2013

From Physical to Digital Textualiity: Loss and Gain in Literary Projects

Peter Shillingsburg
Peter L Shillingsburg, peter.shillingsburg@gmail.com

For The CEA Critic: Journal of the College English Association

Peter Shillingsburg
Loyola University Chicago

From Physical to Digital Textuality: Loss and Gain in Literary Projects

Not all texts were written, edited, and produced for the same purpose, nor are they all used in the same way; therefore, there is no set of rules that will apply to all digitizations of all texts. This essay focuses on so-called literary texts (poetry, drama, fiction, and familiar essays) treated as works of art rather than as cultural documents. The main reason to be so specific is that standards of accuracy and precision in uses of literary texts as art are different from the standards often in play for cultural documents or corpora of texts for linguistic analysis. Accuracy is good for all uses, but some users dispense with it more readily than others. As has now often been said, "Any text might do, but no two texts of a work will do the same thing." Hence, if what is to be done with a text depends on which text it is, on when and where it was created, and on who its authors, revisers, censors, or producers were, then there is no substitute for knowing precisely which text one is using and how it differs from other texts bearing the same title and purporting to be the same work. The standards for digitizing such texts are high; good enough isn't.

Furthermore, not all literary works studied today were written and published digitally--most in fact were written with the anticipation of print production, or, if long enough ago, for preservation and use as manuscripts. What happens to textuality when such texts are digitized? Was their original physical form a significant indicator of meaning, purpose, or status of the text? Will digitization destroy clues to the work's significance? What standard of accuracy and comprehensiveness matters? How should those who digitize texts address what is lost during the process?

Whether a digital textual project is a professional scholarly edition or a student project, makes no difference with regard to the facts with which one works nor to the ideals for which one strives. Whether on stone, clay, papyrus, vellum, parchment, or paper, the surviving material texts are what we have. They are the primary materials; they constitute the evidence upon which all subsequent uses depend. One cannot go behind these basic physical objects in search of the answer to the question: "Where did this come from?" The physical stuff is where, for us, it comes from; it forms the basis for answers to the question: "And what happened to it on its way to its present form?" The question is important because no two copies of a text are identical and their variances can or should affect how we formulate statements

like: "The text means X" or "It shows that the author thought Y" or "It reveals a delicate sensitivity that was senselessly trampled by the printer or the censor" or even "Printing techniques at the time did not allow for W". Such statements cannot be made without direct analysis of the documentary record of textual history for the work one is reading and studying.

Students trying to research literary works usually find that a semester is not long enough and that oceans or continents separate them from primary documents. These difficulties do not justify the notion that it is okay to do primary research on derived materials of questionable accuracy and provenance. One still wants to know, Where did that come from? What level of so-called scholarship is okay with the fact that available copies are derivative and simply don't bear the evidence that allows one to ask the question: "What happened to this text on its way to this form?" And just to be clear all digital surrogates for the originals are derivative; they are not “the real thing”. And yet, there are good reasons to create surrogates, but no excuse for not knowing the consequences of the decisions made about how to construct the surrogates.

Digital, virtual, representations, accessible in every library, including Podunk U itself, is desirable--so desirable, in fact, that naive enthusiasts with access to scanners and computers are everywhere offering textual goods at discount prices or for free. Who wants to look this gift horse in the mouth? Can we dispense with the questions: "Where did this come from and what happened to it on the way?" Most literary texts available on the Internet are unfit for serious scholarly use because they fail to say which source texts were used, where they came from, or how they were produced and/or failing to be properly proofread. Actually, from the point of view of anyone looking for accurate reliable digital surrogates for literary texts, the first error most digital enthusiasts make is to misconceive what a literary text is; the second is to misconceive what is lost in translation from physical to virtual form; and a third is to assume that every text of a work is more or less identical to every other and that differences are negligible. The gains in digitization or literary texts are obvious and energizing; the losses are often overlooked.

