TRUSTWORTHY 100-YEAR DIGITAL OBJECTS

SYNTAX AND SEMANTICS—TENSION BETWEEN FACTS AND VALUES

H.M. Gladney

HMG Consulting
Saratoga, CA 95070

17 November 2003
updated 8 January 2004

Prior *Trustworthy 100-Year Digital Object* articles describe a method for preserving digitally represented information. Trustworthy Digital Object (TDO) representation and packaging makes any digital content reliably meaningful to consumers, no matter how distant these are in time, in space, and in social affiliation from their information sources. The current article focuses on digital document authenticity and on evidence a consumer can use to decide whether to trust the content.

Such considerations are necessarily epistemological. Arguing the issues must start by conveying as unambiguously as possible what we mean by words like 'authenticity' and 'evidence' and by distinguishing between statements that are 'objective' and those that are 'subjective'. Our analysis applies Wittgenstein's teaching to pictorial models of digital and conventional communication.

This analysis leads us to identify an ethical imperative for digital preservation, and to suggest that the TDO method defines a quality standard against which any method of digital preservation should be judged.

Table of Contents

Language disguises thought. So much so, that from the outward form of the clothing it is impossible to infer the form of the thought beneath it, because the outward form of the clothing is not designed to reveal the form of the body, but for entirely different purposes. Wittgenstein, Tractatus 4.002

Wittgenstein's radical separation of facts from values can be regarded as the terminus of a series of efforts to distinguish the sphere of natural science from the sphere of morality, [by] Kant, … Schopenhauer, and … Kierkegaard. … [T]he model theory of the proposition became the basis upon which Wittgenstein could give scientific language a sure basis, while drawing an absolute distinction between what language *says* and what it *shows*—that is, what is 'higher'." [Janik, p.197]

1 INTRODUCTION

Written records, multi-media recordings, and data collections are the product of the mind's work, and are the principal legacy of those who live by and for work of the mind. They include information critical to democratic institutions and to our well-being. Much of this content now begins in digital form and some is available only in digital form. It includes additions to every branch of knowledge and culture.

The people who support the information infrastructure deserve assurance that its best holdings will survive reliably into the future along with their social security records, building permits, and other practical records. Absent sound procedures not in place today, their expectations will be disappointed.

The ***Trustworthy 100-Year Digital Object*** articles [#1] treat digital records whose falsification might damage eventual users. They describe a way of protecting information consumers from technological obsolescence, disappearing expertise, and deliberate or accidental misinformation. The current article argues that the **TDO** (Trustworthy Digital Object) method is sound and achieves as much protection as clerical procedures and machines can, in principle, deliver.

1.1 What is the Digital Preservation Problem?

Consider what someone a century from now might want of information stored today. This might be a scholar who wants to interpret our writings and to decide whether to trust them, a businessman who needs to guard against fraud, or an attorney surveying fiduciary records. For some applications, users will want, need [Graham], or even demand **evidence that information used is authentic**.

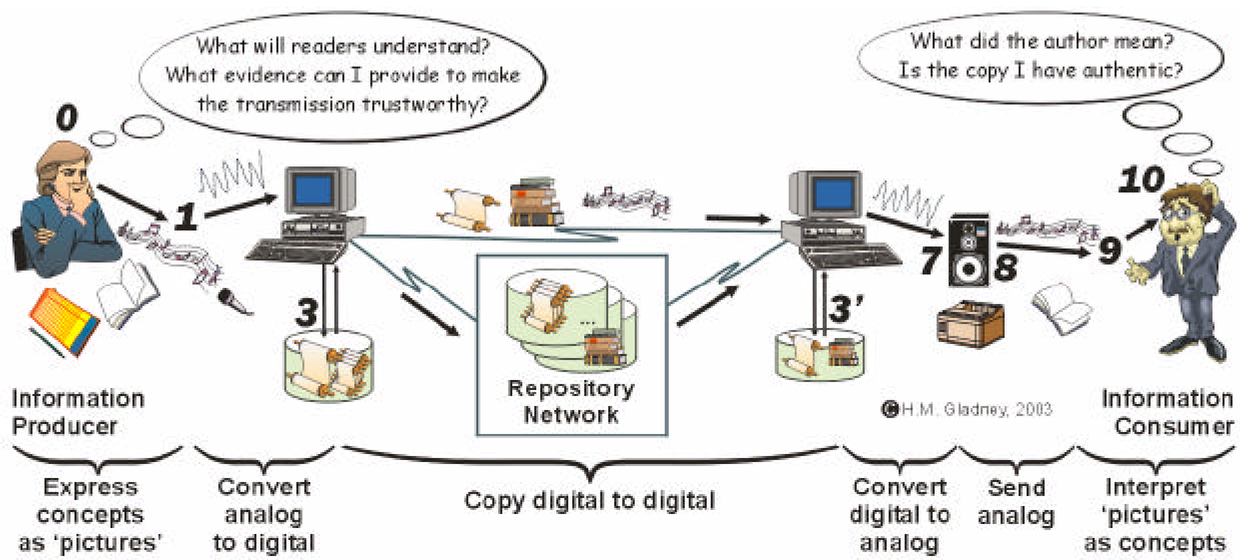


Figure 1: Information transmission channels, identifying intermediate object copies.
0 through 10, the names for document representations, reappear in other figures.

Figure 1 helps us discuss communication reliability challenges. The technical challenges of digital preservation, excluding content management challenges that would exist even if no erosion over time occurred, have for some time been known to include:

- Ensuring that a copy of every document survives as long as it might interest someone;
- Ensuring that consumers can use any document as its producers intended, avoiding errors introduced by third parties that include archivists, editors, and programmers;
- Ensuring that any consumer can decide whether any document is sufficiently trustworthy for his use;
- Hiding information technology complexity from end users (producers, collectors, and consumers);
- Minimizing labor costs by automating clerical steps;
- Empowering editors to package information to relieve overloading of professional cataloguers, and
- Exploiting already deployed technology without disruption.

Since eventual users of preserved information might suffer harm or loss if they are misled by misinformation, we are interested in the **reliability of, and distortions in, the communication channel that transmits an input** 1 **to become a replica** 9.[1]

1.2 Why Write this Article?

> "Concerns about authenticity … are not new. … It is not unusual to find resources purporting to represent the same thing at many different sites. But we are without widely understood and employed methods of assessing and establishing authenticity of digital sources. It is crucial for … knowledge in many fields that it be possible to determine the relative authenticity of a number of different representations, ... [and] to establish the integrity of a particular digital copy." [Bearman]

Archivists and research librarians have started to grapple with these issues.[2] For instance, both the U.S. National Archives and the Library of Congress are working to extend their collections to digital records that will sometimes be the only complete carriers of the information they represent. Investigators are finding that straightforward extensions of institutional practices developed over the past century are insufficient.[3] This is partly for technical reasons, and partly because methods used for works on paper require more human attention than is likely to be affordable for the flood of digital content people will want to preserve. Similar concerns arise in the commercial world, but are not yet as well focused as they are in some cultural institutions.

[#2] and [#3] describe what makes a data object also a **TDO**—schema constraints and tools that conform to evolving and *de facto* standards for content management. Readers might inquire, "Can we do better than the TDO method?" or, **"Is the TDO method optimal for its intended applications?"** [4]

Part of this question is, "Does the **TDO** method automate everything possible?" This is motivated by two quantitative factors. Applied technology prices have decreased dramatically in the last decade and are likely to continue decreasing, whereas labor costs are high. Secondly, the number of digital documents is much larger than the number of paper-based documents that collection managers have needed to consider.

Thus, the current article and [#5] focus on answering:

- Does the **TDO** method use machines for everything that machines can do, without attempting to use machines beyond what they can do? Are the tasks it assigns to human beings reasonable?
- Are the **TDO** method and its justification in the ***Trustworthy*** articles as precisely described as they can be? Is ambiguity minimized?
- Does the **TDO** method meet every pertinent known requirement?[5]

---

1 No producer knows the risks that users of his information might face. The marginal cost of a most reliable solution is not much greater than that of any solution at all. The simplicity of the [#3] solution makes satisfaction of this premise likely.

2 [Steemson] illustrates confusion between objectives and mechanisms.

3 "Although many aspects of digital preservation have received attention since the mid-1990s, most of the presentations and papers on the subject have ended with little more than general comments about the complexity and expense of the tasks, and ambiguity about responsibilities and roles." [Marcum]

4 Here, by 'optimal' we mean at least 'most convenient', 'most reliable', and 'least expensive'. This question implies critical comparison of the TDO method with other initiatives.

5 "What happens in the digital case if there are no stable, enduring digital objects? One possibility is that we will find a way to *create* them. In one current view, objects are at least in part socially constructed; they are bounded and stabilized through social interaction." [CLIR 92, Levy]

This is close to what the **TDO method** accomplishes, with part of the stabilization method being technical.

1.3 Preservation Based on Trustworthy Digital Objects

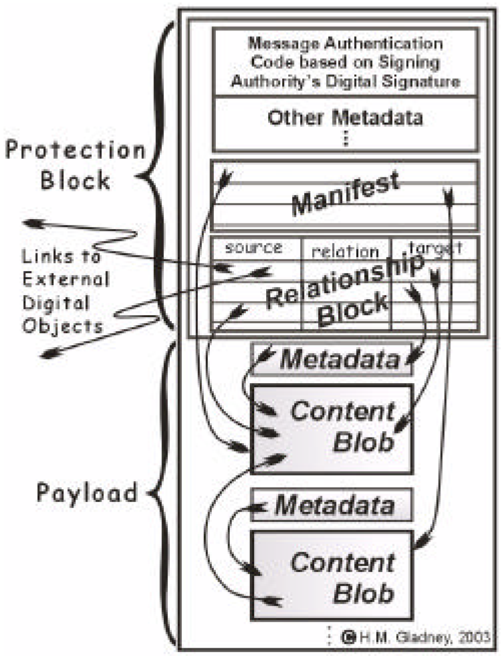


Figure 2: Trustworthy Digital Object structure depicting encapsulation of content with its essential metadata as a trustworthy digital object (TDO), and suggesting internal and external links. See [#3 §3.2].

Our methodological core is record schema for long-term holdings. To prepare the object set that makes up a work, an author or editor causes conversion of each blob into a durably intelligible representation [#2] and collects the results, together with metadata prescribed by proposed standards (e.g., [METS]) for each content blob, to become the payload of a new **TDO**. (Figure 2) In addition to its payload, this **TDO** is given a ***protection block (PB)***. Into this **PB** the editor loads metadata described in [#3], and can also record relationships among parts of the new **TDO** and between this **TDO** and other digital objects.[6] The final construction step, executed at the command of a human agent only after all other additions and changes, is to seal this bundle with a *message authentication code* signed by the agent.

Nesting **TDO**s is often useful. In a valid **TDO,**

1) The blob set that represents a work is XML-packaged with registered schema.
2) These blobs and metadata are encoded in platform-independent representations.
3) These metadata include identifiers for both the instance and the work's version set.
4) Every critical link to another **TDO** is secured by a message authentication code.
5) The package includes or links reliably to all metadata needed for interpretation and as evidence.[6]
6) All these contents are packaged as a single blob sealed using public key message authentication.
7) The cryptographic key management is grounded in keys that widely trusted institutions publish and periodically replace. Such a key is evidence for the approximate date of signing.

