Archival Research Theory: Putting Smart Technology to Work for Researchers

Kenneth Thibodeau
National Archives and Records Administration (retired)
Denver, Colarado, US
ORCID 0009-0002-9652-082X

Alex Richmond
Bank of Canada
Ottawa, Ontario, Canada
ORCID 0009-0008-7856-3758

Mario Beauchamp
Carlton University
Ottawa, Ontario, Canada
ORCID 0009-0008-5636-5566

***Abstract*—This paper describes a research project that aims at extending the domain of archival theory and science so that it can actively support researchers attempting to exploit on the informative potential of archives. It builds on the conceptual foundation of Constructed Past Theory, semiotics, type theory and the polymorphic entity relation attributed data model.**

***Keywords*—archival bond, archival theory, Constructed Past Theory, semiotics, type theory**

I. INTRODUCING ART

We introduce a research agenda that aims to expand the domain of archival theory to support the conduct of research in archives. We distinguish archival theory from archival science by designating as theory the assumptions and articulations that constitute the conceptual foundation underlying the methods, techniques and exercise of archival functions. Archival science collects empirical evidence to test and validate, correct or enhance archival theory.

Archival theory has been developed, predominantly by and for archivists, as a conceptual basis for the functions that that govern creation, organization, disposition, preservation and access to both current and historical archives. Thus, it can be accurately described as archival management theory (AMT). We propose to extend the domain of archival theory to facilitate and enrich research in archives. Hence, the extension is called archival research theory. This article presents only a high level overview of Archival Research Theory (ART).

AMT does not ignore research in archives. It provides information that enables researchers to identify records relevant to their interests, to assess their authenticity, chain of custody, chain of preservation, etc, as well as ensuring enforcement of access rights and restrictions. But AMT leaves the actual conduct of research in archives to each researcher. That is unavoidable for archives in hard copy for the simple reason that once researchers have opened a box or folder of paper documents, photographs, et al., they are on their own. But for born digital and digitized archives, archival programs can support the conduct of research. Moreover, for digital archives some support for research is necessary because they require appropriate platforms for access and use. Given the open-ended challenges of digital preservation, researchers may not have access to the technologies needed to access authentic digital records. Given that archival programs require access to appropriate technologies to perform their functions, we propose that they include support for the conduct of research in their systems. The additional functionality would also be useful in the work of archivists. Support for research in digital archives should be guided by the principle that archival programs should not bias or interfere in the process of finding and exploiting data that are responsive to a researcher's objectives.

ART does not intend to replace or conflict with AMT, but to supplement it and to contribute to improving it. Nevertheless, expanding the theoretical domain will inevitably engender some revision or specialization of concepts as well as of methods and techniques. For example there is an apt bias for standardization in description. It serves both professionals and archival institutions via consistency in managing holdings and reducing risks. It also benefits researchers interested in different, even unrelated, ensembles of records in one or more archival institutions. In contrast, support for research in archives has to prioritize applicability and adaptability to diverse researcher interests, and fidelity to variety within archives. ART should not limit what researchers can find in archives or constrain the level of specificity they need in characterizing it.

ART goes beyond AMT to encompass anything that researchers may be interested in, assuming that there is more interest in what archives can tell about the people, events, actions, and other things in the past than about the characteristics and relationships of records as such.

The articulation of ART begins with the precondition that a researcher has identified records of interest and been granted access to them. Facets of the articulation that are most directly related to core archival concepts will be described below, but first we present the motivations for ART and criteria that guide its development. Ideally, ART applications should be available to researchers to use at their discretion.

The primary motivation for ART is its potential benefit first of all for researchers, but also to archival programs. Facilitating research in archives and enhancing its fruitfulness directly benefits researchers and it should increase support for archival programs because the value of historic archives is realized when they are used. Smart technologies offer the greatest potential for realize such benefits. Smart technology includes not only artificial intelligence (AI), but also other well-grounded and well-proven technologies that avoid AI's problems like hallucination and accountability. Our strategy is to facilitate the use of different types of AI appropriate for exploration and discovery complemented by other technologies that ensure the quality, appropriateness and coherence of

discovered data, enable sophisticated analysis, and offer possibilities for cumulative and collaborative benefits both among researchers and between researchers and archival programs. This reinforces ART's focus on digital archives.

