AnthroDataDPA Report

Anthropological Data Digital Preservation and Access (AnthroDataDPA)
Report from an NSF/Wenner-Gren supported workshop, May 18-20, 2009

For more than a century, anthropologists and other observers have been collecting data about the human experience. These data include the details of human history, the characteristics and evolution of the human species and other primates, the variety of languages spoken and written, and the cultural features of the world's societies. Unfortunately, many data already have been lost to us and will not be available to future generations. Failure to record data properly, failure to store it appropriately, and failure to sustain our ability to "read" the data with changing technological platforms are the principal causes of data becoming compromised or lost. The profession is concerned with the possibilities of using new digital and Internet technologies to save anthropological data – in archaeology, biological anthropology, cultural anthropology, and linguistics. If we are successful in this great enterprise, we can stop such information about our cultural heritage and human biodiversity from being destroyed, lost, or so poorly maintained as to be worthless to future generations of scholars and communities in the U.S. and around the world.

On May 18-20, 2009, a workshop was held in Arlington, Virginia to evaluate and potentially decide on the basics of a strategic integrated four-field plan for digital preservation of and access (DPA) to anthropological research materials (AnthroDataDPA for short). The workshop was funded by the National Science Foundation (NSF) and the Wenner-Gren Foundation in a grant to the Human Relations Area Files at Yale University1. The principal investigators, Carol R. Ember, Eric Delson, Jeff Good, and Dean Snow, each respectively represented one of the four traditional subfields of anthropology—cultural anthropology, physical anthropology, linguistics, and archaeology. Three groups of people participated: 1) individuals actively involved in and/or planning the creation of digital object repositories for anthropological data; 2) individuals from institutions involved in the creation of relevant international standards and metadata to enhance interoperability and long-term preservation; and 3) representatives of organizations that represent the various fields of anthropology in the United States. Also attending were observers from political science, NSF, NEH, and Wenner-Gren (view attendees). We had nine breakout groups at the workshop. Each breakout group was charged with discussing key issues and then their discussion was summarized by the breakout chairs. After incorporating points raised in the discussion period, the chairs put together reports of their breakout groups . The following is a summary report of the workshop put together by the PIs.

This overview lays out our vision, goals (both long and short term), general principles (or strategic decisions), as well as more specific issues and concerns. The report concludes with some possible next steps for the anthropological community to begin to comprehensively address DPA issues. The PIs are moving forward on applying for grants to continue this effort. In the meantime, we encourage those of you who want to digitize your data to follow some of the guidelines in this report.

Vision Statement

Our overall goal is to stem the tide of the loss of precious anthropological data comprised of qualitative and quantitative research materials, both digital and non-digital. We aim for a discipline-wide plan for

digital preservation and access (DPA). This includes gaining acceptance in the anthropological community for common reference standards and metadata.

Long-term Goals

- Advance digital re-use and interoperability of data within and between the four broad divisions of anthropology to encourage integrative research.
- Stimulate future continuity, adoption and cumulative improvements of DPA by developing open-source tools and online services that build on state of the art technologies to assist anthropologists in applying accepted DPA standards for legacy conversion and future data ingestion.
- Establish a community of anthropologists engaged in finding solutions to digital preservation and access problems in anthropology
- Establish a network of trusted digital open-access archives for the anthropological community designed for interoperability and long-term preservation following recommendations of a anthropological standards body.
- Integrate individual "silo" projects of digital data preservation and access both within and across the subfields.

Short Term Goals

- Promote the outline of a strategic plan through the web, list-servs, newsletter reports, conference papers, liaising with anthropological organizations and with additional organizations engaged in or planning DPA activities.
- After obtaining commentary and suggested revisions, post a revised plan online in the spring of 2010.
- In the absence of a coordinated network of trusted digital archives, promote better practices for digitization and preservation using existing resources.
- Apply for funding to advance this agenda.

Data and Metadata

Two terms are essential for understanding this report. The first is data; the second is metadata.

Data

There are many types of anthropological data. The table below represents some of the most important types:

Type	Examples
Images	Photographs, maps of excavation sites, biomedical images (e.g., radiographs)
Texts	Field notes, annotations, excavation plans, manuscripts
Audio	Recordings of songs, conversations, oral histories
Video	Recordings of cultural events, conversations, archaeological excavations
Databases	Database of measurements, lexical items, locations
3-D scans	Scan of fossil or artifact

Table 1. Common types of data in anthropology

Any of these types of data may be stored in digital form. In the broadest sense, digital "data" are thus simply electronic coded forms of information. For anthropological purposes, a more pragmatic definition of data are measurements, observations or descriptions created or collected by a researcher. Different subfields vary widely in the types of things described (referents). For example, in cultural anthropology, the units or referents might be observed events, informant interviews, households or communities. In archaeology, the units might be settlements, quadrants of a grid, or artifacts. In linguistics the units might be lists of vocabulary items, recorded and transcribed texts, or grammatical patterns. In physical anthropology data units might be measurements, character states, scans, images, or even the fossil on which those were taken, genetic sequences or bases, behavioral observations, sonograms, phenological observations, or radiometric dates.

Throughout this report, for purposes of exposition, we will assume that anthropological data are collected by anthropologists, given that the anthropological community is our intended audience. However, many of the points made here will center around the digital preservation and access of anthropological data generally, whether collected by professional anthropologists or coming from some other source.

But storing data is not sufficient without the preservation of their context. Attention to metadata is essential to DPA.

