Library collaboration with large digital humanities projects

William A. Kretzschmar Jr and William Gray Potter
University of Georgia, USA

Abstract

The sustainability of digital humanities research projects is a pressing issue for humanities computing. Currently, even well-established large digital projects like the Linguistic Atlas Project (LAP) are at future risk because funding and other resources are contingent on grant funding or faculty status of the director, neither of which will necessarily be available to maintain the project over time. The mission of the university library, however, includes archiving and dissemination, now increasingly of digital materials as well as traditional paper. Collaboration with the university library is the only realistic option for long-term sustainability of digital humanities projects in the current environment. Unlike paper collections, which only require secure storage, digital projects also require the means of adaptation to new electronic media and operating environments. Even data storage requires that materials from digital projects be included in library media refresh cycles, which will include transfer of old data to new media as technology develops. Projects like LAP should provide resources to assist the library in starting the project archive, including staff time, and funding for equipment. Project metadata must be provided and, to the extent possible, integrated with library systems and finding aids. Project staff will also need to maintain a Web presence and tools developed for the project. Such cooperation leads toward the development of a digital institutional repository, in which research results and tools may be maintained in the library, not just for the humanities but across many disciplines.

Correspondence:
William A. Kretzschmar,
Department of English,
University of Georgia, 254
Park Hall, Athens, GA
30602-6205, USA.
E-mail: kretzsch@uga.edu

The phrase 'if you stand still you die' has nearly become proverbial in American English. The typical use of the phrase refers to business. In his famous 1970 book called *Up the Organization*, Robert Townsend (1970) wrote that 'Almost everybody subscribes to the myth that a company has to keep growing. "If you stand still, you die," they say. I don't know which idiot first carved this imperative on the tablet' (p. 151). These days, it is possible to find business users of the phrase who agree with Townsend, but also many writers who accept the idea that you have to keep moving in business to stay alive—perhaps now more than ever. The phrase has now become popular, too, in reference to computer role-playing games where it is literally true that, if you stand still in the game, you die. It turns out that this phrase also refers to digital humanities projects. At DH2007 a special session on 'finishing' large humanities research projects (now a cluster in *Digital Humanities Quarterly*, 2009) suggested, in part, that particular stages of such projects might be completed, but that

doi:10.1093/llc/fqq022 Advance Access published on 19 October 2010

continuing institutional support was important for the long-term sustainability of the projects and their products. At DH2008 a special session was devoted to 'Aspects of Sustainability in Digital Humanities,' in which technical, organizational, and scholarly dimensions were discussed with reference to a museum project, along with metadata and the issue of portability in other settings. In this article we would like to continue the theme of sustainability. We will discuss issues of institutional support for a large digital humanities project and then propose that collaboration with the university library is the only realistic option for long-term sustainability in our environment. If digital humanities projects stand still, they will indeed die, and the library is the only part of our institutional structure that can keep them moving enough to save them.

The stand-still-and-die problem for digital humanities projects is two-fold. Our first big problem is that our digital environment keeps changing around us and shows no sign of stopping. The invention of printing on paper provided a technology with which information could be stored over the long term. We have nothing like that for digital storage. We have progressed from punch cards to tape to magnetic and optical spinning disks to non-volatile flash memory, all in a few short decades. Moreover, the operating environment in which these media are used has also been changing and at any one time there have been multiple popular operating systems in common use. For digital humanities, we therefore face the problem of continually having to change media and operating environments just to keep our information alive and accessible. Our second big problem follows from the first; we need to have access continuously to new resources, not to create new programs or products, but just to keep up with changes in media and operating environments. Even the seemingly most secure digital humanities projects, say icons like the *Oxford English Dictionary* or the TEI Consortium, face difficult budget challenges—the Arts and Humanities Data Service, another central resource in digital humanities, has now actually lost its funding (ahds.ac.uk). The fact is that most digital humanities projects, even famous ones, are managed by just one developer, or by a small working group, with inconsistent and unreliable funding. We do not have access to grant funding in the same way that our colleagues in the natural and physical sciences have it, with renewable grants that can keep laboratories running for long periods. The sustainability of a project relies on continuity of both the developer as a human resource and funding as a material resource. Of course, the stand-still-and-die problem includes the certainty that the original developer of any project will be lost at some point, whether to a change in interests or research focus or to retirement or even death. The obvious implication for sustainability is that somebody else besides the original developer will have to keep the digital project alive. Our digital projects will die without adaptation to changing media and environments, and the adaptation will not happen without continuing sources of new financial and new human capital.

