Breaking Down the Invisible Wall to Enrich Archival Science and Practice

Kenneth Thibodeau
National Archives and Records Administration (retired)
Washington, DC, US
email: KThibodeau@Fordham.edu

Abstract—**This article reviews the state of archival science where basic concepts have been subject to a long stream of criticisms without satisfactory resolution of the issues identified. It establishes a ground for progress by articulating criteria for evaluating archival concepts and proposes a path forward by enriching archival science with concepts and methods from systemic functional linguistics and graph theory. Finally, it demonstrates how borrowing from these fields would satisfy the proposed criteria.**

Keywords—Archival Science; systemic functional linguistics; graph theory.

I. INTRODUCTION

While we don't think about them much, walls are one of the most important inventions in human history. They define the places where we live and work. Walls also serve important purposes of protection. But defensive walls have negative consequences. Internally, they constrain the possibilities of action and, externally, important developments can arise for which existing walls have no protective value and indeed for which they can be impediments in the ordinary course of affairs.

Archival literature for over a half century has seen a continuing stream of arguments to the effect that key concepts in the field, such as provenance, original order and archival fonds, effectively constitute an invisible wall that unnecessarily constrains archival practice and inappropriately excludes many aspects of the creation, use, and management of information, especially as enabled by digital information and communication technologies.

II. THE INVISIBLE WALL

Inappropriate concepts create two types of impediments to archival science and practice. First, Information and communications technologies are creating new forms of information and enabling new ways of producing, acquiring, and using information. Established concepts are not adequate for all new types of information objects nor for all the ways information is being used. Second, technology offers opportunities for improving archival practice, but inappropriate concepts can hinder our ability to profit from such opportunities. This discussion focuses on digital records, but conceptually it applies to physical records as well.

The criticisms of established ideas have been so extensive that it is impossible to summarize them in a short space. The range of these criticisms of basic concepts that have been put forward is reflected in these examples:

- Archival provenance does not encompass many factors that influence the creation of records.
- Series may have multiple provenance.
- An original order may never have been imposed or, if imposed, may have been altered and be unrecoverable.
- Rigid distinctions between "artificial" and "organic" collections are unsupportable.
- The aggregation and arrangement of records in a record keeping system may not represent all significant relationships among the records.
- The relationships among records are dynamic and change over time.

While there are many issues regarding traditional concepts, there are also problems with the criticisms that have been raised. The criticisms tend to be more argumentative than constructive, creating problems across a broad spectrum.

At one end of this spectrum, many criticisms take an overly narrow "either/or" approach. For example, the Series System is described as contradicting the traditional view of provenance [1]. At a conceptual level, the idea that a series of records may be created and maintained by different entities over the course of time does contradict the view that a series can only one records creator. However, at the empirical level, the Series System expands the range of our conception of records creation to include multiple provenance. It does not negate or even diminish the fact that many series have only a single creator.

At the other end of the spectrum, many criticisms of traditional views take an expansive approach, expounding ideas that are complex, vague and difficult to apply. This problem is aptly described by Jennifer Douglas, who found in reviewing the literature on provenance that arguments have been made for loading the term with so may facets of context that it would be impossible to apply [2].

III. OPENING ARCHIVAL SCIENCE

The problems with both traditional views and the criticisms of them are unlikely to be resolved by further dialogue in the argumentative mode that has characterized the debates to date. It would be preferable to overcome these tensions by reformulating archival science so as to encompass what is valuable in

both traditional concepts and recent criticisms. Doing that requires clear, objective and executable criteria for formulating and evaluating concepts in archival science. Three appropriate criteria are applicability, implementation and enrichment of understanding. Ideally, archival concepts should be universal in applicability, but they should also be supple enough to reflect the immense diversity and complexity of records, actors, activities and their relationships. Archival concepts also should provide a sound basis for archival practice, and hopefully be easy to implement. Reformulated concepts should enable improvements in practice. Finally, and most importantly, archival concepts should enrich the potential for understanding records, their creation and use; enhancing the understanding both of archivists and those who use the records they preserve.