Several criteria that follow from the orientation towards digital empowerment can also improve the quality of research results. Criteria that guide the development and application of smart technologies to support research include:

Clarity: definitions of terms of ART should be unambiguous, univocal, applicable to any data found in or generated from research in archives and readily implementable in software. Clarity requires not using the term, archives, as used in AMT where it has four different meanings: (1) the archival fonds produced by a records creator in its activity, (2) a building where the records are stored, (3) an organization responsible for them, (4) the totality of the holdings of such an organization, even when they include multiple fonds. Things are even worse with AMT concept of 'provenance' whose meaning has been the subject of debate for decades and where proposed alternatives would bring so many disparate factors under its umbrella that, if adopted, would make the term unworkable [7]. Provenance, however defined, has shown its value in managing archives, but it is not needed in research in archives because ART researchers have already identified records they are interested in, and ART leaves it up to each researcher to determine what types of things that could influence the production of archives are relevant in a given research project and to explore whether and how these factors shaped the records that are valuable for their research objectives.

Comprehensive: production and management of research data should be grounded in a compressive schema that encompasses all the types of data that researchers may look for. Doing this requires many terms that are not necessary for management functions. Given that archives can document any and all human activities, a comprehensive schema has to be articulated at a high level of abstraction. We are doing this in a way that enables both extension for increased specificity and adaptation for different needs.

Consistent: empirical, logical and functional consistency is necessary to support application of smart technologies and to generate high quality results in generating and extracting data from research in archives and in subsequent analysis.

Cumulative: it should be technically possible to accumulate data found in or generated from archives in a single structured data store. The compilation of data should not be constrained by organizational boundaries; that is, it should be possible for archival institutions to complement their linked open data with data generated in archival research. Given ART's precondition of approved access, a researcher should be able to compile data acquired from different archival portals whose use is required because of variant access rights and restrictions. It would also be valuable for researchers if the data store could encompass data from sources other than the records themselves, such as archival descriptions and appraisal reports and also other sources, such as histories, linguistic analysis and findings of social sciences.

These four criteria are technical in nature. Others relate to the substance of research in archives:

Implementation of ART should facilitate collaboration. It should be possible for researchers to access and use data that resulted from prior research in the same records, provided the earlier researchers agree to sharing. This entails some management challenges in that a researcher who agrees to share data should have the option of stipulating that the sharing not reveal the researcher's identity. However, the data would have to be linked to the researcher to enable that researcher's work over time. Ideally, researchers should be able to share selectively in order to protect original research findings.

Another criterion addresses a fundamental issue in research in archives: discerning meaning. It should be possible for researchers to recognize, determine and utilize differences between what things in archives mean to them and what they meant to the parties involved in producing the archives. Appreciating what things meant in the past cannot achieved solely via definitions that were contemporaneous with the documents in archives. Consider named entities. AMT needs to address the issue that an entity designated by a name, such as a nation, can change over time. But ART needs to recognize that what a name meant to someone creating or using a record could vary significantly between instances and in different contexts. What "America" meant to someone deciding to emigrate to the U.S. was probably very different from what it meant after several years living there, and between what it meant as an epithet for the general situation in which a person found herself and in different interactions with government agencies.

Closely related to discernment of meaning is complexity. A researcher may apply concepts that are posterior to, different from and even unknown or unknowable to people involved in the production of archives. Nevertheless, such concepts can be appropriate to and even enrich interpretation of historical materials. Examples include applying modern medical knowledge to understand the morbidity and mortality that resulted from the introduction of contagious diseases in the colonization of the Americas and using techniques such as ground penetrating radar or electron microscopy in archaeological discovery and analysis.

II. The Aim: Maximizing Informative Potential

An archive can be seen one dimensionally as an information store. ART looks at an archive from a two dimensional perspective in which an archive is not seen as a static repository of information but as having informative potential, the ability to provide information responsive to specific interests [18]. Informative potential is determined by the intersection of what an archive is and contains with what someone wants to learn and how they go about learning. ART distinguishes three dimensions of the informative potential of an archive. The first is what can be learned from the face of records: their explicit form and content. The second comes from their interrelationships, both the archival bond as a whole and the subgraphs and edges within it. The third comes from the context of archive production, not only the activities of the archive producer, but also the roles of interacted in those activities or reacted to them.

III. Terms of ART

ART's formal terminology is extensive, comprising all terms defined in the ART schema. This section only addresses terms that have a broad scope and correlate with, but differ, from major terms in AMT. Other terms that are not common in

AMT, but are taken from other disciplines and used as defined in those disciplines are note defined in this article.

Archive: ART departs from common archival terminology English by using 'archive' in the singular. The multiple senses of 'archives' mentioned above are only part of the motivation for this departure. The main reason is ART adopts Cencetti's concept of the archive as a "universitas rerum," a closed cosmos or ensemble that differs from other collections, such as those of libraries and museums, in that all its constituents are related by their use by an agent in an activity [4]. In mathematical terms, an archive is a complete non-uniform hypergraph. It is complete in that there is nothing (no node) in it that is not related to at least one other node by its actual or expected use by an agent in an activity. It is non-uniform in that there is no reason to restrict the number of nodes that can share a relation (edge) a priori. For example, a record series can contain any number of dossiers and a dossier any number of records.