Metadata

Metadata are comprised of descriptive documentation essential to informing the process of data creation, collection, management and preservation. Metadata provide information about the original referent, the collection processes, rules of collection, as well as descriptions of data management processes and provisions for access and use of the data (such as licensing of data to specify permitted uses). Metadata provide key contextual information to facilitate understanding and are intended to assist research within known and predictable scientific domain(s). As research questions in anthropology evolve, metadata may also enable discovery and use of archived data in as yet unanticipated fields of research. Thus, careful effort should be made to make the descriptive content of metadata intelligible to scientists beyond a very limited scientific expertise. Because new technology allows for reuse and expansion of archived data, as well as the creation of new persistent tagging, metadata creation is an ongoing process not a single event, metadata usefully may grow over time by accretion, asynchronously, by the efforts of properly qualified contributors. We anticipate that new data will be linked to older archived data through a continuous process that updates metadata and creates new metadata to inform evolving and expanding datasets.

Digital Preservation and Access (DPA)

There have been great strides made with regard to creating digital object repositories—that is collections of different kinds of digital content—and moving toward interoperability between repositories outside of anthropology. It is prudent to build on rather than reinvent these developments. The best way to do this is to work with experts who are familiar with the accomplishments from these fields. For a review of some of these efforts, click here. For a review of developments in anthropology see below.

Why DPA is vital to anthropology

1. Background materials provide the context for understanding the research undertaken, whether qualitative or quantitative research. The appropriate analog is the "lab notebook" in the physical sciences. These are critical for evaluating published research. But other information about the observer is also important and certainly critical for evaluating any biases. So, preservation of any associated materials (dairies, correspondence, etc.) is also of intellectual value.

2. Physical archives have only stored a very small portion of the anthropological corpus. For example, Robert Leopold of the National Anthropological Archives estimated that 500 anthropologists retire each year, but the NAA only acquires 6-8 major collections each year[1]. And universities, with limited funding, always make choices about which collections they will take and process. Participants in the workshop on which this report is based speculated on why potential donors have been reluctant to give their materials to archives to date (click for details). Understanding these reasons may suggest how digital preservation may play an important role in future preservation efforts.

3. Many of the anthropological data now being accumulated are "born-digital" and physical repositories will find it difficult to preserve this material in a form that will be accessible in the future. It will be necessary to migrate date from old formats to new ones over time. It is likely that new tools will be invented that will allow updates and data migration to be managed automatically by repositories.

4. Digital preservation can lead to more open access and to productive repurposing of old datasets. Legacy data are particularly important in all subdisciplines of anthropology. Exceptions are to be found in techniques such as three-dimensional modeling and scanning, where researchers are likely to prefer new scans over archived old ones. However, this presumes that the specimens will be preserved for reanalysis as necessary. In cases where the original specimens have been destroyed or are inaccessible, archived scans might be the only option available.

5. Access increases research potential

Background in Anthropology

In anthropology, digital preservation of scientific data is a relatively new enterprise, but as early as 2001 plans were underway to create distributed digital archives of anthropological material[2]. Table 1 above lists the various types of anthropological material that lend themselves to preservation in a digital archive.

Anthropology has taken some steps to encourage scholars to preserve research data. For example, the American Anthropological Association, at its annual meeting in November 1968, adopted a resolution urging the preservation of anthropological field materials and consideration of the National Anthropological Archives as a suitable repository for materials not committed to other institutions. The need for preserving the anthropological record was clearly stated in 1992 when the Wenner-Gren Foundation sponsored a symposium, "Preserving the Anthropological Record[3]" . Papers discussed existing archives, preservation issues, and issues of how to preserve and archive the records. The results of the symposium included the passing of a number of resolutions and the creation of the Council on the Preservation of the Anthropological Record (CoPAR). This council meets at the American Anthropological Association, has workshops, and from time to time posts bulletins on the Smithsonian Institution web site. As of 2005, the NSF programs in archaeology and physical

anthropology as of 2005 require detailed plans for data sharing as a condition of funding, and NSF's Documenting Endangered Languages program has instructed applicants to discuss plans for archiving data since its inception in 2004.

Some DPA and interoperability efforts have already been initiated in the individual fields of anthropology. Perhaps linguistics, physical anthropology, and archaeology have talked more about interoperability than cultural anthropology, but there have been no large-scale accomplishments within each subdiscipline and no overall anthropological efforts.

Umbrella digital projects in linguistics include: The Open Language Archives Community; the Rosetta Project; archiving and tool development activities within the DoBeS Project; the Digital Endangered Languages and Musics Archive Network and associated archives; the Hans Rausing Endangered Languages Project; the Linguistic Data Consortium; TalkBank; a range of projects associated with the Institute for Language and Information Technology, including the E-MELD project and the GOLD Community project, the latter of which sought to enhance interoperability of linguistic data by creation of a formal ontology. In addition, NSF recently funded recent Cyberling workshop, whose goal was to lay the groundwork for the development of a unified cyberinfrastructure in linguistics. Many of these projects have been developed in the context of a rising concern in the preservation and dissemination of data from endangered languages. 4

In physical anthropology, the major digital projects focus primarily on primate morphology and the fossil record including Paleoanthportal with constituent databases called PRIMO—Primate Morphology Online Database, and HOD–Human Origins Database; RHOI—Revealing Human Origins Initiative, an NSF HOMINID project; and NESPOS—Neandertal Studies Professional Online System. For behavioral data there is the Primate Life Histories Database. Finally, there are a number of large biomedical databases that are becoming critical resources to physical anthropological research. These databases include GENBANK, ALFRED — the ALLele FREquency Database as well as dbGaP.