With these criteria in mind, there is significant potential for reformulating and enriching archival science by incorporating concepts and methods from systemic functional linguistics and mathematical graph theory. Why these two disciplines? The archival field can be enriched by applying insights gained from several different disciplines. For example, Angelika Menne-Haritz has used functional systems theory to elucidate the context of records creation and use [3], and Ciaran Trace has argued that ethnomethodology will give us deeper insights into the relationship between people and records [4]. But systemic functional linguistics and graph theory are particularly valuable for archival purposes. In many ways, they are complementary to archival science, which studies the relationships between records and the activities in which they were created or used. Systemic functional linguistics offers a rich conceptual apparatus for investigating and characterizing how written and spoken language is used in interactions among people, while graph theory can be used to formalize and quantify the study of relationships among records, agents and the activities in which records are used. Together they offer the possibility for formulating or reformulating archival concepts in a way that avoids being either overly narrow or excessively vague.

The concepts and methods of systemic functional linguistics (SFL) are intrinsically complementary to the objectives of archives. (i) Both are empirically grounded. SFL's theories are derived from studies of the actual use of language, rather than a priori principles, as is the case with other schools of linguistics. Similarly, archival science studies real documents. (ii) SFL focuses on "text", which it defines as written or spoken "language that is functional, does some job in some context," [5] which corresponds to the focus of archival science on documents that were used to do something in a particular context. (iii) SFL regards text as inextricable and inexplicable apart from its context, while archival science defines a record by the context in which the document was used. (iv) SFL examines context in terms of what is done; who is involved, in what role; and how language is used in accomplishing the action or interaction. Similarly, archival science describes records in terms of the activity in which they were used; the parties involved; and the ways records were used.

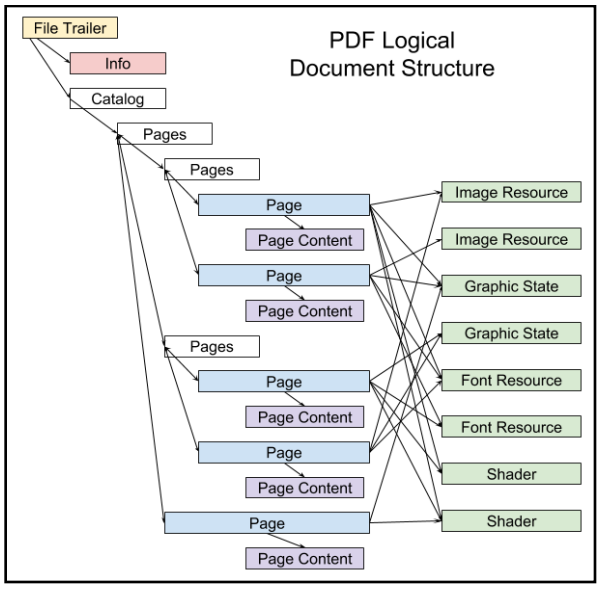
Graph theory offers a rich potential for application in the analysis and communication of data about records and their informative values. This assertion, that it has rich potential for application, naturally raises the question: how can graph theory be applied in archives? Answering that requires first identifying where graph theory could be applied. In mathematics, a graph is a combination of two sets: a set of things, called nodes or vertices, and a set of binary relations between things, called edges or links. Archivists inevitably deal with things that can appropriately be represented and analyzed as graphs, even though they are often not recognized as such.

The structure of individual documents can be depicted as a graph. Digital documents in particular are often articulated on the basis of an explicit graphical structure, as illustrated in Figure 1, which shows the logical structure of PDF files as a tree [6]. Analogously, physical records have basic structures that can be represented in graphs, as illustrated in Figure 2, which analyzes a manuscript from the early twentieth century [7].

