Rhiannon Bettivia
Digital Preservation Final

Preserving Media Commons: A philosophical and occasionally practical investigation

I first became interested in Media Commons after hearing a presentation by David Millman, the head of Digital Library Technology Services at Bobst. From the standpoint of academic literature, the endeavor was interesting in that it proposed to offer real alternatives for scholarly publication for scholars in the humanities, where adoption of digital publications has been generally slower than in the sciences. It was also interesting because the subjects mentioned by David as having recently been discussed in the Media Commons project seemed to focus particularly on communications and mass media, an area that should be embracing multi-media scholarship but often doesn't for fear of copyright issues. In selecting this topic for a research project in the area of digital preservation, Media Commons seemed ideal as it encompasses a large number of information types each with specific preservation challenges.

In beginning the research for this project, I quickly found myself overwhelmed. It was hard initially to grasp what I would be researching: I did not quite understand how one could actually archive, and preserve long term, web pages. My only previous experience in this field was limited use with web crawlers, which thus far I had found woefully inadequate: while they certainly provide an interesting service and capture more or less interactive images of some websites faithfully, I did not see them as a real preservation tool because of the large number of challenges they pose. There were the websites that didn't capture, the difficulty of setting the scope, the changes rendered in layout and design because JavaScript or Flash made capture impossible and even sites that completely blocked efforts at preservation with no robots files. To my mind, web crawlers did not constitute a real preservation option and I had a hard time conceiving of how else a website might be saved.

This difficulty was overcome when I began to think more broadly about Media Commons, and its two major publication projects, *The New Everyday* and *In Media Res*. Shifting focus to thinking about what the projects actually consisted in considerably complicated my research but also offered real avenues for pursuing. This change in thinking brought me closer to the process that scholars like David Millman and Brian Hoffman must undergo in practically addressing challenges like preserving the Media Commons projects, and helped dictate the course of my research. This shift also brought about a considerable change to how my own project unfolded: I had initially envisioned rather technical findings that would speak about the tools for web preservation in great detail. In actual fact, the project turned out to be more philosophical, probing the question of "what are we actually trying to save here."

First things first: What is Media Commons?

The first thing was to look into the Media Commons project. My initial understanding was vague yet exciting: scholars tossing around contemporary topics of interest like Glee and Harry Potter in a rapid publication venue that allowed for almost instantaneous peer review and response. More radical, to traditional academics, was the opportunity the public had to weigh in on the topics and the scholarly articles themselves. That such an endeavor had intrinsic worth was immediately apparent to me: I have been using university library facilities as long electronic journal services like JSTOR have been around,

and in a decade and a half there seemed to have been little advancement in the way we think about publication in academia: the content and way in which it was presented had remained largely the same, with the only change coming in the method of delivery. Now texts and their catalogs showed up on a computer screen instead of on shelf in paper. The content, in the form of highly structured papers that followed general trends within disciplines, remained the same and so did the process by which scholars actually published their work. There was still the process of submitting initial versions, having it tossed around to peer academics for review, validation, commentary, and eventual publication several months down the line. Even with the seemingly faster capabilities of e-publishing, academic literature still took months to make it from professor's desk to journal database. The public is, as ever, closed off from this process and its products: academics don't write for the public and their works usually are not available for those outside universities, whose libraries pay a premium to carry e-journal subscriptions that local public libraries or even individuals can't or don't want to afford.

Media Commons seeks to change all this. It seeks to create a truly new form of academic publication that utilizes the tools provided by the Internet like speed and universal delivery. Media Commons is a collective endeavor between NYU and the Institute for the Future of the Book, with funding from the National Endowment for the Humanities.[i] The Institute for the Future of the Book is associated with the University of Southern California but is an independent think tank that works on issues of publication in the era of the Internet and new technologies[ii]. It engages in numerous projects, and receives funding from the MacArthur Foundation. Its offices are located in Brooklyn in the US and London in UK.

The initial goal of this collective project was to create new modes of thinking about academic writing. First was the goal to utilize the speed of the Internet by creating scholarship in real time: rather than wait for the lengthy process of writing a paper, submitting it to journals, making changes and edits, then shopping it around for peer review and ending with publication months or more later, the idea with Media Commons was that a call would be put out for papers or ideas around a certain topic and materials would be made instantly available when received: publish first, edit later[iii]. By posting the works immediately and then opening them up for commentary from peers in the field, a second goal was to make the process of scholarly publication more transparent. Now, rather than a journal shopping a work around to other academics who would work on reviews and responses isolated from each other, this process takes place visible to all on the web. Works are posted, and peers comment blog-style. Authors can respond, discussion can be had, but all of this is done openly in a forum that can be viewed both by other scholars and by the public. In a sense, the mystery is removed: we see the inner workings like putting a clear container over a piece of machinery rather than an opaque cover that obscures the mechanisms. This second goal provides for a third aim: by opening and demystifying the process of publication, academic publication is opened to others, primarily students entering the fields. Students are able to both see the process unfold on Media Commons as well as engage in the process by submitting comments and even papers of their own. This provides an entry point for students into scholarly publication: it teaches them how to write by providing examples and experience, and for universities and forward thinking fellowships, it provides a list of publications that a student can reference on a CV.