Let us turn now to a particular example of a digital humanities project in its institutional setting. The Linguistic Atlas Project began in 1929, and its digital conversion started in the 1980s, first at the University of Chicago and then from the mid-1980s at the University of Georgia (http://www.lap.uga.edu). We have been interested to digitize our large archival body of interview data, on paper and audio tape, at the same time that we continue to collect new information about American English in native digital format. As treated in some detail in *DHQ* (Kretzschmar, 2009), the Atlas has gone through several generations of programming and computer tools, each of which required special funding. The Atlas is lucky in that it has a small endowment. However, the greatest continuing resource for the Atlas has been the faculty contract for its developer (Kretzschmar), which has also provided office space. The need for funding has become ever greater as we succeeded in our digital stages—the work exceeded the capacity of a single developer and Kretzschmar had to spend more and more time soliciting grant funding just to keep the operation alive. Under such circumstances, which we believe to be typical of successful projects, without the faculty contract and continuing fund raising, the digital Atlas would die.

Then, five years ago, an institutional option emerged. When Georgia created a Research Computing Center (RCC) as a service in addition to the institution's other computing resources, the Atlas was one of the first clients so that we could avoid the continuing expense of system administration. However, over the course of several years the RCC changed from an essentially institutional budget to a fee-based service for researchers with annual external funding. This meant that humanities projects like the Atlas without consistent external funding, while not excluded from the RCC, had to hope for sufferance from the paying customers, our friends the biologists, chemists, and physicists. This option did not appear likely to sustain our Atlas digital products and archives over the long term.

Enter the University Library. At the University of Georgia, as at most research universities, the Library is faced with the problem of archiving materials in many formats, ranging from paper to film to a wide range of digital formats to ensure that the content is preserved for future generations. Among these, Georgia has a large collection of film and magnetic tape, built around the entries to the Peabody Awards, one of the premier awards in broadcasting (now rather 'in electronic media,' as the term *broadcasting* is becoming dated). The Library has developed a plan to archive and care for this collection that involves creating the best possible digital archival copy of an object, usually a radio or television program, while also deriving a viewing copy at the same time. The digital archival copy is then stored on offline magnetic storage. The best available technology at this time is Linear Tape-Open (LTO), a cartridge of magnetic tape that resembles an old eight-track cartridge and that adheres to an open standard supported by members of the LTO Consortium, which includes Hewlett-Packard and IBM. LTO replaces the many proprietary formats previously in use, which each required specific proprietary hardware, software, or both to recover data from the medium. The LTO cartridge has remained the same in appearance and function, but the capacity has grown over the years. LTO version 4 can now hold up to 1.5 Tb of data per cartridge. As the format evolves, the capacity of the cartridges will continue to grow. At some point, they will be replaced by another format but we are confident that an upgrade path will be provided given the dominance of this tape format.

The physical and paper archives of the Atlas are currently stored in the Library's special collections facility. Therefore, it was fitting that Potter, the University Librarian, offered to incorporate the multimedia collections of the Atlas there. With the entries to the Peabody Awards and other materials, the current media archive consists of over 100,000 radio and television programs. The current 20 Tb of audio files in the Atlas archive do not compare in scale to what the Library needs for the Peabody materials, and even our hoped-for addition of even larger collections of image files made from the Atlas field records on paper (perhaps 100 Tb) will not put that big a dent in the Library's multimedia archival space. Unlike the RCC devoted mainly to high performance computer processing, the mission of the Library includes both archives and dissemination, now increasingly of digital materials as well as traditional paper. We have now agreed on cooperation not just for a multimedia archive, but also for dissemination of Atlas products and information through Library facilities in the context of continuing scholarly activity. We hope to convey here the central points of our agreement in terms of the sustainability problems we have already highlighted: How will we deal with changing media and operating environments? Who will pay for it? And who will do the work?