The concept of archive as closed cosmos is the basis for the concept of the archival bond. The archival bond has often been equated with the filing system in which records are organized. ART recognizes that an arrangement of records in a filing system expresses the archival bond but, it is not equivalent. For example, in organizations the existence, documentary characteristics and content of many records conform to directives, but copies of directives are not included in the case files that implement them. Furthermore, the archival bond includes relationships establishes by inclusion explicit references and hyperlinks between records.

Archive producer: On its face, 'records creator' is a misleading term. An agent does not create any received message. In the InterPARES glossary, something becomes a record when the agent sets it aside for reference or use Fidelity to the concept of the archival bond does not even require that the agent keep a record. But it is necessary that the agent keep it for it to be part of that agent's archive (see discussion of 'record' below) [28]. The archive producer is the agent responsible for assembling and keeping the records that constitute an archive.

Record: Given Cencetti's definition of the archival bond, it is ontologically prior to and independent of whether an information object is subjected to a records management regime. The threshold criterion in ART is that a record is used by an agent in an activity. In this sense ART's definition of record is closer to the common sense use of the term than to archival definitions. Consider a case where the archives of two individuals include their correspondence. Suppose one archive includes a letter received from the producer of the other, but that letter is not in the author's archive. Provided the archive that does contain it is seen as trustworthy, the received letter does relate to its author's activity. Not only does it satisfy the definition, but it would be reasonable for a researcher to consider it a record of the author's activity.

The keeping and organization of an archive by its producer increases its informative potential. ART accommodates this by defining two dependent types: kept record and managed record. They are both subtypes of record. A kept record relates it to an action the archive producer took to have physical possession or to control the retention of the record. A managed record relates it to a records keeping regime. The kept record type allows ART to accommodate the concept of record as sediment that is mentioned in the introduction to the Dutch manual and has been developed more formally in Italian archival tradition [10] [16]. Managed record also allows ART to treat information that an archive producer does not keep, but regularly uses in its activity, such as a proprietary online information resource that the agent subscribes to in order to obtain information critical to its activity. Knowing how an agent selected, obtained and used such data would be essential to a researcher trying to understand how the agent carried out its activity.

Archival ensemble: ART uses ensemble, rather than aggregate, to designate sets of records at any level within an archive. A pile of rocks is an aggregate. A stone wall is an ensemble. Ensemble better indicates that series, sub-series et al. are put together by the archive producer and that action increases their informative potential.

IV. THE ART APPROACH

ART builds on concepts and perspectives from other relevant disciplines, especially semiotics, Constructed Past Theory and type theory.

A. Semiotics

The challenge that researchers face in recognizing and respecting differences in what archives mean to them and what they meant to their producers entails a challenge for archival programs that support research: respecting variant approaches and meanings of different researchers. ART address such variations via semiotics, the discipline that studies how meaning is determined. Traditionally, semiotics has focused on the sign as the basic unit of meaning. In the dominant strand of semiotics, a sign is something that stands for something else in some way for someone. For example, the color, red, can represent anger in some cultures and good fortune in others. Although the terminology varies, the elements of a sign are often call the object, what it refers to, the sign vehicle, which stands for the object, and the interpretant, how the other two are related.

Bio-, cognitive, and computational semiotics, substantially enrich possibilities for discerning and respecting differences in meaning. Biosemiotics expands the focus of traditional, philosophical semiotics from humans to all living organisms and from the sign to how meaning is determined in the context of organisms' interactions with their environment [13] [20].

Cognitive semiotics combines the theoretical sophistication of semiotics with the empirical methods and findings of cognitive neurology and psychology [23] [32]. Its differentiation and integration first, second and third person perspectives on meaning can contribute to distinguishing what records meant to the persons who produced them, persons to whom they were addressed and persons who were the subjects of record content, as well as to researchers and others, who interpret records after the fact [5] [17].

Computational semiotics applies semiotic concepts to systems, applications, and human/computer interactions. Like biological systems in biosemiotics, computational semiotics treats systems as operating in sign-based environments where meaning emerges through interaction [11]. Computational semiotics supports use of semiotic concepts in development of applications for ART as well as in understanding the impacts of information and communication technologies on how humans

determine meaning; for example, how computer applications can serve as proxies for human cognitive actions. This insight can be very helpful in research not only in born-digital archives but also to any case where computers have a significant impact on activities, whether documented in digital or analog form [8].

