Dream the Impossible Dream: Born Digital Stewardship

Bradley J. Daigle; University of Virginia; Charlottesville, Virginia

Abstract

In 2009, The Andrew W. Mellon Foundation funded a research project that sought to steward born digital archives as well as provide a methodology for others to do the same. Born Digital Materials: An Inter-Institutional Model for Stewardship (AIMS) is the result (http://www2.lib.virginia.edu/aims/). A partnership among the Universities of Virginia, Hull (UK), Stanford, and Yale, this international partnership had thee main areas of focus: process eleven hybrid collections, cultivate a new community of digital archivists, and create a methodology that others could adapt to their local practices. This comprehensive methodology broke the complex workflows down into four main parts: collection development; accessioning; arrangement and description; and discovery and access. Each of these parts represents a highly complex and involved set of services. Each partner used the methodology on the various collections that were identified and also consulted with other libraries and archives across the globe. The result is a flexible framework that can be adapted to any level of organization. It presents a detailed decision tree that allows an archivist to work through the daunting task of stewarding born digital content.

This past summer, the University of Virginia Library was able to test much of our methodology in our approach to capturing the historical events related to the resignation and reinstatement of UVa's President, Theresa Sullivan (http://sullivan.lib.virginia.edu/about/). These actions provided a unique and compelling opportunity for the library to demonstrate its leadership in this environment and forge new relationships with units across the university—all working together to provide a comprehensive archive of events as they unfolded.

Born digital materials offer a unique challenge to any organization's digital preservation strategy and infrastructure. For example, there are serious tensions among keeping the original physical media, a forensic or logical disk image of its content, and what is ultimately archived. This content creates a labyrinth of ethics and infrastructure that anyone dealing with born digital materials must navigate. With the Sullivan Archive, we can add to that problem set, the myriad of third party agreements that must be taken into account before any of this content can be made available. In sum, born digital stewardship poses a series of questions that deeply disrupt the role of libraries and archives and the future of the historical record.

This brief paper serves as an overview of what the AIMS (Born Digital Stewardship: An Inter-Institutional Model for Stewardship) Project has accomplished as well as a case study of an historical event that swept over the University of Virginia in the summer of 2012. The main purpose of this document is to underline the need for a vetted and fully tested methodology for managing born digital materials. Knowing that the management and stewardship of digital content is highly iterative, this is not something that can be fixed. Rather, it is one that is revisited often to take advantage of new technology and experience. In many ways, this paper also serves as a cautionary tale for the need for advanced planning and a solid understanding of the roles required for the successful implementation of a born digital team. The world of born digital materials is not a futuristic landscape that exists in the mind of some Hollywood writers. It is the world of the here and now—our current reality.

One of the plain facts of managing archival materials that are both analogue and digital is that they both have significantly different business models. We have been dealing with the world of paper for centuries. Magnetic disks, optical media, flash drives have not been around nearly as long and we have yet to develop a singular strategy for their triage. If this were not complicated enough, ever-evolving trends continue to add new technologies to the trash bin of history and has archives and repositories scrambling to horde obsolete hardware and software in order to unlock the content from their physical constraints. It comes down to how we hope to preserve the historical record when much of this content is digitally created. Consider your own current digital profile. How much of what you would deem as your personal papers is digital? What is on your computer (if you have one), phone, or other portable device? How much of what you would like to pass along to future generations is managed by online services like Facebook, Tumblr, and various individual apps? Do you keep a list of your logins and passwords for your loved ones in case something happens to you? What about your friendly neighborhood archivist? Without such account access for many people, only a fraction of what represents your digital legacy may be available to the future. Some might find that consoling in the world of "the internet never forgets" but in reality, it means that someone else owns your data. That should be a far more disconcerting thought.

Turning back to collecting repositories and the stewardship of born digital materials, creating and maintaining a well-tested strategy for their management is paramount. Why a strategy before a workflow? It may seem obvious but many collecting repositories are doing just that: taking in physical media and other born digital materials without fully understanding what the implications of such actions truly are. This reality is what the AIMS group set out to address.

What were AIMS's objectives? Simply put, the sought to:

- Create a framework for stewarding born digital materials.
- Process fourteen collections that were either born digital or hybrid collections of digital and analogue content.
- Foster a community of digital archivists.