Not everyone gives the same answer to the questions, What is a literary work? or What is significant about any given form of a work? One can dispute whether it is important that not all texts of a work say the same thing. Does it matter that the words and punctuation in the copy one has in hand were changed by multiple persons in unrecorded ways after it left the author’s hand? If the person digitizing the text of a work assumes no one cares about some aspect of the original, then anyone who does care about it will find the “virtual” result unusable--or, worse, will be misled by it. I am assuming that "Yes, it matters; we

want to know the facts of the case." That being so makes digital textual projects both important and very hard to do well. It is possible (and I know from experience probable) that student textual projects conducted in a semester end up being more useful as events that taught valuable lessons than as products that are ready for prime time and reliable as surrogates for the physical originals. It isn't just that semesters are short and that original documents are far away; it is also that we do not have decent tools for creating digital projects--we have tools, yes, but not decent ones. And yet every project has to deal with the same kinds of facts and presumably aims for a goal worth reaching--they are not right till they are right. I don't think anyone starts a digital humanities textual editing project by saying "Oh, heck, a lick and a promise will do"; and one hopes that no one begins such a project believing that computers will do all the work and ensure the quality of the product. It may be true that a semester is not long enough and the standards are very high, and the results not terribly satisfactory in and of themselves. That is no reason to despair, no reason not to try. Failure attends every project worth undertaking; it is a source of lessons to be learned. Furthermore, if the project is well-thought out, the results might be useful down the road. That is because the processes of improving work begun but not finished remain open and available digitally in ways that were clumsy and forbidding in written and printed work. Finally, if a knowledgeable thoughtful transcriber / editor / archivist can identify a discrete part of a larger project to undertake and finish properly, then a piece of the whole is done and ready to join subsequent pieces as they are completed. The fear, however, is about a well-intentioned project begun by someone (an idealist?) who has not worked out what happens to textuality when a text is digitized. I contend that only well-informed, thoughtful, intelligent, *human* diligence can produce a digitized text worthy of further use.

So, what are the facts of textuality; what are the goals worth achieving; and what conditions and methods will lead to success?

First, and most obviously, a virtual collection of transcriptions without images of the material archive is not a virtual archive--is not a surrogate in any sense whatsoever. An "archive" relying entirely on transcriptions is stuck with the fact that transcribing is a process that produces a new edition. Transcriptions are useful and actually necessary, of course, but no textual scholar would be satisfied with relying on it alone, for there is no way, within such a representation, to verify anything. Transcriptions without images scream out, "Trust me." With an image, they say, "Trust and verify."

Alternatively, a collection of images without transcriptions has only one advantage over its physical originals: accessibility by way of the internet. That is not trivial, but it isn't enough, if for no other reason

than that the power of digital media is not fulfilled in images alone. Transcriptions are necessary because textual analysis requires them.

Furthermore, even an accurate and comprehensive virtual archive made of both images and transcriptions entails inevitable losses: images have no weight, depth, texture, or smell and, thus, fail to give a palpable sense of structures, substance, and production; and transcriptions cannot capture every significant aspect of originals, particularly when the originals are manuscripts. Nevertheless, digital archives and editions are definitely worth creating and can be reliable as a basis for most, if not all, scholarly engagement. Minimal essentials for images include high resolution, full, un-cropped images of the document (not just the text written or printed on the document) with standards for size and color. The loss of substance, weight, texture, thickness, or smell can be compensated only by description. In short, there is never a full virtual surrogate for the material archive--not that anyone worth listening to ever claimed that there was.

Misconceptions about transcriptions can also cripple a digital project. Transcriptions add nearly inexpressible value to the virtual archive: digital searching and analysis would alone make transcription worthwhile. Searching and text analysis do not require images, but, unfortunately and inevitably, all digital searching and text analysis is done on new editions created by transcription. Textual errors, regardless of who made them, will affect the results of searches and analysis; therefore, images are important as checks on what is found in the transcriptions; unfortunately there is no electronic solution to the fact that transcription error can prevent searches from finding everything. Fuzzy searches help but tip over when false matches start crowding out the discovery of misspelled search terms. Years of experience have led me to the conclusion that a reasonably good typist is always a better transcriber than is the very best OCR. Shudder.

How much do we care?