The ART schema, as illustrated in figure 1, below, incorporates semiotics' distinction between type and token, where type is abstract and token concrete [14]. ART uses 'semiotic object' instead of 'type', reserving that term for the supertype of all other types as it is used in type theory. An obvious example of the relevance of this distinction to the archival domain is that of directives. Every organization has directives and typically a system which identifies all the basic directives they need. For example, DoD 5015.2 designates the Department of Defense's records management directive. Directive is an abstract subtype of document. DoD 5015.2 is a still abstract subtype of directive, unique and global for the DoD directives system, but instantiated in tokens that are successive versions. ART defines two subtypes of token: impression and expression. An impression is an instantiation within a system, such as a brain or a computer, and is correctly interpretable only within the confines of that system or other interoperable ones. An expression is an instantiation outside of a system, such as printed on paper, displayed on a screen or made audible through a speaker, that can be interpreted by different systems or agents. In this, ART follows the pattern of functional programming languages where input and output are pushed to the periphery.

The use of semiotic object also serves to differentiate between a sign, which has only one meaning in any context, and constructs with multiple meanings. Sign is a subtype of semiotic object that relates a single referent to a single object. A semiotic object can be a composite in a reflexive relation, where parts themselves can be semiotic objects. Parts of a composite semiotic object can have meanings that differ from that of the whole as well as from those of other parts.

B. Constructed Past Theory

Constructed Past Theory (CPT) uses semiotic concepts to address the complications of discerning and respecting differences between what things meant to people in the past and their meanings for those who want to learn about the past. CPT echoes the findings of cognitive neurology and psychology that our memories are not things we store in particular locations in our brains, like documents in file folders, but are embodied in neural networks that are subnets of the entire brain and that memories are not simply recalled, but dynamically reconstituted by activation of neurons in those networks in combination with current perception, cognition and even emotion [3] [24]. Analogously, CPT asserts that anything we say or write about the past is a product of cognitive construction. The product is a constructed past. CPT enables fine tuning ART to support an open ended variety of research questions and approaches. [27].

CPT defines the things a researcher is interested in learning about: the who, what, when, where and why, as the sphere of interest. Different researchers can apply different approaches in the way they formulate questions, the criteria that characterize suitable and sufficient answers and the way they analyze data. CPT designates the application of a particular approach to a sphere of interest as determining an intentional domain that shapes the conduct of research. CPT characterizes research with any significant level of complexity as a dynamic process. Findings or insights gained at any stage of the research can lead the researcher to modify the intentional domain.

To facilitate clear distinction between what something means to a research and what it meant to someone involved in the activities that generated archives, CPT distinguishes information found in archives, as information from the past, from information produced in interpreting archives or other sources as information about the past [27].

C. Type Theory and TypeDB

To satisfy the criteria guiding the development of ART, CPT grounds its data model in the mathematics of type theory and implements it using TypeDB data management software. Type theory provides unrivaled expressiveness and logical cogency for modeling complex domains through its ability to encode both data and logic within a single coherent framework. Its suitability and advantages for developing computer applications is evident in the use of typed lambda calculus as the basis for functional programming languages. Application of type theory in linguistics demonstrates its suitability for analyzing the natural language documents in archives [1] [15] [26].

TypeDB is open-source database management software that embodies a new paradigm for data management grounded in type theory. This paradigm extends and enriches the entity-relation model in a polymorphic entity-relation-attribute (PERA) model. The TypeDB data model establishes entity, relation and attribute as root types, offering substantial advantages for elaborating a high level schema of the variety of things relevant to different research agendas. The PERA model provides expressivity and coherence for modeling complex domains as well as for relating data obtained from a variety of sources [19] [21]. TypeDB facilitates articulating schemas that more fully and faithfully reflect domain knowledge than other alternatives. Its expressive schema model and intuitive query language make it ideal for representing the interconnected, evolving knowledge generated in constructing a target past [2] [25] [30] [31].

The advantage of using type theory and TypeDB is evident in the basic task of categorizing things of interest. Systematic classification in classical ontology and taxonomy is binary and permanent. In contrast, type judgements are contextual [24] [29]. The same thing can be said to inhabit different types depending on context. One obvious example is that the output of one system can be input to another. This example can be generalized to all communications. A message sent by one agent is a message received by another; moreover, a received message can be cast in the role of impetus for a response by the recipient. The PERA model also provides clear criteria for differentiating entity from relation and both from attribute. An entity is something that exists independently, whereas a relation is a dependent type. A relation can only exist if there is at least one entity or other relation that has a role in it. An attribute, also a dependent type, characterizes something, in TypeDB either an entity or a relation.

V. The ART Schema

Figure 1, The ART Activity Niche, shows a small part of the ART schema, but one that illustrates how we approach the

challenge of creating a schema to encompass as wide a scope as possible of what can be learned from archives; how doing this entails differences from AMT; and also demonstrates merits of technological choices we have made for developing ART. Besides the selective view of the schema presented in fig. 1, the schema as a whole is a work in progress. Even things articulated at this point are subject to revision.