Use of this scheme, laid out in detail in [#3], can provide durable authenticity evidence that is as good as is feasible in principle. To justify this claim is an objective of the current paper.

The TDO method merely handles clerical and mechanical details of digital preservation, thereby mechanizing procedural work that might distract information producers and custodians from **preparing durable evidence—an ethical imperative that requires the exercise of judgment and taste.**

2 EPISTEMOLOGY: *PICTURES* AS *SIGNS* AND *SYMBOLS*

> "If I am asked 'What is good?' [, or] 'How is good to be defined?' my answer is that it cannot be defined ... And … nobody can foist upon us such an axiom as that 'Pleasure is the only good' or that 'The good is the desired' on the pretence that this is 'the very meaning of the word.'
>
> "... My point is that 'good' is a simple notion, just as 'yellow' is a simple notion; that, just as you cannot, by any manner of means, explain to any one who does not already know it, what yellow is, so you cannot explain what good is. Definitions of the kind that I was asking for, definitions which describe the real nature of the object or notion denoted by a word, and which do not merely tell us what the word is used to mean, are only possible when the object or notion in question is something complex." [Moore, I,6]

We want to help any information consumer decide whether documents he receives are authentic. We propose to do this by providing him evidence. Given technical means for preserving evidence indefinitely (§1.3)—means whose adequacy can be objectively assessed, what remains is an ethical task. For best execution of that task, and for retrospective assessments of the quality with which it has been achieved in

6 "Every computer science problem can be solved by adding a layer of indirection." — author not known

any instance, we need to be clear about the boundary between objective facts and subjective values, opinions, and choices. **We need to choose what's *good,* both technically and ethically.**

For that, we call back into service basic epistemology (the philosophy distinguishing what can be known as facts and analyzing how we deal with knowledge), starting by separating issues of information representation (syntax) from those of meaning and interpretation (semantics), and subjective from objective aspects. Our arguments are framed in terms of fundamental limits for human communication and by reference to the Figure 1 physical communication model.

Ideally, we should need to say nothing new—we want the best shared basis and have found that this is provided by Wittgenstein. However a spate of articles from the information science and research library community suggest that his and related insights have been forgotten. The current article can be read to be almost entirely an application of Wittgenstein's techniques and ideas to a modern challenge.

2.1 *Objective* and *Subjective*: not a technological issue

> "What roused [Karl] Kraus's resentment ... was the mingling of opinion and fact involved in presenting news slanted by political interest. ... specifically distorted in the free mingling of ... rational objectivity and subjective reaction, that was the deliberate aim of the feuilleton. ... it was a subjective response to an objective state of affairs, … laden with adverbs and especially adjectives; so much so, that the objective situation was lost in the shuffle. Objective facts were thus viewed through the prismlike emotions of the writer. Success in this essay form was open only to those narcissistic enough to regard their own emotional responses as having a universal perceptiveness and quality.
>
> "For the bourgeois Viennese, with their passion for the arts, the feuilleton was the high point of all journalism, ... in the *Neue Freie Presse.* To Kraus, however, the feuilleton destroyed both the objectivity of the situation described and the creative fantasy of the writer ... So it ... prevented the reader from making any rational assessment of the facts of the case." [Janik, p.79]

The difficulties of communicating meaning reliably have little, if anything, to do with computers or other machinery. Consider Figure 3, which parallels Figure 1, except that it portrays no digital technology. The sign 1 might be deficient, omitting details needed to disambiguate the conceptual object 0 from similar objects in the conversational domain. Even if the speaker crafts 1 well enough to avoid this difficulty, it will contain accidental information that the listener has no certain way of distinguishing from what the speaker considers essential.[7] Thus, the listener neither has, nor can be given, a sure prescription for constructing his own conceptual object 10 to be identical to the conceptual object 0. **Communication is always imperfect in this way**, even if it is not mediated by machinery.

We regard communication machinery as perfect only if what the consumer perceives as 9 is a faithful copy of what the producer creates as 1 (i.e., only if 1=9). Whatever errors make a representation 9 differ from the original input 1 might be such small changes that consumers do not notice them. If the errors are larger, a consumer might still understand enough to discount and tolerate the errors, or might reject the communication as insufficiently trustworthy (deserving of trust).

We cannot know with certainty what people think. We can trust at most what we ourselves perceive, or what someone tells us if we trust that person for the kind of information at issue [Denning], [Jordan]. We call the opaque steps 0➔1 and 9➔10 **'subjective'.**

Because we cannot find non-trivial examples that avoid this kind of difficulty, we believe that communication steps illustrated by the 0➔1 and 9➔10 steps in the figures of this article inevitability embed uncertainty that cannot be removed by technology. If this is indeed correct, **machine assistance can be applied at most to steps starting with information objects 1 and ending with information objects 9.**

[7] For instance, communication by speech always has specific pitch (tenor, soprano, …), but pitch is usually irrelevant to the conceptual message. Pitch would be relevant if the communication were a symphonic performance, or a song.

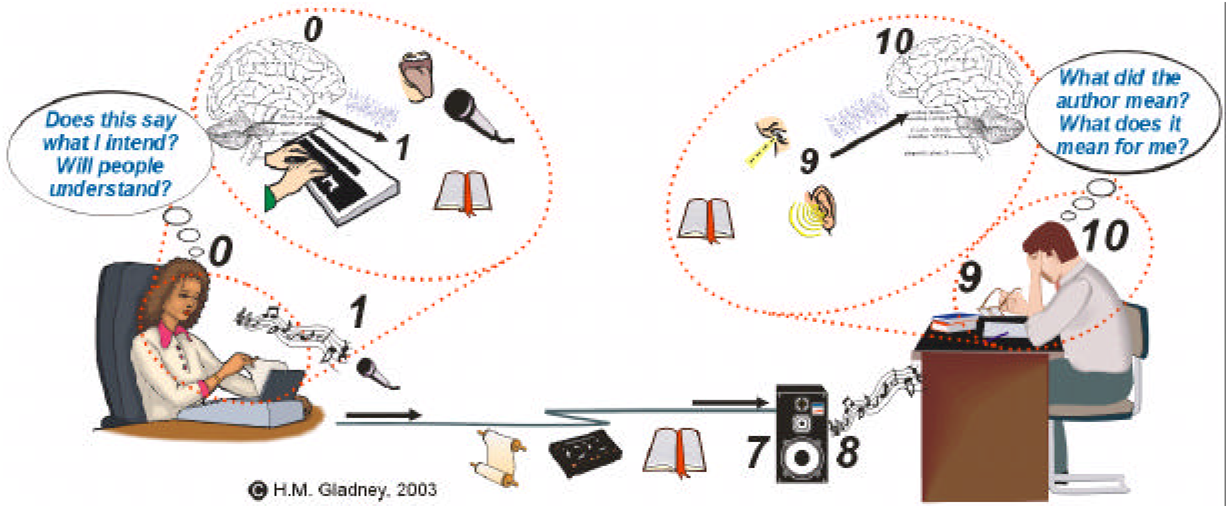


Figure 3: Communicating meaning is difficult even without digital machinery!
Human beings choose how to accomplish every transmission step, including choosing what technology or clerical method to use for the mechanical steps (from 1 to 9). These implementation choices are based on human intentions and expertise that are mostly tacit and subjective, but **every consequence of subjective choices can, in principle, be described objectively.**

2.2 *Facts* and *Values*: how we can distinguish which is which

> "What is the case—a fact—is the existence of states of affairs." [Tractatus 2]
>
> "Wir machen uns Bilder der Tatsachen." (We make for ourselves pictures of the facts.) [Tractatus 2.1]

In Wittgenstein, a *fact* is a statement about an observation of the world, or a statement about an observation that might have been made, or might in the future be made if someone wants to do so.[8] A value is an ethical opinion that might motivate a particular choice made by some human being. Here, *ethical* does not necessarily imply *moral.* Instead it has to do with distinctions of the kinds discussed by [Moore].

If we were to eavesdrop on audience conversations following a performance of Beethoven's Violin Concerto, which exposes its soloist as much as any musical composition does, we would probably hear similarities of critical opinion. What makes these remarkable is that they come from speakers who would insist on their lack of musical expertise. What might induce such similarities?

What is probably occurring is that the opinions, themselves in the domain of value judgments, are based on shared observations that range from quite obvious facts, such as that the soloist played no incorrect notes, to facts close to the boundary (§2.3) with values, such as that the soloist controlled his bowing technique so as to be highly expressive. It is irrelevant to the point being made that the speakers might themselves be unable to identify the facts that influence their evaluations.

2.3 Representation Theory: meanings of signs and sentences

> "... the idea of regarding language, symbolisms and media of expression of all kinds as giving us "representations" *(Darstellungen)* or "pictures" *(Bilder)* had by 1910 become a commonplace in all fields of Viennese cultural debate. Among scientists, this notion had been in circulation at least since the time of [Heinrich] Hertz, who had characterized physical theories as providing just such a *Bild* or *Darstellung* of natural phenomena. At the other extreme, it was equally familiar among artists ... Wittgenstein's achievement ... was not to initiate this discussion, but to draw the threads finally together by providing a completely general and definitive analysis of the issues involved. And ... it enabled him to undercut some pressing intellectual problems about his own ethical position—not by equipping that position with intellectual foundations ..., but by giving seemingly unanswerable support to his own Kierkegaardian view that, over moral issues, no question of intellectual foundations can properly arise." [Janik, p.31]

[8] Facts are the starting point and central to *Tractatus.* Nevertheless, even Wittgenstein is not entirely consistent in his use of the word 'fact'. [Glock, pp.115-120]

> "The fact that the elements of a model are related to one another in a determinate way represents that things are related to one another in the same way." [Tractatus 2.15]

Today's difficulties in finding a sound basis for long-term digital preservation have their origin partly in how people use words. Building on Fritz Mauthner, Wittgenstein points out that the deepest confusions and misunderstandings are rooted in confusions about language—about what words mean. His *Tractatus Logico-Philosophicus* [Tractatus] teaches the boundary between *Natural Philosophy* and *Ethics,* which is much the same as the boundary between *facts* and *values.*

Any discussion of document quality[9] depends on what language symbols mean. As well as managing the boundary between facts and values, we must avoid common confusions. Misunderstanding contributes significantly to failures of trust and security—perhaps more than deliberate falsification.[10] For instance, a problem with relationships suggested by Figure 4(a) is that, if you say, "X is authentic", I might infer that X was also useful for whatever we were talking about. In contrast, Figure 4(b) suggests that *authenticity* is made up of *integrity* and *provenance* qualities, that a useful object might not be authentic, and that an authentic object might not be useful.

Figure 4: Possible relationships of meanings; (a) is unsatisfactory, and (b) works well.

In Figure 5 an inscribed sentence, or any other inscribed pattern, is a special kind of 'picture' or 'model' [Tractatus 2.1] that Wittgenstein calls a 'sign'. [Tractatus 3.32] This extends to other kinds of physical signals, such as spoken language and electronic signals. Any sign can be a 'symbol' for something other than itself. [Tractatus 3.31] Specifically, a 'symbol' is a sign that refers to a conceptual object [Tractatus 2.1-3.33]—it stands for an object or a concept. Today we call a sign used this way a 'representation'.