In archaeology, the major digital projects are: Chaco Digital Initiative in cooperation with the National Anthropological Archives; The Digital Archaeological Record (tDAR), which is the core element of the Digital Antiquity Project; ArchaeoInformatics; ArchSeer, a specialized archaeological search engine; ADS (Archaeological Data Service), an on-line service of York University.

Cultural anthropology is characterized by many individual "silo" digital projects, many self-created and others part of university efforts to digitize faculty material. Some of the larger projects include: Tibetan and Himalayan Digital Library; Melanesian Archive at Virginia and Oceania Digital Library; Digital Himalaya project (Cambridge); American Philosophical Society digital collections; the American Museum of Natural History/Digital Library Project, and the digital projects at the National Anthropological Archives. Other projects representing different types of efforts are The Virtual Institute of Mambila Studies and Robert Kemper's work as literary executor for George Foster, who is digitizing George Foster's extensive material from Tzintzuntzan. Some scholars who have substantial digital material from a variety of data types include: Michael Agar, Janet Bagg, Brent and Elois Berlin, Neville Colclough, Nick Colby ,Roy D'Andrade, John Davis, Jim Dow, Roy Ellen, Michael Fischer, Joel M.Halpern, Eugene Hammel, David Kronenfeld, Alan Macfarlane, A. Kimball Romney, Henry Selby , Paul Stirling, and David Zeitlyn. While its primary digital databases (eHRAF World Cultures and eHRAF Archaeology) are designed for rapid retrieval of mostly published ethnographic and archaeological descriptive materials, in 2005 HRAF began planning a separate database (called the

Culture Conservancy) involving 20 individual collections of fieldnotes and photographs and began looking for DPA startup funding. In the interim, HRAF will incorporate some of this material into its eHRAF Collections. In 2009, HRAF put its first field research photo collection (from Joel M. Halpern) online and will follow with Melvin Ember's collection.

General Background

There have been great strides made with regard to creating digital object repositories and moving toward interoperability between repositories. It is prudent to build on rather than reinvent these developments. (Click here for an overview.) The best way to do this is to work with experts who are familiar with the accomplishments from these fields.

1. Schmid, Oona. 2008. Inside the National Anthropological Archives: An Interview with Robert Leopold. Anthropology News, January: 32-33. [↩]
2. Clark, Jeffrey T., Brian M. Slator, Aaron Bergstrom, Francis Larson, Richard Frovarp, James E. Landrum III, William Perrizo. 2001. "Preservation and Access of Cultural Heritage Objects through a Digital Archive Network for Anthropology," Virtual Systems and MultiMedia, International Conference on, pp. 28, Seventh International Conference on Virtual Systems and Multimedia (VSMM'01). [↩]
3. Silverman, Sydel and Nancy J. Parezo editors. 1995. Preserving the anthropological record. Papers presented at a symposium : Preserving the Anthropological Record : issues and strategies / sponsored by the Wenner-Gren Foundation and held February 28 – March 4, 1992 in Rancho Santa Fe, California. Contents: Introduction / Sydel Silverman — The National Anthropological Archives / Mary Elizabeth Ruwell — Discipline history centers in the sciences / Joan Warnow-Blewett — The Melanesian Archive / Donald Tuzin — Preserving the archaeological record / Don D. Fowler and Douglas R. Givens — The records of applied anthropology / John van Willigen — The role of museums in preserving the anthropological record / Thomas H. Wilson and Nancy J. Parezo — Saving the past for the future: guidelines for anthropologists / Nancy J. Parezo, Nathalie F.S. Woodbury, and Ruth J. Person — The physical preservation of anthropological records / Mary Elizabeth Ruwell — The potentials and problems of computers / Robert V. Kemper — The future uses of the anthropological record / Shepard Krech III and William C. Sturtevant — The next steps / Sydel Silverman and Nancy J. Parezo. New York. Wenner-Gren Foundation for Anthropological Research. [↩]
4. Bird, S. and G. Simons. 2003. Seven dimensions of portability for language documention and description. Language 79:557–582; Gippert, J., N. Himmelmann, and U. Mosel. 2006. Essentials of language documentation. Berlin: Mouton de Gruyter. [↩]

Essential Elements for Effective DPA

In addition to persuading the profession of the importance of DPA, certain major issues have to be addressed to have effective AnthroDataDPA. If these issues are not resolved, plans have to be in place for how to address those issues. Breakout groups addressed the following topics:

- Data Preservation Issues
- Access Issues
- Metadata
- Digitization Issues
- Storage/Backup and Long-Term Preservation
- Depositors to Archives
- Privacy and Ethical Issues
- Copyright
- Funding and Sustaining Support for Long-Term Preservation

There are some general strategic principles that the group agreed upon which we will summarize first and then move on to more specific issues and decisions. Other decisions had to be deferred because they could not be made within the context of a two-day workshop.

General Principles

1. Whenever possible physical records (e.g., notebooks, photographs, artifacts) should be physically preserved rather than discarded after digitization. Digital preservation, on the other hand, with migration strategies, may be best for other material such as tapes and objects on computer disks that have shorter life-spans. Some professionals believe that if done properly, digital object repositories can act as long-term preservation strategies and have the advantage of allowing multiple copies to be "housed" in different places (decreasing the risk of destruction from physical or social disasters/ upheavals). However, many digital projects do not have plans for long-term preservation in place. If there is any doubt about long-range preservation, both strategies should be pursued.