The partnership was among institutions in both the United States and the United Kingdom: the University of Virginia, Stanford University, Yale University, and the University of Hull

(UK). This broad partnership allowed for a diverse group of practitioners to come together and determine whether or not a shared methodology was possible. One of the initial challenges we faced was our organizations' highly disparate approach to managing archival materials in general and the individual ability to manage born digital materials in particular. Each partner's infrastructure posed unique problems to a shared strategy for born digital materials. This first takeaway from our project was dramatic: if the four partners could not agree on a single workflow, how could we expect to create a best practices document that would be useful to the international archival community? From that point, we decided to take broader view of what we were trying to accomplish. Instead of trying to create a single, monolithic, and complex workflow perhaps we could all collectively discover where we were making key decisions and begin documenting those moments. This, as it turns out, was a much more successful approach and provided us with the ability to craft a shared framework that could incorporate the idiosyncrasies of local practice as well as highly disparate infrastructures. As a result, we were able to create a framework that could take into account all of these factors. We broke this work out into four main components:

- Collection Development
- Accessioning
- Arrangement and Description
- Discovery and Access

Each of these sections goes into greater detail—documenting the decision points and issues that the AIMS group encountered. As this article is simply an overview, the entire with paper can be viewed here: http://www2.lib.virginia.edu/aims/whitepaper/ .

To give a brief breakdown of each section, collection development deals with the actions and policies of any given organization's strategy to acquire materials for their collections. These are the necessary steps needed to accept stewardship for and legal ownership of materials from a donor, seller, etc. Collection development policies help guide an organization to acquire either certain kinds of objects or materials centered on specific subjects. A large part of this process would be the early inclusion of a donor survey. There is a detailed sample in the AIMS white paper and it is important to understand that this process helps clarify what the materials might be and how they are disposed. It asks critical questions such as: the creator's work habits; how does the potential donor organize his or her files; what types of digital materials have been created (particularly MIME types); how are they organized; whether the donor possess any mobile devices; multiple email accounts; and general practice when it comes to the donor's digital footprint. One important aspect of the donor survey is to establish the ground rules for what content is to be transferred and/or included in a donor's "collection". This is perhaps the most important step in the process as it sets the parameters for everything else that follows. It helps the archive and the archivist to navigate specific ethical issues such as the stewardship of the physical media. What does that mean exactly? Take this example: a donor is very clear that she only wants specific materials to be part of her archive—say, digital photos, word processing documents, media files, etc. However, since she donated her computer to an archive that has the technological means to scan her hard drive for other things such as web browser history, online site passwords and activities, this too could be programmatically added to the donor's archive. Without proper documentation, this clear violation of a donor's intention is avoidable. There are many careful steps that need to be part of the discussions with any potential donor. These interactions can also guide the donor to work more closely with an archive to ensure the proper capture of the appropriate digital materials.

The second component in the AIMS framework is accessioning. Accessioning has always been a central function for archives. Accessioning actions relate to the organization taking physical and legal custody of the materials. The receiving repository also documents this transfer in the appropriate manner for the institution. In other words, these are the processes that establish physical, administrative, and intellectual control over transferred materials. It can also take into account any assessment and documentation of future needs. Accessioning plays a significant role in the future disposition of the materials. For many institutions, this is also the point where any restrictions that a donor may place on a collection are recorded. This is particularly necessary for planning out the future access strategy for the content and in some cases, restrictions can be more stringent than copyright law. This part of the AIMS framework might also consist of pulling the files off of physical media and transferring them to a preservation environment pending further processing. This would be a critical step if the archive itself has no real means to process the digital materials. At a minimum, the materials themselves have been stabilized.