---

9 By 'document quality' we mean whatever attributes of digital content are important for the conversation of the moment, including at least security, reliability, historical validity, beauty, intelligibility, and usefulness.

10 "In everyday language it very frequently happens that the same word has different modes of signification—and so belongs to different symbols—or that two words that have different modes of signification are employed in propositions in what is superficially the same way." Tractatus 3.323

"In this way the most fundamental confusions are easily produced (the whole of philosophy is full of them)." Tractatus 3.324

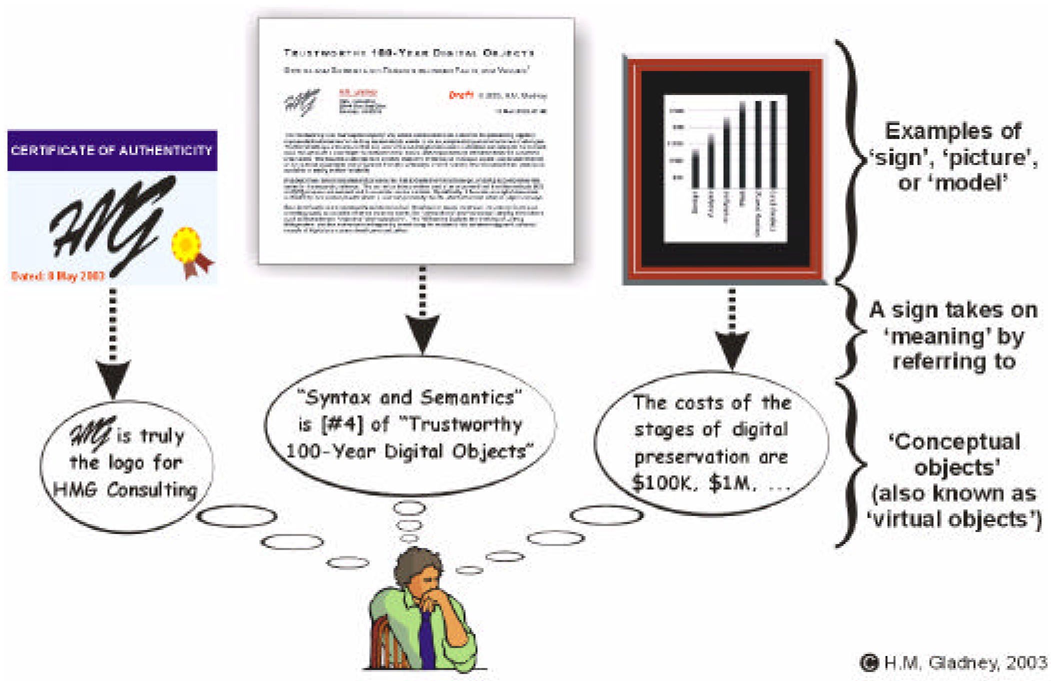


Figure 5: Elementary epistemology from *Tractatus Logico-Philosophicus* [Tractatus].
[Janik] suggests that what Wittgenstein intended would be easier to understand if we translated 'das Bild' as 'model' rather than as 'picture', as is conventionally done.[11]

Wittgenstein's *Tractatus* asserts that **whereas facts can be spoken, values can only be shown**. His friend Paul Engelmann provided an island analogy to explain the boundary between facts and values: by talking about the limits of factual reasoning—the island's edge, the *Tractatus* shows the extent of values—the ocean's coastline. This analogy helps us organize Table 1 distinctions for exploring whether the **TDO** method is sound and complete within its sphere.

Table 1: Engelmann's island/ocean analogy [12]

objective	⇔	subjective	§2.1
Natural Philosophy	⇔	Ethics	
facts	⇔	values and opinions	
die Darstellung (representation)	⇔	die Vorstellung (idea)	*Tractatus*
evidence	⇔	opinion	§2.6
form or syntax	⇔	meaning or semantics	§2.4
action	⇔	intention	[LW 39]
… …	⇔	… …	…
Procedures, rules[13]	⇔	intuitions[14]	[LW 39]
trustworthy	⇔	trusted	§4.4.4

Imagine having to write an essay comparing Shakespeare's *Hamlet* to that of Lucasfilm's *Star Wars.* Your statements might oscillate from side to side of the objective/subjective boundary—the Table 1 boundary between what can be mechanized and what must forever remain a human value decision or judgment. Readers are well served by authors who recognize

[11] "British and American interpretations of the *Tractatus* have suffered as a result of difficulties over the German word *Bild* and its associated word forms, *abbilden* et cetera. When writing in English, philosophers have tended to discuss Wittgenstein's 'picture' theory as though it invited us to think of 'propositions' as providing, so to say, snapshots—or even mental images—of 'facts.' … we understand Wittgenstein's aphorisms better if we think of linguistic *Bilder* as "deliberately constructed verbal representations," instead of misleading ourselves by the use of the much looser English term 'pictures.'" [Janik, p.182]

[12] "Positivism holds—and this is its essence—that what we can speak about is all that matters in life. Whereas Wittgenstein passionately believes that all that really matters in human life is precisely what, in his view, we must be silent about. When he nevertheless takes immense pains to delimit the unimportant [i.e., the scope and limits of ordinary language], it is not the coastline of that island which he is bent on surveying with such meticulous accuracy, but the boundary of the ocean." [Janik, p.191]

[13] "Although I myself have completed only finitely many sums in the past, **the rule** [for addition] determines my answer for indefinitely many new sums that I have never previously considered. This is the whole point of the notion that in learning to add I grasp a rule: my past intentions regarding addition determine a unique answer for indefinitely many new cases." [Kripke, pp.7-8]

[14] Knowing "or not is simply a question of whether he does it as we taught him; it's not a question of intuition." [LW 39, II]

that human communication is limited to the facts that language **can** convey, and signal what is objective and what is opinion.

The careful reader might, in the light of the foregoing paragraphs, object to 'meaning' in "content reliably meaningful to consumers" in our abstract. We would remind him that we do not deny that language conveys meaning, and that, as far as we know, neither does Wittgenstein deny this. (Such an assertion might quickly lead to a conclusion that language is useless!) Instead, we assert that no speaker can be certain that what he intends to imply is precisely what some listener infers.

2.4 *Syntax* and *Semantics*: ambiguity and the meaning of 'meaning'

> "For a *large* class of cases—though not for all—in which we employ the word 'meaning' it can be defined thus: the meaning of a word is its use in the language.
>
> "And the *meaning* of a name is sometimes explained by pointing at its *bearer*." [15] [PI §43]

This kind of meaning of 'meaning', illustrated in Figure 5, is the one most useful in the current analysis. To reduce the potential for confusion, we must mention two other meanings for 'meaning'.

When a computer scientist refers to the meaning of a program, it is likely that (s)he has in mind something different than what Figure 5 suggests. This might be that the 'meaning' of a program is the set of values that it can generate—a functional table in Figure 6. The figure illustrates an instance of a rule (an 'intension') and its associated 'extension'; such rules are discussed at length by Wittgenstein in *Philosophical Investigations* [PI] as later analyzed by [Kripke].

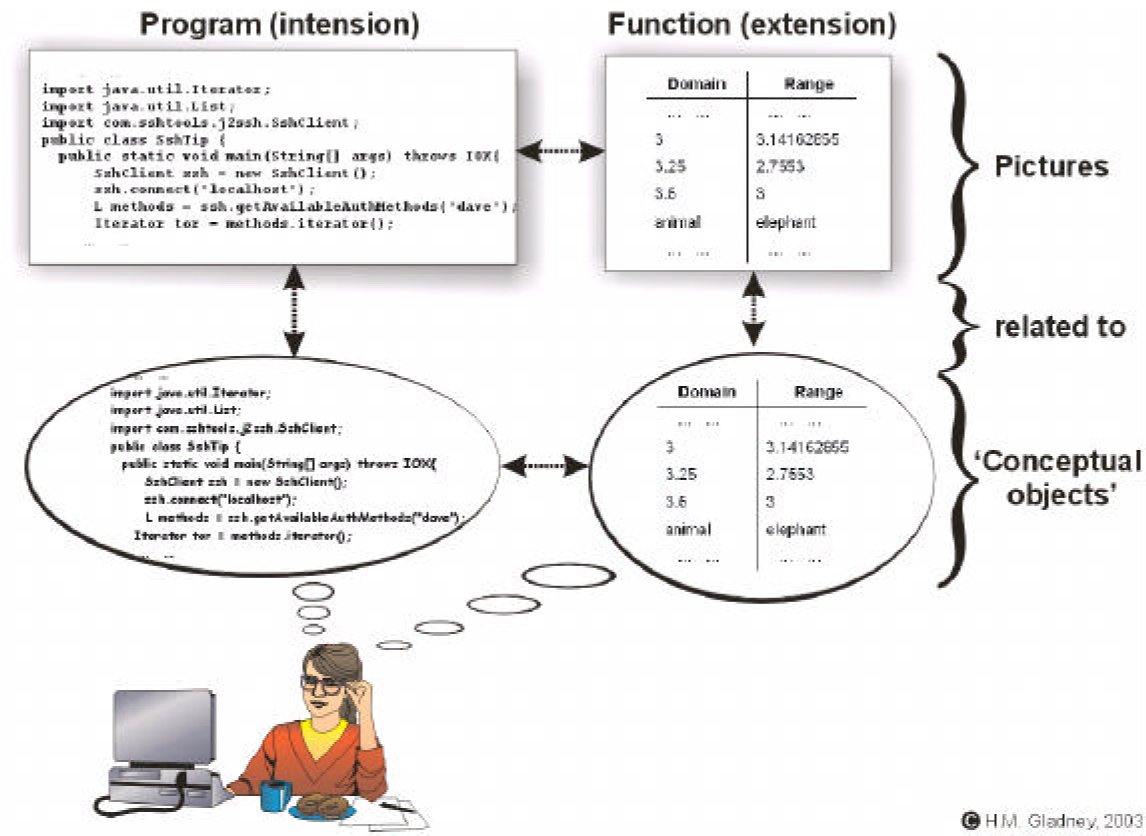


Figure 6: Semantics (a.k.a. 'meaning') of programs
and illustrating the relationship between intensions and extensions

Among computer scientists, another meaning of a program's 'meaning' is its articulation in a formal semantics language, such as denotational semantics [Gordon]. To explain a program's meaning in this way is similar to explaining the meaning of a Sanskrit passage by giving its English translation. This kind of

15 Wittgenstein began his *Blue Book* with the question, "What is the meaning of a word?"

meaning articulation is not prominent in the current article, which concerns itself mostly with the language(s) which a speaker uses as the basis for all explanations. For this, we choose a combination of English and simple mathematical forms—a common social convention.

Syntax is the relationship of sign elements to each other; a syntactic mapping is the set of relationships in a sign to those in some other sign. *Semantics* is the set of relationships of a sign to a conceptual object. In Figure 1, the relationship of the n^{th} intermediate representation to its successor is, for $1 \le n \le 9$, syntactic after the communication has occurred. However its choice before the transformation is accomplished is a *semantic* decision, just as a spoken sentence is a *semantic* choice until it is uttered, but can thereafter be described entirely by syntactic observations.