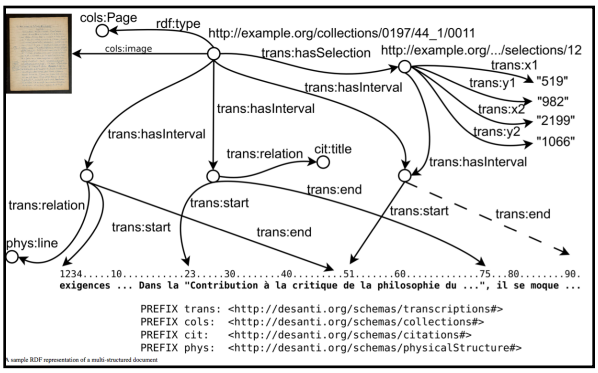
An omnipresent example of an archival graph is the hierarchical description of records, depicted in Figure 3 in accordance with the General International Standard Archival Description. [8] A hierarchy instantiates a subclass of graphs called trees.

Even non-traditional views, such as the Series System, have been depicted in graph form, as shown in Figure 4, which displays the model of in the Archway system of Archives New Zealand [9].

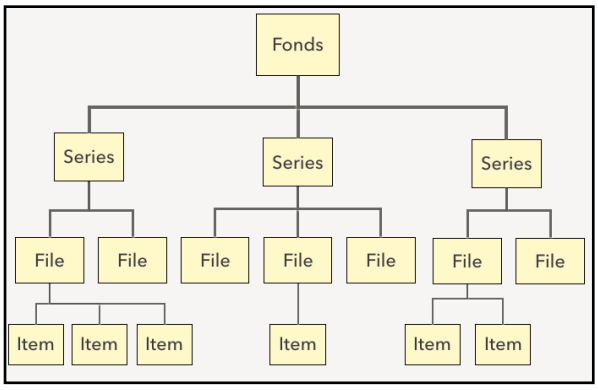
Other obvious cases where the subjects of archival study are appropriately rendered as graphs include the structures of organizations that create records. Recent independent initiatives have defined graphs to indicate



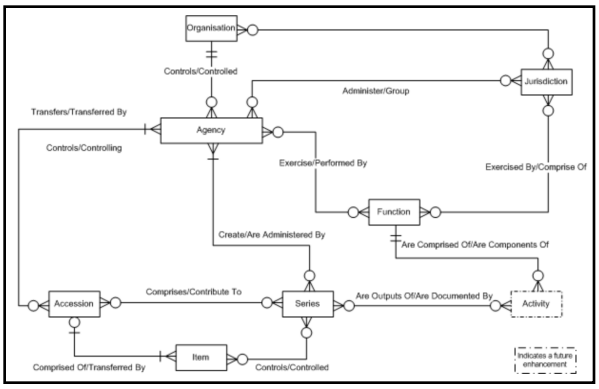
1. Model of PDF Logical Document Structure



2. Graph Representation of a Manuscript



3. Description of a Collection of Records per the ISAD(G) Standard



4. Model of the Archway System of Archives New Zealand

the basic entities and relationships involved in records creation and use. They include the SPIRT Recordkeeping Metadata Project in Australia [10] and the W3C's data model and ontology of provenance [11] [12].

Given the many areas where the objects of archival concern can be depicted as graphs, the question of how graph theory can be applied in archival work can be addressed. Graph theory provides appropriate ways to (i) collect and organize, (ii) analyze, and (iii) present data about records and their relationships. In the current context of rapidly exploding quantities of information being created in new and diverse forms, graph theory is especially valuable because it can easily copy with large quantities of diverse data reflecting multiple types of things and relationships, while remaining sensitive to special cases and nuances.

Let us examine each of the three ways graph theory can be applied. The first way is that graph theory provides rich and varied capabilities for collecting and organizing archival data. In the case of computer systems supporting specific business processes or lines of business, the types of entities involved, including records, actors and actions, and their relationships are often explicit in data models and other artifacts of design, development and implementation of business systems. Even apart from systems designed for specific processes or areas of activity, data about archival objects and relationships exist in many technology applications; such as email, social media, and online collaboration tools. Data relevant for archival graphs can be extracted from such systems and applications automatically. Moreover, there are database management systems available based on graph theory that are appropriate for storage and subsequent use of data about nodes, edges, and their properties.

