th centuries, early bureaucracies began to follow suit. Many of the early bureaucracies began developing records classification schemes that maintained the initial functional approach [78]. Others, on the contrary, while emphasizing a function based approach, did change to reflect the internal structure of the records creating organizations with their hierarchies of departments, divisions, and offices. One of the key figures who next played a central role in our understanding of records classification was Theodore Schellenberg. Schellengerg was the United States (US) National Archivist from 1950 to 1961. As he noted, organizational activities correspond most importantly to function, with public records being the by-products of action and records naturally falling into groups that relate to action [79]. Indeed, as Schellenberg espoused in his 1956 manual, the practice of classifying records by subject matter is something that must be seen as an exception, rather than a rule [80, 81]. In fact, Schellenberg, who was influenced

by the ideas of German archivist Brenneke, strongly rejected the subject-based approach typically used in libraries, noting that this approach would only be suitable for reference and information files. On the contrary, he emphasized the importance of starting from the analysis of an agency's functions, activities, and transactions when classifying records, building a strong case for the functional approach to records classification.

Without question, Schellenberg is an important figure in the discussion of records classification. He is the first author who elaborated a set of principles for classifying records and who highlighted the importance of functional analysis. His classification development rules and, in particular, the hierarchy of functions, activities, and transactions he identified as the basic structure of his functional classification model have become a launching point for discussion and development within the archival community globally. Schellenberg's methods call for one to divide the whole of an agency's functions (defined as "all the responsibilities assigned to an agency to accomplish the broad purposes for which it was established") into two main groups of activities when classifying records [79]. These activities were called "substantive" (i.e., "activities relating to the technical and professional work of the agency") and "facilitative" (i.e., "activities relating to the internal management of the agency [59]. Indeed, as Schellenberg argues, by splitting the structure of the classification into two parts, the first one specific to each agency and the second one potentially shareable among agencies, classification activities avoid redundancy while at the same time establishing a principle of uniformity that enables future interoperability.

While Schellenberg has a significant impact, it is important to also consider the Canadian perspective. In Canada, until the 1980s, most classification systems were subject-based. As Foscarini notes, the first attempt to use a functional approach to classification can be found in the systems developed by the Provinces of British Columbia and Nova Scotia, respectively called ARCS (Administrative Records Classification System), ORCS (Operational Records Classification System), STAR (Standard for Administrative Records) and STOR (Standard for Operational Records) [66]. One Canadian archivist who has played an important role, writing on the role of records classification, and who has reviewed and criticized "functional" records classification methods is Paul Sabourin [49]. At the end of the 1990s, as an archivist of the National Archives of Canada (NA – today Library and Archives Canada – LAC), Sabourin participated in a project for the review of the NA's records classification system known as Subject Classification Guide. Within this review, one of the first challenges that Sabourin, like others, faced, was defining what a function is [1]. As Sabourin & Hurley have noted, a function can be defined as:

- (1) any high level purpose, responsibility, task, or activity which is assigned to the accountability agenda of an institution by legislation, policy, or mandate;

- (2) typically common administrative or operational functions of policy development and program and/or delivery of goods and services;

- (3) a set or series of activities (broadly speaking, a business process) which, when carried out according to a prescribed sequence, will result in an institution or individual producing the expected results in goods or services that it is mandated or delegated to provide.

As he and others have argued, determining what record a function corresponds to can be a difficult task [65]. Furthermore, he argued that one of the key shortcomings of the functional approach is the fact that not every single activity behaves as a structured, repetitive process. As he argues, there are many areas of human endeavour (e.g., academic research, teaching, or artistic performance) that, on the contrary, have the characteristic of being creative and unpredictable. In these cases, the relevant activities or functions may not necessarily follow any pre-established, linear, or cyclical sequence of steps, making the classification of a record according to function difficult [59]. What's more, he noted that some offices may find it more convenient for the purposes of their business, for instance, to keep all of the records originated by an entire process, or even an entire function, together in one single file. In such a case, the higher activity level, not the transaction level, should be tagged as the entry point for file creation [59].

C. Functional Records Classification

Function based classification has been pervasive since the 1990s. However, the idea that records should be classified according to functions has been around for many decades [1]. Historically, Theodore Van Riemsdijk, a 19th Century State Archivist in the Netherlands, is seen by Ketelaar (1996) as an early proponent of using business functions to understand records. In 1877 he suggested that archival documents "reveal their nature and meaning best" when they are "placed in their natural and original context" [82]. Following this, in the early part of the 20th century, Van Laer (1910) pushed further, arguing that the archival 'Principle of Provenance' should be used as the basis of classification [83]. In particular, he described the Principle as, "…a system of arrangement of public archives whereby every document is traced to the governmental body, administrative office or institution by which it was issued or received and to the files of which it last belonged" [83, 84]. Through his arguments, he, as well as others, both inferred and argued that functions and organisational structure strongly corresponded and were critical to the proper classification of records.