2.5 Word Distinctions Sought and Distinctions Avoided

We sometimes want to make distinctions that other people overlooked or intentionally wrapped into a more general case. We also sometimes want to avoid vernacular distinctions because we intend to discuss what different situations have in common.

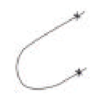
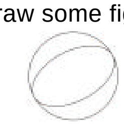
2.5.1 Identifiers and 'the same'

> "The assignment of identifiers to works is a very powerful act; it states that, within a given intellectual framework, two instances of a work that have been assigned the same identifier are the same, while two instances of a work with different identifiers are distinct." [Lynch]

IETF [RFC 1737] reads, "The same URN (Uniform Resource Name) will never be assigned to two different resources." URN applications often assume that blobs associated with the same identifier must be identical, although this constraint is not explicitly called for. The assumption might be appropriate if the referent is a program whose versions behave differently. On the other hand, publishers often use a single ISBN for different printings of a book, even if later versions correct errors and add new material. I.e., **in some social contexts, we use a single sign to signify different things**, even though Tractatus recommends against doing so.[10]

[RFC 1737] does not define 'resource'. Interpreters often ignore common meanings, assuming that 'resource' always refers to a single object that is all 'in one place' even though no such constraint is explicitly specified. A resource might consist of distributed, mobile, and ephemeral pieces, as is the case for 'the British fleet', or 'the *Encyclopedia Britannica'*.

The [Lynch] reminder that "the assignment of identifiers" is a subjective choice circumscribed by complex social conventions teaches that identifiers for digital objects need not follow the same rules as identifiers for books, videotapes, automobiles, or passports. For a resource that encompasses versions or other closely related digital objects, [#3] introduces a new kind of identifier (see §3.2).

> Suppose a man takes two points and draws a line through them:
>
> 
>
> He then shows me two points and tells me, "Do the same"; and I draw this line:
>
> 
>
> Am I necessarily wrong? If he says, "No, I told you to do the same"', I may say. "Well, this is the same."
>
> Or suppose we take a circle and inscribe in it a pentagon, and then a square, and then a triangle. And we now say, "Go on and inscribe a biangle." He might perhaps draw a diameter. "'Now go still *further*, inscribe a monangle."
>
> He might draw some figure:
>
> 
>
> If we said, "But that is different", he might reply, "Well yes, of course. But then the pentagon was different from the square, and the square was different from the triangle."—What is the continuation of that line, and why shouldn't we say that this is?
>
> Or he might say, "There is no such thing"—which would come to "I am not inclined to call anything the continuation of that line."
>
> [LW 39, Chapter 24]

[Renear] reminds us that deciding unequivocally whether two bit-streams represent the same document might be impossible.

Notice the use of 'the same.' [LW 39] teaches how much freedom we have, how subtle the issues are, and how easily we can be misunderstood. Consider 'is similar to' and 'the same' for the simplest pictures (text box), and "I am not inclined to call anything" the same as or similar to some selections you might have made. Even elementary choices are subjective. Do '11001010001' and '11001010001' mean the same thing? [16] The answer can only be, "It depends on the conversation we are having and, even within this context, you might have to interrupt me to ask."

A word or phrase takes on meanings that we share within a context as a matter of social convention. Even context boundaries are social conventions. **We choose particular shared meanings and shared contexts because we find doing so to be useful.**

2.5.2 Synonyms for 'write', 'read', and 'speaks for'

In the current discussion we choose to use verbs denoting communicating, such as 'say', 'write', 'send', and 'sing', as synonyms of each other and of 'transmit'. We do so because, **for the epistemic factors at hand—distinguishing facts that can be objectively known and communicated from concepts that cannot be reliably communicated—the communication modality is irrelevant.**

In a seminal work on computing security, [Lampson] uses 'speaks for' to denote agency. Any written work can be viewed to 'speak for' some authority. Although [Lampson] is analyzing fidelity in the invocation of nested computer programs, its treatment applies also to messages in any social hierarchy.

> "This way of looking at documents also sets up a strong parallel between documents and people. Each in their own way are talking things. This is hardly an accidental parallel. Documents are exactly those things we create to speak for us, on our behalf and in our absence. And in speaking for us, they take on work, they do jobs for us. And not unlike people, they often wear uniforms which broadcast the roles they're intended to play. A newspaper, a cash register receipt, a greeting card, a detective novel—each of these has a distinctive look that's meant to signal what it's for and the kind of content it's meant to carry. ... Each document genre is essentially a specialized form of talk, tailored to operate in particular circumstances ... to do a certain kind of work in the world." [Levy]

The parallel between [Levy] and [Lampson] is striking and purposeful. And Emily Post's *Etiquette* speaks for unnamed arbiters of good manners. How broadly 'speaks' may be used is illustrated by:

> "Heidegger could also be impressively 'silent.' To Mörchen who, alongside philosophy ..., was also studying theology, the lecture on 'existence' had religious significance. He questioned Heidegger, who remained silent—for Mörchen, proof "that nothing speaks more definitely and louder than essential silence." ... In class Heidegger once said: "We honor theology by keeping silent about it." " [Safranski, p.133]

and by

> "French, who sculpted the statue of Lincoln that stares out on the Reflecting Pool, studied ... Lincoln's life mask. ... Embarking ... to begin his first term as president, Lincoln said, "I now leave, not knowing when, or whether ever, I may return, with a task before me greater than that which rested upon Washington." ... [only] when I saw again the look on Lincoln's face that French had captured did I understand.
>
> "French knew Edward Miner Gallaudet, founder of ... the nation's first institution of higher learning for deaf people. Lincoln signed the bill that chartered the college. Look at the statue. Lincoln's left hand seems to spell out in American Sign Language the letter A, and his right hand, the letter L. ... [E]ven if it's just a legend, it's another way Lincoln speaks to us today. [Buckley]

We use as synonyms verbs such as 'see', 'hear', 'read', and so on; nouns for information producers, such as 'author', 'artist', 'speaker', and so on; and other noun sets denoting kinds of information participant.

In ordinary conversation, we often use signs that look different to mean the same thing. How can you know we are doing so? Again, the answer can only be, "It depends on the conversational context."

2.6 *Authenticity*, Documentary *Evidence*, and *Trust*

> "The concepts of reliability and authenticity as expressed in archival discourse are posited on a direct connection between the word and the world and are rooted, both literally and metaphorically, in observational principles. A reliable record is one that is capable of standing for the facts to which it attests. Reliability thus refers to the truth-value of the record as a statement of facts and is assessed in relation to the proximity of the observer and recorder to the facts recorded. An authentic record is one that is what it claims to be and that has not been corrupted or otherwise falsified since its creation." [MacNeil, p.39-40]

Authenticity is a subtle topic made complex by people's tacit expectations. Modern technology supports high expectations.[17] However, there is no *a priori* justification for applying methods used for diplomatic

[16] Even within this article's careful language, we use 'identical' to mean either 'the same object as' or 'equal to', because it is conventional to describe one digital object as identical to another when it might be more accurate to say that the two digital objects are equal!

notes on paper to other kinds of communication or media. Demanding rigidly 'traditional' notions of *authentic* is neither warranted by the facts, nor practical. For instance, some people emphasize 'traditional' authenticity for music; it is, however, ridiculous to insist that the only authentic rendition of a piece would be as it was heard in its first performance.

> "The nineteenth century ... saw compromise as a blemish upon possible perfection ... It was put aside as if, like original sin, it were best ignored; pretending, if it showed, that it didn't exist. All cults, religious and political as well as musical, tend to reject compromise as an unacceptable failing that mars the ideal, diminishes the particularity and weakens the message. ... Churchill is said once to have advised a fellow politician never to abandon his ideals but, equally, never to try to put them into practice.
>
> "Of course, it would have been ... amazing revelation to have heard [Bach's cantata] *Wachet auf* in Leipzig under Bach's direction [in] 1731, but no amount of wishing will make it happen. …" [Leppard]

What's now meant by 'an authentic performance' was socially defined in the latter half of the 20th century, when widespread quality music broadcasts began, and when long-playing records became inexpensive.

We consider a document to be *evidence* if it shows itself to have been indelibly recorded by collaboration of two or more people in a position to know the facts (or if other evidence shows this), and if we further believe that these people have insufficient motive to conspire in falsifying the document at issue.[18] This distills more than a century's deliberation, discussion, and debate by professional archivists.[19] To capture these notions precisely, we chose the following definition[20] for *authentic*. [#9]

---

17 "There seems to be a sense that digital information needs to be held to a higher standard for authenticity and integrity than has printed information. In other words, many people feel that in an environment characterized by pervasive deceit, it will be necessary to provide verifiable proof for claims related to authorship and integrity that would usually be taken at face value in the physical world. For example, although forgeries are always a concern in the art world, one seldom hears concerns about (apparently) mass-produced physical goods—books, journal issues, audio CDs—being undetected and undetectable fakes.

"This distrust of the immaterial world of digital information has forced us to closely and rigorously examine definitions of authenticity and integrity—definitions that we have historically been rather glib about—using the requirements for verifiable proofs as a benchmark. … authenticity and integrity, when held to this standard, are elusive properties. It is much easier to devise abstract definitions than testable ones. When we try to define integrity and authenticity with precision and rigor, the definitions recurse into a wilderness of mirrors, of questions about trust and identity …" [CLIR 92, Lynch]

18 If a scholarly work is sufficiently specialized and difficult, it was surely produced by a scholar—not by a publisher or a librarian. If it conformed in detail to the form of a known periodical, and a copy had been accepted by a reputable library, we would regard it as evidence. The collaboration would be the act of accepting it from the publisher into the library. The publisher and the librarian will have almost surely been motivated against improper collusion.

This definition of documentary evidence itself requires evidence that the collaborating people are in fact the individuals that it represents them to be, and might require evidence that these individuals have sufficient authority to execute the indelible recording whose trustworthiness is in question.

19 See, for instance, [Duranti] on a discipline called 'diplomatics'—content analysis of documents that are bound into document sequences and individually sealed This analytical thread reaches back at least to 16th-century practices in which rulers sealed each important document with a wax impression of a signet ring that never left the ruler's possession.

20 We use mathematical language because doing so helps us communicate succinctly, precisely, and unambiguously. For readers who want ordinary English, the following paragraphs explain the mathematical form.

> The derivation statement R says that some object V is a copy of some object Y. It further states that V is related to Y by some 'copy' function C(y) in "V=C(Y)", but does not in itself say what we mean by C(y).
>
> The copy function C(y) composes transformations $T_k(y)$. $T_k(y)$ describes changes that occurred in the k^{th} transmission step.
>
> The provenance statement S says that there was some person X_k who created some object Y_k in some historical event Z_k. Its practical effect is that we are not willing to accept that the object V is authentic without explicitly identifying the history of V, which includes identification of the authors of the transforms that created V from the original Y.
>
> These statements are merely definitions for V, Y, C, X, and Z. The rest of the mathematical language defines 'integrity', 'true provenance', and finally 'authentic'.
>
> For digital objects Y and V, what we mean by 'having integrity' is that V is identical to Y. For other object types, such as audio sounds and physical artifacts, strict integrity is impossible.
>
> What we mean by 'having true provenance' is that the object V is accompanied by a provenance statement S, and that this statement is true.
>
> What we mean when we say, "V is authentic", is that the integrity and provenance statements are satisfied.

