Part II: Information Object Structure

> [T]he nature of the most important different object types and their distinctions can be characterized on the basis of the constructional system. … Of the autopsychological object type, we consider the experiences, their individual constituents, and the qualities (of sense impressions, emotions, volitions, etc.). Of the physical object type, we consider the physical things. Of the heteropsychological objects, we consider again experiences, their individual constituents, and the qualities; of the cultural objects, we consider the primary cultural objects and general higher-level objects.
>
> Carnap 1928, *LSW* §160

In this book, we look at digital preservation differently than most authors, perhaps unconventionally, building recommendations on a digital object model that exploits constructs suggested by the above *LSW* quotation.[100] For each object, someone must decide precisely what is to be preserved. We treat each repository as a *black box* whose input-output relationships are what we most care about. The book's main thread is based on the model of communication suggested by Fig. 2 and the model of information-carrying objects suggested in §6.3.

Choosing how to accomplish digital preservation without a sound intellectual foundation risks incurring systematic errors that might not be discovered until it is too late to put matters right, and perhaps also errors that are discovered earlier, but not before corrections require expensive rework of the preserved content.

We distrust unsupported common sense. This is partly because the digital preservation literature contains serious confusions and misunderstandings, and more generally because intellectual history is full of common sense assertions that later careful analysis demonstrated were incorrect or misleading. We need to clear this underbrush if we want to discern trunks strong enough to support sound methodological branches. For these reasons, we address digital preservation with a second unconventional way of thinking, building on philosophical theories of knowledge. Such treatment

---

[100] As of late 2005, this approach might no longer be idiosyncratic. Shirky 2005, *AIHT: Conceptual Issues from Practical Tests* includes:

> "[W]e have become convinced that data-centric strategies for shared effort are far more scalable than either tool- or environment-centric strategies. A data-centric strategy assumes that the interaction between institutions will mainly be in the passing of a bundle of data from one place to another—that data will leave its original context and be interpreted in the new context of the receiving institution. Specifying the markup of the data itself removes the need for identical tools to be held by sender and receiver, and the need to have a sender and receiver with the same processes in place for handling data."

responds to the need for an intellectual foundation asserted in a U.S. National Archives call for action:[101]

> The state of affairs [in digital preservation] in 1998 could easily be summarized:
> - proven methods for preserving and providing sustained access to electronic records were limited to the simplest forms of digital objects;
> - even in those areas, proven methods were incapable of being scaled to a level sufficient to cope with the expected growth of electronic records; and
> - archival science had not responded to the challenge of electronic records sufficiently to provide a sound intellectual foundation for articulating archival policies, strategies, and standards for electronic records.

The most troublesome published difficulties with preservation seem to be related to failure to understand the logic of our language. We need to examine the errors made to learn how to avoid repeating them.

For instance, I feel that "knowledge preservation" is too grand a term. Although the phrase has a satisfying ring, I prefer to describe the objective as "information preservation." This is because I do not believe that we can preserve the collective knowledge of any community. What its members communicate, and what can therefore be captured, is only a small portion of what they know. Philosophy teaches us to distinguish between information and knowledge. Information is what we write onto paper and speak into microphones. Knowledge is much more. It is part of what makes each of us more than what he writes or says. We use knowledge to write, to teach, to invent, to perform, to earn our livings, to care for our families, and to accomplish the myriad mundane and almost unnoticed things we do every day for safety, health, comfort, and amusement.

"Getting it right" depends on precision in language and in action. Whenever it is difficult to express an idea both simply and correctly, this book favors precision, accepting the view that simple does not justify simplistic.

Natural language is full of ambiguities. Language problems in digital preservation are a small of example of Wittgenstein's dictum, "Most of the propositions and questions to be found in philosophical works are not false but nonsensical. … Most of the propositions and questions of philosophers arise from our failure to understand the logic of our language." (*TLP* 4.003) It is surprisingly difficult to avoid difficulties caused by imprecise or misleading use of ordinary language. For instance, "knowledge management" suggests a different meaning for the word 'knowledge' than the

---

[101] Thibodeau 2002, *Overview of Technological Approaches to Digital Preservation*.

traditional one. We therefore begin with an introduction to knowledge theory and then apply that theory to digital preservation.

We find it helpful to distinguish three topics conceptually and architecturally as much as possible without introducing absurdities: (1) individual works as the proper targets of preservation attention; (2) collections of works that information providers choose to identify as being closely related, extended by such further works as are necessary to provide technical context for the interpretation of these conceptual collections; and (3) archival and library mechanisms—digital repositories—that are essential parts of the infrastructure for making accessible and preserving individual works and the information that binds individual works into collections.

An objective is to choose a structure sufficiently general to describe every kind of information. We choose the Digital Object (DO) model suggested in §6.3, the digital collection schema suggested in §6.4, and elaborations of these models. In fact, a single schema and model suffices for all DOs and also for digital collections. The chosen DO structure exploits ternary relationships recursively, object identifiers as references, and mathematical values. Our preservation model, the Trustworthy Digital Object (TDO) construction described in §11.1, is a modest DO extension. Our theory of knowledge and the models of its representations, particularly Carnap's *LSW*, support these models as being sufficient for any content collection, without any restriction to the information that can be described.

We believe that the TDO evidence of provenance and authenticity is as reliable as is feasible, as are the trust relationships essential to that evidence. This opinion is based on analyses of the subjective/objective distinction (§3.3), of the ethical value/fact distinction (§3.4), and of relationships (§6.5.2) and identifiers (§7.3).

We further believe that technology for preservation can be designed to be a small addition to software already deployed to support current and future information access services and other aspects of digital repository management. The inherent complexity of digital preservation software can be mostly hidden from its human users. To the extent that these objectives are achieved, the convenience, flexibility, and cost of digital archiving services will be optimized.

As much as I might want to wave some technical wand to preserve knowledge, I do not know how to do that. So I settle for information preservation.