Error can be just noise. We correct it or ignore it just as we talk and listen over the noise of a passing train or when surrounded by cocktail chatter. Error can also be serendipitous discovery, leading us to new insights and allowing for repurposing of old texts. Columbus discovered America while looking for India. We embrace error not only when it is interesting and stimulating, but when it is embarrassing because detected error exposes our own mistakes and misconceptions. But error is error nonetheless, and undetected error accounts for many failures and self-deceptions. In fact, *undetected* error should be, to the extent possible to intelligent academic mankind, separated from *undetectable* error. We must live with undetectable error, having no choice, but why should we also live with detectable error which we

have simply lacked the industry to expose? Shall we leave it to others to crowd-source accuracy? If one's tolerance for noise and inaccuracy is high, error does not matter; but one could say about the toleration of detectable error what Samuel Johnson once wrote about ignorance, which "in other men may be censured as idleness, [but] in an academick it must be abhorred as treachery."[1]

When broken by error or misconceived standards, digital archives and editions fail. But what constitutes breakage? It is not enough to point out that images are not fully adequate surrogates for material documents. That isn't broken, it is just the nature of digital images. Nor is it enough to point out that a transcription is a new creation, not the original, so that one cannot truthfully say of a transcription of Tennyson's 1842 poem "Ulysses" that "this is the 1842 edition of 'Ulysses'", since "this" (the transcription) was created in 2012. That, too, is just an aspect of the nature of transcriptions. Breakage occurs when the image or the transcription fails to do what it can do (fails to fulfill the promise of the digital medium). And of course it fails if it claims to do what it does not actually do or if it gets the metadata wrong, crops the image to text only, cutting away documentary evidence, or is inaccurate--introducing errors or unannounced corrections. Either the virtual archive represents a material archive or it does not; scholars using the virtual archive as an archive are irritated by textual noise; it interferes with their wish to trust the accuracy of the content and of the descriptions of the sources. They check transcriptions against images.

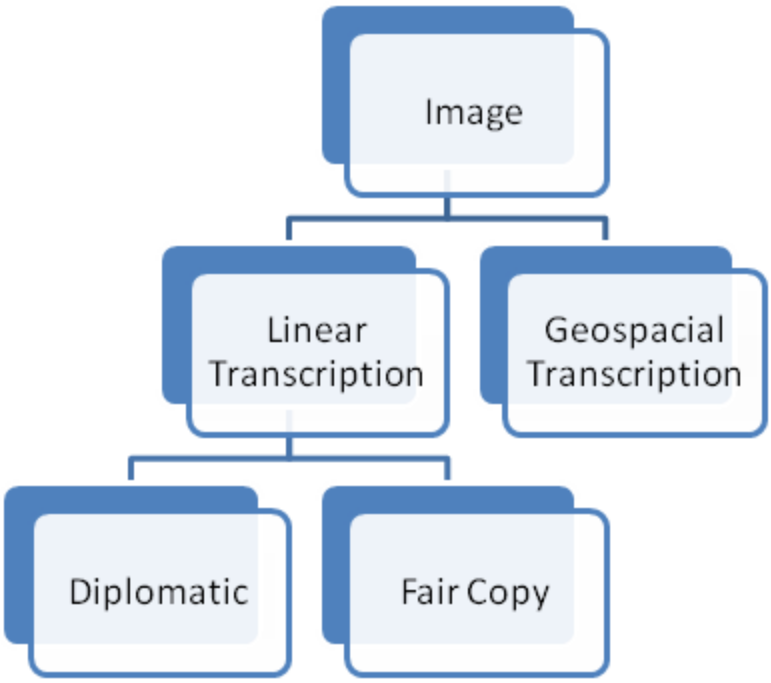
What do we want to see?