> **Given a derivation statement R, "V is a copy of Y (V=C(Y))" based on**
> **a copy function, "$C(y) = T_n(\ldots(T_2(T_1(y))))$" and**
> **provenance statements S_k, "X_k said or created Y_k as part of event Z_k",**
> **we say that V is a *derivative* of Y if V is related to Y according to R.**
> **We say that 'by X_k as part of event Z_k' is a *true provenance element* if R and S_k are true.**
> **We say that V is *sufficiently faithful* to Y if C conforms to social conventions for the genre and for the circumstances at hand.**
> **We say that V is an *authentic copy* of Y if it is a *sufficiently faithful derivative* with *true provenance.***

The function $T_k(y)$ describes signal transformation occurring in the k^{th} step of an information channel. (Figure 1) Although we call C(y) a 'copy function', it might describe a more complicated transformation than people usually understand by the word 'copy'.[21] Each factor $T_k(y)$ is some user's subjective choice, which is limited only by technical feasibility. Anyone may do as he pleases, at the risk that other people do not accept his information deliveries as authentic derivatives of their inputs if his choices are outside social conventions for the circumstances. The descriptions of the Z_k, X_k, and $T_k(y)$ can be transmitted in protection blocks that are part of each **TDO,** as suggested by Figure 2**.**

Thus, the Figure 1 consumer might judge object 9 to be ***authentic*** if it is sufficiently similar to the input 1 and if its purported history carried as **TDO** metadata is sufficiently complete and sufficiently close to apparent truth. Because some consumers might be unable to resolve uncertainties by dialogue with witnesses, information producers (including editors and any agents who modify what is transmitted) should provide sufficient provenance metadata to answer consumers' anticipated questions.

2.7 Protecting Intangibles: a lesson from copyright law

What deserves preservation is often the same as what merits copyright protection. Since copyright law requires fixation (inscription on a physical substratum), a key question is, "To what does copyright apply if every fixed instance has vanished?" In a parable about intangibles, [Nimmer] teaches that **what is essential about a work is the pattern inherent in its reproductive instances.**[22] Such a pattern is not a conceptual object, but rather a representational pattern (within a Figure 5 'picture').[23]

We can relate what's valid for copyright and what's appropriate for digital preservation, as follows:

- A digital representation is a 'sign' in the sense of *Tractatus* 3.11-3.2.
- Every pattern in the sign is eligible for copyright protection (perhaps in every legal regime).
- Signs have essential and accidental features. The distinction between essential and accidental is a matter of author's intent that cannot be known reliably by other people except to the extent that the author articulates it. Copyright coverage extends to the accidentals.
- One way of communicating about what is essential is to point at schema that articulate how to interpret the bit string that represents the document.
- Schema are signs that might themselves require further signs to explain to readers by reminding them about how the words are used. (See early parts of [PI].) The explications of schema are sometimes called 'reference models' or 'ontologies'.

---

21 For instance, the owner of a photographic portrait of his wife might write the date and her name on the front, further information on the back, and enclose the whole in a glassed frame. Even if the photograph is blurred and the date is incorrect, but indicates the correct year, his wife would probably consider this photograph to be her authentic portrait.

22 This notion of 'what is essential' is at the core of the 1990s lawsuits about 'look and feel' of screen presentations by software, such as the Apple vs. Microsoft lawsuit of 1988-1994. The Society of Motion Picture and Television Engineers calls this 'the essence of the work'**,** and often alludes simply to 'essence'.

23 In contrast, a patent protects ideas inherent in the conceptual object related to the patent. Perhaps this distinction is what makes patent litigation seemingly more difficult than copyright enforcement.

3 ANALYSIS OF THE TRUSTWORTHY DIGITAL OBJECT METHOD

TDO architecture specifies schema and tools for creating and managing cryptographically protected evidence that can be constructed to be very strong. It is silent about either corroborative evidence or evidence within any TDO payload because we have nothing new to say about them. [24] We do, however, emphatically reject the notion that cryptographic audit trails will be, all alone, sufficient evidence under all circumstances.

Trust is immensely complicated both in people's interpretation and their expectations [Denning], [Jordan]. The boundary and interplay between objective facts and subjective values is often not exploited to best advantage—each has importance that is best realized when the interplay is managed explicitly.

Figure 7 reminds us that long-term preservation can be considered a special case of communication patterns such as transmission on physical media, sharing by electronic mail, and sharing by way of digital repositories. What makes preservation more difficult than other information services is that we cannot assume that information producers or other witnesses will be available to resolve consumers' questions.

3.1 A Communication Model based on Digital Objects

Figure 1 and Figure 7 help us express distinctions about representations that occur during message transmission. The document **3** might have been generated from any combination of text, analog sound, images, videos, and computer programs. The transmission 0 to 1 involves semantics (expression of meaning with symbols) and transmission 9 to 10 includes interpretation (trying to understand what the authors intended that received symbols were to mean).[25] All the other steps are (or can be made to be) purely mechanical and syntactic (what the signals look like and how they are transformed). After a communication has been completed, we can describe (at least in principle) the transformation that occurred in each part of the 1➔9 transmission channel. In contrast, we can say nothing about which we are objectively certain about the relationship of a conceptual input 0 or conceptual output 10 to any intermediate form in the transmission channel. We therefore restrict **TDO** authenticity assertions to comparing a 9 instance with a 1 instance, or sometimes also to relationships between other pairs of forms in 1 through 9.

3.2 References and Identifiers

Many digital objects will have related versions in the repository network. We signal this kind of relationship by sharing an attribute that has the semantics of an identifier—a *digital resource identifier (DRI)* [#3 §7.3]—which the **TDO** can include in addition to more conventional kinds of digital identifier. Each producer must decide whether to use a new DRI or the DRI of some existing **TDO**.[26] **TDO**s use DRIs heavily to build audit trails into records.

The authenticity evidence of most **TDO**s will be contingent on reliable linking to other **TDO**s with their own authenticity evidence. Evidence that links are correct is achieved by including with each reference the message authentication code of the referent.

These measures effectively bridge over the difficulties described in §2.5.1.

---

[24] [#3] hardly mentions corroboration or internal evidence that is part of a document's content. For old scholarly documents, the place and manner of holding will often be considered supporting evidence, as will the content of the document at issue and related documents. For instance, a World War II history allegedly written in 1935 is almost surely bogus.

[25] Two observers of the same scene might interpret what they see quite differently; for instance,

> "… if you see this leaf as a sample of 'leaf shape in general' you see it differently from someone who regards it as, say, a sample of this particular shape. Now this might well be so—though it is not so—for it would only be to say that, as a matter of experience, if you see the leaf in a particular way, you use it in such-and-such a way or according to such-and-such rules. Of course, there is such a thing as seeing in this way or that; and there are also cases where whoever sees a sample like this will in general use it in this way, and whoever sees it otherwise in another way. For example, if you see the schematic drawing of a cube as a plane figure consisting of a square and two rhombi you will, perhaps, carry out the order "Bring me something like this" differently from someone who sees the picture three dimensionally." [PI, Part I, § 74]

[26] Only producers have the authority and information to estimate when users will appreciate that a new document is sufficiently closely related to a prior document to do this; machines cannot do it without human guidance.

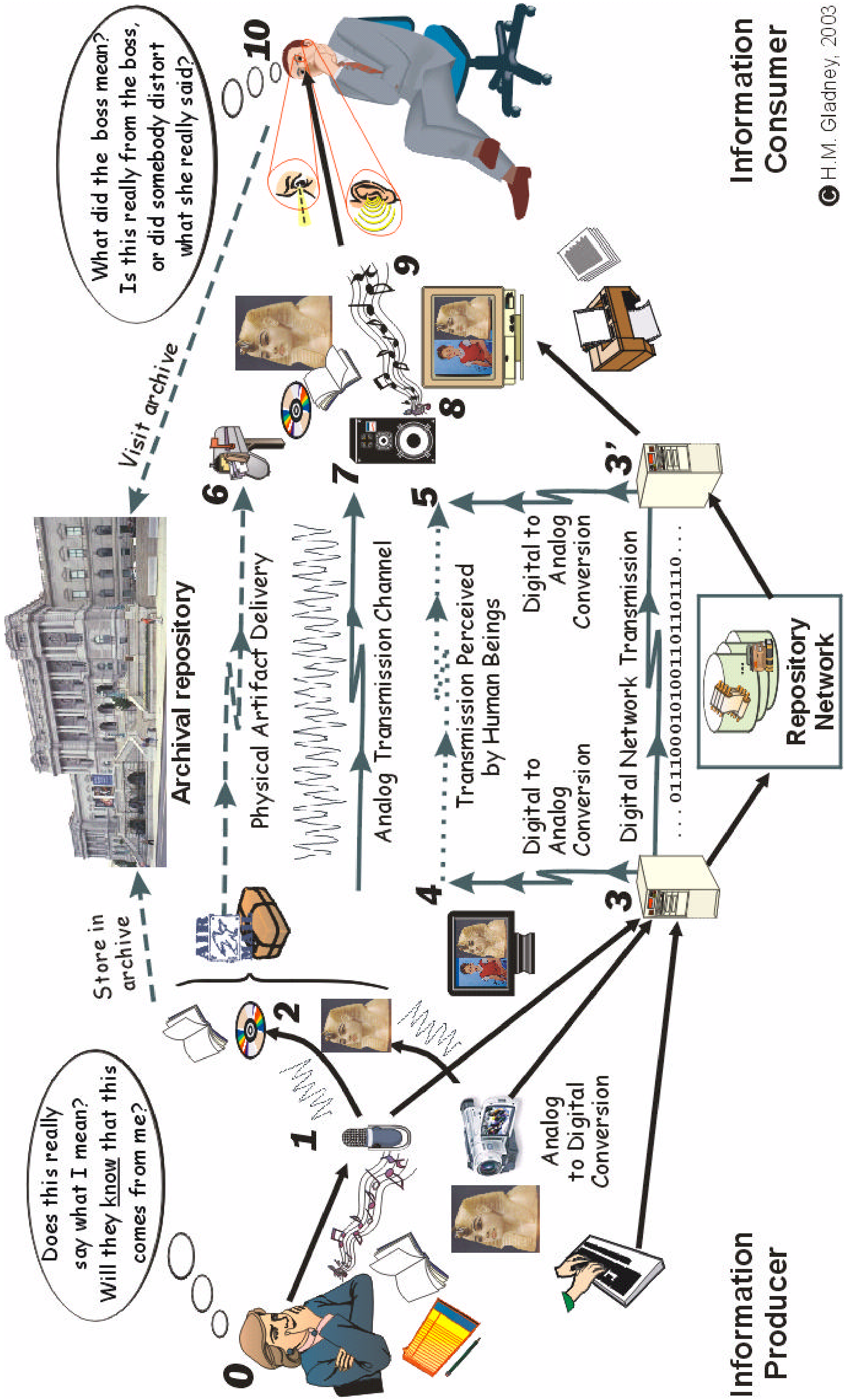


Figure 7: Digital communication steps and intermediate states made manifest, depicting delivery by hard copy, by analog signal, and by digital network. suggesting copying and transformations that might occur.

