

Digital publication for digital libraries

Neel Smith, Nov. 24, 2004

Copyright 2004 Neel Smith and licensed under the Creative Commons Attribution-NoDerivs license. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nd/2.0/> or send a letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.

Contents

| | |
|---|----|
| Introduction..... | 1 |
| Digital incunabula..... | 3 |
| Defining scholarly publication..... | 6 |
| Digital technologies for scholarly publication..... | 8 |
| Defining scholarly libraries..... | 18 |
| Digital technologies for scholarly libraries..... | 20 |
| Postscript: craft and engineering in hardware and software..... | 22 |

Abstract

I define the primary function of scholarly publication to be the creation of a permanent record of reference for scholarly work. Works fulfilling that function must possess an explicitly identified edition and explicitly identified citation scheme, and must be capable of being irrevocably and identically replicated. Based on these requirements, I propose eight requirements for the design of scholarly publications in digital media.

I define a scholarly library as a setting enabling the scholarly use of a collection of published works. This suggests a minimum of four requirements for the design of a digital scholarly library.

Introduction

The immediacy and reach of the internet are reshaping scholarly communication. Alongside traditional venues such as professional meetings, conferences, colloquia or panel discussions, the internet can help create new communities of interest, and offers existing communities new forms of communication including email lists, discussion groups and blogs.

Digital information technologies have equally dramatic implications for the permanent published record intended to communicate to both present and future scholars. My intention is to define as clearly as possible the assumptions and reasoning behind technical work at the Center for Hellenic Studies in support of its electronic publication initiative.

It is vital that we define explicitly what we mean by "scholarly publication," independent of any possible media for publishing, and what we mean by a "library of scholarly publications," independent of any possible media. Otherwise we risk falling into the trap of reshaping our work to the forms of a particular medium, rather than asking how different media enable us to pursue our goals. This is a serious risk for a discipline like Classics, because decisions we make today about the form of our permanent scholarly record will have consequences reaching far into the future.

Normal English usage complicates the job. The word "publication" can refer to the process of publishing something (e.g., "the future of scholarly publication"), to a single act of publishing ("publication is expected by January of next year"), to a product of that act ("The

library has Austin and Bastianini's publication of the Posidippus papyrus"), or to the content of that product ("Petrie's publication first developed the idea of seriation"). For the sake of clarity, I will use the term "publication" only to refer to the process of publishing; I will refer to specific products of that process as "published works," and where necessary further qualify that with expressions like "published works in print media" or "digital published works." I will use the phrase "scholarly work" to refer to the content of a published work without implying that the work is published in any specific form, or even that it is published at all.

I begin this introductory section with an example illustrating why it is necessary to separate definitions of basic concepts from any reference to specific media. The following sections then set out

- a working definition of "scholarly publication"
- consideration of what that definition implies for the design of scholarly published works in digital form
- a working definition of "scholarly libraries"
- consideration of what that definition implies for the design of a digital scholarly library

Finally, a postscript briefly points to analogies between the production of hardware and software on the one hand and the production of scholarly published works on the other.

In considering such a broad topic as "digital publication for digital libraries," I hope that classical scholarship's unusual experience can provide a useful perspective. Classical scholars have seen the practice of recording primary scholarly materials migrate from papyrus rolls to codex manuscripts to printed materials; that experience should help us now as we migrate from print materials to collections of bits capable of being represented in many physical media (magnetic disk, RAM, and laser discs such as CD ROMs, for example).

Digital incunabula

In the transition to printed materials, Classics' track record has arguably been superior to most disciplines, as an important example will illustrate. Consider the consequences of including page numbers in printed works. This would have been pointless in hand-copied codices (since pagination might vary with each copy), and page numbers are not regularly found in fifteenth century incunabula, which tend to perpetuate the physical form of hand-copied texts. In the course of the sixteenth century, however, printers and readers came to recognize the value of labelling pages that would be identical across every copy of a given edition, and page numbers came to be used as a standard navigational device for working with printed books through new inventions like indices and tables of contents.

The temptation that most authors and readers succumbed to was to use this convenient navigational device as a reference system. Page numbers became the normal way to cite a passage in a printed text. This remains today a severely limiting convention, because it ties the logical act of reference—as I will argue below, perhaps *the* most fundamental scholarly activity—to a specific physical edition or version of a work. How frustrating this limitation is will be familiar to any teacher who has had to use a new edition of a textbook with nearly identical content, and found last year's notes out of sync with the current edition.