Media Commons has the additional goal of introducing completely new modalities of scholarly literature with the eventual hope that these will become standardized and accepted ways that scholars can be evaluated by their institutions for tenure or for the receipt of grants and fellowships[iv]. Despite changes in speed and transparency, submitting a paper or abstract to Media Commons is still an old-fashioned notion at its core: academics are submitting papers even though the distribution type looks radically different. Media Commons aims to create entirely new forms of written scholarship formed by the way that articles and commentators interact using the Media Commons web projects. Two important aspects of the Media Commons projects lend themselves to new scholarship: first, the curation aspect and second, the built-in mechanisms for commentary. Discussions around certain themes are submitted together at a certain time and are curated by a scholar: this person might pick the order in which articles are released or the ways in which clips and comments are presented. A hope is that eventually academics could be evaluated not only on their writing but also on their curation skills: in theory, a tenure committee or grant committee could be pointed to Media Commons threads or clusters where they could see the skill an academic has at arranging papers and comments towards a common goal, in a way that furthers learning and discussion and hopefully renders a curated discussion more valuable than simply the sum of its parts. Another aim is that scholars can also be evaluated on their ability to provide commentary on submitted works, again by pointing a committee to Media Commons and body of commentary that has been built on any number of subjects. Again, the idea would be to demonstrate that the commentary provided actually furthers discourse and learning, a skill which a committee would like to see in a candidate.

The two main projects that I explore under the auspices of my research project are *The New Everyday* and *In Media Res*. *The New Everyday* will call for clusters of work around a particular theme and operates as a sort of cross between a journal and blog, allowing people to post and both peers and the public to comment. These clusters are curated, as was mentioned earlier[v] . The other endeavor, *In Media Res*, calls weekly for 5 submissions on a particular topic. These submissions include a video ranging from 30 seconds to 3 minutes in length, with a short essay on how this clip has been recontextualized for the article and subject of the week. The topic and authors are announced at the beginning of the week, and each day a new one is made available for perusal. Accompanying each subject is a Facebook group and Twitter page to encourage active discourse[vi].

Why Preserve Media Commons?

In first hearing of Media Commons, given that it was in the context of a course on digital preservation, my first thought was naturally: *this project should be preserved*. At the time, preservation on this project had not begun, although brainstorming around this idea and first steps were being taken at the Digital Library. To me it was obvious that this content was deserving of a thorough preservation plan: it plays a large role in our cultural heritage and the human record in general. If wide adoption of alternative publications becomes common place in the future, Media Commons will have been a pioneer and have played a role in this. But even if this never comes to pass, given the topics around which the articles are written and the dynamic discourse that takes place between academics and the public alike, Media Commons offers a valuable picture of what is important in society now, and what popular themes pervade and inform contemporary scholarship. A hundred years from now or more,

Media Commons will be a way for future generations to understand that millions of people watched a show called Glee and millions more read books about Harry Potter, in the same ways serials and circulation numbers tell us that, years ago, people were obsessed with chapters of Dickens that appeared in papers episodically.

However, not everyone can take a digital preservation course and thus it may not be clear to everyone that this material is inherently valuable and worth saving. An MLA task force found in 2007 that very few tenure committees look on content like the Media Commons as valuable, and very few considered blogs and wikis, among other items, when making tenure decisions[vii]. In his article in the American Libraries journal, Steven Escar Smith posits that preservation and value go hand in hand: until we value the content, we won't preserve it; the very act of preserving content demonstrates its worth to other academics and tenure committees alike. He also posits that librarians need to lead the charge in demonstrating value through consistent use and preservation: as traditional stewards of the human record, preserving items like the Media Commons materials falls within their mandate (Smith, 2010).

The idea of preserving web content is foreign to many and there could be many possible reasons behind the reticence to adopt alternative web publications within academia: asking around, I heard reluctance borne of fear of copyright issues and reprisal from rights holders to desires not to lose the value of the book in the digital world to the very valid point that not everything that goes on the web can or should be preserved. However, the reality is that web-based materials are quickly becoming the norm. Libraries are spending increasing amounts of money on electronic materials and students and faculty would rather use more loosely related materials drawn from the web than have to trek to the library to seek out a book that more closely matches their needs or interests[viii]. The content in Media Commons, being highly curated and peer reviewed, and thus peer edited and corrected, needs to be preserved just as any other academic electronic journal might: its quality and utility are not in doubt, and its contribution to the future understanding of our contemporary culture is evident.

What Are We Actually Doing?

The fact remains, however, that neither *The New Everyday* nor *In Media Res* are just another electronic journal. They are something different by design, and pinpointing exactly what they are is necessary as part of a greater discussion about what we actually want and need to preserve when working with materials such as these. My initial interest in this topic was partly derived from the complexity of the Media Commons materials: in a sense, nearly all types of web materials can be found in the Media Commons projects and thus it offers broad and myriad avenues for discussion. However, it is in this very complexity that the issue lies: what exactly do we seek to preserve?

In beginning to address this, it was necessary to look at the component parts of any Media Commons cluster or weekly set of articles and see what comparisons could be made to the preservation attempts of other web-based content. As *The New Everyday* sees itself as being part blog and part electronic journal, looking at the preserving of blogs and electronic journals were obvious paths of research to be pursued. However, *In Media Res* offers something slightly more complicated as does the Media

Commons experience as a whole. Towards this end, I looked not only at general web preservation, but also the preservation of time-based media. The latter seemed necessary as capturing the user experience of Media Commons is as necessary as saving the written content, and the work done on capturing user experiences at art installations mirrors this challenge to capture not a simple physical or digital object, but the way in which the users interact with it and how that experience unfolds over time.