3.3 Can Digital Communications be Faithful?

Figure 7 emphasizes details about the handling of derivative objects.[27] Without computed transformations, human users cannot easily interpret digital representations, such as **'…01110001010…'** at 3-3', so we provide computer screen representations of the information. The 3-4-5-3' rectangle suggests machine support that is so ubiquitous that we rarely discuss it.

Suppose that an information consumer talked to its producer about a photograph of a King Tutankhamen alabaster bust—an image that the producer had previously sent. A careful explanation of what might have actually happened is that these people caused a binary image 3 in the producer's PC to be copied into an identical binary image 3' in the consumer's PC.[16] Later they conversed about the screen images 4 and 5; 4 is a derivative object of 3 created by rendering software in the producer's computer and 5 is a derivative of 3' created by similar software in the consumer's computer. If these rendering software packages are not identical in the two machines, different pictures will be shown even though 3=3'. Even if the rendering services are identical,[16] it is most unlikely that color rendering by the phosphors of the two video monitors is identical. Almost surely, the consumer will be shown something different from what the producer is shown. Compare the passage:

> "The ideal preservation system would be a neutral communications channel for transmitting information to the future. This channel should not corrupt or change the messages transmitted in any way. … In abstract terms, we would like to be able to assert that, if X_{t0} was an object[28] put into the box at time, t_0, and X_{tn} is the same object retrieved from the box at a later time, t_n, then $X_{tn} = X_{t0}$.
>
> "However, the analysis of the previous sections shows that this cannot be the case for digital objects. … So, the black box for digital preservation … includes a process for ingesting objects into storage and a process for retrieving … and delivering them to customers. These processes, for digital objects, inevitably involve transformations; therefore, the equation ... $X_{tn} = X_{t0}$ cannot be true for digital objects.
>
> "… Nonetheless, given that a digital information object is not something … preserved as an inscription on a physical medium, but something that can only be … reconstructed by using software to process stored inscriptions, it is necessary to have an explicit model or standard that is independent of the stored object and that provides a criterion … for assessing the authenticity of the reconstructed object." [Thibodeau]

The language in such descriptions (others occur in the literature[29]) suggests that a single object is pertinent in each network location, rather than multiple objects, such as 3 and 4 in the producer's PC. In fact, a clump of information can occur each space/time location in several representations that we might need to mention to describe accurately what is happening. Using the [Thibodeau] notation, $\mathbf{X_5 = X_4}$ would be very difficult to achieve and almost impossible to demonstrate had been achieved. In contrast, $\mathbf{X_{3'} = X_3}$ is both easily achieved and easily demonstrated.

---

27 "In the digital environment, we also commonly encounter the issue of what might be termed 'situational' integrity, i.e., the integrity of derivative works. Consider questions such as "Is this an accurate transcript?", "Is this a correct translation?", or "Is this the best possible version given a specific set of constraints on display capability?" Here we are raising a pair of questions: one about the integrity of a base object, and another about the correctness of a computation or other transformation applied to the object. (To be comprehensive, we must also consider the integrity of the result of the computation or transformation after it has been produced). This usually boils down to trust in the source or provider of the computation or transformation, and thus to a question of authentication of source or of validity, integrity, and correctness of code." [CLIR 92, Lynch]

28 Conventional language for this is 'the state of the object', a representation of the conceptual object, as discussed in Tractatus 2.121 to 3.

High-energy physicists use their notion of a timeline—an object trajectory in 4 dimensions that include time—to talk about a sequence of conceptual objects that is the history of some object under discussion.

29 "For digital copies, however, there is likely to be a much greater range of variability. … Some variability is unintended and is an unresolved problem: digital copies are extremely sensitive to the technical environment, to the point that features we would like to preserve in subsequent copies might be hard (or impossible) to maintain. Displaying the file on a different computer might lead to font substitutions, different line breaks, and so on. These same sorts of variability might even occur on the same computer if, in the interim, the environment has changed in some crucial way. Consequently, two different viewings of the 'same' source might differ in important ways—they might not be 'the same.'

"Under such circumstances of radical variability, there does not appear to be anything like a stable document or object. … The version that ends up on your local computer may have been copied from a server and will likely have undergone further transformation; for example, your local browser or editor may generate other local, and possibly partial, digital sources in the process of creating something you can actually see. What you do see at any given moment will be the product both of the local digital source and of the complex technical environment, which is itself changing in complex and unpredictable ways." [CLIR 92, Levy]

Thus the [Thibodeau] conclusion "$X_{tn} = X_{t0}$ *cannot be true for digital objects*" is misleading, as anyone who uses bootleg software, or who fought for passage of the Digital Millennium Copyright Act, can tell you. Figure 7 suggests how **much care is needed to identify to which objects conversation refers**.

3.4 Shifting the Loci of Trust

To make repositories as trustworthy as would be required to ensure the authenticity of their most sensitive holdings would probably be both difficult and expensive.[30] For that reason, the **TDO** method prescribes how to protect documents against improper modification, rather than prescribing how to build "Trusted Digital Repositories" [RLG]. Its trick is to use cryptographic message authentication. [Feghhi]

Repositories have only two critical output classes: reproductions of the records they hold and calculated search results.[31] What archive users want is trustworthy information. They do not particularly care whether any particular library's holdings are trustworthy.[32] While I would value a reliable British Library signature certifying the authenticity and provenance of a document received, I wouldn't much care whether the document was copied from a British Library repository or from somewhere else.

The **TDO** method is insensitive to potential weaknesses in how repositories manage content. **If a properly constructed TDO reaches its destination, its content will be as reliable as its announced editors have in fact made it.** Authors and editors can, of course, be mistaken or even lie; no mechanical process can protect against misinformation arising this way.

> "If we do not have secure access to an independently maintained, known true copy of the object (or at least a digest surrogate), then our testing of integrity is limited to internal consistency checking. If the object is accompanied by an authenticated ('digitally signed') digest, we can check whether the object is consistent with the digest (and thus whether its integrity has been maintained) by recomputing the digest from the object in hand and then comparing it with the authenticated digest. But our confidence in the integrity of the object is only as good as our confidence in the authenticity and integrity of the digest. We have only changed the locus of the question to say that *if* the digest is authentic and accurate, then we can trust the integrity of the object. Verifying integrity is no different from verifying the authenticity of a claim that "the correct message digest for this object is M" without assigning a name to the object. The linkage between claim and object is done by association and context—by keeping the claim bound with the object, perhaps within the scope of a trusted processing system such as an object repository." [CLIR, Lynch]

This passage above is entirely correct, but ends prematurely. It overlooks that we can transfer the loci of trust from numerous objects that are individually relatively large to a much smaller number of small objects—from the individual document copies of interest to a very small number of short bit strings. These bit strings—the private cryptographic keys of a small number of widely trusted institutions—can be easily and very inexpensively protected against improper disclosure.

The **TDO** method does just this, transferring the protection burden that would occur if the primary loci of trust were digital documents (and their interrelationships) to a burden of ensuring that a small number of cryptographic keys are prudently managed. It facilitates three generic sources of trust—information that a consumer can readily exploit to decide whether to trust the provenance and integrity of a **TDO.**

1) The linked context of cited documents, especially linked **TDO**s, whose contents can be judged for consistency with the content at issue. Other **TDO**s are made more reliable for this than they would otherwise be by binding their message authentication codes into the links to them.
2) Access to prior versions of each **TDO**, either by including them in the **TDO** payload ([#3, §3.5]) or by their availability by Internet searches based on their shared resource identifier (§3.2).

[30] No widely available computing system other than some designed for and managed by the military and intelligence communities is sufficiently secure and reliable to store and share holdings that are likely to attract fraud or other chicanery.

[31] This somewhat oversimplifies, because an archive also provides an institutional record and evidence of the quality with which that institution has performed its obligations. Creation of an archive is not the same task as information preservation.

> "Archives truly exist only when there is an unbroken chain of custody from the creating agency to the archives. …
>
> Why is the authorized transfer of a complete set of records to an archives with an unbroken chain of custody important? First, it helps maintain the evidentiary value of the records. An archivist can be called upon to testify in court about the nature of the records in his or her custody. That archivist would not be expected to testify as to the accuracy of the contents of the records. However, he or she should be able to assert that on the day when the records left the custody of the originating "agency or organization, a particular document was included as part of the records." [Hirtle]

[32] This does not imply that librarians and other information service agents can ignore repository trustworthiness. Their perceived performance is likely to affect the reputation of their institutions.

3) Links to descriptors of each **TDO'**s authors and editors and through them to a network of identity certificates that are rooted in the public keys of well-known and respected institutions. ([#3, §4.3])

Any of these records might itself depend on a record network for its authentication. **TDO** end user services will include tools for rapid exploration of this network.

3.5 Evidence of Reliability

> "[T]he methods for assessing the truth-value of records as evidence are rooted in a particular way of looking at the world and in a particular conception of records as a kind of testimony about that world. The criteria they establish for determining what counts as true are themselves the product of historical, cultural, and political choices …" [MacNeil, p.45]

Who is to decide what evidence is important and what tests are worth applying? That must be each end user of a preserved document. His decision will be economic, weighing his cost of providing or testing evidence against his risks of a wrong decision. Producers can only guess what information consumers will find persuasive. Technology can minimize the cost of testing. **An object's producer can make it *trustworthy*, but only its consumer can test and judge whether it is to be *trusted.***

To help the consumer decide whether documents he receives are authentic, the **TDO** method provides him evidence in the form of metadata sealed with the content at issue. People usually accept what they see 'in black and white' because to be skeptical about everything would be exhausting, tiresome, and counter-productive. They do, however, notice when information they depend upon clashes with their practical and professional experience.[33] That is when they might seek refutation or evidence.

If the risk to the consumer is low and the fact at issue is of a kind for which chicanery is unlikely, one element of strong evidence might be sufficient. If the risk is great, and misdeeds are common (for instance, if a great deal of money is involved), deciding on the basis of a single piece of evidence might be imprudent.

How can we test for authenticity (for a genre [#9])? Books on paper often carry evidence accidentally, because they include a great deal of information not specifically planned by their authors. Furthermore, the social infrastructure for producing paper renditions—social infrastructure that evolved over more than a century of trials that established what 'works'—has not been matched by any counterpart for digital renditions, at least not yet.[34] Since digital documents are relatively sparse in accidental information, their authors and editors should consider deliberately creating evidence similar to that provided by customary procedures and side effects of publication on paper.

4 DISCUSSION

The **TDO** methodology is intended to satisfy archivists' criteria for prudent custody of socially important evidence—criteria refined by years of critical thinking and discussion. Since solutions developed for content that is mostly represented by paper documents lend themselves poorly to adaptation for digital content,[35] [#3] sketched a method well adapted to digital objects and to the evolving international information infrastructure.

---

[33] For instance, even the most exacting scientist accepts hearsay about scientific facts that do not conflict with his education and experience, and makes extra effort only for information that could jeopardize his conclusions if it is flawed. In fact, a speaker's offer to substantiate hearsay is often rejected because he offered, the offer itself being taken as evidence.