In the mid 1950's, Schellenberg, the US National Archivist from 1950-1961 followed this trend, holding considerable influence on records and archives management especially within the area of classification [85, Schellenberg, 1965). In particular, Schellenberg argued that while records could be classified by function, organizational structure or the subject matter of the records [2], "… records, as a rule, should

be classified according to function". From this work, Schellenberg described a hierarchical classification of organizations that was valuable for future work on auto-classification, using the terms 'Functions - Activities - Transactions' (F-A-T) to label the layers within an organization.

Following this, many other authors expanded the work on classification started by Schellenberg. As these authors noted, functional classification provided links between records that originate from the same activity or from related activities. These links could be used to determine where a record should be placed within a larger aggregation of records, to assist users in retrieving and interpreting records, and to assign and control retention periods [5, 86, 87, 88, 89, 90]. Today, most records management resources in the UK and Australia promote functional classification as the primary means of classifying records [91]. Moreover, globally, specifications for electronic records management systems [92] and international records management standards (International Organization for Standardization) advocate functional classification as the best practice for records management. As these resources argue, a functional classification scheme must lie at the heart of any electronic records management system, since such a scheme is linked directly to the business context in which the records were created or transmitted.

D. Understanding Auto-Classification

Automatic classification has been a challenging research issue for decades [6]. In particular, the high cost of manual classification has encouraged the development of new auto-classification methods [93]. Unfortunately, there are many barriers, including weak research as well as uncertainty around how to apply current subject-based methods of automatic classification to functional records classification problems [94, 95].

Functional auto-classification of records refers to the use of computational methods to automatically organize records on the basis of their activity or program (function) and place records into such an organization. To complete an auto-classification task, a computational system must be able to "read" records as they are made or received, categorize those records by distinct attributes, and index record attributes accordingly [6]. A computational device that can auto-classify a record must also be able to manually capture detailed information from various records that can be used to build indexes which support the efficient storage and retrieval of information contained within the record. This includes both content and context. Effective auto-classification applications must be able to deal with binary and multiple classification hierarchies, preserve as much context from a record as possible, have updatable index attributes, as well as the ability to map relationships between records in different areas of the functional hierarchy. Successful auto-classification methods must also be able to differentiate and sort a record between administrative and operational functions, while also supporting timely and accurate record retrieval. Moreover, a robust functional auto-classification system must be able to automatically index a single record, or a group of records, enable mass automated updating of document relationships, contain the ability to use natural language queries which can improve user access to data and information repositories post-process, handle multiple languages, and be platform independent [6].

Over the last 20 years, basic auto-classification systems have evolved, using a variety of computational tools which have improved their classification accuracy rates [96]. Unfortunately, a lack of substantive research in this area has left open the question of the accuracy of these methods, especially relating to functional classification.. Moreover, in addition to the challenges associated with record volumes, other concerns relate to the ability of records managers and archivists to utilize functional classification methods [97]. As a number of researchers have argued [59, 98], significant knowledge gaps exist within the archival and records management communities, with the role of the archivist and records manager evolving rapidly in the last decade. This evolution, along with challenges associated with using functional classification schemes, have made it difficult for records managers and archivists to understand and apply classification schemes manually when time and resources permit [6].

The most simple approach to auto-categorization focuses on decision rules. Decision rule-based systems use rule-based inference to classify documents according to their categories [99]. Algorithms for this type of auto-classification usually construct a rule set that describes the profile for each category. Specific rules are then constructed in the format of "IF condition THEN conclusion", where the condition portion is filled by features of the category, and the conclusion portion is represented with the category's name or another rule to be tested. While decision rule-based approaches are useful they contain a number of advantages that have limited their effectiveness and which merit future research [8]. For example, while association rule algorithms can be formulated to look for sequential patterns which meet the need of users conducting searches in large collections, at high dimensions with large numbers of categories, association rules do not show reasonable patterns and are not able to do variable or categorical selection at scale. The association rule method also fails to produce useful results if the information that the rules are built around does not provide clear support or is easily distinguishable. As a result, as the number of information categories within a repository grows, and the complexity of the queries needed to be used expands, rule based methods typically fall short and are not adequate. Thus, tools that leverage this approach are not currently adequate and need further review, especially when the number of dimensions within the data is high [6, 100].