At this juncture, much work has been done on the preservation of electronic journals, and this was a good starting point since, of the four identified avenues of research, this was in some ways the most straight forward. Electronic journals, whether digitized or born-digital, constitute somewhat concrete objects not so different from their paper counterparts in essence. The challenges lie in how to preserve the digital media in the face of technology obsolescence, the need for storage and storage infrastructure, identifying responsible parties, and granting adequate access. Digitized journals have been around at this juncture for more than a decade, but the real push for a preservation plan for electronic journals began in 2004 and 2005. The call came from the Mellon Foundation and spawned several initiatives still in use today. The Association of Research Libraries and the Council on Library Information and Resources began in 2005 to identify the needs that a preservation plan would have to meet[ix]. This entailed the creation of a repository audit that would allow others to see its trustworthiness, created by the National Archives and Records Administration (NARA) and the Research Libraries Group (RLG) and entailed looking at models both in the United States and those from Germany, Austria, and Australia among others (Kenney, 2006). Among those projects deemed most successful were CLOCKSS and Portico[x] (Kenney 2006; Kirchhoff, 2009).

In writing about lessons learned from the creation of Portico, Amy Kirchhoff points out that preservation is the burden of all stakeholders; in the case of e-journals, this means the libraries and institutions that stock them, the publishers of the journals themselves, and the academic users and contributors (Kirchhoff, 2009). As such, preservation must address the needs of these stakeholders in order to garner interest and financial support. Libraries have a vested interest in preserving digital journals: they are already designated stewards of paper versions, and particular libraries have the additional impetus to stock materials that either reference or are of use to resident faculty. Libraries and users want access, but access needs in addition to preservation needs must address all users: Kirchhoff points out that the digital medium lends itself to easy repackaging and thus e-journal materials remain valuable to the rights holders, publishers in this case, in ways that traditional journals do not (Kirchhoff, 2009). Such needs on the parts of rights holders were the impetus for dark archives like Portico and CLOCKSS that limit access until a certain time or trigger event.

Likewise, the cost of maintaining the repository must be spread across stakeholders (Kirchhoff, 2009). While in some ways, this addresses a needs issue, it speaks more to sustainability. A trusted digital repository must have consistent funding; a repository that relies solely on one source, save perhaps one that relies on its own capital fund, may disappear when an administration or economic change wipes out the budget. By spreading the cost across users in the forms of fees or though mandatory and cost free deposit, in addition to funding from grant organizations like the Mellon or Ford Foundations or the NEH, multiple sources create lifelines in case one set of resources should disappear.

Finally, Kirchhoff points out what I have learned firsthand this semester: in creating a repository that will hold e-journals, migrate them, an ensure their integrity over time in addition to providing access, the repository should aim to serve a number of purposes for the institution that holds it (Kirchhoff, 2009). Rather than create a repository just for e-journals, the repository should also serve to hold faculty materials, other library content like digitized books, and the materials from other projects in which the university engages. This we see at NYU: the repository, designed to be trustworthy, holds faculty materials, library content, and media for the Hemispheric Institute among others.

In analyzing what this might tell us about the preservation of Media Commons content, it becomes necessary to indentify stakeholders since they play such a dominant role in our understanding of preserving traditional electronic journal materials. In this case, the roles are much less clearly defined than those of publisher, libraries as points of access, and users. Like with journals, the users of Media Commons are in many ways the creators as well: academics write the materials and they also use them, both in the sense that they might cite an article or assign it for a class and in the sense that they will utilize the record of publication in furtherance of their career and reputation. However, in this case the point of access and the publisher are the same: Media Commons itself acts in both of these functions. Additionally, I think it acts in the role of partial creator. The Media Commons journals are more than simply a sum of their parts, but rather constitute a total experience of seeing media, reading scholarly writing, and participation in discourse, both academic and pedestrian. In facilitating this total experience, Media Commons itself becomes a creator of sorts. With Media Commons, there appear to be 2 principal parties, instead of 3 or more, and their roles are nebulous and over-lapping. While the financial motive for dark archives is removed, we do see preservation as the burden of all stakeholders. This does not mean that individual users are tasked with a hugely active role in the preservation of content, but instead that they advocate for preservation by demonstrating the value of the project through their use and participation. In this case, the role of active participation falls predominantly on the NYU Libraries, and we can take lessons from Portico on how this should be addressed: namely that we use the infrastructure already in place in the digital repository when finding a place and method of storage, and that the funding costs should be spread as much as possible between the university, primarily for the long term upkeep of a Media Commons archive since it can provide consistent funding, and grants for the initially more costly exploration into methods of conservation and the set up of whatever system is decided to be best.

The preservation of blogs is itself an emerging field of study, because they have only relatively recently become both so ubiquitous and also valued as cultural artifacts. They are also increasingly becoming tools of the academic trade: besides the Media Commons, they are utilized by other academic disciplines in a variety of forms, not to mention the large number of personal blogs held by professors and other academic professionals. Blogs are now an essential part of both the academic and human record, and thus they need to be saved[xi]. Jessamyn West likewise calls blogs “history as conversation,” something that plays an active role in the shape and creation of contemporary society, thus further highlighting the need for seeking means of preservation for some blogs; given the ubiquity of this format, it certainly is not advisable to attempt to preserve all, but there are some that will play an important role in history[xii].

In preserving blogs, there are a number of considerations. First is the fact that blogs often include clips and images whose rights holders are not the authors of the blogs themselves. Depending on the length or types of embedded materials, a responsible repository might need to seek rights clearance if preserving the blog, as preservation necessarily involves some form of duplication; this is particularly pertinent in cases where the repository is desirous of granting access to the preserved blog for public viewing (Hank, Sheble, and Choemprayong). This can be difficult and time consuming depending on the number of blogs an institution may try to preserve and this would greatly affect an institution's ability to cull and preserve blogs through some sort of automated system: it might require a body to look over collected materials in order to discern if there are any items which might need rights clearance. Such a concern would greatly affect the number or types of blogs an institution might choose to save, and it would be unfortunate if material concerns over resources led to some very historically or culturally valuable blogs being overlooked.