Digital transcription serves a number of purposes, each with its own demands, which makes transcription a complex thing, not the simple one it is sometimes taken to be. If the transcription is to represent what the images in the virtual archive "say" then perhaps the transcriptions should be (1) arranged in the same geospatial arrangement of the original with words between lines and in margins and all errors reproduced faithfully--thus representing what the original "looks like" and actually "said". But it might be useful in another way if it (2) represented an "intended linear order"as a diplomatic transcription--the order it would appear linearly to create the desired text. Then there is (3) the compositional order--the order in which the text was inscribed and revised, which is notoriously difficult to establish, though that seldom keeps scholars from positing a most likely order. And of course most readers would would also like (4) a fair copy--a final order without the record of composition and revision. Ideally, all four displays of text would be mined from a single database, so that when transcription errors are detected or when new materials or

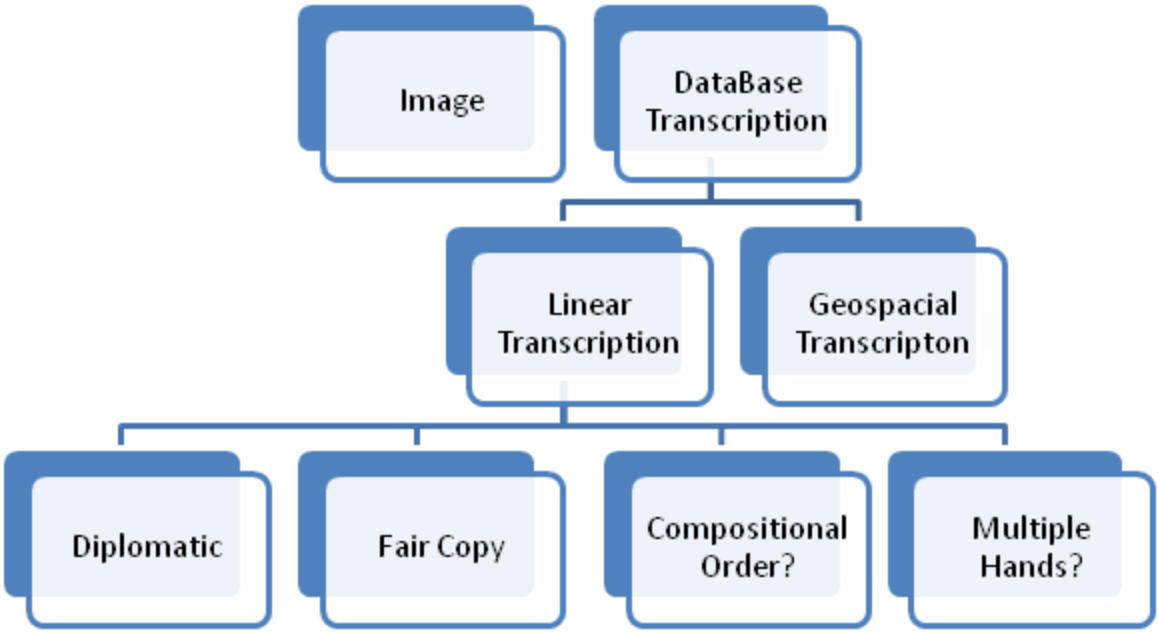
[1] "On the Character and Duty of an Academick" in David Fairer, "J. D. Fleeman: A Memoir," *Studies in Bibliography* 48 (1995), 24.

new display options are added, only one file or database need be revised.

All of these transcriptional orders might be complicated by the intervention of multiple hands. An adequate digital tool to handle this complexity has yet to be developed. And all these transcription goals fail to address a range of enhancement interpretations or commentary that a scholar/critic might wish to add.



Now, wouldn't it be good if instead of four transcriptions we had one database transcription that could be output in four ways:



Markup and Annotation / Text and Analysis

A long-standing dispute about mediated and unmediated data still is unresolved. At stake is the question of whether there can be such a thing as a digital surrogate for the "real" physical originals. Even images are manipulated in color, size, and granularity. But especially transcription--even "text only" transcriptions--involves interpretation (is it an i or e; is it underlined or crossed out; is the obscured letter a k or a t; should that upright have been crossed as a t or is it an l--were the bushes lopped or topped). And these questions about the text can be multiplied if one asks what is the meaning of underlining or italics (is it for emphasis or to indicate a foreign word, title of a book, or name of a ship). And so, it is asked, can the surrogate be unmediated, representing exactly the original, such that the user need not see the physical document? And if transcription is always interpretive, is all the interpretive analysis by transcribers of a piece? Is it futile to distinguish levels of intervention so that the decision about the e/i or t/l, the decision not to include crossed out words, the decision about the emphasis/ship's name, and the decision to add links to related documents just a continuum of editorial intervention from minimal to unlimited?