One method that seems to improve upon the use of rule based methods are vector space methods. Vector space methods create prototype vectors for each class using a training set of records (i.e. the average vector over all training document

vectors that belong to a record class), and calculate similarity between test records and each of the prototype vectors which assign test records to the class with maximum similarity [101]. In general, vector space algorithms are easy to implement, efficient in computation, are fast learners and employ relevance feedback. Unfortunately, the algorithm has low classification accuracy and is low in detail [102]. Moreover, vector space methods often have difficulty dealing with lexical ambiguity and variability, and do not deal well with feature sparseness. Thus, as the number of categories grows within a set, vector space methods will often struggle if the categories present occur at infrequent intervals -- assigning an irrelevant score to texts which may be relevant simply due to sparseness. This can reduce the quality of output, and lead one to advocate to reduce the number of categories that exist within the records system if vector space methods are to be used to analyze the system [6].

Another method that has evolved to improve upon manual classification methods as well as rule based methods is the use of decision trees. Decision trees seek to rebuild the manual categorization of training records by constructing well-defined true/false-queries in the form of a tree structure [103]. In a decision tree structure, leaves represent the corresponding category of records and branches represent conjunctions of features that lead to those categories. Using this logic, organized decision trees can easily classify a record by putting it in the root node of the tree and letting it run through the query structure until it reaches a certain leaf. In general, the decision tree classification method has many advantages which make it widely used in many applications. The main advantage of a decision tree is its simplicity in understanding and interpretation. What's more, the explanation of a given result can be easily replicated by using simple mathematical algorithms, and so can provide a consolidated view of the classification logic. While this is positive, the major risk of implementing a decision tree is that a selected model may overfit the training data. Some of these issues are addressed by using ensembles or random forests. Random forests are a substantial modification of bagging that build a large collection of de-correlated trees, and then averages them [104]. In addition to these issues, the excluding of duplicate information, change in the sequence of analysis of the documents, and/or change in the number of categories (as well as the categories' relationships to each other), can result in either empty branches within a tree, or insignificant branches, overfitting, and other challenges. Moreover, adding more categories can create problems at high dimensions [6].

One final, more recent method that has evolved and could be useful for the record classification task are artificial neural networks. Artificial neural networks are constructed from a large number of elements with an input fan order of magnitudes larger than in computational elements of traditional architectures [105]. These elements, namely artificial neurons, are interconnected into groups using a mathematical model for information processing based on a connectionist approach to computation [106]. Different types of neural network approaches have been implemented for document classification tasks, with little research completed in the area of functional records auto-classification [107]. Today, the drawback of artificial neural networks for the large scale auto-classification task relates to their high computing cost, which consumes high CPU and physical memory usage. Another disadvantage is that the artificial neural networks are extremely difficult to understand, and hard to tune. Moreover, at high dimensions, neural networks that use probability estimation for their posterior probabilities can lack rigor in producing reliable results. What's more, the back propagation algorithm for neural networks often struggles when the number of documents is large, with maximum likelihood classification based approaches -- the most common case - - also struggling [108]. These issues create significant opportunities for future research.

E. Context In Archival Diplomatics

There are many definitions of context from a wide variety of disciplines, but the definition of context that relates to records is that from archival diplomatics, a blend of traditional diplomatic theory and archival theory. Archival diplomatics refers to the study of the creation, form, and transmission of records, and their relationship to the facts represented in them and to their creator, in order to identify, evaluate, and communicate their nature and authenticity [37]. Archival diplomatics gives importance to the broad context of creation by emphasizing the significance of the juridical system (that is, the social body plus the system of rules which constitute the context of the records), the persons creating the records, and the concepts of function, competence, and responsibility; but never distances itself from the reality of the records [109]. Archival science derives its construction of authenticity through the management of aggregates of records with reference to their functional, procedural, and documentary contexts from the principles of archival diplomatics [28].

Within archival diplomatics, the building blocks of context used by classic diplomatists were [37]:

- The **juridical system**, which is the context of records creation
- The **act**, which is the reason for records creation; the persons, which are the agents
- The **procedures**, which guide the actions and determine their documentary residue
- The **documentary form**, which reflects the act and allows it to reach its purpose
- The **archival bond**, which reveals the relationship of a record with all the other records in the same aggregation

Classic diplomatics identified the juridical system as the only relevant context of records creation, and defined it as a social group organized according to shared principles and values that allow it to establish and recognize a body of rules and to give institutions the power to enforce them [110]. Thus, a juridical system is constituted of a community, its institutions, and its legal system. The legal system in this case comprises

positive law and all the other conceptions of binding law (e.g., natural law, ethics, custom) that are held by a community [111].

Today, modern archival diplomatics, having extended the concept of record, has redefined the context of records creation as the framework of action in which the record participates. Moreover, modern archival diplomatics has identified five relevant contexts, proceeding from the general to the specific. They are:

- The **juridical– administrative** context, that is, the legal and organizational system in which the record creating body belongs, as indicated by laws, regulations, etc.

- The **provenancial** context, that is, the record creating body, its mandate, structure, and functions, as indicated in organizational charts, annual reports, the classification scheme, etc.