For the most part, Renaissance classicists and biblical scholars resisted this temptation. They recognized that their texts would be read in multiple versions—different editions and translations—that could all use a single canonical citation system. Phrased differently, they devised a logical reference scheme independent of any specific version of a text, and, because of their long experience with multiple versions of their central texts, avoided confusing the physical form of a new printed book with its logical structure. So today, we can still cite "Herodotus 1.5"

and be confident that it refers to the same passage (book 1, chapter 5) in any version of Herodotus we prefer to read.

There is an audible echo of this decision in a discussion that has developed in the decade since the introduction of the World Wide Web. "Uniform Resource Locators," or URLs, are addresses on the internet providing a navigational scheme, and it is tempting to use them as a form of reference. Like page numbers, however, they are valid for only one version of a document, and in the accelerated time of the internet's world, the limitations of this navigation scheme as a form of citation are already familiar to all users of the WWW who run into dead-end links ("404 limbo"). In many cases, the desired document may still exist, but at a new address.

The Online Computer Library Center developed "Persistent URLs" [1] to implement the notion of Uniform Resource *Names*, or URNs, articulated by the World Wide Web consortium (W3C). URNs are intended to describe logical names rather than specific locations on the internet. The confusion of logical name and physical address has become pervasive enough that the W3C maintains a page clarifying how terms for logical and physical reference are used in different senses in different parts of the W3C's documentation. [2]

In at least one respect, however, the discussion of how to refer to resources on the WWW has not yet recognized a significant implication of classicists' traditional practice of citation by logical reference: namely that the idea of a logical reference scheme follows naturally from a notion of a *logical text*. A canonical reference scheme citing Herodotus by book and chapter is a logical or notional reference in the sense that the organizational scheme is independent of any

[1] PURLs, see <http://purl.oclc.org>

[2] <http://www.w3.org/TR/uri-clarification/>

specific version of the work. By the same token, reference to a text called *Histories of Herodotus* is itself purely logical in the sense that it refers to a notional text independent of any specific version. While URLs and URNs, on the other hand, both assume that documents may move their physical address, their concern is really with a single document. URLs and URNs have no concept of a notional document that would allow readers to treat a Greek text and English translation of Herodotus as two versions of a single logical text. [3] For a classicist, there would have been no incentive to develop a regular, canonical scheme of book/chapter/section citation for one document called *Histories* by Herodotus if references in that scheme only applied to a single edition or translation of the work. The power of the logical reference scheme lies largely in its application to an abstract notion of a work encompassing different editions and translations. That is why the first project undertaken by the Center of Hellenic Studies' technical work group defines a protocol supporting logical citation schemes for logical texts.

This extended story about how to cite a work should serve first as a warning. We still live in a period of experimentation in digital media comparable to the experimentation we see in fifteenth-century incunabula, when early printed works tended to mimic the physical form of hand-copied manuscripts. Our conventions are not yet settled, and the very real risk is that we will uncritically allow the new media to shape those conventions, just as most disciplines uncritically came to use page numbers as an easy substitute for a logical reference scheme. At the same time, the story is a challenge: classicists can draw on the long experience of their

[3] Some applications using URLs approximate this kind of behavior. If a Web server and browser both support content negotiation, for example, the server can recognize a browser's preferences (for example, for one language over another) and serve one version of a document rather than another. This behavior is application specific, however: there is nothing in the URL protocol supporting it.

discipline to inform their thinking about today's digital incunabula.

Defining scholarly publication

We start with a functional definition: the distinctive role of *publication* in the scholarly world is to serve as *the permanent record of reference for scholarly work*. By this I mean that in contrast to other vital but more ephemeral forms of communication, published works constitute the sole record of scholarly work that other scholars, present and future, can reliably refer to and cite.

We habitually imagine this function in terms of published works in print: volumes of ink on paper, physically produced by publishers who market the volumes to scholars and libraries. We may even assume that features of published works in print that directly contradict the aim of a permanent record of reference are not artifacts of a particular system of technologies and institutions, but are intrinsic to scholarly publication. It is easy to assume, for example, that difficult or technical scholarly works will be accessible only in a few major research libraries—because after all it's not feasible to print more than a very small run of such material. Humanists whose work depends on published photographs and drawings will "know" that those published images will be limited in number because of the expenses of "publishing" them—although there is no scholarly rationale for this limit, which is antithetical to the goals of publication. We sometimes accept these limitations on the scholarly record as inevitable, but before we can hope to reimagine the permanent record of reference in terms of a networked world of digital information, we need to set aside assumptions that follow from our well established system of print publication. If the essential *function* of scholarly publication is to serve as a permanent record of reference, we need to identify what features of *form* are essential to scholarly publication in any medium.