The second way graph theory can be used in archival science and practice is for analysis. Graph theory offers powerful analytic capabilities grounded in the fact that graphs can be represented quantitatively. The quantitative expression of a graph enables the calculation of revealing statistics that can summarize very large, complex graphs, even with millions of nodes.

An important area for exploration is determining what types of graphs, what graph theoretical methods of analysis, and what graph parameters, are best suited for different aspects of archival work and the communication of the results. This is an issue requiring extensive in depth investigation in a variety of contexts. It will be touched upon in considering the third and final way where graph theory could benefit the archival endeavor.

The third area where graph theory can benefit archival science and practice is its rich potential for displaying large quantities of related data in a variety of ways that can be optimized for different purposes, coupled with facility to drill down from overviews to details. This potential can readily be seen by summarizing different ways that graph theory has been used to display data about documents, authors and activities. These instances will also illustrate the other benefits of applying graph theory, because visualization requires that the data first be collected and organized and then analyzed in order to generate visualizations.

One preeminent area for graph display of data relevant to the management and use of archives is in the analysis and characterization of documents through techniques such as text mining, named entity recognition, and feature extraction. The value of a graph based approach in this area has recently been recognized by the International Council on Archives in

its release of the draft standard for archival description, the Records in Context — Conceptual Model (RIC-CM) [13]. In contrast to the strictly hierarchical approach of ISAD(G), the RIC-CM allows for much more varied, "multidimensional" description.

A variant technique, temporal text mining, could be used to automatically detect and communicate changes in records over time [14].

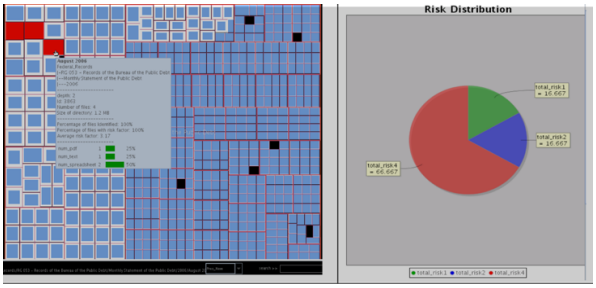
Another relevant approach is a visualization technique called "history flow", which displays relationships among different versions of a document. History flow depicts time along the x-axis and assigns to each author who contributed to a work a different color bar on the y-axis. The variable width of each bar shows the extent of a given author's contribution at any time and the sum of the author bars indicates the size of the document in different interactions. This technique has been used to discover patterns of conflict and collaboration among authors of Wikipedia articles [15].

Cluster analysis is another way to examine graph data. This method has been used to identify and display the frequency of email communications among individuals in the failed Enron Corporation [16]. It has also been used in the analysis of data about use of social media to identify individuals who exercise real leadership in an organization, regardless of their official positions [17]

Maria Esteva and colleagues developed an innovative, graph based tool to assist in preservation management of large collections of electronic records. As shown in Figure 5, in the graph on the left, each rectangle represents directory of electronic files. A rectangle within another rectangle is a subdirectory. Colors indicate the level of risk, and the white border around certain rectangles indicates the level of uncertainty in the risk assessment.. The pie chart on the right shows the ratios of high, medium and low risk files in one of the high risk (red) directors [18].

The exploration of the potential value of systemic functional linguistics and graph theory in archival science and practice naturally leads to the question of wether there is or can be any relationship between SFL and graph theory as applied in archival science and practice. The answer is simply: yes. The potential relationship starts at the fundamental conceptual level of systemic functional linguistics. In addition to the potential for depicting the structure of documents as graphs, SFL describes the generation of a text as a path through the network of potential meanings available in a given language and a specific context. A network is a type of graph and a path is a connected sequence of alternating nodes and edges in a graph. While the potential meanings available in a natural language are perhaps infinite, the rich and nuanced conceptual apparatus that SFL provides for the concomitant analysis of text and context not only makes the exegesis of the generation of a text tractable but also makes the elaboration of the path of selected meaning predictable.

