

Position Paper: Problems and Possibilities of Information Integration in Archives

Kenneth Thibodeau*

Archival institutions face information integration problems and challenges which are distinguished by a long time frame, the nature of the information assets archives preserve, and their uses. For government archives, the issue is compounded by the scope of their responsibilities and the diversity of materials that fall within that scope. The National Archives of the United States is responsible for preserving records generated in any computer application in any agency of the Federal Government, and also the White House, the Congress, and the Supreme Court.

Time Frame: Archival institutions concentrate on materials that will be kept forever. The prospect of information integration is not simply that of realizing potential from a large and highly diverse set of materials, but a set which will continue to accrete members, where the properties of materials accreted at different times will be significantly different.

Nature of the Materials: Archives preserve records, a distinct genus of information objects. To illustrate via contrast, we can describe a spectrum of information objects at one end of which is an archetypical publication: a self-contained object capable of delivering its informational content at any time to anyone competent to read and understand it. At the extreme, the required competence is only literacy and average intelligence. At the other end, a record is an instrument or by-product of an activity intended to communicate to parties involved in that activity or its aftermath. At the extreme, a record is intended to communicate only between parties involved in a single action at the time of that action.

Records are surrounded by what we might call informational dark matter. Records are created in contexts where there are high levels of shared knowledge related both to the specific action at hand and to the framework and circumstances in which the action occurs. A record can serve its purpose when it contains information specific and necessary to the action, supplemented only by contextual information sufficient to situate the action. For a third party to interpret a record correctly, information from outside of the record must be brought to bear.¹ Some of the dark matter exists only in the minds of the actors, while some is recorded elsewhere, but in some effectively inaccessible place.

Archives try to enable the correct reading of a record by supplementing it with related information, with the most natural relationship being that of the *archival bond*; that is, the

* The author is Director of the Electronic Records Archives Program at the National Archives and Records Administration.

¹ An oft cited example is a message from John Poindexter to Oliver North containing only the words "Good work." It is impossible to appreciate what that message meant without knowing that it was sent when North returned from testifying at a congressional hearing on the Iran-Contra affair, for which North was subsequently convicted of obstruction of Congress.

relationship between a record and other records of the same actor.² One of the fundamental responsibilities of an archives is thus information integration: ensuring the integrity of the archival bond over time. This necessity is enshrined in the twin archival principles of provenance and original order: records of the same creator must be kept together and they must be kept with the relationships the creator established. In the past, this was relatively easy because, in dealing hard copy records, there is only one effective means of making the archival bond explicit, physical juxtaposition. In the digital realm, we lose not only the stability that comes from durable containers, like file folders and steel drawers, but we have to assume that the logical means – the data structures and software – used to implement relationships among records and among aggregates of records will become ineffective due to obsolescence. The challenge of integrating records which are naturally related is further compounded in the digital realm because (1) records creators have many more ways – both systematic and ad hoc – of establishing relationships among pieces of information, (2) removal of the collocation constraint should lead us to pay more attention to relationships among records of different actors, and (3) archives have the potential for bringing implicit relationships to light.

A significant example of this challenge is the relationship between records of an activity and the records documenting the rules applicable to it. In institutional contexts, most activities are instances of processes which are executed hundreds, even hundreds of millions, of times. In the hard copy world, policy files are kept separate from case files, and the only effective linkage is through the minds of the actors. In the digital realm, rules can be embedded and operative in the applications that generate records of individual actions. While the relationships between rules and transactions are not implemented in any way that resembles any traditional record keeping system, the possibility for correct interpretation of records would be diminished if the archives does not preserve the nexus between data objects that contain rules and those that document transactions.

The temporal dimension is also important here. Archives collect many series of records that are generated over very long times and transferred to the archives incrementally. A series may undergo substantial changes, but must be managed and accessible as a coherent corpus.³ Moreover, archives should preserve and synchronize relationships which transcend series. If, for example, we cannot link changes in rules with instances of action governed by different versions of the rules, we may not be able to understand individual actions or changes in instances over time.

² The response each of you sent when invited to this workshop has a natural and obvious relationship to the invitation. One step removed, so do all the position papers. Ultimately, the network extends back to all the records generated in planning this conference and will extend to everything that comes out of it.

³ E.g., State Department sent diplomatic correspondence in the form of telegrams starting in the nineteenth century and throughout the twentieth; however, in 1972 the Department started using digital technology for diplomatic correspondence. While the records maintained the manifest form of telegrams, there were three major system modifications between 1972 and 2000, and several changes in the way the records were organized.

Use: Most uses of records preserved in archives are orthogonal to the purposes for which records were created.⁴ Many users of archives are thus not well served by search paths that depend on provenance and original order. Records are typically organized parallel to the structure of the processes in which they were created and used, while search strategies are more often focused on particular subjects. Information about records creators and record keeping enables archives to map from search subjects to activities in which the subjects would have been involved, and then to navigate through the preserved record keeping structure to find information about particular subjects. While this method may be very adept at navigating hierarchical trees, it does not even recognize the forest. The possibility of applying advanced IT to generate on-demand integration of information about a given topic from heterogeneous types of records, produced in different activities, by multiple actors, across broad spans of time – while bringing to light essential informational dark matter – could vastly increase realization of the informational value of preserved records.

⁴ For example, the government officials who created punch cards about Japanese Americans interned during World War II did not foresee that fifty years later that data would be used to analyze the impact of that experience on the works of Japanese American artists.