2. The aim should be to preserve all anthropological research materials. This includes materials in less than desirable formats if that is all there is and "gray" literature (a term widely used for research reports in archaeology produced for contract work) which is not particularly accessible. There was more debate about the need for setting priorities and whether different forms of the "same" material should be preserved. On the one hand, archivists stress that it is not easy to know in advance how information might be useful in the future, and it is not always clear that two forms are identical, so it is preferable to preserve all forms that are available. On the other hand, such a practice might be a waste of resources, such as preserving a fuzzy and a clear picture of the same subject. It is probably more labor-intensive to sort through material to decide what is worth keeping and what is not, so keeping all related materials is probably the best strategy.

3. While there are important exceptions, in general we see no reason to restrict access to anthropological data. The group does not believe that is possible in practice or advisable in principle to use access control to restrict access to prevent uses that we may not like (e.g., by creationists or racists). There are a great variety of possible audiences, with the top three most highly prioritized: professional anthropologists/graduate students; other scholars; informants or subjects and subject communities; government agencies; journalists; advocacy groups; general adult public; college students; K-12 students; commercial interests; and unanticipated users in future generations.

4. Overall strategy must be constrained by considerations of privacy and ethics. As anthropologists working with humans as groups or individuals, there is an implicit trust between research and subject that participation will not cause harm in any way to the individual. We must protect privacy and at the same time remain flexible so that any system can adjust to new concerns or new standards. It will be necessary in the future to provide clear statements of intent, while allowing for evolution of technical and tactical tools to meet them while adjusting to changing conditions. In other words, it is not possible to secure privacy over the long term by simply adopting permanent policies early on, however firm and comprehensive those policies may seem to be at their inception. (More on privacy and ethics.)

The timely generation of appropriate metadata is a professional and ethical obligation. It follows that funders, both private and public sector, must recognize metadata, and data curation more generally, as essential and legitimate expenses that must be adequately supported.

Issues and Problems

We now turn to more specific issues regarding AnthroDataDPA.

Preservation and access

1. Data are rapidly degrading in quality and being lost on a continuing basis. Much has already been lost irretrievably. We badly need functional repositories for digital data as soon as possible. These repositories need to be open to a broad range of depositors and backed up by institutional (including funding agency, university, professional association) commitments.

2. Formal repositories are needed and investigator- or project-oriented data-silos are not and will not be financially or technically sustainable, nor will they likely provide the sorts of access—and access control—that are needed.

3. A major issue is whether preservation and access should be undertaken by means of centralized or distributed repositories. However, a unified repository structure for all anthropology is unlikely to be the best solution. The scope of anthropological repositories should be based on shared needs for functionality and the nature of the data at issue. The fields of anthropology are sufficiently divergent in terms of research goals and the data used to address research questions that trying to unite them now is neither realistic nor necessarily desirable.

4. Data should be deposited in a trusted repository during or as soon after data collection as possible in order that the needed metadata can be accurately and inexpensively collected and that a secure copy of the data is maintained. However the repository should provide the ability for the investigator to have exclusive access to the data (or for the investigator to directly control access to others) for a reasonable period of time to permit publication. What is a reasonable time for investigator control may differ by subdiscipline depending upon the dominant publication modes. Enforced mandates from funding agencies and better guidance from professional societies would be most helpful in defining appropriate limits. With public funding, perhaps 3-5 years after the termination of the grant collecting the data is a reasonable limit, with 5 years for dissertations. In any case, 10 years seemed like an absolute maximum to restrict access to protect the investigator's publication interests.

5. To preserve data for long term use, researchers must ensure long term 'intelligibility' in both human and computational terms. (See technical sections on Maintenance of Data Integrity, Best Practices for Storage Infrastructure). "Human intelligibility," refers to the ability of future researchers to understand the information; this is too often compromised by the lack of documentation accompanying the digital file. "Computational intelligibility" refers to the ability of future hardware and software to interpret the file format; and this can be compromised by the pace of technological change. Since the 1996 report of the Taskforce on Digital Archiving1, it is commonplace to remark on the 'digital dark age,' Preservation is threatened by the rapid obsolescence of physical recording media and the equally rapid obsolescence of operating systems and file formats. Simons noted that physical media have declined in durability over the years, contrasting the long term legibility of inscriptions in stone with the many different types of storage media in use in the past 25 years (5.25" floppies, 3.5" floppies, Zip drives, Memory sticks, CDs, DVDs, Blu-ray discs).2 The obsolescence of operating systems and file formats is even more striking: current version of MS Word cannot read documents created in Word 1.0.

Decisions Regarding Depositors

While the group agreed in principle to the idea that all anthropological materials should be digitally preserved, it was recognized that prioritization of projects is unavoidable. The following criteria should be used to set priorities. The relative importance of each criterion must be determined on a case-by-case basis, considering the nature of the material, the resources available, and the goals of the project. [3] They are listed here in no particular order.

1. Ease of digitization: Some records are 'low-hanging fruit' that may take relatively little effort to digitize because of their condition, organization or description.

2. Format of material: Certain formats (e.g. magnetic tape) are inherently unstable and are likely deteriorate. Material in fragile formats may be prioritized in the interest of preservation.