The third major component of the AIMS framework is arrangement and description, which can be seen as the processes to establish intellectual control of the materials. It also prepares the content for the appropriate level of access. Arrangement and description seeks to preserve the original context as part of that means of discovery. At this stage, any implementation of policies and agreements with donors would take place to position an end user to access the content. It is here in the workflow that the deepest understanding of existing infrastructure is required. The activities related to arrangement and description would be guided by a processing plan and would have to take into account what a given institution's technical abilities might be. In other words, what is the organization's strategy for managing and delivering this content? Creating the metadata is, for the most part, fairly straightforward assuming an adequate ability to read the various file formats. Given the huge range of possible formats, an archivist would need to know whether or not a file format would need to undergo transformation in order to be accessed. This transformation would then need to be confirmed in some manner. Linking content to appropriate rights policies would also be part of this stage as well as a clear understanding of how users will interact with the content. This will be discussed further in the access and delivery section. With respect to description—does your institution have a strategy for making hundreds of thousands of emails searchable? Would this content be added to the general searching from one's catalogue or separated out? How would you manage a collection that has searchable text but by the millions of files? Do you have the ability to check every file for sensitive information? Arrangement and description of born digital materials poses the greatest challenges to an archivist. At the time of writing this article, the archives world still lacks a comprehensive software environment to do this work adequately.

The final component of the AIMS framework is discovery and access, though this is by no means last in importance. In fact, all other stages that lead up to access should take an institution's ability to make content available in mind. This stage refers to the systems (hardware and software) and workflows that make collection materials and their metadata available to users. A solid understanding of what this entails will inform most of the processing of collection materials. In other words, what is your institution's access strategy for born digital content? Does it:

- Create emulation environments that show the digital materials within their original hardware and software environments?
- Transform digital objects for sustainable access. In other words, provide the materials in an updated format (e.g. migrate from a WordStar file to a RTF file). If so, does is still allow for access to the untransformed original?

Does the institution expect to provide all the commensurate functionality for born digital materials or does it expect to take advantage of web services developed outside of your infrastructure? These are all questions related to the relative functionality of digital content. There are an equal number of questions that relate to the rights and intellectual property issues related to the materials that must be part of any access and discovery framework. Can you properly restrict content to the appropriate, authenticated user? Does it need to reside in a different system or architecture in order to do so? Can the materials only be viewed on site or can they all be accessed remotely? Do you share basic metadata with users even though the content is restricted? These are all simple questions that have highly complex answers and undoubtedly a profound impact on the infrastructure required to mange the content.

This brief overview of a series of highly complex issues is more deeply explored with examples and case studies in the AIMS whitepaper. It does underscore the necessity of having a strategy in place before any kind of born digital stewardship can occur. Otherwise there is the risk of potential data loss, donor displeasure, and other forms of mismanagement. It may not be possible to have the entire framework in place but executing a few simple steps in advance can avoid a huge amount of future difficulties. This perhaps, can be best expressed by means of example.

The Sullivan Archive

In June 2012 the University of Virginia (UVa) experienced a series of events that were soon to embroil much of our university as well as most of higher education. The President of UVa, Theresa Sullivan, resigned. Though that does happen at American Universities occasionally, this event was punctuated by contradicting statements and reasons, political backlashes, and, ultimately, a grass roots movement to reinstate her. Founded in 1819, UVa is no stranger to historical events. However, what was markedly different about this situation was the predominance of activities taking place on social media. The debates that followed were carried out both in physical and virtual locations. Rallies were organized and carried out on campus grounds as well as on Twitter, Facebook, and blogs. While our special collections staff quickly began to gather physical artifacts from the rallies: posters, pictures, and other physical objects; the library was slower to realize that much of what was happening in the digital realm was not getting captured. If this oversight continued unchecked, we would not have a full picture of what transpired.

As it turns out, capturing those digital materials posed many more problems than we at first anticipated. Here are some of the major issues that resulted: we had an unclear view of who the main players were who needed to be involved in capturing the digital component of these historical events; we also had no idea there would be so much material that had to be captured; and finally, perhaps most critically, realized that much of the historical record was taking place on sites that required individual accounts and logins. As a result, we had to quickly convene a stakeholders group to identify roles and responsibilities. We had members from our special collections, digital curation services, library information technology unit, university records, and general counsel. In addition to that administrative group, we also enlisted aid from key faculty and faculty groups across the university. The data could be divided into two main groupings:

- University records—materials related to how business is transacted by university employees (memos, emails, other documents)
- Social media—materials related to individuals or groups that function outside of university control (Facebook personal and group sites, Gmail, blogs, etc.) Some of these materials are public (e.g. public tweets) some is private or require a specific users' credentials.