The question is important because of the potential of transcribed texts to be repurposed by other students and scholars. Several questions about repurposing reveal what is problematic about embedding codes in text transcriptions: If a transcription has been well proofread and marked up by a linguist interested in phonetics, marking the silent letters, or by a historian interested in names, or a literary scholar interested in

style and literary allusion--a well-proofed text, encoded for any one of these purposes--what must be done to that text before it can be used for another purpose. Can a stylist use text-analysis software on the file without first stripping it of code? Can a literary editor reuse that text to generate a list of textual variants between that text and another version of the work without first stripping them of code? Will scholars give access to their well-proofed file so that new code can be added by someone else? And if code has to be stripped out of the file before collating to detect variants, what happens to the textual information encoded because italics, indentation, diacritical marks, etc. have to be in code?

In short, when does the useful act of transcribing text (mediated or not) with useful encoding of one kind or another become unwieldy for repurposing? Is or should encoding affect transcriptions in a way that, realistically requires that a second and subsequent project on a given work must each create its own transcriptions? In absolute terms, no transcription (even backed with images) can fully stand in for an original--something is always missing or distorted. Scholars compensate by adding codes for a variety of things. Code tends to interfere with repurposing. What is to be done if transcriptions are to be 1) shared for collaborative development, 2) safe from inadvertent alteration while being further developed, 3) kept clean and simple for repurposing or use with software that requires text only, and 4) actually represent fully the textual record, including those things for which the qwerty keyboard is unable to record with one keystroke?

Consider the question, what markup is essential to represent the source document? and can such markup be distinguished from analytical or enhancement markup or annotation? Is there, for example, a difference between identifying a word as italic and identifying the meaning of the italics (to be emphasis rather than a foreign word or ship's name)? I think it is useful to have a sense of what constitutes the minimum required markup to achieve full representation of "text only". Markup is required to represent any aspect of the text for which there is not a single key on the keyboard. This includes special characters, italics and other font changes, and special uses of blank space. But if that minimal markup is mixed with additional analytical or enhancement markup and annotation, it becomes virtually impossible to repurpose the text along with its minimal textual markup. The question is not about whether to mark up or not mark up a text; the question is where to store the markup. Minimal markup (essential to make the transcription accurate and reusable) needs to be treated in a different way from analytical and explanatory markup.

Markup and annotation beyond the minimum need to be stored separate from minimal textual markup (in standoff mode, for example). There is a second important reason to keep embedded markup to a

minimum, for if all markup and annotation is embedded in the text, the text becomes vulnerable to inadvertent change every time it is opened to additional markup or annotation. Hence, embedding analytical markup and annotation is the enemy of collaboration; no scholar will allow his or her proofread transcription to be annotated or further marked up by others because they might inadvertently also introduce new textual errors. Standoff markup is a good solution for that. By embedding only visible textual markup (minimal markup) and relegating all other markup and annotation to standoff or other separate storage, the transcription can be more easily repurposed in separate projects and can serve as the foundation for multiple developers in a single content management system. Well-proofed text transcriptions stand, in that case, as the foundation for multiple standoff encoding of a variety of enhancements by a variety of people. The text remains untouched. Where is the software that will allow students and scholars other than digital experts to created such transcriptions?

Minimal markup

For clarity, anything that is added to a text because a keyboard cannot accomplish the representation with one keystroke: accented letters, font changes, strikeouts, interlineations, marginal notes, changes of hand or ink, underlining, superscripts, subscripts, centering, in-text images, and a few featured which, though capable of being represented by a single keystroke, usually are not: indentation, line-breaks, and hard-end-line-hyphens. If one leaves any of these things out of the transcription and then runs collation software on two or more such texts to reveal textual differences and the result is that the list of textual difference leaves out what was not coded into the transcription. By contrast, code not only the minimal textual elements but all kinds of other analysis of function and meaning, then run collation software and the result is unusably complicated by non-textual coding (for this purpose, garbage). Strip all the coding out and run collation, and the result lack any indication of italics or special characters. This probably incomplete list focuses on the visible textual features that are or can be semantically meaningful in physical texts. The point is that they describe only the signifying visual features of the source document; they do not describe the meaning or function implied by these features.