There is the additional need to capture sufficient metadata about the blog when capturing and preserving it. While the need for metadata is universal to all kinds of archiving, populating metadata fields can be tricky when culling blogs from the web, either manually or with some kind of crawler. It can be difficult to find information about the blog, and this is especially true when trying to find out information about the blogger: often times, bloggers will write anonymously or under pseudonyms. In recording data about a blog, ideally one would like to be able to include information about the true author. While this may seem trivial, consider pen and paper examples: how would our relationship with literature be different if we did not know that the Bronte sisters were behind Currer and Acton Bell; how would our relationship to philosophy have been changed if we couldn't attribute Kierkegaard's many pseudonyms to his work? The most obvious one is an example in which the author is, to some, still unclear: who was Shakespeare and is that name just a front or pen name for the real genius behind his plays?

Finally, blogs are difficult in that they lack clear and defined boundaries (West, 2007). A blog might link to other pages or link internally to different parts of itself; it may reference other materials or even cull information from outside databases and other sources. This lack of boundaries, which may be more or less complex depending on the particular blog, speaks to the larger difficulty of defining a blog as a preservation object. The content on a blog is dynamic, and capturing the same blog at two different moments will reveal different things about it. Additionally, blogs can be viewed in a variety of ways: we can go to a blog hosted by WordPress, but we might also look at RSS feeds or use Firefox applications like Grease-monkey that alter the way we view a blog (West, 2007). Several people can view a blog in different ways: which of these is the right preservation object? Do we try to capture them all or hope that one is representative? Likewise, as the content changes, how often do we preserve what we see? At what point do intermittent snapshots fail to demonstrate the dynamic nature of the content? Hank et al. speak to the need to ensure that the blog is understood when stripped of its live context: preserving a blog is more than copying its text and images, and must somehow reflect the intended design, the user tools, and its overall functionality for a user (Hank, et al.). Perusing a blog employs several behaviors, and without somehow conveying these, we cannot call the blog preserved. It is easy to understand this with an analog analogy: a book has certain behaviors that make it a book. A random pile of papers with

words is not a book; rather, in a book one should be able to turn pages, and what comes next should follow from what was written previously in some kind of order determined by the author. A blog is much the same: text and images are not the only parts that make it up. Its tools and user experience are integral to understanding what it is.

This speaks directly to some of the difficulties in preserving the Media Commons materials, and these difficulties will recur when the paper moves on to discuss issues of web and time-based media preservation. In terms of rights issues, Media Commons may not have as much to worry about. Given that there is no profit motive, and that the length of the clips is prescribed as such a short amount that fair use could be claimed, Media Commons might be spared all the more in that their current plans for preservation are aiming for a pseudo-dark archive: they are not attempting just yet to create a user interface that would grant access to preserved materials. The issue of anonymity may be one that Media Commons will have to address: they allow for commentary that requires nothing more than a name and an email for parts of the project. Article authors and curators are seemingly better defined, although even these vary in subtle ways that might need to be addressed when preserving materials: sometimes authors' names appear in all capital letters or all lower case, and titles and degrees such as PhD can appear as PHD or phd. These are simple metadata items that it may make sense to streamline. Media Commons also needs to ask itself: is it okay that we know more or less about certain contributors? While it certainly does not make sense to ask for so much biographical data that potential users will be put off, a researcher peeking at these materials a century from now may be dismayed that a particularly poignant comment is attributed only to "anonymous" or an obvious pseudonym. The issue of boundaries and defining the preservation object will be particularly difficult ones for Media Commons, and this issue will be further discussed below. Suffice to say that similar to blogs, we could not call it preservation if all that gets preserved is a series of text files, images, and moving image clips. Rather, Media Commons is a user experience above and beyond the mere sum of its parts, and the preservation needs to convey this somehow.

Web preservation is a concept that stymied me at the outset of this project. I did not really understand how it could be done. Here, I will briefly explore some of the policy issues that arise when considering web preservation. With the breadth and relative evanescence of the content on the web, there is an increasing push to have the creators play a role in the preservation of their materials[xiii]. This means asking those who post content to the web to save copies of their images and writings. Creators must also have a role in creating and attaching metadata to their work: ideally they would fill in any information not automatically populated by the creation device to meet Dublin Core fields at a minimum. In terms of academic web endeavors like Media Commons, or the History of Recent Science and Technology (HRST) project, authors submitting work should provide as much information as possible. The trouble arises when the web designers and managers have to find balance between wanting the metadata and wanting constant quality submissions of work. It becomes a tradeoff: in allowing for flexible submission guidelines, a project might receive excellent work but suffer from a lack of quality metadata (Smith, 2003).

There might also be metadata issues in terms of cataloging the holdings in a web archive- given the overwhelming myriad subjects, how do you create functional file naming systems and functional

controlled vocabularies that encompass everything? In doing a web archive study, researchers at the University of Illinois found that even when only archiving web sites on a particular topic, in this case hate and intolerance websites from Illinois and its immediately surrounding states, they came across terminology that was totally unknown to them and outside any controlled vocabulary[xiv]. This becomes more complicated when looking at broader cross sections of the web: for instance, what happens when trying to develop controlled vocabularies to incorporate works in a variety of languages? With technical and cultural terms in particular, direct translations in other languages often do not even exist.