- The **procedural context**, that is, the business procedure in the course of which the record is created, which, in the modern environment, is often integrated with documentary procedures, as indicated by workflow rules, codes of administrative procedure, classification schemes, etc.;

- The **documentary context**, that is, the archival fonds to which the record belongs and its internal structure, as indicated by classification schemes, record inventories, indexes, registers, etc.; and

- The **technological context**, that is, the characteristics of the technical components of the record system in which the record is created.

III. OPERATIONALIZING CONTEXT

In this paper, the following paradigm is proposed as an approach to operationalize context from the perspective of archival science:

- The **juridical– administrative** context

- The **provenancial** context

- The **procedural context**

- The **documentary context**

- The **technological context**

Figure 1 shows a mapping table. This table maps the traditional archival science definitions of context to clear contextual elements that can be captured during feature extraction.

Archival Science Definition	What To Be Operationalized?	How It Will Be Implemented?
Juridical- Administrative	Juridical Category	Categorical (Public / Private)
	Juridical Relevance	Score (Relevant / Irrelevant)
	Juridical Status	Categorical (Fact / Act)
	Private Act Category	Categorical (Mere / Public)
Provenancial	Record Persons	Array Of Author, Addressee & Writer (Creator & Originator to be considered)
	Place	Categorical (Internal / External)
	Reliability (Writer)	Score (0 through 1)
	Known Source	Categorical (Yes / No)
	Role	Categorical (Scheme Defined)
	Transmission Status (Primitive)	Categorical (First / Not)
	Authenticity	Score (0 through 1)
Procedural	Record Category	Categorical (Dispositive, Probative, Supporting, Narrative)
	Record Phase	Categorical (Initiative, Inquiry, Consultation, Deliberation, Deliberation Control, Execution)
	Simple Act	Categorical (Yes / No)
	Collegial Act	Categorical (Yes / No)
	Contract	Categorical (Yes / No)
	Collective	Categorical (Collective, Multiple, Compound)
	Compound	Categorical (Yes / No)
	Compound Status	Categorical (Continuative, Complex, Act On Procedure)
	Transmission Status (Effective)	Categorical (Yes / No)
Documentary	Index	Unique Record ID
	Temporality (Time)	Record timestamp (UNIX)
	Temporality (Time)	Record day_of_week, hour
	Fond	Email Category (Folder)
	Subject Classification	Subject Header & Theme
	Bonded Elements	Unique Record ID of Related Records
Technological	Message Stability	Evidence Of Mutability (Binary)
	Medium (At Capture)	Categorical
	Medium (At Storage)	Categorical
	Transmission Status (Completeness)	Categorical (Complete / Not)

Fig. 1.   A Mapping Framework For Context Capture

IV. HOW TO OPERATIONALIZE CONTEXT

In order to operationalize this framework, for every record to be classified, an automated feature extraction approach for both standard records meta-data elements and new record contextual elements is recommended to be created and applied. In particular, it is recommended that any new method seeking to utilize this framework for auto-classification leverage the ISO 23081 standard on records management processes and metadata and capture all standard records meta-data elements from the records in our set [112]. Following this, the classification method would then extract new contextual meta-data elements from each record using a novel mapping framework that will be inferred from the records content and context. In particular, rather than using a purely computational

linguistics approach that uses atomic elements of speech, the new classification framework is recommended to use a hybrid machine learning approach that uses the atomic elements of speech as part of a larger machine pattern recognition scheme that includes, and aims to identify, the record's contextual elements. This approach, would seek to build the ontology for the various elements of context as the records sample is processed. Such an approach, would ensure that one does not have to have all variants of classification pre-specified.

For example, for standard record meta-data fields, it is recommended that any future method for auto-classification automatically extract the features from the records content or body, title or subject, as well as other areas, using regular expressions as well as machine learning methods. For fields where the content is uncertain, future methods are recommended to use a probability score and machine learning classifier to classify the unknown fields. All data extracted for the traditional meta-data elements should be stored as accessible meta data tags. Once this process is complete, the method should then scan each record for their contextual features, such as those indicating operational or administrative functions, whether a record refers to a fact or an act, and other dimensions of archival diplomatic context. To do this, the method should create an initial functional ontology for each of the functions from the training set. This initial functional ontology would include an index of relevant terms corresponding to the various functions. The method should also create a primary index of terms and phrases corresponding to acts utilizing computational linguistical theory as a guide. Furthermore, a future auto-classification method should be able to apply natural language processing activities including parsing and stemming to parse record contents and eliminate terms that may not be relevant for the classification process. Latent Semantic Indexing (LSI), principal components for text, as well as artificial neural network methods from deep learning should also be used to parse and extract contextual features from records which are not captured within the standard records meta-data elements.