3. Fragility of material: Records that are damaged or that have been stored in less-than-ideal conditions may be fragile and subject to deterioration.

4. Current level of access: How accessible are the records already, both to potential researchers and to the creators of the records? Will digitizing increase accessibility?

5. Frequency & intensity of anticipated use: Digitization can prevent damage from frequent handling of material. While future use can be difficult to anticipate, factors such as the identity of the creator or interest in the subject matter can be predictive.

6. Rarity or uniqueness of subject matter: If the records document a completely unique subject area (e.g. the only known recordings of an extinct language), they may be given priority. In most cases primary data should be given preference over derivative analysis.

7. Material in finite custody: An archive may wish to digitize material that is to be repatriated or is only in temporary custody, assuming that such digitization does not violate any agreement with the owners of the material.

8. Prioritize value of material within collections: In addition to prioritizing collections, material within collections can be prioritized. In a very large collection, the volume may preclude digitizing all at once. In such cases, a representative sample or a select subset can be digitized first.

Fostering Interdisciplinary Collaboration

Whether it is a committee, a consortium of archives, a series of ongoing workshops or an affinity group, there are several areas of activity that would benefit from central leadership.

- Preparing material to be archived: A central organization can help anthropologists prepare material to be archived. This includes recording information and describing context that could otherwise be lost or recorded inaccurately (such as the purpose of the research project and dates, places and descriptions of each item or file).[4]
- Match material with archives: A central group can help address the problem of 'orphan' archival material (records with no archival home). We can increase the portion of the anthropological record that is archived through outreach and collaboration. For this purpose, it would be appropriate for teams of archivists and researchers to focus on a specific domain.[5]
- Adapt recommendations and standards: There are many existing standards for digital archiving. It is unreasonable to expect individual anthropologists to interpret and implement these standards on their

own. A central group can identify relevant standards, adapt them if necessary to make them relevant within the context of anthropology, and work to encourage their adoption among anthropologists.6

- Identify challenges to digital archiving: What are the challenges or barriers to progress in digital archiving? Are these challenges mainly social (e.g. related to peoples' expectations and conceptions of archives)? Are they technical (related to infrastructure, user interfaces)? What sorts of resources are necessary to undertake a major digital archiving project?
- Develop portals: While it is probably impractical to propose a single digital archive for the discipline of anthropology, it is possible to create portals to data or metadata.7
- Education and Outreach: There is a need for outreach to scholars and other practitioners in the discipline of anthropology to increase awareness about digital archiving. Initial steps to educate anthropologists (such as panel discussions and workshops at regional and national conferences) are within immediate reach and should begin in the next year.8 Also, materials should be prepared to incorporate into classroom curricula, such as Field Methods and Research Design courses.

As we will discuss in the section "Funding and Support," larger-scale efforts will take some planning, including application for funding. Furthermore, if such efforts are to be successful in the long term, anthropology will have to work to develop a sustainable community model bringing together all of the stakeholders in anthropological data DPA.

What to Do About Data in the Meantime?

In the absence of a central coordinating institution, which is the current case, the best solution is to find a trusted repository —perhaps even one's university library—and, if possible, provide copies of data to other institutions. As already discussed, if at all possible, it is wisest to avoid going it alone. If you have not decided on a repository, you should follow the guidelines discussed in this working report. The absolutely worst solution is to store data in proprietary formats without publicly available file format specifications that may not be readable in the future. If the media are not upgraded, the data may also be lost.

Unresolved Issues

The two biggest areas in which the breakout groups did not arrive at a consensus revolved first around copyright, or more broadly, the ownership claims and interests of professional researchers and second, the type of metadata that are needed for searching across platforms. In the latter case, the metadata breakout group simply felt that that the topic was too difficult to tackle within the short time of the workshop.

Regarding ownership claims and interests of professional researchers, there was more genuine disagreement over the degree to which unrestricted, anonymous access to research data should be allowed. Although all agreed on the importance of DPA, the two perspectives can be summarized as:

1. The library perspective—knowledge should be shared as widely as possible. Withholding data works against core scientific principles.

2. Concern over "free-riders"—field researchers and data collectors may suffer because of the significant amounts of time they spend to collect data. Others who "use" their data can publish faster. Any DPA efforts must seriously address credit, incentives for depositing data, and knowing who accessed the data.

The various arguments are summarized in the Copyright Working Group Report .

The copyright working group also discussed the ambiguity of copyright laws with regard to data, datasets, and metadata. For example, in the U.S. copyright does not apply to "facts" but rather to "expressions." Certain forms of metadata, such as metadata describing the meaning, methods, and limitations of a dataset would be likely covered by copyright. Other forms of metadata, particularly technical metadata (e.g., file formats, collection structures) would probably not be covered by copyright. Laws in other locales complicate the sharing of data. For instance, the EU has database protection laws that protect compilations of data. The desirability of some form of standardized licensing, such as Creative Commons, was mentioned.

Other questions that need to be pursued further are:

How do needs vary by subdiscipline? Disciplines vary in the ways they handle location, scale, temporal transgression, and representation in one, two, or three dimensions, not to mention in the kinds of data which are of primary interest. They also vary in the degree to which they have discussed and resolved ethical issues with regard to standards and access.

What is the proper role of universities in preserving and providing access to digital records? What are the current roles and the proper roles of individual researchers, academic departments, university libraries and university presses?

What are the cultural impediments to cyberinfrastructure development? How do we accommodate notions of ownership, senior grumpiness, lack of training, academic competition, fear of contradiction, and fear of preemption.