As a first essential feature, I would argue that the form of a scholarly published work

must be *identically replicable*. This follows from its function as a work of reference: scholars must be able to find identical content in identical copies of a publication. Of course, citation of archival material such as manuscripts inevitably leads any scholar to the same content, but the requirement that a scholar consult a single physical copy is perhaps the most salient distinction between archival material and published works.

Replication also directly serves as the best guarantor of the permanence of a work. If there is any lesson that is beyond dispute in the thousands of years of transmission of classical texts, it is that no single Library of Alexandria (or Library of Congress) can ensure the survival of a work: a fortunate copy anywhere can perpetuate a work beyond the lifespan of any single collection, however imposing. The more widely spread copies of a work are, the greater its chances of survival. If we hope for the preservation of classical Greek literature as far into the future as its creation is distant from us in the past, we should promote its replication, not just its archiving in a few massive repositories.

A second essential feature of scholarly publication is the *alienation* of the work from its author. Its form is fixed and outside the further control of the author once it is published: like evidence in a court case, a published work can no longer be withdrawn or altered once it has been accepted into the record. An author can of course change views or correct errors—but such changes can only enter the scholarly record through a further published work (whether a revised edition, or an altogether separate work). Our first two features together imply that a scholarly published work must have a form that can be irrevocably and identically replicated.

Our third essential feature of a record of reference is that it must be *citable in a fixed version*. Our system of print publication handles this requirement well: editions of a work are generally unambiguously identified, and reference to a specific section of a specific edition of a printed work leads any reader to identical content.

Putting these features together, we can summarize the *form of scholarly published works* as *works possessing an explicitly identified edition and explicitly identified citation scheme, that can be irrevocably and identically replicated.*

Digital technologies for scholarly publication

With these working definitions of both the purpose of scholarly publication and the form of scholarly published works, we can now ask what a published work might look like in digital media. The defining functional characteristics, permanence and reference, have both been elusive in the internet's digital environment. By focusing first on the formal features of scholarly published works, we can treat replicability, alienation and citation in fixed versions as requirements that our technological specifications must satisfy.

Replicability

In any analog medium, reproduction introduces some loss of quality from the original. By contrast, digital data can be replicated perfectly bit for bit, and at minimal cost. It is no surprise that the recording industry, which has relied in part on the qualitative superiority of professionally mastered musical recordings over bootleg copies, feels threatened by file-sharing users who can download from anywhere on the internet recordings that are identical to a track on a professionally mastered CD. If this astonishing characteristic of digital data threatens the recording industry's business model, it obviously presents great opportunities for scholarly publication with its requirement of perfect replication.

Any plan to replicate scholarly published works in digital form must address two important questions. First, a technical question: given the rapid evolution of information technology, how can we ensure that digital data will still be intelligible many years from now? At the beginning of the twenty-first century, our experience with maintaining long-lived

information systems covers only a few decades, but even that comparatively brief time is enough to provide the most basic guideline: to guarantee the accessibility of data into the indefinite future, we must insist on archiving data in openly specified standard formats. I use the term "standard" not in the common sense of "widely used," but in the technical sense of "specified by a standard." In the loose sense, "Word 95" might be called a "standard" format for word-processing documents because it is widely used. In the sense intended here, it is not a "standard," because there is no publicly accessible specification of the format. Presumably, Microsoft has such a specification, but someone who wants to verify that a document complies with the hypothesized specification has no way of doing so.

Any future programmer can translate data that is preserved in a standard format, because any programmer can be confident that the data are stored as the standard specifies, so when new standards inevitably replace older ones, there are no barriers to the migration of data stored in openly specified formats.

The second, more difficult question is essentially scholarly: what, in a digital environment, do we want to replicate? The ultimate motivation for the replicability of scholarly published works is to allow: 1) reliable peer review and 2) reassessment of published scholarly works. This adds a further dimension to the problem of designing replicable digital published works: peer review of a specific publication requires *reproduction* of the original; reassessment requires that the meaning of the original be *reusable* in new analyses or interpretations. In the design of inert print material, this distinction is not significant. Works are identically reproduced by printing ink on paper. The *only* way a printed work can be reused is for a scholar to handle one of these identically reproduced copies and to read ink on paper. Additional "reuse" only occurs as the scholar creates a new print publication that, like its predecessor, is a self-contained and isolated unit. If observations or analyses from an earlier publication are to be

incorporated into new scholarly work, they must be recreated anew and inserted into the new publication; if the earlier publication chooses to present complex data only in a summary, only the summary will be available for reuse.