V. Conclusion

Current methods for auto-classification **focus on a record's content and not its context**. As can be argued, these methods are unable to achieve the levels of precision, accuracy, and recall that match or exceed the levels generated by human classifiers because of their focus on content and not context. Moreover, no methodology exists that specifies how to extract contextual features from a record that can be used to improve the auto-classification accuracy, precision, and recall of records at scale. As can be seen in this paper, by using the diplomatic definition of context to operationalize the extraction of contextual elements from each record, a clear mapping between contextual elements of a record and useful meta-data for auto-classification can be obtained. This mapping, is the first building block of a formal method for functional auto-classification and contextual feature extraction that can utilize a record's context to improve functional auto-classification

accuracy, precision, and recall. This work is essential, paving the way for a long term research roadmap that will enable the development of new methods that extract contextual elements from digital records of all types for a variety of use cases.

VI. Significance

Output from this paper has the potential to make many significant impacts. This paper shares the first example of an attempt to operationalize the definition of a records context that leverages archival diplomatic's definition of context. This paper, also concretely illustrates how to create and operationalize a novel feature extraction methodology that will extract contextual features from a record, thus enabling the auto-classification of a record based on its contextual elements. This work is important, as it represents the starting point for future research around the development of new graphical, deep learning functional auto-classification machine learning algorithms. This work is also critical when it comes to dealing with the challenge of managing records as big data, and is also potentially applicable to the challenge of classifying misinformation to advance social responsibility, global security, threat detection, and crime prevention.

References

[1] Orr, S. (2006). Functions-based classification of records: is it functional?. *Archives and manuscripts*, *34*(1), 44.

[2] Schellenberg, T.R. (1956). Modern archives: principles and techniques. Reprint, Chicago: University of Chicago Press. Schellenberg, T.R. (1965). The management of archives. New York: Columbia University Press.

[3] Guercio, M. (2001). Principles, methods, and instruments for the creation, preservation, and use of archival records in the digital environment. *The American Archivist*, *64*(2), 238-269.

[4] Atherton, J. (1985). From life cycle to continuum: some thoughts on the records management–archives relationship. *Archivaria*, *21*, 43-42.

[5] Shepherd, E., & Yeo, G. (2003). *Managing records: a handbook of principles and practice*. Facet publishing.

[6] Payne, N., & Baron, J. R. (2017, December). Auto-categorization methods for digital archives. In *2017 IEEE International Conference on Big Data (Big Data)* (pp. 2288-2298). IEEE.

[7] Lemieux, V.; Gormly, B.; Rowledge, L. (2014) "Meeting Big Data challenges with visual analytics: The role of records management", Records Management Journal, Vol. 24(2), pp.122-1414

[8] Joseph, P., Debowski, S., & Goldschmidt, P. (2012). Paradigm shifts in recordkeeping responsibilities: implications for ISO 15489's implementation. *Records Management Journal*, *22*(1), 57-75.

[9] Lappin, J. (2015). Evaluating automated approaches against records management principles. *IQ: The RIM Quarterly*, *31*(3), 16.

[10] Sanderson, N. (2019). Records management by stealth: An Australian practitioner's view. *Business Information Review*, *36*(1), 23-29.

[11] Packalén, S., & Henttonen, P. (2016). Recordkeeping professionals' understanding of and justification for functional classification: Finnish public sector organizational context. *Archival Science*, *16*(4), 403-419.

[12] Bellardo, L. J., & Bellardo, L. L. (1992). *A glossary for archivists, manuscript curators, and records managers*. Society of Amer Archivists.

[13] Duranti, L. and Thibodeau, K. (2006), "The concept of record in interactive, experiential and dynamic environments: the view of InterPARES", *Archival Science*, Vol.6 No.1, pp.13-68.

[14] Riles, A. (2006). *Documents: artifacts of modern knowledge*. University of Michigan Press.

[15] Briet, S. (1951). *Qu'est-ce que la documentation?* (Vol. 1). Éditions documentaires, industrielles et techniques.

[16] Clanchy, M. T. (1993). From Memory to Written Record: England, 1066-1307 (London, 1979). *ClanchyFrom Memory to Written Record: England*, 1066-13071979.

[17] Duranti, L. (1995) "Reliability and Authenticity: The Concepts and Their Implications." Archivaria 1 (39) (1995): 5– 10

[18] Duranti, L., Franks, P. (Ed.) (2015) Encyclopedia of Archival Science. Rowman & Littlefield Publishers

[19] Schamber, L. (1996). What is a document? Rethinking the concept in uneasy times. *Journal of the American Society for information Science*, *47*(9), 669-671.

[20] Lund, N. W., & Buckland, M. (2008). Document, documentation, and the Document Academy: introduction. *Archival Science*, *8*(3), 161.