We have agreed to create a permanent digital archive for Atlas materials. Our first decision was to follow the Library's plan to base its multimedia archive on LTO-4 computer tape. The current price for an LTO-4 cassette is about $80, much less than equivalent storage at the RCC. However, it is necessary to refresh tapes periodically: LTO-4 tape is rated to last up to 30 years, but the tape drive mechanisms will only be usable for about 10 years. The Library has agreed to include Atlas tapes in its regular tape refresh cycle. Moreover, the Library will have the opportunity across refresh cycles to update its technology, so that Atlas data can be copied next time on newer tape drives, or to another medium that we have not yet imagined. The refresh cycle will also need to accommodate changes in operating environments; the data on the tapes can be converted to new formats as required when the tape is refreshed. Thus the Library can provide both the new technology and the continuing resources for a sustainable archive as part of its mission at the university.

The Atlas project, on the other hand, must provide the resources to start the archive. Owing to our need to protect the privacy of research subjects, digital copies of original interview sound recordings will be preserved in a 'dark' archive, not for public access. Each recording will also be preserved in a public version, with sensitive information edited out. This practice mirrors the Library's plan to create two copies of multimedia holdings, one archival and one for public access. Original LTO-4 tapes of both the dark and public versions will be created in the Atlas office for deposit to the Library multimedia archive. Atlas staff will manage distribution of the public data, so as not to add work for Library staff. The Atlas also has grant funding at the moment to help procure some of the archival equipment needed in the Library. In this way the Atlas and the Library share the burden on resources; while the Library provides long-term sustainability, the Atlas can contribute human and financial resources during the times that it has them. The archive focuses on content, something that can be sustained without use of Library staff beyond the archival refresh tasks that they would normally carry out.

Another aspect of sustainability for Atlas content is management of our metadata. The Atlas keeps metadata about our research subjects, and we also keep operational metadata about the processing we do. On our existing Web site, we maintain information about many regional Linguistic Atlas projects, and we maintain highly interactive information about the Middle and South Atlantic States (LAMSAS)—a model for the new site that we will build at the Library. For recordings of speakers to be located on the Library site, we will keep track of audio processing performed on each interview in addition to the information we keep on the LAMSAS interviews (which exist only on paper). Again, while this kind of metadata is important for research purposes, it remains beyond the scope of the Library's mission. It will remain the responsibility of the Atlas staff to generate and store these kinds of metadata, and to create means by which users can find the Atlas data they want to use. A third kind of metadata is required for participation in the Open Language Archive Community (OLAC), the international effort to collect and share Dublin Core metadata about language archives. Our participation in the OLAC network will be a primary means of providing Web notice of our collections. Notice of the Atlas materials will be also integrated with records of other Library electronic resources. The burden for creation and maintenance of in-project metadata will remain with the on-going research project, not the Library, and we will cooperate on integration of externally directed metadata. Again, inclusion within the Library system meets the need for integration, by redirecting work that we normally do for the Atlas without imposing an extra burden on Library staff.