This comes across clearly in a work I admire for its reassessment of earlier scholarship, Robin Osborne's *Greece in the Making, 1200-479 B.C.* [4] Osborne repeatedly pulls together data scattered across numerous published works, and identifies patterns that would not be apparent from any one of them individually. In discussing how population distribution changes in the eighth century B.C., for example, Osborne presents four maps, based on a variety of sources, summarizing settlement patterns in three different regions of the Greek world. [5] In this reuse, the earlier research contributes to a breadth of interpretation that is not possible from the individual earlier published works alone.

With an apparent rise in numbers of eighth-century settlements, Osborne then compares the rise in numbers of burials from the same period, and is caught by the limits of working from print sources. Two graphs [6] show two very different patterns: a graph based on Anthony Snodgrass's work shows a sudden, sharp rise in burials in both in Athens itself and in the surrounding countryside. The simplest interpretation would be that increasing numbers of graves reflect an increase in population. A graph based on Ian Morris's work, however, shows that the numbers of adult burials in Attica increase notably before the number of child burials increases. This suggests that the changing burial patterns are not simply due to population increase, but

[4] Published by Routledge: New York, 1996.

[5] Figures 16-18b on pages 71-76, with bibliographic note for Chapter four, on page 362.

[6] Figure 19, page 79.

must at least in part reflect changing burial practices, perhaps reflecting changes in social organization.

The difficulty for Osborne, or anyone else who wants to reassess the work of Snodgrass and Morris, is that we have no way to get behind the particular presentation of material that Snodgrass and Morris choose for their works published in print: we cannot see the underlying data. The perfect reproduction of print media stands in the way of effective reuse. In a digital environment, we might conceive of published works differently. As in print media, reviewers must be able to assess a specific version of a published work in its entirety: they must be certain that they are evaluating a reliably identical copy of that edition. But to remain with our example, if we could reproduce exactly the view that a Snodgrass or Morris chooses by algorithmically generating a graph from an equally accessible data set, then later scholars could reuse the data and subject them to new analyses. This is a very different kind of scholarly reassessment from reviewing a fixed edition, and it is central to the scholarly endeavor. Here, too, digital media offer possibilities that printed works can never approach, for digital works not only allow perfect replication, they also make it possible to automate selective, semantically meaningful access to material.

For static published works, the key to enabling reproduction and reuse simultaneously rests in the *semantic description* of a document; specification of surface appearance should be secondary and subordinated to this goal. The principle of semantic markup is by now well established in the scholarly world, and codified in standards such as the *Guidelines for Electronic Text Encoding and Interchange* developed by the Text Encoding Initiative, an international scholarly collaboration that since 1987 has worked on how best to mark up many standard categories of texts. [7]

For scholarly publication, we have to consider more than just producing identical, semantically explicit copies of static documents, however. While works published in print are inherently inert, digital published works may be interactive. Replication of these works implies that we must be able to replicate their *functionality*, not merely a sequence of bits. Moreover, we must be able to reproduce that functionality into an open-ended future.

As open standards provide long-term security for our data, openness and specification are crucial to ensuring the long-term viability of a published work's functionality. In any specific electronic published work, the final definition of its functionality is the source code. This must be openly available for inspection and reuse, and must be written in non-proprietary languages. Programs written either in languages with open formal specifications (like C, defined in the ANSI standard X3.159-1989 "Programming Language C", and ISO/IEC 9899:1999), or in openly implemented languages (like Perl, which is defined by a particular release version of the freely available perl source code) will be the most widely portable in the present and immediate future: programs can be written in these languages to run unchanged on all of today's widely-used operating systems (GNU/Linux, the Unix family of operating systems including Mac OS X, the various Microsoft operating systems), while programs in a proprietary language like Visual Basic cannot. Moreover, just as data in openly specified data formats can be confidently translated to future formats, openly available programs in these languages can be rewritten in the future if necessary, since the programmer has access to the program, and can be confident of the semantics of the language.

Just as our concern with reuse of static documents implies that we need to expose their semantics as part of our routine process of publication, reuse of dynamic works requires us to

[7] <http://www.tei-c.org>

expose the semantics of their operation. Free access to source code may be sufficient for *reproduction* of a work, but still make *reuse* difficult.

In designing digital scholarly published works, therefore, we should where possible encourage reuse by coding to openly specified protocols and standard interfaces. A very simple example can illustrate this. A scholarly project producing a work written as a Java servlet might choose to use the Apache project's freely available servlet container, Tomcat, as the engine to run the servlet. To reuse the servlet does not specifically require Tomcat, however: any servlet container that follows the Java Servlet specification [8] can be substituted. Modularizing functionality of the publication in this way means both that reviewers can test and isolate portions of the work, and that pieces can be reused for other purposes in other works.