[34] "… technological solutions … will probably emerge that would obviate the need for trusted third parties. Such solutions might include, for example, embedding texts, documents, images, ... with various warrants (e.g., time stamps, encryption, digital signatures, and watermarks). The technologists replied with skepticism, saying that there is no technological solution that does not itself involve the transfer of trust to a third party. Encryption ... and digital signatures are simply means of transferring risk to a trusted third party. Those technological solutions are as weak or as strong as the trusted third party." [CLIR 92, Smith]

[35] Archives of physical objects usually require a client to come to their facilities to inspect any holding, and often impose constraints that help make the object safe. For instance, visitors to the U.K. Public Record Office (PRO) sit in a room under the surveillance of a uniformed overseer. The PRO also records who handled each holding. Such practices might be what lead archivists to state their requirements in terms of the physical integrity and history of a document instance and of the premises and procedures of the archival institution.

4.1 Challenges: Justification in Lieu of Proof

We want confidence in our preservation methodology long before its potential beneficiaries are likely to discover any errors or omissions. Part of achieving this is painstaking analysis of the epistemology, standardization, and procedures for packaging and administering digital documents.[36] Since the cost of such analysis is tiny compared to the effective costs of delayed discovery of technical design or implementation errors, the most thorough possible analysis is not only affordable, but also a practical and ethical imperative.

Logical proof is probably impossible for the assertions in this article, because there exist neither proof rules nor an axiomatic basis for epistemology. Plausibility relative to common experience and basic ethics is as much as we can manage. On the other hand, anyone who contests one of our assertions has the opportunity to provide a counter-example that persuades independent observers. Absent a counter-example in an environment that has encouraged scholarly and public criticism for some years, it would be prudent to accept assertions from a careful analysis.

We believe that our analysis justifies the soundness of the **TDO** methodology to the extent possible by logical reasoning. However, until the ***Trustworthy 100-Year Digital Objects*** article series has been subjected to searching peer review we cannot be as confident as we would like to be.

4.2 Reliable Documents vs. Reliable Repositories

Most efforts toward long-term digital record-keeping have focused on the repository, rather than on the structure of preserved digital information. For instance, [RLG]'s declared approach to authenticity is to:

> *"Develop a framework and process to support the certification of digital repositories.* ... their absence has been an impediment to assigning trust. Model processes, including checklists for certification reviews, should be developed incorporating the community-approved attributes of trusted digital repositories, …"

How can a repository be built so that an audit trail cannot be faked? This cannot be done without special technology and business controls which few customers will be willing to pay for or will understand how to manage properly. **Trying to construct a 'trusted' digital archive is a waste of money.**

For long-term preservation of digital content, **it's even the wrong objective!** In a century or so, no one will care about the capabilities and weaknesses of today's repositories. Instead, what people will care about is whether digital content they can fetch into their own workstations is authentic and tightly bound to reliable evidence for its authenticity.

The **TDO** mechanism does not ensure the survival of the world's last copy of any document. This risk can be inexpensively mitigated by combining managed replication [Reich] and distributed Internet crawlers that test for the continued existence of multiple copies of any document instance of interest, creating fresh copies of instances at risk.

The [RLG] recommendation has inherent weaknesses including that:

- It depends on an unexpressed premise—that exposing an archive's procedures can persuade clients that its content deliveries will be authentic.
- It does not identify objective certification criteria—only subjective guidelines. It does not identify risks and failure possibilities, much less mechanisms to minimize the frequency and impact of each.
- Repository reliability would be insufficient. It's essential also to block chicanery in the Internet.
- Certifying institutional procedures is more expensive and less reliable than collecting stored objects bound to their own audit trails. Occasional audits cannot prove content authenticity.[37]

4.3 Summary: How TDO Methodology Uses Epistemology

> What am I believing in when I believe that men have souls? What am I believing in, when I believe that this substance contains two carbon rings? In both cases there is a picture in the foreground, but

[36] Some authors call this 'knowledge management'. We prefer 'information management'.

[37] The reader of a century-old document would have to assume that, during the entire time the archive held the document, no lapse in the exercise of management controls permitted the key content to be improperly modified. For information of critical importance to this consumer, this might not be a prudent assumption.

> the sense lies far in the background; that is, the application of the picture is not easy to survey.
> [PI, I, 422]

Careful attention to rules and procedures has its origin in formal logic. Wittgenstein's discussions of **rules** [Kripke] help towards understanding how his teachings apply to digital preservation methodology**.** We cannot convincingly **say why** Wittgenstein's teachings are the best available ground, but we can **show what** use we make of them. Specifically:

- Words and sentences take meaning only from the contexts of which they are parts. What is important is how words and sentences **are used**—and **what we find useful**.
- Misleading language impedes achieving the trust essential for effective communication.[38] We look at how authors use critical words, and decide whether each sentence expresses a fact or an opinion. Such distinctions help the *Trustworthy 100-Year Digital Objects* series stay within the engineering role, avoiding infecting technical solutions with value judgments that readers might find inappropriate.
- The limits of what can be automated, or specified as clerical tasks, are rules[39] whose execution requires no more than a finite number of steps.[10] [40]
- We cannot with words **define** a boundary between what's objective and susceptible to automation and what's subjective and cannot be provided by machines, and must therefore be left to human judgment and taste. We can **illustrate** the boundary with examples.
- The meaning and social significance of any information object is contextually rooted in other information objects. Its producers can (and should) identify the most important of these for eventual consumers. **TDO** structure supports doing this with reliable references.

Nothing in our application of epistemology above has been intended to say anything scientific (facts about the world) or prescriptive (how anyone should behave). Instead, we have used a language game[41] to learn and show how to describe abstractions such as 'evidence' and 'authentic'. The language game is based on analyzing message transmissions into pieces whose potential imperfections we can describe.

4.4 Near-Pedantic Care—An Ethical Imperative

The end to which we should apply technology is not a technical issue, but rather an ethical question.

In what follows, we abandon ethical neutrality to assert essential digital preservation objectives.

4.4.1 Facts and Values

> "Schopenhauer's attack on the Scholastic residue in Kantian thought ... makes morality immediately dependent upon feelings and intentions. The sphere of facts and the sphere of values, which in Kant were distinct but by no means separate, come to be separated widely in the philosophy of Schopenhauer. In the thought of Soren Kierkegaard, this separation becomes an unbridgeable chasm." [Janik p.156]

> "Let your light so shine before men, that they may see your good works, and ..." Matthew 5:16

Just as Karl Kraus "resented the mingling of opinion and fact" "because it prevented the reader from making any rational assessment of the facts" (§2.1), we believe that **everyone who participates in information packaging for preservation should do his best to signal whether each statement expresses a subjective opinion about values or an objective fact about the world or an event.**[42]

---

[38] Inattention to the meaning and intent of key words, phrases, or sentences can make consequent reasoning and conclusions incorrect or misleading. The risk is greatest when we include ill-founded assumptions about notions that are 'obvious'. Potential misunderstandings between different professions figure prominently.

[39] An early attempt at this kind of automation is *The eXtensible Rule Markup Language (XRML)*. [Lee]

[40] The reader might recognize this to be an application of the Church-Turing Thesis. See the *Stanford Encyclopedia of Philosophy* at http://plato.stanford.edu/archives/fall2002/entries/church-turing/ (visited 8 Jan 2004).

[41] Here 'games' are not frivolous, and certainly not 'spin doctoring'. Instead this is a reminder of *Philosophical Investigations,* which includes*:* "We can also think of the whole process of using words … as one of those games by means of which children learn their native language. I will call these games 'language games' …" [PI, Part I, §7]

[42] The distinction might not matter for the manner in which or the purpose for which much content is used. However, careless treatment of this distinction will inevitably damage some consumers.

4.4.2 Objective and Subjective

One can always construct an objective fact that corresponds to a subjective value, opinion, or choice. For instance, the claim, "On 6/1/01, John Doe said, 'I believe X.'" is objective, even if it happens to be false. The metadata that testify to an object's origin and historical significance might include such assertions.

In fact, the provenance statements of §2.6 ought to conform to the paradigm, "**X said S as part of the historical event Z.**" What we call for as 'an unbroken chain of custody' is that **the creation of a record and every subsequent change should be addressed by such a provenance statement.**

4.4.3 Syntax and Semantics

Words, pictures, models and signals represent things other than themselves. **Computers manipulate symbolic surrogates for meanings**. A digital model is useful if its pattern conforms to the pattern of what it stands for.[43] [LW 39] illustrates the relationship of language to meaning, teaching that language consists of symbols taking meaning from how they are used. [PI] discusses what language cannot convey at all (ethical values) or express unambiguously about intended meaning (semantics).

The **TDO** method pays explicit attention to the boundary between syntactic transformations and semantic interpretations. Its prescriptions are constrained to objective fact management that decomposes information transmissions into steps whose inputs and outputs can be compared with one another.

The current article identifies minimal acceptable quality for digital preservation.

4.4.4 Trust

'Trusted' should not be used when what is meant is 'trustworthy'. §3 evaluates an alternative to the 2002 preservation strategy of NARA and SDSC [SDSC]. Relative to the question "how objective and precise can we be?", [Thibodeau] fails to consider who decides to trust and when the decision is made as implied by Figure 7.

Is there a reasonable of extension of the human notion of trust to inanimate systems?

> "Human trust relationships come into existence only when one individual knows another sufficiently to be confident that some limited responsibility will be faithfully discharged ... or that some defined risks will be avoided ... . In most cases of trust conferred, there are adverse consequences to a breach of trust; these may be explicit, but are often implicit (e.g., if the babysitter ignores my wailing child, I will not employ her again ...). Trust relationships are extended from individuals to corporate entities by agreements made with human agents of those corporations and tend to be more explicit both in their scopes and their damage commitments than are those between individuals, ... " [HMG]

Notwithstanding such published cautions, [RLG] focused on archival procedures that require certification. Its very naming, 'Trusted Digital Repositories', sets a trap for the unwary—a sort of false advertising.[44]

5 CONCLUSIONS

Originators, editors, archivists, and librarians can make any record **trustworthy** by binding metadata and test procedures to it durably and reliably, i.e., to be firm over time and proof against otherwise undetectable modification. Whether or not the document will be **trusted** by its eventual user will depend on its quality, the quality of producers' metadata, the application at hand, and the user's judgment. *Trustworthy 100-Year Digital Documents: Syntax and Semantics—Tension between Facts and Values* shows that:

---

[43] Jargon for this includes sentences like, "Information is **represented by** an XML document." A model that faithfully represents what's intended is said to be ***isomorphic*** to the facts.

[44] This use of the term 'trusted' originated in 1975 in the Trusted Computing Base (TCB)—a security processor kernel whose behavior is effectively protected against alteration by applications—intended for military command and control applications. This designers' jargon—calling the kernel 'trusted'—was appropriate because they also built the trusting entity—an operating system that needed the TCB to meet defense security criteria. This operating system was itself protected against modification by the kernel, and was certified as secure by audit teams. [NCSC] In contrast, the internal workings of an archive cannot be engineered so that a cautious user could reliably test the authenticity of information he receives.

The [RLG] authors did not invent sloppy use of 'trusted'. By 1990, several authors had apparently forgotten that 'trusted' made sense only for systems whose user agent was within the control perimeter.

- Current uncertainties and confusions about digital preservation can be reduced by clear distinctions between assertions that are facts and assertions that are opinions about values.
- Sufficient fundamentals are known for making any digital document survive forever in useful form. However, further engineering will be needed to provide tools for repository managers and end users.
- Achieving trust by audit trails built into digital documents will be more reliable than seeking it by institutional procedures certified by independent auditors.
- Consumers' decisions about digital content authenticity and completeness can be made entirely by analysis of the evidence inherent in properly constructed **TDO**s and their referents.

