Digital Scriptorium: Ten Years Young, and Working on Survival

Consuelo W. Dutschke

The editing of texts has always demanded a comprehensive array of hands-on skills, minute, precise, unforgiving of lapses; it has never been enough to fully know the author and to fully comprehend the text (even if those two goals are possible in reality). The editor must also understand and work with the mechanisms that have ensured the text's survival until today, and that encourage continued work on that text. Digitization is one such mechanism that offers a new tool to the editor. The topic of my paper is focused on one digital project, Digital Scriptorium[1], and in particular addresses a number of questions surrounding its life in the coming years.

Texts, whose survival up to the era of print depended upon transmission in the form of manuscripts, and for which there is no one authorial copy, present additional complications to their editors. The first step, necessarily, to such an editor is to assemble the body of evidence from which he will work; that first effort is one to which Digital Scriptorium contributes in more than the obvious manner. DS (as we will refer to it in the course of this paper) may be of service to the future editor, in that it assembles in a single point the indexed holdings of multiple libraries; even by virtue of a library's decision to participate in DS, that library will often have made a more concerted drive to identify its manuscript texts than it might have without the impetus of participation in DS. Thus, the potential for any one text to come to the surface is that much greater, as is the ease in discovering that one text.

DS also serves the cause of the editor in allowing him a first glimpse of the world that a given manuscript occupies: the other texts with which it circulates; the miniatures, if any, which always imply interpretation; the level of expense that went into its production; early and late owners with their notes and their bindings, each bringing a historical glimpse of that manuscript's value – both semantic and financial – to the whole. Leonard Boyle reminds us that no text exists without its physical means of transmission[2], and DS significantly aids the editor in building an understanding of the physical and intellectual environment of the chosen text. The editor's understanding grows intensively in each single instantiation of the text, in viewing the resultant collaboration of the team that physically constructed the one manuscript (its scribe, its artist, the person who did the penflourishing, and the more shadowy entrepreneur, the person who organized the team, who may have been the book's first owner, or who may have been the scribe, or yet again someone who left no visible traces in the book). The editor's understanding grows horizontally, as he examines other contemporary codices of the same text, and even other codices of different texts but of similar place and date of origin. And of course the editor's understanding grows vertically, as he follows the characteristics of one

collection or another, or as he perceives the pattern of later reception of the text, via datable reader notes in the margins.

DS accomplishes these tasks through its mission as an online image database of medieval and renaissance manuscripts, combining into a single resource the scattered holdings of libraries across the country. At the present writing, approximately thirty libraries participate in the consortium, offering jointly some 5,000 records and some 25,000 images (costs of digitization have usually precluded full reproduction of each manuscript; the average number of images per codex is six). Libraries in the consortium are private research foundations (such as the Huntington Library in southern California), major academic research libraries (such as the University of California, Berkeley, Harvard University and Columbia University), libraries of liberal arts colleges (Oberlin), libraries of religious instititutions (such as the Jewish Theological Seminary of America and Conception Abbey), as well as large and small public libraries (the Free Library of Philadelphia and the Public Library of Providence, Rhode Island). There has also been a concerted effort to connect the geographical spread of libraries, from Fordham University on the East Coast, to San Francisco State University on the West Coast, with Notre Dame and the Universities of Kansas and Missouri in the middle.

As all researchers know, what matters is that the library happen to have the one text, the one author, the one scribe, the one artist on which the research focusses; the total size of that library's holdings becomes irrelevant. Thus, the University of California, Davis has a rare life of Petrarch, composed by his contemporary, Sicco Polenton (it is the only copy in the United States). The University of Kansas holds a *laudario* of Italian religious poetry, with some poems of praise in unique copy. Which is not to say that major libraries don't also hold materials equally of interest to Italian literature (chosen as examples given the present venue): at the Huntington Library is an autograph manuscript of one of Tasso's discourses; at the Houghton Library one can read a *cantare* by an almost unknown author, Teo da

Perugia. All of these texts, or rather images of the codices that carry these texts, are retrievable through and represented by DS.[3]

To some extent, everything said up to this point is known in principle or in fact: that the challenge of locating all copies of the editor's chosen text can be arduous in the extreme; that editors of texts must tabulate and position their sources in an initial survey; that images, even select images of every text, are a significant boon, allowing the editor to form his own judgments as to place and date of origin of the manuscript (and occasionally even to form an initial determination of recension, since the recensions often circulate with set rubrics).