Designing modular, standard interfaces for a digital published work requires a deep familiarity with the technologies being used, but is at base a scholarly problem. It will strongly influence how readily reusable the publication is, including how readily reusable it is for *purposes not foreseen by the original author*. In this respect, the design of the published work, independent of the research leading up to it, is more fundamentally a scholarly problem in a digital environment than in the print world. In some cases, scholars may find that basic functionality of their disciplines has not yet been abstracted in the form of protocols or standards. This is especially likely for scholars in the humanities, whose work may be less familiar to the commercial, government and academic institutions that tend to drive the development of information technology standards. In this case, scholars will have to address the issue by developing specified standards on their own. As the Text Encoding Initiative has defined semantic markup for many categories of static documents, scholarly collaboratives will need to

[8] <http://java.sun.com/products/servlet/reference/api/index.html>

define the protocols for the functional semantics of many categories of electronic published works.

To summarize, then, the implications of the principle of replication for the design of our scholarly digital publication, a digital publication must

- represent the semantics of static documents in openly specified formats
- implement the functionality of interactive works in freely available and reusable source code, written in openly defined languages
- write source code where possible to open standards and protocols with the goal of defining and modularizing the semantics of the work's interaction

Alienation

Designing digital publications so that other scholars can reuse and extend them is one of the most difficult challenges of the new medium. The principle of alienation of the work from the original author complements this: since authors cannot retire a published work, others can rely on the continued availability of a given edition or version when they choose to reuse pieces of it.

Beyond underscoring the importance of designing published works for reuse, the principle of alienation implies that digital published works must be made available under a license that permits all forms of reuse necessary to further continued scholarly work. For static data, a license like the Creative Commons Attribution-ShareAlike license requires that authors receive credit for their work, while allowing others to reuse it. [9] For source code, the classic General Public License ensures that "you have the freedom to distribute copies of free software

[9] You can read version 2.0 of the Creative Commons Attribution-ShareAlike license at <http://creativecommons.org/licenses/by-sa/2.0/>.

(and charge for this service if you wish), that you receive source code or can get it if you want it, that you can change the software or use pieces of it in new free programs; and that you know you can do these things." [10]

Some print publishers today are experimenting with various forms of subscription services offering access to their "published works." In some licenses, if an institution or individual does not maintain the subscription, access to the entire service is discontinued. While most of these services fall short of our definition of published works on many other grounds, licensing of this kind means that a scholar's access to the work can be withdrawn, and for this reason alone the work cannot serve the fundamental function of scholarly publication. Other licenses might allow access but place restrictions on the reuse of electronic material. This, too, is contrary to the principle of alienation. In a very real sense, publication makes a work the property of the scholarly community; authors and publishers alike can no longer exercise control over a scholarly work once it has been published to the community.

In the world of published works in print, we have become accustomed to accepting limitations on the principle of alienation, because they seem inevitable or necessary. We accept, for example, that a work might be unavailable to us because it is out of print, and there is no practical way to reprint it. This of course is another potentially enormous advantage for the scholarly world of publication in digital media: infallible replication at minimal cost should open our eyes to the fact that when a publisher announces that a work is "out of print," it represents a breakdown of our publication system: a work that is no longer replicable is not just "out of print," it is "out of publication."

[10] Quoted from version 2 of the GPL, see <http://www.gnu.org/licenses/gpl.txt>.

Citation in fixed version

The function of published works as a record of reference requires that the works have an explicitly defined citation scheme, and that they be published to the scholarly community in fixed versions.

"Publication in fixed versions" of course does not imply that a digital published work need be static: the possibility of dynamic interaction is the single most important difference between digital and print publications. But since we have already seen that replication of a digital work must include replicating its functionality, a "fixed version" is simply one that has a defined and replicable interaction. It is nothing more than a fixed version of a piece of software, and of course software developers have long since developed work routines and supporting tools to document and produce specific release versions of their work.

In the feverish experimentation with new forms of communication that the internet has made possible, this principle has sometimes been lost sight of, and on occasion people misleadingly refer to any digital resource that is accessible on the internet as "published." Fluid or non-fixed forms of communication can play a valuable role in scholarship, but that should not obscure the necessity of releasing *published works* in fixed versions: otherwise, replication, alienation and any hope of coherent reference must fall by the wayside.