The previous two concerns about documenting the web assume that you have already dealt with the most pressing issue: selection. How do you choose what to archive? First and foremost, there is the sheer number of websites that grows every day. While the Internet Archive attempts to survey the entire web and projects like Australia's PANDORA attempt to save all websites in the .au domain, for an institutional repository there must be narrower guidelines. For some, the selection is easier than others. Library of Congress' Minerva seeks to capture websites about those events most important to U.S. history, like presidential elections and catastrophic events like the attacks of September 11, 2001. California Digital Libraries, which is currently still experimenting with its web archiving service as I understand it, will crawl website collections specified by its partner institutions[xv]. An institution wishing to start a web repository must choose which websites are important to its mission.

The other related issue, once you have chosen which sites to crawl, is when and how often to crawl them. Few sites are completely static; some like popular news sites change every few minutes, while others change daily or weekly. How do you decide when and how often to crawl? Such a question is intimately bound with the purpose of the archive: when attempting to document an event like Hurricane Katrina, it might make sense to capture as many copies of websites as frequently as you can to show events unfolding. If the purpose of the archive is simply to give a snapshot of contemporary society at a particular point in time, a single capture of even the most dynamic website might be sufficient. Magali Haettiger posits that to truly capture a particular website, a single snapshot is never sufficient: the site is dynamic and so must its preservation be and as many as possible versions, if not all, are necessary for preservation with integrity[xvi].

The bigger problem is in defining what is meant by a website. The boundaries, like with blogs, are unclear. Sites reference each other—this could go on for hundreds of iterations. Do you preserve every site linked? How are boundaries determined? Like with blogs, websites are also more than the information they contain. In order to preserve with integrity, its functionalities needs to be preserved in addition to forms and content (Haettiger, 2003).

Finally, given the complex and sometimes vernacular nature of the topics pursued on Media Commons, issues of controlled vocabulary could become problematic. Schmidt and the other researchers at Illinois were stymied by terms they had never heard before, like the number 88 continually used; it turns out that this is a reference to "Heil Hitler" as the letter H is the 8th letter of the alphabet (Schmidt, et al. 2008). The Media Commons covers esoteric topics ranging from Glee to Harry Potter, and such topics could raise issues in terms of cataloging and descriptive elements when words like *horcruxes* and the like come up—the term is common enough now Potterphiles, but outré to others and it may fade into

history as time passes. Likewise, with metadata overall, the submission guidelines for Media Commons are purposely flexible to encourage submission. David Millman admits, however, that the submission process may change in subtle ways to reflect metadata needs upon the creation of an archive.

The difficulty in defining the preservation object led me to research the preservation of time-based media art as well. This type of artwork is part performance, and part sculpture[xvii]. It is not one static object like a traditional sculpture or painting. Rather, it involves many component parts, such as materials, the space in which the work is proscribed, and the experience over time that observers have with the art. There is no one piece that can simply be put on a shelf till next time: the work evolves and unfolds over time, so time itself is a part of the exhibit[xviii]. The additional challenge in preserving these art pieces is that they often involve media components such as film or video clips. The difficulties arise from the fact that equipment to play the media not only breaks down over time but becomes obsolete so that replacements are not available; the mere wear and tear of being on a loop for an exhibit can also compromise materials, with new film masters being needed just about every 3 weeks if film is being played constantly as part of a piece, for example[xix].

When challenged with such new forms of art, Pip Laurenson contends that our very notions of what conservation means must change (Laurenson, 2006). We are no longer talking about something that can simply be boxed and stored, but about learning to preserve the integrity of an idea. Preservation is no longer a single conservator wearing white gloves acting in isolation to care for a work: it is now become a social process that requires give and take between artists, conservators, and other staff at a gallery or museum (Laurenson, 2006). Laurenson cites that New Zealand Professional Conservators Group as saying that conservation "is the means by which the true nature of an object is preserved"[xx]. And even more encompassing definition is:

"all efforts designed to understand cultural heritage, know its history and meaning, ensure its material safeguard and, as required, its presentation, restoration and enhancement."[xxi]

This perhaps gets us closer to what we need to do in preserving time-based media art and in working with something like Media Commons. Javier Pes says that our preservation efforts must be proactive and preventative (Pes, 2008). They involve intimate conversations with artists as the art is in creation, and artists may proscribe as many or as few guidelines on how the work should be conserved and accessed in the future (Laurenson, 2006). In practice, working out a plan that suits the artists and the receiving institutions is tricky. At this point in time, the Guggenheim, Berkeley Art Museum, Pacific Film Archive, Rhizome, The Franklin Furnace Archive, Walker Arts Center, Cleveland Performance Festival and Archive, Tate, SFMoMA, and MoMA are all engaged in projects around developing standard practices for working with artists to develop plans to preserve and redisplay time-based media art[xxii].

The problem of displaying this artwork anew is equally bound up in its conservation. Howard Besser talks about the translation problem in preserving and recreating complex media art[xxiii]. As was mentioned earlier with blogs, these pieces exhibit certain behaviors that must be conveyed when preserving and granting access. The work is not comprehensible without its expected behaviors, just like pages are not a book until they behave a certain way. Additionally, some types of media and

artwork are too complex for every element to be saved: preservation must be done selectively with careful documentation about what is saved, what isn't, and why.