When XML was created (and TEI followed suit) someone decided that one hierarchical structure would suit all users--a generic structure of works with headings, sections, paragraphs, sentences, and lists. It was decided that what the text looked like (how long the lines were, where the page-breaks came, how documents were printed and bound physically, how white space was deployed) was less important. That decision entailed the belief that transcriptions should be informative about functions and meanings, not

about appearance. It assumed that format was to be flexible, not representative. That fundamental decision meant that no XML (or TEI) transcription would ever serve as a digital surrogate for an original document. It also meant that if anyone did want to included codes to indicate appearance, it would just be additional code, embedded along with function and generic structure codes. The result is texts that cannot be repurposed and files that are jealously protected by their creators. Generous people share copies, not the master file, which at best leads to multiple projects on the same work that must be consulted separately.

One defense for marking generic and function features rather than visual ones is that digital texts are malleable, capable of assuming many formats. It is said that marking the visual elements of a text assumes that it should always take that visual form. Marking generic features, instead, assumes that the text can take any form a user wishes for it. That sounds as though such markup makes a text more adaptable, but in fact, when visual features with semantic force are eliminated, and function markup is embedded in a text, the transcription becomes dedicated to a single use--the one chosen by the transcriber. Flexibility of format is purchased at the price of inflexibility of purpose for the text. Another way to see this is to ask if the transcription is to serve in any way as a representation of a historical text? or is it only to serve future purposes. If the latter, then generic coding is more important, but something significant will be lost. If the transcription is to represent its source, as, in an archive, it must, it must represent what the source looks like, not what the transcriber thinks it means. That is extraordinarily difficult to do as a few examples will show. Compromises are going to be necessary, but embedding generic and functional interpretations in the text is not one to be tolerated any longer.

Manuscript and print provide very different kinds of transcription problems that transcribers could identify in a variety of ways. Peter Robinson asks, for example, when is an 'i' and 'i'? In typography we have a roman 'i', an italic and a bold 'i', but generally speaking an 'i' is an 'i'; Robinson identified in medieval manuscripts a host of 'i's with different meanings or implications.[2] In an essay on punctuation and the multiple meanings of a capital letter, John Lawler writes:

> "(The convention of starting the first word of a sentence with a capital letter may be considered part of punctuation, spelling, diacritic marking, or even grammar, depending on how one defines each term; in any event letter case distinctions are also of recent origin in European writing, dating, like spelling standardization, from the widespread establishment of printing.)"

[2] "What text really is not, and why editors have to learn to swim," *Literary and Linguist Computing* (2009) 24(1): 41-52.

("Punctuation," John Lawler, University of Michigan, From: *The International Encyclopedia of Language and Linguistics*, 2nd edition, Elsevier, 2006 (http://www-personal.umich.edu/~jlawler/IELL-Punctuation.pdf (6 May 2013)

One could go on because the varieties of ways to mark up the meanings of visual elements is vast. distinguish markup of meaning from markup of appearance. One is text; the other is analytical, interpretive enhancement. When a visual mark has multiple meanings, which of them should the scholarly editor encode? should all of them be encoded? if so, what happens to our concept of text that can be shared and repurposed?

No need to rehearse more examples because it is already clear that transcription at its most basic visual (supposedly uninterpreted) level is mediated, or subject to mediation. And yet it is so at a level that falls short of most interpretive, enhancement markup relating to generic structures, bibliographical structures, indications of supposed functions, and explanations of any kind.

Undertaking a digital project, even if only for a semester, should start with a consideration of all of its potential parts (images, transcriptions, and analysis). Before determining a plan of action, ask what happens, beyond your immediate use, to the products of your efforts. How big a bite is one going to take at any one time out of the very large and diverse project that a digital project can be? To what level of accuracy and adequacy will the project be brought? Then, do not be daunted and back off; just do it.