Designing a reference scheme for a digital work is a challenging and important requirement. Since works may be replicated and read or made available at many physical locations, a scheme of physical addressing such as URLs is inadequate. Instead, some form of logical reference must be defined, that can in turn be translated into an appropriate physical address at the time the publication is read or used.

In the case of a static publication, the reference scheme might be very simply realized in

the structured markup of a text that viewing or reading software can interpret. To take a simple analogy, this could be comparable to inserting named anchors in an XHTML document. Interactive publications will require programs to transform logical references into physical addresses of some kind, but fortunately for scholars, this is a generic problem, and a variety of options are available. [11]

In designing a citation scheme, the publisher should aim for a scheme that corresponds as closely as possible to the semantics of the document; where a well designed interactive publication makes fundamental functionality available through standard protocols, the citation scheme should be capable of exposing this functionality to a direct, citable reference. (This will be particularly important when we turn our attention to the larger environment for digital published works: digital libraries.) In any case, a primary function of a digital published work is to permit identification and retrieval of sections of published material from citations, so the logical reference scheme must be supported by some retrieval mechanism in the digital published work.

Summary: digital publication

Our argument began with two essential qualities of scholarly publication, permanence and reference. We saw that these qualities require scholarly published works to be identically replicable, alienated from the original author, and citable in a fixed version. Taking these formal

[11] The Cocoon document management system enables a remapping of logical references to URLs through its "site map" file, for example, and web servers like Apache have long supported remapping of one (potentially logical) set of references to another. At a different level of functionality, Web applications packaged in a Web Applications Archive file include a "web application deployment descriptor" (the `web.xml` file) allowing the developer to map the logic of the application's classes to URLs.

features as requirements, we defined the following technical specifications for a digital scholarly published work. It must;

- represent the semantics of static documents in openly specified formats
- implement the functionality of interactive works in freely available and reusable source code, written in openly defined languages
- write source code where possible to open standards and protocols with the goal of defining and modularizing the semantics of the publication's interaction
- be available under a license that protects scholarly reuse, such as the Creative Commons Attribution-ShareAlike license for static content, and the GPL for source code
- be published as a clearly identified edition or release version
- include a logical reference scheme designed to function within the published work wherever it is installed or read
- match the reference scheme as closely as possible to the semantic organization of protocols
- implement retrieval, or function with external systems that allow retrieval, of sections of the publication from citations in the logical reference scheme

Defining scholarly libraries

A repository might comprise nothing more than a collection of published works, but a scholarly library is not just a repository. The function of a scholarly library is to provide access to a collection of published works in a setting that makes it possible to use them for scholarly work. A library providing access only to printed works will still include some kind of catalog for finding works, and will set aside space, such as a reading room with tables, where readers can bring together and work with multiple published works simultaneously. In brief, we define a

scholarly library as a *setting enabling scholarly use of a collection of publications*.

John Unsworth has suggested that for the humanities, at least, this use is founded on "scholarly primitives." [12] As a provisional list for consideration, Unsworth suggests that scholarly primitives include the actions of discovering, annotating, comparing, referring, sampling, illustrating, and representing. Unsworth intentionally leaves unspecified exactly what might constitute the direct object of all of these verbal expressions: it is after all central to his argument that these primitives cross disciplines, and therefore will be applied to the distinct subject matter of each.

The distinct subject matter of different disciplines will still be represented in the published record, however. An editor of a text presents a citable version of that text to the scholarly world, an archaeologist who publishes finds from an excavation creates a citable representation of those finds, a reviewer of a unique dramatic performance might create a citable representation of the performance. The meaning of these representations will vary widely from discipline to discipline, but their representation in the scholarly record suggests that we can further qualify Unsworth's idea: scholarly primitives, as they appear in published work, operate on other published material. Even if an archaeologist compares two newly discovered artifacts, the publication containing the comparison will have to present (and represent) those artifacts before comparing them, so it remains true that any reader of the publication can consider the comparison as working with two objects citable in published material.

[12] His contribution to a symposium on "Humanities Computing: formal methods, experimental practice" sponsored by King's College, London, May 13, 2000, and entitled "Scholarly Primitives: what methods do humanities researchers have in common, and how might our tools reflect this?" is available at <http://www.iath.virginia.edu/~jmu2m/Kings.5-00/primitives.html>

If we accept the suggestion that in scholarly publications, we can only compare, annotate, illustrate, etc., objects represented in the published record, this suggests that Unsworth's categories of discovering and referring are functionally and logically prior to his other "primitives." The specific functions the scholarly library must support are first and foremost the actions of identifying and citing published work.

