Posts from LIS 539 – Metadata and digital preservation

Gabriel Chrisman

I'm going to stick my neck out here to defend and support the life-cycle view of records and archives rather than the continuum view. I also have reservations about and disagreements with post-custodialism. My objections break down into two major points:

After listening to the lecture and readings the articles provided, I still feel that there is a distinct and useful discontinuity that occurs when records cease to be in current-use. Before this point, the vast majority of the use of the record (occasional and irregular public use or access excepted) will be by members of the originating agency or closely allied organizational units. The organization of these records can and should be streamlined to optimized to maximize the efficiency of such use – and not public or researcher access. After the record ceases to be current, the vast majority of use will come from exactly these outside groups, and the occasional or irregular users will be those from the originating agency. This distinction was recognized in a quote from Gerald Brown included in Atherton's article, though it is couched in these terms: "the records manager is basically a business administrator and the archivist is basically a historian." This identifies the different roles played by the two groups of record-keepers, but the distinction in the community served is equally clear and equally important. Atherton disagrees, because he sees considerable overlap between the archvists' and records managers' duties and potential efficiency increases from combining their domains. Despite the obvious potential for increases in efficiency by collapsing these positions, I still would maintain that the distinction in the user groups is not one which can or should be ignored – precisely because the organization of the records can and should be optimized for the majority of the users in each case. Government organizations shouldn't have to invest large amounts of time and money to set up records-keeping regimes which prioritize public access when this represents a tiny fraction of the users of these records, and archives should be able to add to or modify the organization of non-current records to allow for easier use and access by researchers less familiar with internal organizational systems used by the originating agency. Attempts to generate organizational systems which would serve both groups equally well (which I think would be tremendously difficult, given their radically different purposes and experience) would still leave records professionals with little flexibility. In order to have the freedom to serve different user groups and respond appropriately to their expectations, it makes sense to separate the functions and maintain (or create, if it is not clear enough as in the case of digital records) this discontinuity in records. Even if the methods they choose are similar, there are other reasons why it is preferable to separate the functions of record management and archives.

Which leads me into my next point: not only do records managers and archivists have different user groups they serve, they also have accordingly different loyalties. This forms the basis of my objection to post-custodialism as well. I feel that records managers, in their appropriate role as being focused on optimization of records within an organization or department, will feel strong loyalties to that unit or department, and that department must trust them completely. Asking them to split their loyalties and focus equally on serving the public or posterity seems misguided and damaging to their role, where archivists would equally be crippled or limited by trying to

both serve the public effectively as well as the departments which supply them with records. In last quarter's archives class, Scott Cline (the Seattle city archivist) described vividly the split loyalties he feels and the suspicion he has to deal with by being an city archivist who frequently provides information to both city agencies and people who are prosecuting lawsuits against the city – imagine if he also was trying to suggest to these same agencies which records to dispose of when, or how they could be more efficient? I don't imagine this would work out well in terms of trust or efficiency. Along the same lines, it makes much more sense from an accountability standpoint to have an organization external to the originating organization which holds the archival or long-term records (again, as distinct from the current-use records – a distinction which should be made). Speaking as a historian, I can say that I would place much more trust in records (especially digital, more easily modified records) that were offered up by an impartial or at least external source when compared to those which were held by the originating agency itself. As easy or appropriate it might seem to leave such records in the hands of those who created them, it seems to me to require a leap of faith on the part of the public and represents a potential loss of accountability. I know that Upward tried to minimize this loss in our reading, but it still seems a very real problem to me.

There are other points made in the readings which I would take issue with, but this is already way too long...what does everyone else think?

I agree with Kylie that more cooperation (or at least communication) between records managers and archivists would be valuable, at least in terms of cooperation to create such useful things as metadata standards and so forth. I'm still a little bit leary of 'trying to make sure organizations create the right records' in the first place. I guess as much as I can understand this in a certain way, as such standards are necessary to fully explicate what is meant by 'accountability', it seems this should rightly be the domain of the lawmakers rather than the archivists/records managers. In other terms that have less to do with accountability, I believe that this would be the province solely of the records managers, as these other reasons seem like they would be based on increasing efficiency for the organization (which to me has little to do with eventual archival use). As far as archivists are concerned, trying to control the types of records which are produced seems to be straight out of documentation strategy theory, and I have my doubts about the efficacy of this in practice. I think of archivists in more of a Jenkinsonian role, as neutral preservers of what is left to them by other organizations. I know, I know, this is outdated and archivists really have a much more active role - but I still see this as the primary or ideal model.