This has many implications for a preservation plan for the Media Commons materials. The Media Commons project is in many ways like a time-based art piece in that it is not simply a journal or an academic publication, but an interactive experience that unfolds over time. In some ways, the most valuable lesson learned from preserving complex media is that the preservation process needs to be dynamic and social: it cannot simply be undertaken by the Digital Libraries Technology Staff, no matter how competent they are in the fields of digital preservation. Rather, this process must involve a conversation between the creators who are also users: the academics and public who contribute most must engage in conversation with the digital conservators to identify what exactly needs to be preserved. The conversation needs to start by identifying those behaviors which make the Media Commons projects what they are and best describe the user experience. Even if access is not the goal of the preservation copies that are made, they still need to be understandable outside their original context.

Additionally, we can learn from previous experiences in selecting which aspects are important to conserve and which must be let go. Media Commons might seek to use a system like the one Chris Lacinak and Brian Hoffman created to track and reference data for the Merce Cunningham project: master diagrams that reference the database from which information comes to indicate typically what kinds of data are placed into the preservation repository and what kinds are not.

How Do We Do It?

In talking about how we preserve, we also need to talk about who is going to do it. There is a growing trend toward pushing conservation upstream; rather than have a conservator struggle to deal with the final project, the creator must take an active role in preservation (Smith, 2003 among many others). In relating this idea to Media Commons, we first have to be clear about who the creators are. As was mentioned earlier, given the collaborative nature of the Media Commons projects, roles such as publisher, user, creator, and distributor are not clear cut. In terms of creators, two sets of people can be seen to fulfill this role. First, there are those who create the articles and commentary around which the websites are based. These are creators in the most basic sense: authors. Does it makes sense in this context to ask these creators to play an active role in preservation, by saving offline copies of all their submitted work on a hard drive where they migrate the format as need dictates or as a paper copy in a file cabinet? I would say no, in this case. This is not to say that academics and people in general should not strive to maintain personal archives of their published and unpublished materials. I only mean that in this case, because there are so many creators involved in one theme or cluster of works on Media Commons that it would be impractical to go back a second time to all contributors and ask for copies of their articles, media clips, or comments. Additionally, this is simplifying too much what preserving the Media Commons would actually entail. I think it has been sufficiently demonstrated that simply saving the text and information is not preservation, even if that text and information lasts 2000 years past the end of the republic. Recollecting the articles will not recreate the Media Commons experience. In this case, we need to talk about a different creator, the minds behind the Media Commons and people that

make the site itself happen. These people are responsible for creating Media Commons as a total experience: they collect the content and arrange for it to behave in proscribed ways, and these behaviors are what make the projects so unique and distinguish them from e-journals found on JSTOR, for example. These creators are the ones who are in the position in terms of capabilities and resources to preserve the authenticity of Media Commons, and lucky for us the Digital Library Technology Services staff are committed to doing so.

How then does this preservation get accomplished? Given that Media Commons is web-based, the first thing that comes to mind is web crawlers, designed to capture web content in snapshots that can be preserved. For this project, I looked into 5 crawlers: Internet Archive's Heritrix, CDL's Web Archiving Service, Epicware's Web Grabber, SiteSucker, and HTTrack. For the Internet Archive, I attempted to look at crawls that a group had done of Media Commons pages for the digital preservation group projects on web crawlers, in addition to looking at archived versions of the Media Commons web pages, particularly of *In Media Res*, in the Wayback Machine. After our experiences with Internet Archive this semester, documenting the issues that people encounter capturing both very basic HTML and Flash content and the troubling decontextualized video player, I was tempted to dismiss this option out of hand. However, the contributions that Brewster Kahle has made to web capture cannot be overlooked, despite the difficulties that exist with the crawler (Schmidt, Shelburne, Vess, 2008). In their exploration of web crawlers and documenting hate literature, the researchers from Illinois encountered troubles with Internet Archive nonetheless, chiefly ghosted texts and broken links (Schmidt, et al. 2008). CDL's web crawler is not currently available for use on an institution's home network; likewise with Internet Archive's crawler, this does have the advantage of not using your own bandwidth to crawl websites for hours on end; this is more or less important depending on the size of your network: Time Warner Cable got mad at me several times during the course of this project for running web crawls from my home network and I was kicked off the internet several times. The drawbacks to using either of these is that they work automatically and not locally, meaning that your direct control over what is being archived and when is somewhat limited. Additionally, it means that you aren't personally storing the content for preservation, even if you can ask for a copy. CDL's web archiving service only works for its partner institutions and for particular collections or topics. Its scope does not seem to encompass Media Commons-like materials, so this too was not a good choice.

Web Grabber, SiteSucker, and HTTrack are all programs that are available for free download onto personal computers. Katharine Dunn recommends alternatives like these for do-it-at-home web crawling and preservation or for preservation in smaller institutions[xxiv]; SiteSucker and WebGrabber work best on Macs, while HTTrack (and Heritrix) work best on Linux machines and PCs (Schmidt, et al. 2008; Dunn2009). With these tools, any user can crawl and capture data from websites of interest. Schmidt and the other researches at Illinois preferred the somewhat manual technique required in using these harvesters as opposed to Internet Archive because they felt active and manual control was necessary to process incoming data, make decisions about scope, and create catalog records with controlled vocabulary (Schmidt, et al. 2008). Dunn recommends version control using simple tools like Time Machine and File Merge, standard or free on most Macs. Time Machine allows the content to be stored on an external hard drive, important as the files accrued in a single web capture can be large, and