Moreover, linking SLF and graph theory is not only conceptually sound, but also has been demonstrated empirically in the use of semantic tagging, based on concepts from SLF, to enhance automated text mining [18].



5. Preservation Risk Assessment of Electronic Records

IV. Conclusion

How well do graph theory and SLF satisfy the criteria for archival concepts proposed above?

Systemic Functional Linguistics satisfies the first criterion, that archival concepts should be universally applicable. SFL encompasses all text, both spoken and written. While this suggests that the scope of SFL is limited to textual records, even if it only applied to textual records, it would be universal for that class, which covers a significant portion of records. Moreover, the proponents of SLF recognize that digital technologies are blurring the distinctions between language and other modes of communication.

Because SFL is empirically grounded, based on analysis of the actual uses of language to accomplish social purposes, it is elastic enough to deal with an unlimited variety of situations. For this, it offers a rich palette of analytic constructs that are highly sensitive to variations in content, context, and structure.

SLF's empirical orientation also makes it well suited to satisfy the second criterion for archival concepts, the potential for implementation in archival practice. SFL enables nuanced and detailed analysis of records, thus providing a basis for improving practice, notably in the areas of information extraction and information retrieval. Its application to digital record is facilitated by the possibility of automated semantic tagging coupled with machine learning.

The use of SFL offers ample opportunity for enhancing understanding records, their creation and use. For example, in contrast to the morass that has emerged from the archival literature about the principle of provenance, SLF provides a rich conceptual framework for enriching the understanding of provenance through the empirical analysis of context. Starting with Malinowski's concepts of the context of culture and the context of situation [20], SFL analyzes the job being done by text in any particular situation in term of the Field of Discourse, what is done, the Tenor of Discourse, who takes part and what are their relationships, and the Mode of Discourse: how text is used. These concepts give us an objective and normalized basis for determining the genesis and use of records.

Focusing on graph theory, it is well suited to the first criterion, that archival concepts should be universally applicable. Any type of thing, such as a record, agent, or actor, can be a node in a graph and relationships between things can be expressed as links between nodes. Further under applicability, archival concepts should be supple to cover very diverse cases. Graph theory is well suited to reflect highly variable empirical situations. Even subtle differences among things can be captured as properties. Different types of relationships can be expressed as different types of links and relationships can be differentiated on the basis of direction and weight, as well as properties. These criteria can be used to extract subgraphs from larger graphs. Graph theoretical analysis provides a variety of parameters that can be used to characterize both graphs as a whole and particular parts of them.

The second criterion is that archival concepts should provide a sound basis for archival practice. The ability of graph theory to address very large, complex and heterogeneous data is well suited to the diverse realities of records, their creation, use, and relationships among records and with agents and actions. The criterion of practicality also values ease of implementation. Digital records offer ample opportunities for automated capture and generation of graph data related to records. Moreover, as recognized in the ICA's draft standard, the Records in Context — Concept Model, the use of a graph model in description even of physical records facilitates data linkage across institutional boundaries. Both automation and data linking provide opportunities from improving archival practice. Graph theory offers the possibility of capturing rich, diverse relationships among records and between records, actions, and actors.

Concepts from graph theory also offer abundant prospects for improving the understanding of records, their creation and use. Its diverse capabilities for data visualization facilitates seeing the big picture while enabling detailed exploration of subsets.