[21] Livelton, T. (1996). MCKEMMISH and PIGGOTT, The Records Continuum: Ian Maclean and Australian Archives' First Fifty Years. *Archivaria*, *41*.

[22] Yeo, J. (2011) Rising To The Level Of A Record? Some Thoughts On Records And Documents. *Records Management Journal*. 21(1)

[23] Penn, I. A., & Pennix, G. B. (2017). *Records management handbook*. Routledge.

[24] Yeo, G. (2007) "Concepts of record (1): evidence, information, and persistent representations", *American Archivist*, Vol.70 No.2, pp.315-43.

[25] Lemieux, V. (2001). "Let the Ghosts Speak: An Empirical Exploration of the 'Nature' of the Record". Archivaria(51): 81-111

[26] Cox, R. J. (2001). *Managing records as evidence and information*. Greenwood Publishing Group.

[27] Yeo, G. (2018) Records, Information and Data: Exploring The Role Of Record Keeping In An Information Culture. Facet Publishing. London, England.

[28] Schwartz, J. M., & Cook, T. (2002). Archives, records, and power: The making of modern memory. *Archival science*, *2*(1-2), 1-19.

[29] Brothman, B. (2001). The past that archives keep: memory, history, and the preservation of archival records. *Archivaria*, *51*, 48-80.

[30] Brothman, B. (2006). Archives, life cycles, and death wishes: a helical model of record formation. *Archivaria*, *61*, 235-269.

[31] O'Shea, G. (1996). "Keeping Electronic Records: Issues and Strategies," *Provenance* 1, no. 2 (1996)

[32] Bearman, D., Trant, J. (1997). Electronic records research: working meeting: May 28-30, 1997: a report from the archives community. D-Lib Magazine, 3(7-8)

[33] Hedstrom, M. (1996) "Cohesion and Chaos: The State of Archival Science in the United States," in *The Concept of Record: Second Stockholm Conference on Archival Science and the Concept of Record, 30–31 May 1996*, ed. Kerstin Abukhanfusa (Stockholm: Riksarkivet, 1998).

[34] Duranti, L. (2001). Concepts, Principles, and Methods for the Management of Electronic Records. The Information Society: An International Journal, 17(4), 271

[35] Duranti, L. (2002) "The Concept of Electronic Record," in Luciana Duranti, Terry Eastwood, and Heather MacNeil, *Preservation of the Integrity of Electronic Records* (Dordrecht: Kluwer, 2002), 9.

[36] Duranti, L. Endicott-Popovsky, B. (2010) Digital Record Forensics: A New Science And Academic Program For Forensic Readiness. Journal Of Digital Forensics, Security and Law, Vol 5(2)

[37] Duranti, L. (1998). Diplomatics: New Uses for an Old Science. Lanham, MD, and London: The Scarecrow Press. Archivaria 28. Part 1.

[38] Eastwood, T. (1994) "What is Archival Theory and Why is it Important?" Archivaria 37 (Spring 1994): 122-30.

[39] Eastwood, T. (2006) Building archival knowledge and skills in the digital age. Arch Sci (2006) 6:163–170

[40] Eastwood, T. (2010). A contested realm: the nature of archives and the orientation of archival science. Currents of archival thinking, 3-21.

[41] Duranti, L. (1997) "The Archival Bond," Archives & Museum Informatics 11, nos. 3-4: 213-18

[42] Duranti, L. (1998b). "Archival Science," in Encyclopedia of Library and Information Science, edited by Allen Kent, vol. 59. New York, Basel, Hong Kong: Marcel Dekker, 1996: 1-19

[43] Lemieux, V. L., & Sporny, M. (2017, April). Preserving the archival bond in distributed ledgers: a data model and syntax. In *Proceedings of the 26th International Conference on World Wide Web Companion* (pp. 1437-1443). International World Wide Web Conferences Steering Committee.

[44] Duranti, L. (1997) "The Archival Bond," Archives & Museum Informatics 11, nos. 3-4: 213-18

[45] Menne-Haritz, A. (2004) Business Processes: An Archival Science Approach to Collaborative Decision Making, Records, and Knowledge Management (Dordrecht: Kluwer, 2004), 11

[46] Jimerson, R. C. (2003). Archives and memory. OCLC Systems & Services: International digital library perspectives, 19(3), 89-95.

[47] Buckland, M. (1998) "What Is a Digital Document?" In Document Numérique Paris. 2(2).

[48] Higgs, Edward. (1996) Record and Recordness: Essences or Conventions?," in *The Concept of Record*, 105.

[49] Sabourin, P. (2001). Constructing a function-based records classification system: Business activity structure classification system. *Archivaria*, *51*, 137-154.

[50] Ilerbaig, J. (2019). Records, Information and Data: Exploring the Role of Record-Keeping in an Information Culture. *The American Archivist*, *82*(1), 221-225.