Dana, you are totally right to call me out on my views of government and other bureaucratic organizations - I guess I do view these organizations' activities as generally non-cultural, or at least totally different and separate from other aspects of culture (by which I guess I mean the sum of fundamentally individual actions, such as academic research). I'm not sure whether its because I just don't want to 'own' our government and acknowledge its influence, or because I am just such a certified individualist that I simply don't want to deal with organizational 'culture'...well, obviously I should think that one out further. My primary point in this context, though, is that its (government's) needs and methods for access are fundamentally different from those of non-

governmental researchers. Is this valid? Whatever my motives or reasons for seeing this distinction, I still think that the two sets of users (government or other organizations, and academic or private researchers) have different needs. These different needs can be best served by different professional roles and different records organizations.

Dana, I also like your point about the continuum being a model for understanding records as opposed to a programmatic system for dealing with them. You're right, and this is probably part of my objection to it - while it implies a great deal, it says very little straight out about how to deal with these implications. What you see as flexibility, I see as vagueness. I like the comparitively programmatic simplicity of the life-cycle model, because it forces managers of records to make some binary decisions that I believe need to be made, and help to create clarity where vague situations exist. I see this clarification as one of the duties of information professionals: people (hopefully) will pay us to make decisions and create clarity.

From a historical perspective not based in records management, and from my own experience researching historical topics in archives in widely varied media environments, I can say that I have made frequent reference to aspects of records which have only survived because the original artifacts were preserved. Acid stains in books from items which were left in place, bookmarks in financial ledgers, and slight changes in handwriting are all highly interpretable evidence which could easily have been lost if there was any conversion of the media involved in the preservation process. I am very hesitant to endorse any process which loses significant information - and how can we tell what is going to be significant, given the many ways in which documents, even digital ones, could convey information to different audiences? Also, look and feel is very influenced by the hardware and software, and even emulated versions lose something of the initial context. I am with the audiovisual people you mentioned that a videotape conversion of an old film is not the same thing. It might be useful as a use copy, for ease of access, but for serious researchers I would expect that examination of the original would be indispensible.

My worry is that in making decisions about what parts of records are important, we may essentially cripple the future hypothetical digital 'bibliographers': the ones who study the contruction and design of the more artifactual components of digital records. It is possible that if this data is transmigrated, this sort of field could essentially disappear (or might never exist at all). This aspect of digital records is not particularly interesting to me, but I would hate for it to become impossible for future researchers who are interested in the original architecture of the files, or the intricacies of the display standard in some program to find or view the information which would give them insight into their chosen aspect of human social activity.

The SAA-UW (Society of American Archivists, UW Chapter) visited the Microsoft Archives yesterday, and we viewed their methods for preservation of their fairly large digital collection. It

was very simple: they preserved as much old hardware and software as they could, and set up specific hardware/software packages in their reference room on request in order to view old documents or use old software. They also used commercially available emulation software to provide environments for which they currently lacked the hardware. And yes, they had Altairs with 8" floppy drives in working condition. As for media migration, so far they have managed to avoid doing it for the most part. Admittedly, they have a short history (30 years or so), and have a long way to go to see how long they can keep this up. However, to my mind, they are on the right path: preservation of the original source with the hardware/software that can provide the original experience of the record, down to the last detail (the noise of the floppy drive, the hum of the computer fan, the original color balance in the monitor).

Even if we expect the hardware to eventually wear out, we should credit future generations with the ability to rebuild compatible hardware: an example related to the audio examples given would be the Archeophone. This is a modern device, recently built, which allows for playback of the Edison wax cylinder audio recordings made between the 1880s and 1929. Similar devices could be constructed in the future to allow for 'playback' of obsolete media.

To add a final note, I guess that part of my issue with selective alteration or migration of digital files is the claim that it is still a complete, authoritative copy with all of the information contained in the original. This seems to be the claim that is being made, or at least the claim that would be desirable to make. I can perhaps accept that files need to be converted, etc, but can't agree that they still contain all of the information of the original. As such, I don't agree that they could represent a full and faithful copy. Whether or not they are legally acceptable as documentary evidence is a matter for the lawmakers, but as an archivist or a historian, I wouldn't accept it as a substitute for the original.

Nor can I really imagine that metadata can fully solve the problem - even if it can document every change made along the way during the object's preservation, it cannot preserve or even accurately identify the information that was lost in these multiple changes (otherwise, why not just keep the information in the file itself during the conversion). Metadata may be useful in preserving the legal authority of the file (or in discovery), but this is not the same as preservation of the information itself.

What Cate describes with the birth certificates is exactly the issue that concerns me - that records can in fact be fully authentic without preserving all of the original informational content. In other words, preserving authenticity is not the same as preserving all of the original content. With digital records, most of the standards seem to be assuming that much of this 'incidental' or contextual information will be lost, and are content with simply preserving authenticity and legal

aspects. I would prefer if these standards' authors concentrated more on preserving all of the information, even what they may see as unimportant, and less on the legal issues around records - but their choices are understandable given the people who are generating these standards.