V. References

1. Adrian Cunningham, "Series System," Encyclopedia of Archival Science. Luciana Duranti and Patricia C. Franks, eds. Lanham, DM: Rowman and Littlefield, 2015, pp. 380-385.
2. Jennifer Douglas, "Origins: Evolving Ideas about the Principle of Provenance," Currents of Archival Thinking, Terry Eastwood and Heather MacNeil, eds. Santa Barbara, CA: Libraries Unlimited, 2010, pp. 33-44.
3. Angelika Menne-Haritz, Business Processes: An Archival Science Approach to Collabortive Decision Making, Records, and Knowledge Management. Dordrecht: Kluwer, 2004.
4. Ciaran Trace, "Ethnomethodology: Foundational Insights on the Nature and Meaning of Documents in Everyday Life ," Journal of Documentation, Vol 72, Jan. 2016, pp. 47-64. DOI: 10.1108/JD-01- 2015-0014
5. M.A.K Halliday and Ruqaiya Hasan, Language, Context, and Text: Aspects of language in a Social-semiotic perspective. Oxford: Oxford University Press, 1989.
6. Google, Skia Graphics Library. PDF Theory of Operation (https://skia.org/dev/design/pdftheory)
7. Pierre-Édouard Portier and Sylvie Calabretto. "Multi-structured documents and the emergence of annotations vocabularies" Balisage: The Markup Conference 2010. Proceedings. (http://www.balisage.net/Proceedings/vol5/print/Portier01/BalisageVol5-Portier01.html).
8. International Council on Archives, ISAD(G): General International Standard Archival Description - Second edition. Sept. 1, 2011. (http://www.ica.org/en/isadg-general-international-standard-archival-description-second-edition).
9. Archives New Zealand, Archival System. (https://www.archway.archives.govt.nz/ArchivalSystem.do).
10. Recordkeeping Metadata Research Team. SPIRT Recordkeeping Metadata Project. Conceptual and Relationship Models: Records in Business and Socio-legal Contexts. (http://www.infotech.monash.edu/research/groups/rcrg/projects/spirt/deliverables/conrelmod.html).
11. World Wide Web Consortium. PROV-DM: The PROV Data Model. W3C Recommendation 30 April 2013. (https://www.w3.org/TR/2013/REC-prov-dm-20130430/).
12. World Wide Web Consortium. PROV-O: The PROV Ontology. W3C Recommendation 30 April 2013. (https://www.w3.org/TR/prov-o/).
13. International Council On Archives, Experts Group on Archival Description, Records In Contexts - A Conceptual Model For Archival Description, Consultation Draft v. 0.1, September 2016.
14. Qiaozhu Mei and ChengXiang Zhai, "Discovering Evolutionary Theme Patterns from Text - An Exploration of Temporal Text Mining," KDD-2005 - Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, R.L. Grossman, R. Bayardo, K. Bennett, and J. Vaidya, eds., 2005, pp. 198-207.
15. Fernanda B. Viégas, Martin Wattenberg, and Kushal Dave, "Studying Cooperation and Conflict between Authors With History Flow Visualizations," CHI 2004, Conference on Human Factors in Computing Systems, April 2004, Vienna Austria. pp. 575-582. (http://www.ifs.tuwien.ac.at/~silvia/wien/vu-infovis/articles/Viegas-CHI2004.pdf).
16. Jeffrey Heer, UC Berkeley Enron email analysis. (http://hci.stanford.edu/jheer/projects/enron/).
17. Michael Fire and Rami Puzis, "Organization Mining Using Online Social Networks," Networks and Spatial Economics, June 2016, Volume 16, Issue 2, pp. 545–578. doi: 10.1007/s11067-015-9288-4.
18. Maria Esteva, Weijia Xu, Suyog Dutt Jain, Jennifer L. Lee, Wendy K. Martin. "Assessing the Preservation

Condition of Large and Heterogeneous Electronic Records Collections with Visualization." International Journal of Digital Curation. 2011, Vol. 6, No. 1, pp. 45-57. doi:10.2218/ijdc.v6i1.171.

19. Astika Kappagoda. The Use of Systemic-Functional Linguistics in Automated Text Mining. Defence Science and Technology Organisation. Department of Defence. Government of Australia. 2009. DSTO-RR-0339.

20. Bronislaw Malinowski, The Language of Magic and Gardening. Bloomington: University of Indiana Press, 1967.