Arguably, the topic at hand is most of the history of the future. This expectation justifies our making every practical effort to ensure that our methodology is sound and widely understood. For this, we have returned to fundamental epistemology that teaches us (1) to share definitions that are sufficiently precise and distinct from each other to explain what we are doing and to minimize community confusion, and (2) to articulate the boundary between what can be specified precisely and automated and what must forever remain issues of human values, opinions, and imperfectly communicated intentions. We have applied these distinctions to a pictorial digital communication model illustrated in Figure 1 and Figure 7.

Such philosophic distinctions are critical to issues of trust and of preserved information authenticity. We believe that **the current article identifies essential criteria for any approach to long term digital preservation.**

Acknowledgements

Once again, it is a pleasure to acknowledge that discussions with John Bennett, Tom Gladney, Raymond Lorie, Peter Lucas, and John Swinden were extremely helpful towards creating this article.

Bibliography

[Bearman]	D. Bearman and J. Trant, *Authenticity of Digital Resources: Towards a Statement of Requirements in the Research Process*, *D-Lib Magazine,* June 1998.
[Buckley]	C. Buckley, *Mr. Lincoln's Washington*, Smithsonian Magazine 34(1), 76-84, April 2003.
[CLIR]	C.T. Cullen et al., *Authenticity in a Digital Environment*, CLIR pub92, 2000. ISBN 1-887334-77-7.
[Denning]	Peter J. Denning, *Accomplishment,* Comm. ACM 46(7), 19-23, 2003.
[Duranti]	L. Duranti, *Diplomatics: new uses for an old science,* Scarecrow P., 1998. ISBN 0-810-83528-2
[Feghhi]	J. Feghhi, P. Williams, and J. Feghhi, *Digital Certificates: Applied Internet Security*, Addison-Wesley, Reading, MA, 1998. ISBN 0-201-30980-7
[HMG]	HMG and J.B. Lotspiech, Safeguarding Digital Library Contents and Users: Storing, Sending, Showing, and Honoring Usage Terms and Conditions, D-Lib Magazine 4(5), 1998.
[#1]	H.M. Gladney, *Trustworthy 100-Year Digital Documents: Summary of a Digital Preservation Proposal,* preprint July 2003.
[#2]	H.M. Gladney and R.A. Lorie, *Trustworthy 100-Year Digital Documents: Durable Encoding for When It's Too Late to Ask,* June 2003.
[#3]	H.M. Gladney, *Trustworthy 100-Year Digital Documents: Evidence Even After Every Witness is Dead,* May 2003 preprint.
[#5]	H.M. Gladney, *Trustworthy 100-Year Digital Documents: What's Authentic? Essential and Accidental in Documents,* planned for December 2003.
[#9]	H.M. Gladney and J.L. Bennett, *What Do We Mean By Authentic? What's the Real McCoy?* D-Lib Magazine 9(7), July 2003.
[Glock 96]	Hans-Johann Glock, *A Wittgenstein Dictionary*, Basil Blackwell, Oxford, 1996. ISBN 0-631-18112-1
[Gordon]	M.J. Gordon, *The Denotational Description of Progamming Languages: An Introduction*, Springer-Verlag, New York, 1979.
[Graham]	L. Graham and P.T. Metaxas, *"Of Course It's True; I Saw It on the Internet" Critical Thinking in the*

Internet Era, Comm. ACM 46(5), 70-75, May 2003.

[Grandison] T. Grandison and M. Sloman, *A Survey of Trust in Internet Applications*, IEEE Communications, Surveys & Tutorials, 4th Quarter 2000.

[Janik] A. Janik and S. Toulmin, *Wittgenstein's Vienna,* Ivan R Dee, Inc., 1997. ISBN 1-566-63132-7

[Jordan] K. Jordan, J, Hauser, and S. Foster, *The Augmented Social Network: Building identity and trust into the next-generation Internet*, First Monday 8(8), August 2003.

[Kripke] S.A. Kripke, *Wittgenstein on Rules and Private Language,* Harvard U.P., 1984. ISBN 0-674-95401-7

[Lampson] B. Lampson, M. Abadi, M. Burrows, and E. Wobber, *Authentication in Distributed Systems: Theory and Practice,* ACM Trans. Computer Sys. 10(4), 265-310, 1992.

[Lee] J.K. Lee and M.M. Sohn, *The eXtensible Rule Markup Language,* Comm. ACM 46(5), 59-64, 2003. See also http://xrml.kaist.ac.kr/download/XRML-Ver05.doc

[Leppard] Raymond Leppard, *Authenticity in music,* Faber Music, 1988. ISBN 0-571-10088-0

[Levy] D.M. Levy, *The Universe is Expanding: Reflections on the Social (and Cosmic) Significance of Documents in a Digital Age*, Bull. ASIS 25(4), 17-20, 1999.

[Lynch] C. Lynch, *Identifiers and Their Role in Networked Information Applications*, ARL: A Bimonthly Newsletter of Research Library Issues and Actions 194, October 1997.

[MacNeil] H. MacNeil, *Trusting Records in a Post-Modern World,* Archivaria 51, 36-47, Spring 2001.

[Marcum] D. Marcum, *Research Questions for the Digital Era Library*, Library Trends 51(4), 636-651, 2003.

[METS] *Metadata Encoding and Transmission Standard* (METS), 2001-2.

[Moore] G.E. Moore, *Principia Ethica,* Cambridge UP, 1903 and 1999. ISBN 0 521 44848 4

[NCSC] National Computer Security Center, *A Guide to Understanding Audit in Trusted Systems*, NCSC-TG-001, 1 June 1988. Part of the *Rainbow Series Library*. Also Guidelines for Writing Trusted Facility Manuals, NCSC-TG-016, 1992. Part of the *Rainbow Series Library*.

[Nimmer] D. Nimmer, *Adams and Bits: of Jewish Kings and Copyrights*, 71 S. Cal. L. Rev. 219-245, 1998.

[RDF] W3C (E. Miller, R. Swick, and D. Brickley), *Resource Description Framework (RDF)*, Nov. 2001.

[Reich] V. Reich and D.S.H. Rosenthal*, LOCKSS: A Permanent Web Publishing and Access System*, D-Lib Magazine, June 2001.

[Renear] Allen Renear and David Dubin, *Towards Identity Conditions for Digital Documents*, Dublin Core Conference: Supporting Communities of Discourse and Practice—Metadata Research and Application, 2003. **annotated**

[RFC 1737] K. Sollins and L. Masinter, *Functional Requirements for Uniform Resource Names*, IETF RFC 1737, 1994.

[RLG] N. Beagrie, M. Bellinger, R. Dale, M. Doerr, M. Hedstrom, M. Jones, A. Kenney, C.Lupovici, K. Russell, C. Webb, and D. Woodyard, *Trusted Digital Repositories: Attributes and Responsibilities,* RLG-OCLC Final report, May 2002.

[Safranski] R. Safranski, *Heidegger: Between Good and Evil,* Harvard UP, 1998. ISBN 0-674-38710-4

[SDSC] R. Moore and A. Merzky, *Persistent Archive Basic Components,* Global Grid Forum; Jul 27, 2002.

[Steemson] M. Steemson, *Digital Experts Search for E-Archive Permanence* in *Integrity and Authenticity of Digital Cultural Heritage Objects*, DigiCULT Thematic Issue 1, August 2002.

[Thibodeau] K. Thibodeau, *Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years,* in CLIR and the Library of Congress, *The State of Digital preservation: An International Perspective*, April 2002. ISBN 1-887334-92-0

[Tractatus] L. Wittgenstein, (trans. D.F. Pears and B. F. McGuiness), *Tractatus Logico-Philosophicus*, *with Introduction by* Bertrand Russell, Routledge, London, (1921 and 1997).

[LW 39] L. Wittgenstein, ed. C. Diamond, *Wittgenstein's Lectures on the Foundations of Mathematics*, *Cambridge, 1939*, Univ. of Chicago Press, 1976. ISBN 0-226-90426-1.

[PI] L. Wittgenstein, tr. G.E.M. Anscombe, *Philosophical Investigations: The English Text of the Third Edition,* Prentice Hall, N.J., 1958. ISBN 0-631-23127-7

Glossary (Conventional Definitions) [45]

authentic

1. entitled to acceptance or belief because of agreement with known facts or experience; reliable; trustworthy: *an authentic portrayal of the past.*
2. not false or copied; genuine; real: *an authentic antique.*
3. having the origin supported by unquestionable evidence; authenticated; verified: *an authentic document of the Middle Ages; an authentic work of the old master.*
4. *Law.* executed with all due formalities: *an authentic deed.*

fact

1. piece of information about circumstances that exist or events that have occurred; "first you must collect all the facts of the case"
2. statement or assertion of verified information about something that is the case or has happened; "he supported his argument with an impressive array of facts"
3. event known to have happened or something known to have existed; "your fears have no basis in fact"; "how much of the story is fact and how much fiction is hard to tell"
4. concept whose truth can be proved; "scientific hypotheses are not facts"

a thing done; an action performed or an incident happening; an event or circumstance; an actual occurrence; an actual happening in a time or space or an event mental or physical; that which has taken place. A fact is either a state of things, that is, an existence, or a motion, that is, an event. Black' Law Dictionary, 6th ed. (St. Paul, 1990).

value

2. quality (positive or negative) that renders something desirable or valuable; "the Shakespearean Shylock is of dubious value in the modern world"
6. an ideal accepted by some individual or group; "he has old-fashioned values"

datum, data

1. collection of facts from which conclusions may be drawn; 'statistical data'
1. item of factual information derived from measurement or research

information

1. message received and understood that reduces the recipient's uncertainty
2. collection of facts from which conclusions may be drawn; "statistical data"
3. knowledge acquired through study or experience or instruction

integrity

1. property of a system, component, or data structure of being protected from inappropriate modification;
2. for a passive data structure, a synonym for consistency (q.v.);
3. for a set of data and processes, the property of conforming to some external set of rules defining correctness. Commonly these rules constitute a contract or promise of a guaranteeing agent to some set of human beings considered as customers;
4. the property that an object meets an a priori of expectations. One example of integrity is that changes rnust be accomplished in a specified and authorized manner.

intensionally

Sets and sequences are communicated either *intensionally*—"(3n+1) where n is a natural number"—or *extensionally*—"1, 4, 7, 10, …".

knowledge

cognition, knowledge, noesis—the psychological result of perception and learning and reasoning

objective

1. nonsubjective, undistorted by emotion or personal bias; based on observable phenomena; "an objective appraisal"; "objective evidence"
3. expressing things as perceived without distortion of personal feelings or interpretation; "objective art"
4. belonging to immediate experience of actual things or events; "concrete benefits"; "a concrete example"; "there is no objective evidence of anything of the kind"

subjective

1. taking place within the mind and modified by individual bias; "a subjective judgment"
2. immanent (vs. transeunt), of an act performed entirely within the mind; "a cognition is an immanent act of mind"

semantics

relationship of meaning to signs, pictures, models, or language (Figure 5)

syntax

set of rules that define what it means for a language (or a sign or a model) to be well-formed

[45] These definitions of key words were mostly selected from WordNet.