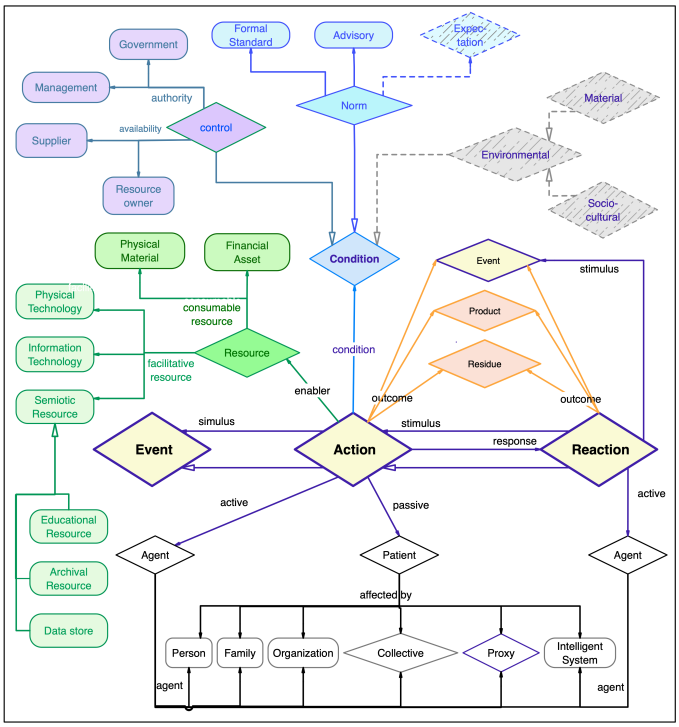

Fig. 1. ART Activity Niche

A. *Iconography*

In Fig,1, rounded rectangles represent entities and diamonds represent relations. Arrows point from relations to things that have roles in them. Labels on these arrows identify the roles. Arrows with hollow heads point from subtypes to their supertypes. Dashed lines indicate placeholders for elements that have not yet been formally defined in the ART schema. Groups of things that are directly connected are color coded for visual clarity.

B. *Typology*

The figure illustrates the expressivity of a type theoretical approach and its implementation via TypeDB. It shows how both entities and relations can have roles in relations. For example both event and reaction have roles related to action. Technically, in a TypeDB schema, a role is an interface.The six types of roles in the action relation (active, passive, stimulus, response, enabler, condition and outcome) demonstrate support for n-ary relations. The condition relation illustrates multi-level inheritance, with control, norm and environmental control as direct subtypes and each of these subtypes having their own subs. Norm illustrates polymorphic interfaces: two of its role players are entities while the third is a relation. Fig. 1 also illustrates that type casting is contextual rather than categorical. The same types of entities that could play the role of agent in an action might also play the role of patient, a party affected by an action. A patient in one action can be an agent in a reaction to it.

Similarly, the material environment includes fossil fuels, and fuels can be related to some actions as products and others as material resources. This example shows another major area where ART goes beyond AMT. AMT only addresses material things related to records, such as medium, storage, movement and preservation.

Another aspect of polymorphism is overwriting. Control is a subtype of condition; therefore, it, like every subtype, inherits the condition role. However, control overwrites that role by specifying that the controlling role is one exercised as authority by government or management, while resource owner and supplier exercise control over availability.

Fig. 1 also illustrates that there can be reciprocal relations. A action is the stimulus for a reaction to it, while the reaction is a response to it.

All the elements in figure 1 are abstract. Concrete instances belong to subtypes, often several levels lower than those shown. For example, in the figure a data store is a subtype of semiotic resource. Data store has subtypes like data warehouse, data lake and database, and database has its own subtypes, such as relational and graph oriented.

C. *Substance*

Fig.1 focuses on action because an archive documents and is a product of action. Both the content and structure of an archive correlate with the archive producer's activity. Having a detailed and rich model of activity effectuates the concept of the archival bond. While not explicit in fig. 1, given that the formation, management and use of records are actions, the top level action schema can be and is being extended for actions related to the production, use and management of records and archives.

Figure 1 is labelled as activity niche because in the ART schema all action occurs within an activity niche which encompasses not only things that have specific relations to the activity of an agent, but also everything that might be involved in it. Activity niche is an adaptation of the biosemiotic concept of semiotic niche. Biosemiotics asserts that meaning is determined not by a living creature acting autonomously, but in dynamic interaction between itself and its environment. This interaction generates a semiotic scaffolding in the niche and the scaffolding shapes how the organism reacts to or interprets things it encounters [12] [20] [22].The meaning of things involved in action are determined by their situation in the scaffolding, including how structures and patterns in the niche shape an agent's understanding, expectations and decisions. The semiotic niche generally, and the activity niche in particular provide a framework for discerning what things meant to people involved in an action. For research in archives, the activity niche is the immediate context in which the archive producer's actions take place It encompasses not only things that have a direct filiation to those actions, but also things related to others who contribute to, react to or otherwise are involved in the archive producer's activity.