How do we treat sharing? Should prepublication sharing be encouraged or merely facilitated? Less controversially, how do we treat post-publication sharing? "We recommend that it becomes mandatory for scientific papers to explain where and how to access data and resources generated as part of the investigation. We are aware that some journals already have strong policy positions in this area, insisting that large data sets must be deposited in public databases, and that all reasonable requests for materials from other researchers must be fulfilled. There is however, heterogeneity with both policy and enforcement; surprisingly, many journals have no written policy on the availability of either bioresources or primary data9"

How does replicability influence best practices? How do we accommodate differences between fields that advance by generating new databases to replicate research as opposed to fields that advance through the accumulation of shared data. Should even replicable data be preserved?

5. Garrett, John, and Donald Waters. 1996. "Preserving Digital Information: Report of the Task Force on Archiving of Digital Information commissioned by the Commission on Preservation and Access and the Research Libraries Group.Washington, DC: Commission on Preservation and Access." http://www.rlg.org/ArchTF/tfadi.index.htm [↩]
6. Simons, Gary F. 2006. Ensuring that digital data last: The priority of archival form over working form and presentation form. An expanded version of a paper originally presented at the: EMELD Symposium on "Endangered Data vs. Enduring Practice," Linguistic Society of America annual meeting, 8-11 January 2004, Boston, MA. http://www.sil.org/silewp/2006/003/SILEWP2006-003.htm [↩]
7. ViPIRS (http://library.nyu.edu/preservation/movingimage/vipirshome.html) is an example of a tool that tracks assessment data for audiovisual preservation projects. [↩]
8. Digital Antiquity provides a model for the recording of collection-level metadata when depositing data. [↩]

9. A collaborative, strategic approach to documenting specific topical domains is reviewed and critiqued in Malkmus, Doris. 2008. Documentation strategy: Mastodon or retro-success? American Archivist 71(2): 384-409. [↩]
10. Digital Antiquity provides a model for the recording of collection-level metadata when depositing data. [↩]
11. Portals can take many forms; examples include the Digital Archive Network for Anthropology and the Open Language Archives Community. [↩]
12. The field of Linguistics has been successful in increasing awareness about archiving and can provide models for educational efforts. See, for example, the E-MELD school of best practices: http://emeld.org/school/index.html. [↩]
13. Schofield, et al 2009. [↩]

Impediments to DPA

There are a number of technical impediments to the effective adoption and use of digital repositories. The main ones are cost/time impediments and the technology-related impediments. These will affect the scope of the data that is deposited for a given project or endeavor. Investigators are sure to contemplate the tradeoffs between the costs in time and money of depositing a given set of data and the benefits to the investigator and to the field more broadly. We believe that these tradeoffs are likely to be evaluated differently by subdiscipline.

To the extent that these tradeoffs are actively evaluated we need to change reward structures (e.g., though grant or publication incentives or requirements) to encourage deposit for data. More broadly we need to change disciplinary norms about what constitutes responsible professional behavior with respect to depositing different classes of data. Professional societies can play an active role in this regard. Other ways of encouraging deposit will be to require attributions of credit—or better, formal citation—of deposited data and professional valuation of these citations as we value ordinary publication citations.

Diminishing the disincentives to deposit would be accomplished by maximizing ease of use and by low cost. However, even with software tailored to streamline use, there will be a necessary tradeoff between the time investment required and the quality of the metadata and data obtained. Finally, prominent and compelling examples will be invaluable in demonstrating the scholarly value of deposit.

In this context, it is important to distinguish between “new” and “legacy” data. For projects that are just starting, digital archiving is a much simpler problem. The costs of archiving can be built into the project as well as the procedures, metadata standards, and the identification of the ultimate repository. Projects that are complete or that are on-going present a very different set of problems. The data were not collected with digital archiving in mind and often the investigators are dead or incapable of placing the data in acceptable formats or creating the needed metadata to make them useable. Even in cases in which the investigator is willing to invest the time and energy, there is great difficulty obtaining financial support. The two situations are qualitatively different and require very different solutions. Solving the archiving issues for new projects is simpler and easier and should proceed first. Professional societies and funding agencies should set guidelines for new projects and begin to enforce them at the same time they tackle the much more difficult issues involved with legacy data.

Repositories must have secure platforms with strong safeguards to prevent access to sensitive materials by individuals who should not be authorized for access. This demands not only a login but

also ways of reliably authenticating user credentials. It was generally but not universally accepted in the full group that a login should be required even for access to material that is not in some way restricted. User agreements, informed by professional ethics, will need to be established by the repositories.

As noted in the OAIS standard (CCSDS 650.0-B-1) for a digital repository and reference model for a digital information object, storage, is one of six interconnected components (Ingest, Administration, Data Management, Access, and Preservation Planning) of the reference architecture. No component stands alone, and it is important to approach this subject as an interconnected web linking various issues.

There is a steep learning curve to understand these technologies and the cost to hire developers is very expensive. One way to overcome these challenges is to appeal to granting agencies to provide additional support to build specialized systems based upon open source technologies that could be leveraged by other anthropological research projects. Although repositories have mostly the same functionality there are important differences in how the systems represent stored data that is technically referred to as a data model. Just as the ability to search and discover is tightly bound to the representation of data the ability to preserve data is tightly coupled to a data model that facilitates preservation planning and preservation treatments.