The remainder of this paper will address the topic of its title: survival of DS a digital project. Digital Scriptorium began its life ten years ago, with a joint grant proposal from Berkeley and Columbia to the Andrew W. Mellon Foundation, written by Prof. Charles Faulhaber, Director of the Bancroft Library at Berkeley. The proposal was initially drafted over the Christmas holidays in 1996, and became a reality in 1997. In retrospect, two decisions of those early days may be partially credited with the success of DS today: establishment of standards for data collection and for photographic capture; documentation of those standards.

Because Berkeley and Columbia are 3000 miles apart, and suffer from three hours' time difference, and yet were committed to share this deadline-driven project, they came to working agreements very quickly. The agreements were modified, reconfigured, massaged, and then accepted in an essentially final form in an invitational meeting in November 1997.[4] Thereafter, once there was a consensus on the intellectual content of the database, work intensified on compilation of the accompanying documentation. There was some time lag in compiling equivalent documentation on the photographic standards, but by the end of the initial grant period, this, too, was in place. To this day standards and documentation of the standards are iteratively updated, while the core remains constant. To this day, the documentation is

been posted publicly on the DS website.

Existence of bibliographic and photographic standards, with relevant documentation has enabled other institutions to examine DS as a project, and make appropriate decisions on joining the consortium. I will cite an Italian proverb, since we're meeting at the Italian Academy: "Patti chiari, amici cari." The rules are clear; there are no surprises; an institution understands what it is committing to, and what it can expect as a result. During production phases, as a new DS partner accomplishes its data entry and its photography, it continuously refers back to the documentation to resolve uncertainties, with the result that the final version of the new partner's data and images are on a par with those of the other contributors. In a consortial project, such as DS has been since its inception, parity is the sine qua non. Had the unlucky chance been that the two original partners lived next door to each other, the demand for documentation might never have been so strong. In retrospect, it seems certain that the consequent weakness in documentation would certainly have been deleterious, and possibly fatal.

Indeed, the standards and their documentation have become crucial to DS in another way, in that they play a role towards ensuring the technical sustainability of the program. "Sustainability" is the one word in digital projects that stands for a very large array of issues, some of them as yet only twinkles in the eye, and some of them ignored elephants in the room. Fundamentally, the word raises the question of finances: how will we find the money to keep this digital program alive in two years, in five years, in fifteen years?

This inexorably looming question implies several slightly more answerable questions: what does it cost to keep DS alive over the coming years? are there steps to take now that will pare down future costs? In terms of technology, one of the major cost areas, DS can and has accomplished certain actions that will pay off, very concretely, in years to come.

The first is documentation. As pointed out above, documentation is what holds the system-wide but independently implemented standards on a sufficiently level ground to minimize costs of copy-editing data and of reshooting images. Human ingenuity being what it is, there is always room for misinterpretation of the standards, no matter how well they are documented, but misinterpretation is contained within acceptable deviations from the norm.

A banal, even silly example of an acceptable deviation (now in fact corrected) occurred when one DS partner input "yes" in a database field labelled "Music." The field, as explained in the documentation, is intended to contain a description of the form of musical notation, as in "Black square notation on 4-line red staves." While "yes" is not a description of notation, it can still alert researchers to the presence of notation in that manuscript, and with the images of the manuscript, the researcher himself will determine the format of the notation.

A second means of constructing today's technology in such a way as to limit tomorrow's expenses lies in the transparency of the technology itself. The simpler that it is for a computer analyst to open up a database and identify problem areas or areas that require updating, the less it will cost to hire that computer analyst. Again, an example makes the point: a previous version of the DS database had named one of the program's crucial tables, "tblLink3"; today that table is called "tblTexts" and even the inexpert among us begin to fathom its function. The database configured for DS use currently has 23 tables and 43 queries that support 37 forms and subforms, as well as 16 reports; a limpid naming structure cannot but help future manipulation of the

database.

File-naming of images is another area where simplicity should reign, and when it does, costs all along the line are reduced. There is an unfortunate tendency to want to make the file name carry verbal meaning, to allow humans to understand how it refers to the real-life object. An image of a fifteenth century leaf from a calendar of a book of hours might be (and has been) given the unwieldly monniker:

"1_1A-Calendar_Leaf_December_1460_BACK"

and yet cataloguing information on this leaf occurs elsewhere, more fully and more appropriately, in the tables of the database. The purpose of a relational database is to tie categories of information together when they need such connections (such as between the bibliographic description of an item and the photographic depiction of the item), and to refrain from duplicating information. In the present example, the cataloguing information is duplicated to no advantage, and to possible disadvantage, should scholars later determine that the leaf in fact dates from the mid fourteenth century, or represents the month of April, or that it isn't the verso of the leaf, but instead its recto. Thus, simplicity, if applied now will trim costs associated with updating of semantic values that are no longer correct, and will help avoid the costs of mis-typed file names (with spaces, underscores that disappear from human view in an underlined URL, slashes in either direction, mixtures of weighted upper and lower case letters). DS recognizes that unreliable data at any level is an expense that is better controlled at its source, and not in mopping-up operations.