The core of fig. 1 comprises the complex of event/action/ reaction relationships. Everything else in the figure relates to this complex; however, fig. 1 is a selective view that does not

include everything related to the core. In ART an event is something that happens, that changes something that was, a prior state, to something different, a post state. Fig. 1 does not address cardinality, but a single event may relate changes in several things.

Action is a subtype of event in which something capable of acting with some degree of autonomy does something — in the role of agent — that causes an event to happen or influences it in some way so that it is different from what would have been the case had the entity not acted. Thus, the agent role is causal.

The third element of the core, reaction, is a subtype of action. It is distinguished from action in general in that a reaction must have an action in the role of stimulus. The connections between action and reaction are reciprocal. Action is a stimulus for reaction and reaction is a response to action. A reaction might have other stimuli besides action, as illustrated by the role of event as stimulus for a reaction.

An agent has an active role in action. The figure shows five types of things that can play the actor role. Person is an individual. Family is a socially defined group of individuals connected by relationships of kinship, marriage, partnership, or adoption. Organization is a group that has a persistent identity, structure and norms. An organization, or some other entity that has authority over it, chooses its members. In contrast, a collective is a group that interacts voluntarily and ad hoc, such as commenters on a social media channel. In an organization, actions are determined by membership status while in a collective, membership is entailed by action. A proxy acts in the place of another. It could be a spokesperson, an official of an organization or even a computer system, such as a financial system that accepts payments and makes outlays. While a proxy should act only as directed, an intelligent system has some degree of autonomy and may act in ways not strictly controlled or intended. Currently, intelligent systems are limited to AI.

As fig. 1 shows, the entities that can be agents can also be patients. Patient, a term imported from philosophy and linguistics, is somebody or something, such as an intelligent system, that is or might be affected by an action. Illustrating the contextual nature of type judgments, something can be an agent in one action and a patient in another, notably as shown when one action is a reaction to another.

In addition to the agent role, ART recognizes two other types of causality shown in fig. 1: resource, which enables an action to occur, and condition which shapes or constrains it. Here too the contextual nature of type casting can be found. For example, roads and airport landing strips are elements of the environment. In travel and transportation, they are also enablers.

Resource relates something that enables an action to an action it enables. There are two roles in which something can enable an action: consumable resource and facilitative resource. A consumable resource is something that is depleted or used up in action, while a facilitative resource contributes to the accomplishment of an action but is not necessarily depleted or altered in that role. For example, many types of information technology support transactions, but are unlikely to change unless a bug or other problem surfaces in use.

A condition is a relation in which something shapes an action either positively or negatively. It has three subtypes: control, norm and environmental condition. Controls can operate either though authority or by determining availability of resources, as show by the two control relations that overwrite the condition role of the condition relation. Controls often have negative repercussions imposed by the entity exercising control if not respected. In contrast, adhering to norms is basically voluntary, unless an entity exercising control makes it mandatory. A formal standard is one established by a formal standards organization for a domain of action. In contrast, an advisory is ad hoc and addresses a specific action or type of action, such as a written or oral recommendation from a consultant. Expectation is not yet formally defined. It could be particular to an agent, such as what the agent expects to be the result of an action. It might also be what a social group expects.

As stated above, fig. 1 does not include everything related to action. Some types not included can be readily inferred, such as multiple layers of subtypes of resources and conditions. Some can be guessed as obviously missing, such as what constitutes an expectation relevant to an action. Some might not be obvious, such as the multiple subtypes of agent and patient roles recognized by ART. The ART schema is a work in progress, most obviously apparent in fig. 1 by the environmental elements, which have dashed lines to indicate that they need further refinement.

Action in figure 1 is generic, not specifically about actions related to records, which are only represented in the figure under the general rubric of archival resource.

VI. PROSPECTUS

ART research has been underway for over three years. Many challenges remain in the agenda. When fully developed, starting from the precondition that a user has found and been granted access to records of interest, ART could be implemented in different configurations according to administrative and technical choices.

Administratively, an archival program could implement an ART system whose use is tailored for and limited to that organization. But several organizations might collectively share a system. A shared system could have full functionality or each program might have a different front end that assists researchers, and staff as well, in exploring its holdings and a common back end for compilation, organization, analysis and sharing of data. Each program could choose front end technology that is particularly apt for the types of records and ensembles involved. For example, a program whose holdings are principally scientific data sets might choose a front end solution best suited to structured data. A program whose holdings have a large percentage of scanned textual documents might prefer a front end solution that could readily accommodate new technologies that offer advanced ways, beyond OCR, to mine such records. Organizations, such as national archives, that have a variety of holdings might pool resources to have a common front end that offers a corresponding variety of functionality.