Finally, we have agreed on terms for creation of an Atlas Web presence on Library servers. The existing Atlas Web site, which has already been transferred from the RCC to a Library server, is heavily interactive for the LAMSAS materials (http://www.lap.uga.edu). Users can make maps and conduct complex searches of our data, as well as download data files. One problem with any interactive Web site is the need to work on the scripts or programming that enables the interactive functions. It is a different kind of problem entirely just to provide system administration, to keep the site current and secure with patches and updates. Having the Atlas site at the Library will solve the second problem but not the first. There is already a Digital Library facility at Georgia that carries out its own digital projects, but those projects have to solicit their own external resources, just as the Atlas has to solicit external funding for every digital advance. The Library just does not have IT staff sitting around waiting to take on programming work for faculty projects, whether for new projects or programming maintenance on existing projects. We are aware, for instance, that at one time there was a plan for the well-established Carolina Digital Library and Archives (CDLA) at the University of North Carolina to receive and care for faculty digital research projects. But CDLA, like our own Georgia Digital Library, relies on external research funding for its operations, and it turned out not to be practical just to add this new function to its agenda. There is now a pilot project to create a new, separate Carolina Digital Repository as part of their Systems Department. The Georgia Library also maintains its own servers and so, like the Carolina pilot program, we plan in the first instance for the Library to provide system administration as the host for the Atlas Web site in order to provide persistent access to the Atlas materials. This common experience demands that we make a strong distinction now between content and tools; the Library can archive content, but its staff cannot at present take responsibility to maintain tools, such as an interactive Web site that makes content and visualizations of that content available to users. We expect that Atlas staff will have to maintain our Web site and any tools we make. To that end, we should build our new site in what seems to be the most sustainable way possible, which to us means separation of basic Web access, as for file downloads, from more highly scripted interactive functions like GIS for mapping.

Currently, the Atlas has engaged in a very successful collaboration on the creation of a user interface with faculty and staff members at the University of Oulu (Finland). The LICHEN system (http://www.lichen.oulu.fi), written in Java as a toolbox for different software tools for display and analysis of multimedia information, has already been implemented for Atlas data in the *Digital Archive of Southern Speech* (DASS), a collection of sixty-four interviews amounting to about 400 hours of audio files distributed on portable USB drives. We have agreed that the modular, self-contained LICHEN system will also form the basis for a user interface on the Atlas Web site at the Library, alongside a simpler, separate system just for download of files from our file structure. There will be times when the Atlas does not have resources to be very active in site building and maintenance, and knowing that, we must prepare for it, by creating the most stable forms of user access in addition to interactive tools. For the Atlas, this means creating a strictly defined file structure that can be accessed by minimal standard HTML means as well as by much more advanced interactive tools.

The Library can offer us a very convenient and effective Web service environment. We plan to use a Virtual Machine (VM) instance for the site, so that anything we do for the Atlas site will be quite separate from other Web services of the Library. This means that we will be able to create logins and means of access limited to the VM, so that we do not create security problems in the Library systems. We could theoretically install whatever operating environment and software we wanted in the VM, but we still plan to integrate our environment with what the Library is already using. Thus, the Library has not implemented Flash server before, and so we will not use Flash server to build our new site. The Library has had problems with security with one or two kinds of software, and so we will not use them. The Atlas can, however, install specialized software when the need arises, such as the LICHEN toolbox. And when we do have the resources, as with our current grant funding, we can offer the Library access to some external funding to acquire server equipment or software on which the Atlas site, and potentially others as well, can be run.

The current solution for the Atlas and the Georgia Library involves archiving the data in a way that can be refreshed and maintained indefinitely for persistent access. However, as we have suggested, providing the tools to access and use that data 50 or 100 years from now presents a more cumbersome problem. The Georgia archive of radio and television programs is less of a problem because they fall into a handful of formats and can be used by a variety of general viewing software as long as they are maintained in a readable format. However, research data often relies upon specially developed tools and, in many cases, ad hoc solutions that are not easily maintained. In the case of the Atlas, the audio files will be maintained in a standard format that should be easily accessed. However, the analysis tools and the presentation of the data will be difficult to maintain if the Atlas office at

Georgia were to cease to exist. Although our solutions in the first instance have not addressed it, the Library, as the archival agent of the University, has an interest in maintaining not just the data but the tools to use that data.

Many libraries and universities have established institutional repositories (IRs) as a way to gather and make available the results of research conducted at their institutions. (http://www.arl.org/sparc/repositories/). IRs collect, organize, and archive the intellectual output of the University (Lynch 2003). The focus has been on documents prepared to report the results of research. Doctoral dissertations are usually included, as are conference proceedings, working papers, pre-prints, and other materials that have long been generated by Universities but never adequately organized and maintained. In recent years, IRs have become a place where faculty can place copies of published papers to make them available beyond the journal in which they were published. Authors often reserve the right to do this or are required to do so by the funding agency, such as the National Institutes of Health (http://www.arl.org/sparc/bm~doc/NIH_Copyright_v1.pdf).