Best practices for storage infrastructure

Best practices emerge over time as a result of a deeper understanding of a problem and outcomes from pilot projects or test beds established for experimentation. While the anthropological community is just beginning to explore storage solutions for LTP (long-term preservation) the Digital Library community has for nearly a decade explored the principal issues and challenges that surround storage and backup of digital data. The principal problems that need to be addressed are well known and include (1) technological obsolescence; (2) media decay (3) replication, and (4) evolving standards to manage large storage pools or networked storage grids.

As already discussed, the worst-case scenario for storage and backup is locally managed storage. This modality is associated with a high probability of data loss over time. In this mode, best practices followed by traditional data centers to protect data and secure unauthorized access to data is nearly impossible to maintain. The challenge is to educate the community on the need to abandon this practice and adopt alternative solutions such as participation in grid storage networks.

At the opposite end of the spectrum and across the Atlantic the European community has successfully demonstrated the efficacy of grid storage for LTP of digital data. The infrastructure for grid storage has trusted governance, which establishes best practices to deal with data management problems, associated with the aforementioned problems inherent in storage hardware and software used to manage storage. One might characterize grid storage as “being alive”, continuously being refreshed and secure since access and replication where an integral part of the management functionality of the grid. In addition, participation in the grid also relieves the student or researcher with the responsibility to plan and manage his or her own media migrations. (More)

Optical disk, magnetic disk and tape have all been successfully used for data storage and backup. In most instances these media are combined to form a hierarchical storage system. Typically these systems deploy magnetic disk for fast online access to data and tape or optical disk to store off-line data

that is infrequently accessed. The goal is to build a configuration that satisfies LTP requirements at a price performance that is affordable and sustainable. Finally the group unanimously recognized that storage and backup did not equate to long-term preservation of digital data. In the absence of a logical layer, such as PREMIS (PREservation Metadata: Implementation Strategies) to overlay storage, over time digital data would become more difficult to: discovered, search, accessed or understood as hardware software and community standards evolved and made older storage and access system obsolete.

Maintenance of Data Integrity

To address the threat of technological obsolescence, Simons (2006) recommends that researchers create an archival master in an enduring file format and deposit the archival master in a preservation archive. A preservation archive is an established institution committed to long-term preservation of the digital object; a distinguishing characteristic is that a preservation archive will have a technology migration plan on which to found its claims of long term digital accessibility. Thus it contrasts with a 'web archive,' which is often only a website serving information from a database or file directory. Web archives rarely serve genuinely interoperable material, and they regularly disappear in response to changes in institutional servers or in the responsibilities of the archive creator.

Enduring File Format

What is an "enduring file format"? In the acronym created by Simons, it is a file that offers LOTS. In other words, it is Lossless, Open, Transparent, and Supported by multiple vendors. Each of these desiderata deserves some discussion.

Lossless. A lossless file format is one in which no information is lost through file compression. It is uncontroversial to say, for example, that an archival master should be uncompressed and unedited.[1] However, copies may, of course, be made from the archival file, and these can be altered to serve as working or presentation copies[2]. Professional archivists usually recommend that the archival master be copied once, to make a 'presentation master,' and that compressed and edited copies be made from the presentation master, not the archival master. Although digital copying does not harm the original file if done correctly, use of a presentation master is probably good advice: some media programs compress automatically when they save a file;[3] and to find this out too late is to irrevocably lose part of the information on the archival master. Although uncompressed file formats are preferable to even those with lossless compression,[4] lossless compression is an option if uncompressed files are so large (e.g., video) that their storage is impractical. Lossless compression algorithms typically remove only redundant information (e.g., pixels of the same color in an image) and allow the full content to be recovered through the use of a decoding algorithm. 'Lossy' compression, on the other hand means that the so-called 'irrelevant' information can never be recovered; thus it is to be avoided for highly valued material. Although the difference between a compressed file and an uncompressed file may be indistinguishable to human ears and eyes, in creating a scientific archive of irreplaceable material (e.g., songs and ceremonies of a vanishing culture), we should remember that the scientific instruments of the future may be able to extract more information from the 'noise' on an uncompressed file than we are currently able to perceive. Table 2 shows some common extensions of uncompressed file formats and formats employing lossless and lossy compression.

Type	Uncompressed	Compressed (Lossless)	Compressed (Lossy)
Audio:	.wav, .aiff, .au (pcm)5	.ape, FLAC, TTA	.mp3, .aac6, .wma
Images:	.bmp, tiff w/o LZW	.tiff (or .tif) w/LZW .png .gif (grayscale)	.jpg
Video:	Rtv	JPEG-2000	MPEG-2, DV, MPEG-4
Text:	.txt	.zip	NA

Table 2: File extensions of compressed and uncompressed formats (Aristar-Dry, 2008)

Openness refers to the fact that some file format specifications are publically available; for example, html, XML, pdf, and rtf are all 'open standard.' This means that any software engineer can develop programs that can read these file formats. By contrast, information in proprietary file formats will be lost when the vendor ceases to support the software. "Open standard" is different from "open source," i.e., software whose source code is publicly available. Examples of open source software include Open Office and Mozilla Thunderbird. Open source software usually creates files in open standards. And proprietary software usually doesn't (though there are exceptions, e.g. Adobe pdf). But for long term intelligibility, open standards are more important than open source software. Table 3 below lists some open and proprietary software. Note that some of the most commonly-used software (e.g., Microsoft Word, Excel and PowerPoint) is proprietary and commercial and therefore the least likely to be preserved in the future.