The database referred to is the mechanism employed for gathering information, and for limited-time storage. Microsoft Access was chosen because of the widespread familiarity with Microsoft products and use; in addition, there was the conviction that the company would continue to support the product, as has been the case. Periodically, each DS partner exports the information specific to its own collections to XML, and forwards

the XML-encoded data to the central organization. It is on the XML-encoded data that technology experts write the applications that make the data useful to scholars, via meshing the data from multiple partners, searching it, retrieving it, displaying it. Because XML-encoded data is non-proprietary and platform-independent, it will enjoy a longer and more productive life; thus, the DS decision to employ XML as its data-transport, long-term storage and manipulation system projects lower costs in this sector in the future.

Interoperability, however, is a larger issue than the interconnections between institutions and systems adumbrated above. In some not too distant future will DS technology crosswalk easily not just between its own database (however many instances of that database come into existence) and its own XML schema, but also back and forth from the library world's ubiquitous MARC format, and from the archivists' usual EAD? Lack of immediate compatibility derives from absence of shared standards, but if the areas of no-match are clearly delineated, it may become more efficient (and thus cost-wise more feasible) to work towards wider interoperability.

Mass storage is another technology-based issue that has yet to receive sufficient attention. It is a commonplace that costs of data storage have decreased drammatically since DS began ten years ago; but are we facing the proportionately larger increase in numbers and sizes of digital files? Size is a factor; method of storage is a factor; even the fundamental question of value of the digital file is a factor. Because this paper intends to examine DS with regard to sustainability, and because DS has not fully addressed this issue, I will only say that duplication of storage (i.e. on the home institution's servers and on DS servers located centrally) provides a certain security to the files. Security of the files has a direct financial impact on DS, whether we are considering the costs of the security itself or the costs of loosing the files due to poor security.

The preceding comments have examined the role that a viable technology plays in the sustainability of a digital program. The most important question

in addressing sustainability, however, is political: do the DS partners want DS to live onwards? does the larger community see a value in DS? These seem self-evident or even self-serving questions. Clearly when Berkeley and Columbia began DS in 1996/97, the two institutions thought that DS was a good idea; clearly the number of new institutions asking to join DS think it's a good idea. Again, there was something of a stroke of genius at the inception of DS: libraries are in the business of shared and disseminated information, so that the goals that libraries have always striven towards matched the goals of the new digital project. Implementation of the goals was, of course, very different, and implications of the delivery system reverberates on the scholarly community in a very different way. Nontheless, DS did not have to combat a deeply ingrained reward system such as that faced by individual faculty members who imagine and attempt to build bright new digital projects.

In a concrete move towards stabilization of the goals of the DS consortium, DS is in the process of forming a governing body with mechanisms for developing policies and strategies and for the daily management of the group's decisions. The structure was built to respond to the shared needs of the consortial group, and to recognize the particular needs of the group's management host (the institution that handles DS finances) and the group's technology host (the institution that keeps DS alive on the web); the structure must ensure continuity, responsibility and flexibility at the same time.

DS remains conscious, however, that its users ultimately make the choice as to DS's survival. In an effort to identify the demographics of DS users, and to assay their opinions of DS, we posted a survey on the DS website in early spring of 2007, and publicized it via printed newsletters, listservs, classroom presentations and direct emailing to selected medievalists. Some 200 people responded; of these 43 offered name and email address for further discussion; criticism took the form of imaginative suggestions for improving technology, and an overwhelming demand for ever more content. This paper

isn't the appropriate forum in which to analyse the results of the survey; the point is that without a solid present and even prescient understanding of user demands a digital project may not succeed in sustaining itself. The will to sustainability lies not only within the project and its creators/partners; it also lies with its users.