frequent crawls of dynamic content can quickly lead to large storage requirements. Versioning is necessary as you capture as many versions either as possible or as you feel you need of a dynamic site to properly portray its user behaviors (Haettiger, 2003; Dunn 2009). File Merge will allow for quick comparisons between files so a user can see how much differs on a site from capture to capture. In their project capturing materials to be included in the Ewing C. Baskette special collection at the University of Illinois, the researchers elected to use Web Grabber. They found that this program captured some websites well, while struggling with more complex media. Particularly, blog sites and other sites with live content written in PHP were not captured well, as well as sites that relied heavily on databases for their content. The additional obstacle they came across was the crawlers unearthed materials that were not actually present for public viewing on the website; I believe a similar issue arose during the Internet Archive project when a group archiving gallery websites came up with a video whose source on the web they could not locate. Schmidt et al. occasionally turned up private web journals in their searchers, and struggled ethically to decide what to do with such content (Schmidt, et al. 2008). Such concerns need to be addressed when potentially choosing a web crawler and more pertinently, when deciding what you are going to do with the information captured. If the idea is to preserve it in perpetuity for use in later generations, then the issue of non-public materials decreases in importance; if you seek to grant immediate access to your captured materials, this needs to be addressed.

I tested SiteSucker and HTTrack on my own computers at home, doing partial crawls of the *In Media Res* site, as well as 2 others for experimental purposes. I tried HTTrack, which Chris Lacinak recommended as accurately portraying older websites in his experience, on my Dell Mini, which runs Windows 7. I tried SiteSucker on my Macbook Pro, which runs Snow Leopard, which I believe is OS X 10.6. I found the two programs to work similarly well in producing a workable version of the *In Media Res* site. However, the accuracy of the experiment was somewhat compromised by the fact that Time Warner did not like the amount of bandwidth I was using while crawling and frequently shut off the internet on most if not all of my computers when I attempted to crawl. A glance at crawls done by other digital preservation students of Media Commons using Heritrix revealed that those crawls took upwards of 6 hours. I had no such luxury and aimed to crawl the website for approximately 45 minutes on each computer. Both produced fairly faithful versions of the homepage and I was able to click on several links and see movies in both. I found the user interface on HTTrack to be much more user-friendly. HTTrack encountered 63 errors in attempting to mirror the site; SiteSucker captured considerably more files and errors in the same period of time. Just for interest, I also attempted to crawl two additional sites: Sean Shepherd.com, a site that I know from a previous project has a no robots file, and the website for DBGB's chef, danielnyc.com, which relies heavily on Flash. Crawls of Sean Shepherd's website revealed a completely black page, while DBGB's site was missing images, movies, and sounds but was still somewhat workable.

Probably using manual crawlers is not the ideal tool for Digital Library Technology Services to use in preserving the Media Commons sites. It is not entirely effective, breaking down the deeper the links go and struggling with PHP and Flash content, which are only becoming more and more common. The addition challenge is that some of the Media Commons topics and clusters link to Facebook and Twitter pages, password protected sites that involve filling in small menus even if not inputting a password, and

sites like these cannot be reached by crawlers. Using crawlers in the way described by the researchers at University of Illinois is too labor intensive to be a regular and large-scale practice, and the recommendations of Katharine Dunn are for a smaller-scale institution (Schmidt et al. 2008; Dunn 2009). Given the other difficulties already mention in this paper and in our digital preservation class, I do not think that web crawlers are sufficient to document and preserve Media Commons as they currently function; perhaps, as Howard Besser suggests, they will be capable of better preservation several generations down the line. I do not discount their importance or the effect that Internet Archive has had on the landscape of web preservation, but I do not believe that they are currently the way to go.

Fortunately, I discovered in researching this project that the Digital Library staff are already hard at work in finding a solution to this preservation problem, as they too have recognized the unique value of the Media Commons projects. They have also recognized the inadequacies of contemporary web crawlers to preserve their content. As such, they have come up with an alternative preservation plan that they are currently submitting in proposal form for grant money. While the solution is not concrete, the idea is one that I found echoed elsewhere. I met with David Millman and Brian Hoffman at the Digital Library to discuss their current plans, and they generously shared with me their ideas[xxv].

They are beginning by moving the preservation process upstream, to the creators in a sense. Because the content that makes the Media Commons website is drawn from a database, the idea is to create an API that will draw information from the database, much the way that the website does when someone uses it, to create an offline version of the Media Commons experience using the same data. An important part of this endeavor would be to first break down the components that make up Media Commons to their most granular level and define the atomic parts. Once the parts are defined, a taxonomy can be created to relate these parts to one another in a way that mirrors the behaviors of the website itself. From this a model can be created that saves the component parts and their relationships to each other to the NYU repository on a weekly basis perhaps, although the time frame is not clear. The dynamic content of the web sites still pose problems: as the content changes daily in some places, the team must figure out how often to capture the content to give a picture of the website that accurately conveys its behaviors and the user experience. They are now hypothesizing that a weekly cull could be sufficient, but this is not set in stone.

I ran across an early article that suggested a similar idea in the early days of web preservation. Carol Casey recommends saving an "offline" version in a catalog that would point to both the data sources and the online version[xxvi]. Presciently, the article also recommends time-stamping archived website versions, much the way the Wayback Machine and Wikipedia currently do.

Challenges

This solution of APIs seems to me a better one than web crawlers did, but there are still many questions to be asked. The Digital Library has realized this and one of the things they called for in their grant proposal was to convene a conference of academic users and technicians to discuss the answers to many of the philosophical questions posed in this paper. Some of the more pressing concerns I have are mentioned below.