[51] Brown, C. (2018). *Archival futures*. Facet Publishing.

[52] Moore, R. (2008). Towards a theory of digital preservation. *International Journal of Digital Curation*, *3*(1).

[53] Tough, A., & Moss, M. (Eds.). (2006). Record keeping in a hybrid environment: managing the creation, use, preservation and disposal of unpublished information objects in context. Elsevier.

[54] Yeo, G. (2007) "Concepts of record (1): evidence, information, and persistent representations", *American Archivist*, Vol.70 No.2, pp.315-43.

[55] Lemieux, V., & Limonad, L. (2011). What 'good'looks like: understanding records ontologically in the context of the global financial crisis. *Journal of Information Science*, *37*(1), 29-39

[56] Wand, Y.; Weber, R. (1995) On the deep structure of information systems. *Information Systems Journal*. 5(3) 203–223

[57] Cook, M. (2017). The management of information from archives. Routledge.

[58] Bak, G. (2012). Continuous classification: capturing dynamic relationships among information resources. *Archival Science*, *12*(3), 287-318.

[59] Foscarini, F. (2009). Function-based records classification systems: an exploratory study of records management practices in central banks (Doctoral dissertation, University of British Columbia).

[60] Gilliland, A. J., & Caswell, M. (2016). Records and their imaginaries: imagining the impossible, making possible the imagined. *Archival Science*, *16*(1), 53-75.

[61] Cook, T. (2011). The archive (s) is a foreign country: historians, archivists, and the changing archival landscape. *The American Archivist*, *74*(2), 600-632

[62] Alberts, I., Schellinck, J., Eby, C., & Marleau, Y. (2013, October). Bringing together Functional Classification and Business Process Analysis: Growing Trends in Records Management. In *Proceedings of the Annual Conference of CAIS/Actes du congrès annuel de l'ACSI*.

[63] Lemieux, V. (2016). Building Trust in Information: Perspectives on the Frontiers of Provenance, Springer

[64] Atherton, J. (1985). From life cycle to continuum: some thoughts on the records management–archives relationship. *Archivaria*, *21*, 43-42.

[65] Hurley, C. (1993). What, if anything, is a function? *Archives and Manuscripts 21*(2), 208-18.

[66] Foscarini, F. (2006). Records classification and functions: an archival perspective. *Knowledge Organization*, *33*(4), 188-198.

[67] Foscarini, F. (2012). Understanding functions: an organizational culture perspective. *Records Management Journal*, *22*(1), 20-36.

[68] Joseph, P., Debowski, S., & Goldschmidt, P. (2012). Paradigm shifts in recordkeeping responsibilities: implications for ISO 15489's implementation. *Records Management Journal*, *22*(1), 57-75.

[69] Seol, M. W. (2012). Research trends and issues of records and archives classification in Korea. *Journal of Korean Society of Archives and Records Management*, *12*(3), 203-232.

[70] Henttonen, P., & Kettunen, K. (2011). Functional classification of records and organisational structure. *Records Management Journal*, *21*(2), 86-103.

[71] Bak, G. (2012). Continuous classification: capturing dynamic relationships among information resources. *Archival Science*, *12*(3), 287-318.

[72] Franks, P. C. (2013). *Records and information management*. American Library Association.

[73] Ernst, W. (1999). Archival action: the archive as ROM and its political instrumentalization under National Socialism. *History of the Human Sciences*, *12*(2), 13-34.

[74] Berger, S. (2013). The role of national archives in constructing national master narratives in Europe. *Archival Science*, *13*(1), 1-22.

[75] Lytle, R. (1980). A national information system for archives and manuscript collections. *The American Archivist*, *43*(3), 423-426.

[76] Müller, P. (2013). Archives and history: Towards a history of 'the use of state archives' in the 19th century. *History of the Human Sciences*, *26*(4), 27-49.

[77] Fritzsche, P. (2005). The archive. *History & Memory*, *17*(1-2), 13-44.

[78] Campbell, E. G. (1941). Functional classification of archival material. *The Library Quarterly*, *11*(4), 431-441.

[79] Schellenberg, T. (1961). Archival principles of arrangement. *The American Archivist*, *24*(1), 11-24.

[80] Schellenberg, T.R. (1956). Modern archives: principles and techniques. Reprint, Chicago: University of Chicago Press. Schellenberg, T.R. (1965). The management of archives. New York: Columbia University Press.

[81] Schellenberg, T. (1968). Archival training in library schools. *The American Archivist*, *31*(2), 155-165.

[82] Ketelaar, E. (1996). Archival theory and the Dutch manual. *Archivaria*, *41*.

[83] Van Laer, A.J.F. (1910). 'The work of the international congress of archivists and librarians at Brussels, August 28-31, 1910', in American Historical Association. Annual Report 1910. Washington, pp. 282-292.