Digital technologies for scholarly libraries

Given our previous specifications for digital published works, how should we design a library for their use? First, we can note that the digital library does not need to be a single physical location. The collection of digital published works it makes available for scholarly work might be distributed; many physical collections might even be treated as a single virtual library.

While it would be easy enough to imagine a distributed digital repository where we store digital works, the more difficult question remains: how should we enable scholarly use of these published works? How should a digital library support activities like John Unsworth's scholarly primitives of discovery, referring, or comparing?

The question is much simpler for a world of print publication: if the only means of using published works requires physical access by a human reader, then reading rooms with generous tables will solve some of the most basic problems. But we have emphasized in our design of digital works the importance of exposing their semantics for reuse by other software. To create an environment for working with digital published works, a digital library will have to enable digital works to build on previous digital published works by interacting with them through these interfaces.

John Unsworth's concept of "scholarly primitives" is helpful in thinking about this challenge, although I believe his analysis of how we should design software to address scholarly

primitives is misdirected. His valuable insight is that higher-order scholarly activities depend on prior primitives. Unsworth imagines implementing these primitives in software that an end-user would consult, such as a web page. A human intermediary is of course always necessary when working with print media, but the promise of a digital library is that these primitives can *also* be implemented as methods for digital publications to interact with each other.

If we try to imagine a broad architecture for permitting interactions among potentially distributed objects, we need look no further than the internet. The internet depends for its most basic functionality on protocols defining how software should transfer information from one location to another. The fundamental job of converting names to addresses in the internet's numeric format, for example, is carried out by a service, a program following a protocol for exchanging information between a client and a server. A "domain name service" fields queries about names like "chs.harvard.edu" and returns information like a numeric address of 207.188.245.132.

We can organize our digital library similarly as a network of services permitting interaction among objects using specified protocols. Services corresponding to Unsworth's primitives will be foundational for other, more complex operations. First and foremost, a digital library's services will have to include services permitting the discovery of published works and their citation schemes, and reference services enabling retrieval by the defined citation scheme. Other services will be built on top of this foundation. A "difference service" might take two references and apply an appropriate algorithm (specific to the type of objects being retrieved) for defining how they differ. The results of a difference service might in turn be used by higher-order functions, say a statistical analysis of results from a series of "diffs". Human users could of course inspect results at any stage of this pipeline, as they must do in print, but the distinctive power of the digital library lies in its potential for augmenting this intervention with

interaction among automated services.

As an initial specification for our scholarly digital library, we can say that it should:

- be organized as a suite of network services
- include services exposing the interfaces to important semantic components of publications
- at a minimum include services supporting discovery of and retrieval by reference to published works
- include any other services dictated by the nature of the particular collection treated by the library

Beyond this initial list, experience with digital libraries may suggest higher-order scholarly activities that need to be supported on top of the minimal primitives in the form of additional services. These services could be a mix of further generic services and discipline-specific services: since the library is after all a virtual construct, it would even be possible for different digital libraries to provide different kinds of scholarly use of an overlapping set of published works, provided that standards for discovery and reference are accepted.

Postscript: craft and engineering in hardware and software

The American Precision Museum in Windsor, Vermont, tells the story of the remarkable changes that overtook production of machinery in the early nineteenth century, most vividly illustrated by the rapid changes that took place in the production of firearms.

An eighteenth-century gunsmith might devote an entire month to producing a single rifle. This highly skilled craftsman would have to fashion the individual parts, from the wooden stock to the numerous metal pieces of the lock mechanism, and then, judging by eye, carefully work

Copyright 2004 Neel Smith and licensed under the Creative Commons Attribution-NoDerivs license. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nd/2.0/> or send a letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.

each piece until it fit precisely with the other individually shaped pieces. The resulting masterpiece of craftsmanship was a valuable and unique artifact—unfortunately for hunters or soldiers in the field, who could neither interchange parts among weapons, nor themselves repair the custom-crafted rifle.

The revolution in gun production depended on the use of machine-made parts. At the London Crystal Palace Exhibition of 1851, the Robbins and Lawrence Armory of Vermont could exhibit a series of rifles made of interchangeable parts. Each of the separate parts of the rifle was manufactured to a highly precise specification; at the Vermont factory, machine operators who needed only modest skills could manufacture the parts in quantity, and in some cases even operate two or more machines simultaneously. To the amazement of visitors to the London exhibition, the rifles could then be assembled from the perfectly interchangeable parts with only a screw driver. In like fashion, soldiers in the field needed no more than a screw driver to swap out damaged parts with parts from other guns. Interchangeability of precisely specified parts effected a revolution both in how guns were manufactured and how they could be practically used and maintained. Mass production did not eliminate the need for highly skilled labor, however: it organized it and distributed it differently. The entire range of the old gunsmith's craft had to be analyzed, each part precisely specified, and machines capable of producing each part to the required degree of precision had to be designed and built. The engineers who accomplished these impressive feats had a different set of skills from the old gunsmith, but developed and perfected a set of talents at least as rare as the gunsmith's. The product of the gunsmith's month of specialized labor could serve one hunter or soldier (at least, until the gun failed); the product of the nineteenth century inventor and machine builder could keep a factory running around the clock producing rifles.