Last in this examination of DS sustainability is what might have seemed the first question: what does it cost to keep DS running? Here, too, however, the question multiplies under our microscope: are we tabulating costs for stasis? for moderate growth? for significant growth? And how many years outwards can we predict costs in each of these progressively larger areas? Actual answers aren't relevant to the purpose of this paper; what matters is the fact that DS is attempting to form answers. What also matters is that DS is asking the questions within the framework of a group of institutions, all of which have a stake in the outcome. "In the late twentieth and early twenty-first centuries, the most significant impact of information technology may be increased collaboration," as Daniel Pitti has pointed out.[5]

Why are we asking about sustainability of a digital project in the first place? It's not simply that digital projects cost money; all human endeavour falls into that category. It's that digital projects remain so new to us that we, as a nation and even as a world-wide community of scholars working in the humanities, haven't fully understood the costs nor factored them out across appropriate bodies. The steps DS has taken towards a more reliable ad efficient technology, and the steps it has not taken reflect growth and uncertainty in the field overall. DS and the digital world as a community still lack a cyberinfrastructure not simply in terms of hardware or software, but even more importantly as a shared and recognized expertise and mode of operation. Definition of a cyberinfrastructure and recommendations on building it are laid out in the 2006 report to the American Council of Learned Societies that was crafted by an impressive committee of thinking people, and financed by the Andrew W. Mellon Foundation.[6] The report is addressed more to the signficant funders of digital projects than to the

managers of projects. Nonetheless, its five proposed goals towards an effective cyberinfrastructure outline areas of active concern to DS, to wit: that such a cyberinfrastructure should:

- be accessible as a public good
- be sustainable
- provide interoperability
- facilitate collaboration
- support experimentation

DS, in its striving, tills these same fields, and plans to grow a good crop.

Note

[1] Digital Scriptorium is a consortial program, uniting at present (but continuing to grow) the medieval and renaissance holdings of multiple libraries into a single tool; it relies on minimal cataloguing and sample images from all manuscripts held by an institution to effect timely deployment of the material on the web. It is available at: http://www.scriptorium.columbia.edu

[2] He explains the point in his article, *Epistulae venerunt parum dulces: the Place of Codicology in the Editing of Medieval Latin Texts*, in Richard Landon (ed.): *Editing and Editors: A Retrospect. Papers given at the twenty-first annual Conference on Editorial Problems, University of Toronto 1-2 November 1985*, NewYork 1988, 29-46; Father Boyle concludes, «[Codicology] is a simple and necessary recognition of the fact that texts have survived because of codices, and that each codex in turn carries a text in its own unique fashion».

[3] Because I am personally acquainted with the scholar who discovered and published these finds, and thus because it is relatively easy for me to cite the call numbers and the relevant bibliography here, I will do so, in alphabetical order according to the institution: DAVIS, Shields Library, UCD D-041:23: D.

Dutschke, *Census of Petrarch Manuscripts in the United States, Supplement I*, in: *Petrarca, Verona e l'Europa*, Padova, Antenore, 1997, 457-465, esp. pp. 461-463. KANSAS, Lawrence Spencer Research Library, MS D113: D. Dutschke, S. Kelly, *Un ritrovato laudario aretino*, «Italianistica», 14 (1985), 155-183; D. Dutschke, *The Translation of St. Antony from the Egyptian Desert to the Italian City*, «Aevum», 68 (1994), 499-549. HARVARD, Houghton Library, Ms. Riant 87: D. Dutschke, *The Classical World in* La caccia *by Teo da Perugia*, in: *Vestigia: Studi in onore di Giuseppe Billanovich*, Roma, Edizioni di Storia e Letteratura, 1984, 221-245. HUNTINGTON, HM 884: D. Dutschke, *Il discorso tassiano 'De la virtù feminile e donnesca'*, «Studi tassiani», 32 (1984), 5-28.

[4] The meeting was sponsored by another Mellon-funded project, the Electronic Access to Medieval Manuscripts; EAMMS was directed by the Hill Museum and Manuscript Library (its current name) at St. John's University in Collegeville MN, and in the end produced both a MARC format application and a XML schema for electronic cataloguing of medieval manuscripts, the former called AMREMM, and the latter in use by Digital Scriptorium when the data originates as encoded prose (rather than as a database, for which see below).

[5] D. Pitti, *Designing Sustainable Projects and Publications*, in: S. Schreibman, R. Siemens, J. Unsworth: *A Companion to Digital Humanities*, Blackwell Companions to Literature and Culture v. 26, Oxford, Blackwell, 2004, 471-487, this statement on p. 485.

[6] M. Welshons (ed.), *Our Cultural Commonwealth, The report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences*, American Council of Learned Societies, 2006.