First is the issue of access. When I asked David and Brian about this, their answer followed a moment of silence. At this point, they have not though about a method of granting access to the preserved versions of Media Commons. The question that David pointed out is that it is hard to envision what this access would look like and what its use would be. This is not to say that he is calling access to old websites useless. Rather, the discussion is around how to present old content to new users a decade or more down the line. Does it make sense to show a researcher a version of a website from 10 years ago as it looked 10 years ago? If a researchers is doing work on what old websites looked like, perhaps. But if they are trying to peruse the content and see how the discourse around a past event unfolded, does it make sense to show them the website as it looked 10 years ago or to render it as it would look should it be created in the here and now (or in this case, as it would look 10 years into the future)? This is an interesting consideration that I don't have an answer for yet.

I also have questions about the form of the offline file that is created using information drawn from the database by the API. How will this be saved? It seems to me to create a dangerously high number of dependencies if the files that are saved are merely maps that point to certain items and relational taxonomies defined in other places. This is how the preservation copy will be created, but will it also be saved this way? Or will the link-free text and images be saved? If this is the case, it brings us back to earlier discussions about adequately translating the original object and contextualizing the content with its related behaviors—either answer seems to pose problems, so it is somewhat of a catch 22.

Finally, there is the concern that if the structure of the database that hold the materials ever changes, the API will have to change and this could have implications for the readability of old files and create a window wherein materials are lost as a new API is created.

Conclusions

Challenges aside, I think this idea of creating an API to recreate the Media Commons websites offline makes sense, especially as many of the particular functions of the API and the look and feel of the preservation copy will be determined in part by long discussions with the creators and users of Media Commons. The project itself is user generated and sustained in so many ways, and thus so must the preservation be user informed. I initially worried about submitting a paper that raised more questions than it answered. However, if such questions can be turned to the focus groups that will discuss how this project will actually unfold, then identifying the questions that needed to be asked was a good and even necessary use of time. I feel that David and Brian, in conjunction with a panel of users and technicians, provided with time, climate control, and catering, could definitely work out the answers to the queries that I have posited today.

---

[i] http://www.mediacommons.futureofthebook.org; retrieved 11/2010
[ii] http://www.futureofthebook.org; retrieved 11/2010
[iii] Santos, Avi. MediaCommons 2: Renewed Publics, Revised Pedagogies. 2006. Retrieved 11/2010 from http://www.futureofthebook.org/blog/archives/2006/07/mediacommons_2_renewed_publics.html
[iv] Millman, David and Brian Hoffman. Personal Interview. 12/07/2010
[v] http://mediacommons.futureofthebook.org/tne/how-it-works; retrieved 11/2010
[vi] http://mediacommons.futureofthebook.org/imr/about; retrieved 11/2010

[vii] Smith, Steven Escar. It's The Content, Stupid. 2010. American Libraries journal
[viii] Kirchhoff, Amy J. Expanding the Preservation Network: Lessons from Portico. 2009. Library Trend journal
[ix] Kenny, Anne R. Surveying the E-Journal Preservation Landscape. April 2006. Association of Research Libraries
[x] Smith, Abby et al. "Sustainable economics for a digital planet: Ensuring long term access to digital information" (often referred to as Blue Ribbon Task Force on Digital Preservation). 2010; retrieved 11/2010
[xi] Hank, Carolyn, Laura Sheble and Songphan Choemprayong. Considerations for the Preservation of Blogs. Digital Preservation Europe. Retrieved 11/2010 from http://www.digitalpreservationeurope.eu/publications/briefs/preservartion_blogs.pdf

[xii] West, Jessamyn. Saving Digital History. Spring 2007. Library Journal/part Net Connect
[xiii] Smith, Abby. New-Model Scholarship: How Will It Survive? 2003. CLIR; retrieved 11/2010 from http://www.clir.org/pubs/reports/pub114/contents.html
[xiv] Schmidt, Karen; Shelburne, Wendy Allen; Vess, David Steven. Approaches to Selection, Access, and Collection Development in the Web World: A Case Study with Fugitive Literature. July 2008. Library Resources & Technical Services v. 52 no. 3

[xv] http://webarchives.cdlib.org/institutions; retrieved 11/2010
[xvi] Haettiger, Magali. Ver la conservation des sites web régionaux. 2003. Bulletin des Bibliothèques de France v. 48 no. 4
[xvii] http://www.tate.org.uk/research/tateresearch/majorprojects/mediamatters/

[xviii] Laurenson, Pip. Authenticity, Change and Loss in the Conservation of Time-Based Media Installations. Fall 2006; retrieved 11/2010 from http://www.tate.org.uk/research/tateresearch/tatepapers/06autumn/laurenson.htm

[xix] Pes, Javier. Time-based art needs plenty of tender, loving care. April 2008. Art Newspaper 17 31
[xx] In Laurenson: New Zealand Professional Conservators Group, *The Code of Ethics,* Wellington and Auckland, New Zealand 1991, amended 1995, p.6.
[xxi] In Laurenson: International Council for Monuments and Sites, *The Nara Document on Authenticity*, Nara, Japan 1995
[xxii] Schwadron, Terry. Preserving Work That Falls Outside the Norm. March 29,2006. New York Times. (Late Edition (East Coast)). New York, N.Y.pg. G.12

[xxiii] Besser, Howard. Many places but currently retrieved in 11/2010 from http://besser.tsoa.nyu.edu/howard/Talks/10stuttgart-media-art.pdf

[xxiv] Dunn, Katharine. Web Archiving for the Rest of Us: How to Collect and Manage Websites Using Free and Easy Software. 2009. Computers in libraries

[xxv] Millman, David and Brian Hoffman. Personal interview. December 7, 2010.
[xxvi] Casey, Carol A. The cyberarchive: a look at the storage and preservation of Web sites. 1998. College & research libraries