This seems to me to offer a close analogy to our current situation in scholarly publishing.

The challenge of designing digital published works for digital libraries is precisely to break down the old tasks into constituent parts, specify them, and design machines (in this case, in software) to carry out those tasks. We could transform the hand-tooled publication process resulting in a unique piece of craft work into a system for producing published works from precisely specified, interchangeable parts. We will not need less scholarly expertise to accomplish this: but we will need to direct scholarly talent away from crafting unique objets d'art, and towards contributing to a system that can change both how we produce publications, and what kinds of secondary uses can be made of them. We will need no less imaginative creativity to accomplish this: to the imagination of the author, we will have to join the creativity of the programmer.

When I have presented some of the ideas in this paper to other humanists, one common response has been, "But I've never seen a digital published work that fits your definition." Often the implication is clearly that no digital work satisfying the definition offered here could possibly exist.

While I can only agree that I have not yet seen a digital published work in the humanities that meets all the criteria that I have proposed, the suggestion that it is impossible to create such works organized in digital libraries of the kind described here is disproved by the counterexample of a very conspicuous community: the global collaboration known as the free software movement. Every guideline we have suggested for the production of scholarly published works is represented by the best free software. Software is distributed in fixed versions, written to open specifications, with freely available source code distributed under licenses allowing full reuse. Projects can design collections of such published software that interact predictably—that is, they can create an environment for working with free software analogous to the role a library plays for published works.

Compare, for example, two "digital libraries" designed to create a working environment

using the Linux operating system, the Knoppix project and the Gentoo project. The Knoppix project [13] puts together an operating system and suite of programs on a single CD. Users can boot a computer from the CD and work with applications like Web browsers, office suites, and many other "published works." The Knoppix project creates a compact, portable working environment that a user can carry on a single CD.

The Gentoo project [14] corresponds to what I have earlier called a "virtual library." The physical repository is distributed at mirror sites across the internet; you can "check out" a work to use on your local computer using a logical reference that might or might not include specifically requesting a particular "edition" or version. You "check out" a published work using logical references; the Gentoo system converts these to physical locations at one of a series of mirror sites, and the material is downloaded to your computer. The Gentoo system tracks "cross references", or dependencies, among different published works in the library, and acquires new "editions" as needed or requested, so if you wish to use a new version of a Web browser that requires "updated editions" of any further "published works," the Gentoo environment sees to it that the necessary "cited works" are available. The Gentoo project allows you to create a highly customized working environment tailored to your interests, and as up to date as you like.

It should not be surprising that the free-software community has evolved a "publication" system that so closely satisfies the needs of the scholarly community, for the operation of this self-organized community may more closely model the scholarly ideal than any other community anywhere. Published contributions can be seen by anyone, and the resulting rigorous scrutiny of

[13] <http://www.knoppix.org/>

[14] <http://www.gentoo.org/>

the source code in free software projects is a major reason for the high quality of projects like the Apache web server, the most widely used web server on the internet. Status in this community can only be gained by submitting work that your peers value.

The work of the highly skilled (and often highly individualistic) programmers who contribute to this community could not be further from the routine of a machine operator in a nineteenth-century factory. They work with the gossamer fabric of ideas, rather than molten metal or wood, but the community's productivity has, like that of the nineteenth-century factory, experienced explosive growth as these software engineers have developed a carefully designed system of interlockable parts made to precise specifications. Engineering ideas in this way is as natural to a community sharing the scholarly world's values as it was to the engineers who transformed the building of hardware in the early nineteenth century.

Our digital incunabula are not yet recognizable as digital published works. Modern scholarship has seen the craft of a skilled manuscript copyist give way to the industrialized production of printed works. This mass production helped effect new uses of writing throughout society, and gave rise in turn to a new set of specialized crafts built around publishing in print. These crafts will now yield in importance to the a new set of crafts centered around the production of digital publications. It is our responsibility as scholars both to learn from this history, and to realize that there are already functioning examples today showing how we might begin to move beyond digital incunabula into an era of true digital publication for work in digital libraries.