[84] Sweeney, S. (2008). The ambiguous origins of the archival principle of" provenance". *Libraries & the Cultural Record*, *43*(2), 193-213.

[85] Cappon, L. J (1970). 'Theodore R. Schellenberg 1903 – 1970: prodigious worker and archival envoy'. American Archivist, 33 (2), pp. 190-1.

[86] Schellenberg, T.R. (1975), Modern Archives. Principles and Techniques, The University of Chicago Press, Chicago, IL.

[87] Smith, K. (2007), Public Sector Records Management. A Practical Guide, Ashgate, Aldershot.

[88] Williams, C. (2006), Managing Archives. Foundations, Principles and Practice, Chandos Publishing, Oxford.

[89] Tough, A. (2006), "Records and the transition to the digital", in Tough, A. and Moss, M. (Eds), Record Keeping in a Hybrid Environment. Managing the Creation, Use, Preservation and Disposal of Unpublished Information Objects in Context, Chandos Publishing, Oxford.

[90] Findlay, C. (2008), "Digital recordkeeping", in Bettington, j., Eberhard, K., Rowena, L. and Smith, C. (Eds), Keeping Archives, 3rd ed., Australian Society of Archivists, Brisbane, pp. 477-507.

[91] Orr, S.A. (2005), "Functions based classification of records: is it functional?", School of Informatics, Division of Information & Communication Studies, Northumbria University, Newcastle

[92] Archives New Zealand (2005), "Electronic record-keeping systems standard", June. Archives New Zealand

[93] Golub, K. (2006). "Automated subject classification of textual web documents". Journal of Documentation. 62(3):350-71.

[94] Santangelo, J. (2009). RISE OF THE MACHINES. *Information Management*, *43*(6), 22.

[95] Büttner, G. (2019). Auto-classification in an international organization: report from a feasibility study. *Comma*, *2017*(2), 15-26.

[96] Walsh, G. M. (2019). The disconnect between archival descriptive technique and records management taxonomies. *Comma*, *2017*(2), 35-42.

[97] Lappin, J. (2015). Evaluating automated approaches against records management principles. *IQ: The RIM Quarterly*, *31*(3), 16.

[98] Gunnlaugsdottir, J. (2006). The implementation and use of ERMS: A study in Icelandic organizations. Tampere University Press.

[99] Chidanand A., Damerau, F., and Weiss, S.M.. (1994). "Automated Learning of Decision Rules for Text Categorization," ACM Transactions on Information Systems (TOIS), 12 (3), pp. 233 – 251.

[100] Sundaram, S. (2012) Comparative Study of Data Mining Algorithms For High Dimensional Analysis. International Journal of Advances in Engineering & Technology, 4(2), pp. 173-8.

[101] Esteva, M., Xu, W., Tang, J. F., & Padmanabhan, K. A. (2013, November). Data mining for big archives analysis: a case study. In *Proceedings of the 76th ASIS&T Annual Meeting: Beyond the Cloud: Rethinking Information Boundaries* (p. 73). American Society for Information Science.

[102] Khan, A., Baharudin, B., Lee, L.L., and Khan, K. (2010). "A Review Of Machine Learning Algorithms For Text-Document Classification. Journal Of Advances In Information Technology." 1(1), pp. 4-21.

[103] Hastie, T., Tibshirani, R., Friedman, J., & Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, *27*(2), 83-85.

[104] Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

[105] Biddle, O. C. D., & Reath, L. L. P. (2017). Auto- Categorization Methods For Digital Archives.

[106] Zurada, J. M. (1992). *Introduction to artificial neural systems* (Vol. 8). St. Paul: West publishing company.

[107] Marinai, S., Gori, M., & Soda, G. (2005). Artificial neural networks for document analysis and recognition. *IEEE Transactions on pattern analysis and machine intelligence*, *27*(1), 23-35.

[108] Priyadarshini, R., Dash, N., Swarnkar, T., & Misra, R. (2010). Functional analysis of artificial neural network for dataset classification. *Special Issue of IJCCT*, *1*(2), 49-54.

[109] Duranti, L. (2010). From Digital Diplomatics to Digital Records Forensics. *Archivaria, 68*, 39-66.

[110] MacNeil, H. (2013). *Trusting records: legal, historical and diplomatic perspectives* (Vol. 1). Springer Science & Business Media.

[111] Macneil, H. (2000). Creating and maintaining trustworthy records in electronic systems: archival diplomatic methods. In *Trusting Records* (pp. 86-112). Springer, Dordrecht.

[112] International Organization for Standardization (2006), ISO 23081-1. Information and Documentation. Records Management Processes. Metadata for Records. Part 1. Principles, ISO, Bonn.