I wonder what academic historians and cultural archivists would come up with when presented with the same task? Are there any groups doing this work? My worry is that if these standards exist without any alternative concepts of how to preserve more of the 'minor' incidental and contextual information present in and around records, such inappropriate standards will become the default option for academic and cultural institutions who don't have time or money to generate their own systems, even though they are designed around the legalistic and fundamentally limited requirements of government and business. If this happens, our historical record will become substantially less rich, and future researchers will have much less to work with.

I completely agree - we are generating more and more information, and I certainly don't expect all of it to be preserved. Even if we could, I don't think it would be a good idea because of the incredible burden it would place on future researchers. I would even argue that in terms of percentage, probably even fewer modern records deserve archival preservation when compared to periods that didn't create as much documentation. My only real concern is that the information which is deemed worthy of preservation be preserved in as complete and accurate a way as possible.

I also agree that the best approach to take would be to continue discussions about how to preserve more of the content of these records, for a variety of different purposes and different possible users/uses. I just feel that the standards we've looked over so far for this class take a different approach than would be appropriate for cultural repositories. Even if they are extensible systems, time again is the issue. Developing and communicating standards is clearly time consuming, and it won't be easy for non-profit, educational institutions to coordinate the sort of work and extension that would be necessary.

Metadata standards and digital archiving concepts of migration and conversion would certainly aim to preserve the explicit informational content (the text), and they would preserve information about the authority and origination of the record (generating department, etc). The preservation metadata would record the history of its format conversions, ideally since its creation. However, there are still components that would be (or have been) lost along the way in this particular conversion of a physical document to an electronic record, even one fully documented in a metadata standard - the wax seal she mentioned (it might be identified as [seal] or something equally un-

descriptive or unhelpful to a future researcher), the typeface in the original document, perhaps the layout/spacing of design elements, the composition and grain of the paper and ink, and probably other 'minor' elements which I'm not even thinking of here. I'm sure that the same minor elements exist in born digital records as well, which may well become the subjects of future study (as studying the composition of paper has). Certain layouts or typefaces in word processing documents, or certain user interfaces in the case of databases or more complex digital files would fit into this category, most of which I see no way to preserve in the context of the existing standards (short of preservation of compatible physical computer systems and refreshing of media, as at the Microsoft archives).

I agree that metadata standards are merely standards- as I pointed out above, I just hope that people (academics) invest the time and energy into doing this work and standardizing the extensions they create so that the metadata standards can provide better service for cultural research institutions. I still have doubts that all of the information and context I would like to see preserved can be preserved in this way, but it would make sense to try...you are correct that some information preserved is better than none. However, we should have better ideal goals, and we should consider the potential future cultural researchers.

I, of course, like this distinction between documents and records. It goes further than the particular targets or preservation activity as well: to me, people who use the term records almost always approach the use of records from a solidly institutional perspective, while documents are very much the realm of the humanities scholar and the manuscripts archivist.

The archives profession (especially in the published literature) seems heavily invested in the records-management, institutional side, while I would like to think of (or imagine) the broader, extra-institutional possibilities of archives. Records, as Jessica describes above, are just so much less than documents - by their own admission, they contain less information and are of a lesser order in some ways. This will be especially true if the definition of what constitutes preservation of records is limited by some significant properties, limited by the time, money, and ultimately the imagination of the standard developers.

Of course, the fact that these authors deny the importance of other aspects of these records in their existence as full-fledged documents (and they can be viewed that way) only makes those aspects more interesting to me as a researcher. 8-) The unintentionally preserved information is usually so much more revealing and honest...and I agree that we need ways to preserve digital documents as well as records.

I fully agree with Andrew that preserving the bitstreams is vital for any digital preservation program. Despite the huge volume of digital records, this should be a priority - and media are likely to expand in capacity rather than shrink.

I think that the procedures described by Andrew to preserve bitstreams represent the current 'best practices', given our technology options. I agree that they aren't currently perfect or foolproof, but I can't see any other options at this point. I don't know much about the nitty-gritty of operating system design, but my hope is that some people on the more technical side of things can develop better copying algorithms (perhaps ones which include more checksum processes during the copying procedure?). It is also clearly important to develop better/more durable media types, which will be less likely to degrade the bitstream. Since checksums can't provide exact information about the location of the corrupted bits, how about comparison between multiple copies of the preserved bitstream to identify corrupted areas, at least for flagging purposes in the metadata? If you had a copy with a passed checksum test, it would make an easy comparison for a file with a failed checksum test. Any different bits would be easily identified by subtracting one bitstream from the other, and then corrected if possible.

This seems to me like the easy part of digital preservation, since it just boils down to mathematics and physical media design...