While the focus has been on using institutional repositories to distribute the results of research, they can also be used to store and disseminate other digital objects, including the data upon which the research is based (muse.jhu.edu/journals/library_trends/v057/57.2.witt.html). Data set curation is thus the next stage for institutional repositories (the University of North Carolina has offered IMLS-sponsored training in digital archive curation, http://www.ils.unc.edu/digccurr/fellows.html). The next and necessary stage beyond that is to preserve the tools used to access and manipulate the data sets. This becomes most important in the humanities where the data and the access tools are often specific to the project.[1] With the Atlas project, for example we would need to ensure that the Web site and its associated applications like LICHEN are maintained over the long haul of decades. To do otherwise risks the loss of the scholarly context for the data that has been accumulated for the project. We do not yet have a solution for the curation of tools by the Library, but that does not mean that we can afford to drop the issue.

As we have presented the different aspects of our cooperation, several themes have emerged. First, we want to highlight the central importance of integration. In order to be the most sustainable, we have to align the digital presence of the Atlas with the mission of the Library. Instead of being free agents with our own servers and software, the Atlas and other digital humanities projects should plan for integration with the Library as a host for Web services and as the archival site for content. We have raised the distinction between content and tools, because the implications are quite different for the resources to sustain them. Information should last forever under the care of Library staff, but particular tools, for now, can only be temporarily maintained given the inconsistent access to resources in the digital humanities. Finally, we have to respect what the resources of the Library and the Atlas will actually bear over the long term. The Library has a continuing budget for its mission, and so it is appropriate to fit the Atlas into such activities, but whenever the Atlas does have resources, we should use those to supplement what the Library has and to perform the kind of work, like tool building and maintenance, which so far is beyond the Library's archival mission. We think that the answers to the questions we have posed lead us toward a model for an institutional repository, a means for many faculties across the university to store the fruits of their research for the long term. We know that persistent access to Atlas materials will require the kind of integration, the content/tools distinction, and the resource sharing that we have addressed here, and we think that any institutional repository will require them as well. If we can do these things, digital humanities research does not have to stand still and die.

References

Digital Archive of Southern Speech. (2009). Athens: Linguistic Atlas Project, American Dialect Society. [200Gb+, 400+ hours of 64 digital audio interviews sampled from the Linguistic Atlas of the Gulf States, with LICHEN finding aids; released on portable USB drive].

Digital Humanities Quarterly (*DHQ*): Special Cluster: Done. http://www.digitalhumanities.org/dhq.

Kretzschmar, W. A. Jr (2009). Large-Scale Humanities Computing Projects: Snakes Chasing Tails, or Every End is a New Beginning? *Digital Humanities Quarterly* 3 n2 (Spring 2009), http://www.digitalhumanities.org/dhq/.

Linguistic Atlas Project: http://www.lap.uga.edu.

Lynch, C. A. (2003). Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age. *ARL: A Bimonthly Report,* no. 226 (February 2003): 1–7. (http://www.arl.org/resources/pubs/br/br226/br226ir.shtml).

Townsend, R. (1970). *Up the Organization: How to Stop the Corporation from Stifling People and Strangling Profits*. New York: Knopf.

University of Georgia Library: http://www.libs.uga.edu.

Note

[1] In the sciences, efforts are underway to standardize data gathering and reporting so that results can be packed, unpacked, and analyzed many years hence. To some extent, this is also true in the social sciences where data gathering is usually built around a few standard packages, usually statistical software, and where there are organizations like The Interuniversity Consortium for Political and Social Research (ICPSR) to centralize and maintain data sets (http://www.icpsr.umich.edu). Efforts in the humanities are not as well advanced. The TEI Consortium (http://www.tei-c.org), of which Georgia is a charter member, is one example of standardization of text archiving practices in the humanities, but TEI does not address the problem of tool maintenance.