Whether a comprehensive ART system or differentiated front and back ends are chosen, there are distinct requirements for data mining of archival resources and for making sense of the resultant data. The ART back end will be a system with a

comprehensive and coherent semantics and syntax optimized for realizing the informative potential of the data, as well as for sharing data among researchers and programs within appropriate access controls. The foundation for the back end is the ART schema. The back end should include functionality for data validation and quality assessment, such as type assignments, and verification of logical relations, such as start dates being earlier than end dates.

Interoperability of front and back ends could be achieved by enabling front ends to output data that can be ingested into the back end. It might require intermediaries, such as smart applications that could convert front end tags into the ART data model. In an ideal scenario, an intermediary would be able to extend or overwrite elements of the ART model in order to respect the special characteristics of data that emerges from research in archives.

Along with development and implementation of the ART schema, the current research agenda includes experimenting with technologies for front end functionality. We are currently exploring deep reinforcement learning for this purpose. We intend to evaluate other options as well, for the intermediary and back ends as well as the front end of ART. Exploration of current options is primarily for proof of concept. Future developments in IT may offer better options than those currently available. Hence, the priority in our research is to develop a comprehensive and durable conceptual foundation to support research in archives. This foundation is being laid as a TypeDB schema.

References

1. N. Asher, “Types, meanings and coercions in lexical semantics,” Lingua, vol. 157, pp. 66–82, Apr. 2015, doi: 10.1016/j.lingua.2015.01.001.
2. O. Bánki et al., “Catalogue of Life,” Catalogue of Life. [Online]. Available: https://www.catalogueoflife.org/data/metadata.
3. L. F. Barrett, “The theory of constructed emotion: an active inference account of interoception and categorization,” Soc Cogn Affect Neurosci, vol. 12, no. 1, pp. 1–23, Jan. 2017, doi: 10.1093/scan/nsw154.
4. G. Cencetti, “Sull’archivio come ‘Universitas rerum,’” Archivi, vol. 4, pp. 7–13, 1937.
5. V. Colapietro, “Habit-change, heightened consciousness, and agential ‘crises’: impersonal mechanisms, personal agents, and their complex entanglement,” Conitive Semiotics, vol. 14, no. 1, pp. 9–28, May 2021, doi: 10.1515/cogsem-2021-2036.
6. M. T. Do and K. Shin, “Improving the core resilience of real-world hypergraphs,” Data Min Knowl Disc, vol. 37, no. 6, pp. 2438–2493, Nov. 2023, doi: 10.1007/s10618-023-00958-0.
7. J. Douglas, “Origins: evolving ideas about the principle of provenance,” in Currents of Archival Thinking, T. Eastwood and H. MacNeil, Eds., Santa Barbara, California: Libraries Unlimited, 2010, pp. 23–44.
8. J. Goguen, “An Introduction to Algebraic Semiotics, with Application to User Interface Design,” in Computation for Metaphors, Analogy, and Agents, C. L. Nehaniv, Ed., in Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 1999, pp. 242–291. doi: 10.1007/3-540-48834-0_15.
9. A. Gomes, R. R. Gudwin, and J. Queiroz, “On a computational model of the Peircean semiosis,” IEMC ’03 Proceedings. Managing Technologically Driven Organizations: The Human Side of Innovation and Change (IEEE Cat. No.03CH37502), pp. 703–708, 2003, doi: 10.1109/KIMAS.2003.1245124.
10. J. Ilerbaig, “Archives as sediments: metaphors of deposition and archival thinking,” Arch Sci, vol. 21, no. 1, pp. 83–95, Mar. 2021, doi: 10.1007/s1002-020-09350-z.
11. M. Irvine, “Semiotics in computation and information systems,” in Bloomsbury Semiotics, vol. 2, 4 vols., Pelkey, Jamin, Ed., London: Bloomsbury Academic, 2022, pp. 203–37.
12. J. Hoffmeyer, “The semiotic niche,” Journal of Mediterranean Ecology, vol. 9, pp. 5–30, 2008.
13. K. Kull, “Catalysis and Scaffolding in Semiosis,” in The Catalyzing Mind: Beyond Models of Causality, K. R. Cabell and J. Valsiner, Eds., New York, NY: Springer, 2014, pp. 111–121. doi: 10.1007/978-1-4614-8821-7_6.