These are gross categories, yes, but they serve to show some basic division of content. Different strategies were required for each type of content. Given that we have robust records management processes in place, the group decided that gathering materials that fell under that rubric was not of immediate importance. We knew that we could circle back afterwards and harvest that content. The second category was where we focused our energies.

One of the most important realizations we had related to social media is that each external site that had content we desired was governed by very different terms of service. So in some cases, even where we had permission from the account owner, we could scrape the content we wanted but would be unable to deliver it any way based on the licensing. Or for others, even doing the scraping would be a violation. Here is a specific example: many individuals were uploading videos to YouTube but according to the terms of service for that site, one is not able to download copies of these videos and redistribute. Therefore, when we were asking individuals for content, they would simply point us to third party sites where basic archiving would be a use violation on our part.

As these issues were arising, we discovered another roadblock to moving forward. Given that we are a public university, much of the content that would fall under university records could be made available under the Freedom of Information Act. That meant that members from our University Records unit and General Counsel were swamped with processing such requests and had limited time to delve into the complexities of third party license agreements. In addition to that hurdle, we elected to put up a "user-contribution" site since so many individuals now have portable devices to capture events in real time. We wanted to provide a means to capture that user content. Therefore we enlisted the aid of one of our developer teams to mount an Omeka site. This is an intuitive exhibition site that is very lightweight. (http://sullivan.lib.virginia.edu/about/). However, since we did not have one ready to go, there was some delay before it was made live. As a result, we lost several key days' worth of content. It went live on the day Sullivan was reinstated and hence the amount of user-generated content was greatly reduced as events ceased. Of particular interest to us were all the ephemera within the social media sphere. In particular, there was a huge influx of Twitter "spoof" accounts—ones that were quickly identified and shut down by Twitter. Unless you were capturing tweets in real time, these would be lost and were an integral part of the social media conversation. Other questions centered around local news media as well. Should we grab copies of what they broadcast or do we just assume that that content will be around and remain accessible to users forever. Finally, with respect to blog posts, when do you choose to archive? Given that a lot of the conversation happens in the comments, once you harvest the posts, you don't get any future comments. Do you bookmark and hope the post is still available after a given period or do you just note in the metadata when you stopped harvesting. There are no easy and unilateral answers to these questions. They have to be taken on a case-by-case basis. In the end, we gathered a total of over 80,000 tweets, 572 news articles, 147 blog posts, 243 Twitter pictures, 69 videos, 21documents and 118 user contributions. All told given the short duration of the historical events, this is large amount. The total collecting period was less than two weeks. UVa Library is currently evaluating what our options are for managing and creating access to this content.

Given what both the AIMS partnership as well as the Sullivan events, there are some critical lessons that we have learned and bear repeating. It is paramount that every collecting repository has an articulated strategy for managing born digital materials. This needs to be done prior to accepting any digital content. Even if there are key elements that are missing, that knowledge of what is needed is crucial. Here is a brief overview of key points:

1. Take some time to develop a framework that works for your particular institution.
2. Identify the key stakeholders in that process and the roles associated with each.
3. If you already have born digital media, do an inventory. If possible, and based on your methodological framework, move the files off of the physical media.
4. Understand that history is happening now. Have a social media strategy codified by your decision makers—*in advance!* Don't wait until you need to have one.
5. Map any of these activities to existing infrastructures whether they are intellectual property, user access, or preservation.
6. Talk to you donors as early on in the process as possible. They may need to change their behavior for digital object management. For example, saving versions of their documents rather than overwriting the same copy.
7. There is no normal. There is very little predictability in the world of born digital materials.
8. Revisit your strategy often. Roles change, people come and go. Like all good policy—don't let it get out of date.

These events taught us a great many things. I cannot recommend enough treating social media the same way we approach the stewardship of our physical collections. Both need plans formulated and tested in advance. Think of your social media strategy as a born digital disaster plan. You do not want to miss out on key events because of logistical uncertainties. Ultimately, we do the best we can but these are ways to stack the deck in your favor. Here at UVa, we grabbed as much content as we could in advance, knowing that we would have to work through the logistics of delivery and rights later on. Like the Sullivan event, there will always be such historic activities where you are. These moments and exhilarating and terrifying at times—your best bet is buckle up!