14. R. Lawes, “Big semiotics: Beyond signs and symbols,” International Journal of Market Research, vol. 61, no. 3, pp. 252–265, May 2019, doi: 10.1177/1470785318821853.
15. Z. Luo, “Formal Semantics in Modern Type Theories: Is It Model-Theoretic, Proof-Theoretic, or Both?,” in Logical Aspects of Computational Linguistics, N. Asher and S. Soloviev, Eds., in Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2014, pp. 177–188. doi: 10.1007/978-3-662-43742-1_14.
16. M. Mata Caravaca, “The concept of archival ‘sedimentation’: its meaning and use in the Italian context,” Arch Sci, vol. 17, no. 2, pp. 113–124, June 2017, doi: 10.1007/s10502-015-9256-2.
17. J. C. Mendoza-Collazos, “On the importance of things: a relational approach to agency: Review article of Malafouris, L. 2013. How things shape the mind: A theory of material engagement. Cambridge: MIT Press.,” Cognitive Semiotics, vol. 13, no. 2, Nov. 2020, doi: 10.1515/cogsem-2020-2034.
18. A. Menne-Haritz, Business Processes: An Archival Science Approach to Collaborative Decision Making, Records, and Knowledge Management. Springer Science & Business Media, 2004.
19. V. Pavlyshyn, “TypeDB: A Polymorphic Database for AI Agent Memory and Complex Ontology,” Medium. Accessed: Nov. 17, 2025. [Online]. Available: https://ai.plainenglish.io/typedb-a-polymorphic-database-for-ai-agent-memory-and-complex-ontology-4854c439dfd6
20. J. V. Peterson, A. M. Thornburg, M. Kissel, C. Ball, and A. Fuentes, “Semiotic Mechanisms Underlying Niche Construction,” Biosemiotics, vol. 11, no. 2, pp. 181–198, Aug. 2018, doi: 10.1007/s12304-018-9323-1.
21. T. Sabat, “Building A Knowledge Graph With TypeDB | PDF | Databases | Conceptual Model,” Scribd. Accessed: Nov. 17, 2025. [Online]. Available: https://www.scribd.com/document/620751658/Building-a-Knowledge-Graph-with-TypeDB
22. A. Sarosiek, “The role of biosemiosis and semiotic scaffolding in the processes of developing intelligent behaviour,” Zagadnienia Filozoficzne w Nauce, no. 70, pp. 9–44, 2021.
23. G. Sonesson, “Still do not block the line of inquiry: On the Peircean way to cognitive semiotics,” Cognitive Semiotics, vol. 7, no. 2, pp. 281–296, Dec. 2014, doi: 10.1515/cogsem-2014-0090.
24. S. Sridhar, A. Khamaj, and M. K. Asthana, “Cognitive neuroscience perspective on memory: overview and summary,” Front. Hum. Neurosci., vol. 17, July 2023, doi: 10.3389/fnhum.2023.1217093.
25. J. Sterling, “Type Theory and its Meaning Explanations,” July 14, 2016, arXiv: arXiv:1512.01837. doi: 10.48550/arXiv.1512.01837.
26. P. R. Sutton, “Types and Type Theories in Natural Language Analysis,” Annual Review of Linguistics, vol. 10, no. Volume 10, 2024, pp. 107–126, Jan. 2024, doi: 10.1146/annurev-linguistics-031422-113929.
27. K. Thibodeau, “The Construction of the Past: Towards a Theory for Knowing the Past,” Information, vol. 10, no. 11, Art. no. 11, Nov. 2019, doi: 10.3390/info10110332.
28. K. Thibodeau, “Discerning Meaning and Producing Information: Semiosis in Knowing the Past,” Information, vol. 12, no. 9, Art. no. 9, Sept. 2021, doi: 10.3390/info12090363.
29. K. Thibodeau, “Archival Theory for the Information Age,” in International Conference on Artificial Intelligence and the Development of Records and Archives Management, Tianjin Normal University, Tianjin, China, June 2022. Accessed: Jan. 05, 2024. [Online]. Available: https://www.academia.edu/87444099/Archival_Theory_for_the_Information_Age
30. R. H. Thomason, “Type Theoretic Foundations for Context, Part 1: Contexts as Complex Type-Theoretic Objects,” in Modeling and Using Context, P. Bouquet, M. Benerecetti, L. Serafini, P. Brézillon, and F.

Castellani, Eds., in Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 1999, pp. 351–360. doi: 10.1007/3-540-48315-2_27.

31. TypeDB, “TypeDB Philosophy: Why We Built TypeDB.” Available: https://typedb.com/philosophy

32. J. Zlatev and A. Mouratidou, “Extending the Life World: Phenomenological Triangulation Along Two Planes,” Biosemiotics, vol. 17, no. 2, pp. 407–429, Aug. 2024, doi: 10.1007/s12304-024-09576-9.