Development	Open	Proprietary
Open	.txt, .html, .xml, .odf, .csv	NA
Commercial	.rtf, .pdf	.doc, .xls, .ppt

Table 3: Open and proprietary standards (Aristar-Dry, 2008)

Transparency. The file format requires no special knowledge or algorithm to interpret, because there is a one-to-one correspondence between the numerical values sent to the computer and the information they represent. Plain text, for example, has a one-to-one correspondence between the characters and the computer-readable binary numbers used to represent them. Similarly, the PCM (pulse code modulation) codec, which is employed by .wav, .aiff, and cdda files, has a one-to-one correspondence between the numbers and the amplitudes of the sound wave. Thus plain text files (.txt) can be read by any software program that processes text. And PCM signals can be interpreted by virtually all audio programs. By contrast, .zip and .mp3 files require implementation of a complex algorithm to restore the original correspondences. Today many programs provide automatic decoding of the common encoded formats. But we cannot be certain that these programs will not become obsolete. In the distant future, some of the encoding algorithms may be lost; and, at that point, interpreting compressed and opaque files will become a costly scientific endeavor.

Note that transparency is not possible with some advanced visualization techniques (e.g., 3-D or CT scanning, GIS).

Support by multiple vendors: Just as lack of compression and transparency are paired in file formats, use of open standards and support by multiple vendors go together in software development. Open standards are more likely than proprietary standards to have wide vendor support, because development using open standards is typically less costly. If a file format is open, there is no inherent barrier to creating another program that handles it. It is not necessary to reverse engineer the format or purchase the specification from the developer. The more software applications that handle a file format, the less likely that format is to fall victim to hardware and software obsolescence.

Best versus good practices. Ideals or "best practices" are not always obtainable; researchers may need to consider "good practices."

Technical recommendations are a moving target. Because technology changes rapidly, regular consultation of up-to-date websites is recommended. See some general resources worth investigating.

14. Arts and Humanities Research Council, (2009). [↩]
15. If the working copy is the primary copy—as, for example, during the ongoing creation of a database—it is important to export the information regularly into an enduring file format. For databases (which are usually managed by proprietary software) this means to export the data regularly into properly documented plain text. A .txt file with informative XML markup is ideal, but often the XML automatically output by a program will be only minimally helpful to someone trying to make sense of the file. In that case, a file including metadata identifying the fields and tables should be created and stored with the database output. [↩]
16. For example, Acrobat 7.0 will automatically compress large pdf files (see: http://www.planetpdf.com/forumarchive/166948.asp). Most importantly, however, as of this writing, most video capture programs automatically compress the audio track along with the video when it is downloaded to a computer. For that reason, linguists and musicologists are advised to make a separate audio recording, using a device like a hand-clap at the beginning to aid in synchronizing the files later on. See: http://emeld.org/school/classroom/video/field.html#1006 [↩]
17. As noted by a Senior Media Specialist at the Getty Museum, "Uncompressed data is trivial to decode, compressed data often is not. This makes for easier long-term viability of the file . . . . " Furthermore, uncompressed data is less prone to loss: "Lossless compression means that a single bit in the compressed file may represent multiple bits in the uncompressed version. This magnifies potential damage caused by bit corruption. In an uncompressed file a single flipped bit will have little overall impact on the renderability of an image. In a lossless compressed file depending on whether the corruption is in the dictionary (in the header) or in image data it can have a larger effect. And in a lossy compression scheme a single bit corrupted can be extremely noticeable." (Howard, 2003). [↩]
18. Technically, .wav and .aiff are container formats, file structures which allow combining of audio/video data, tags, menus, subtitles and some other media elements. They could theoretically contain compressed audio formats, but in practice they usually contain PCM (pulse code modulation) data, which is an uncompressed format. [↩]
19. Apple audio codec (.aac) and Windows media audio (.wma) both have a lossless version. Confusingly, both the lossless and the lossy compression formats use the same file extension. [↩]

Recommended Next Steps

How do we ensure long-term preservation and access in the context of rapidly evolving technology? Obviously, there are major challenges in this endeavor. Not the least is to obtain funding to move this project forward. However, there are a large number of discussions and is a good deal of planning that is needed. The group identified the following needs:

- Some kind of entity, perhaps comprised of multiple institutions and individuals of stature to cooperate in initial round of short-term proposal(s) and project(s). These institutions and individuals would serve as "champions" for the project(s). Their participation would ensure the persons at nongovernmental organizations, governmental agencies, and other relevant institutions that an anthropological DPA project is of critical importance to the physical sciences, the social sciences, as well as the humanities. Exactly what kind of entity is needed, centralized or uncentralized, was not decided. However, the group felt that a centralized entity was probably not possible to achieve.
- Short-term funding to develop ideas for interoperability, long-term planning and further discuss controversial issues. (Note that we have applied for funding from the NSF INTEROP program.)
- A task force to suggest a long-term plan and business model for funding and sustaining DPA specific to anthropology. Identify projects and/or institutions that might be shovel-ready or be appropriate for demonstration projects.
- Create a standards body that will review proposed standards for DPA of anthropological data across the sub-domains. Because standards need to change with technological developments, the standards body needs to have individuals who are familiar with anthropological needs as well as changes in the technological forefront.
- Encourage leveraging the technical infrastructure of both commercial organizations and sister disciplines to promote DPA.
- Anthropology should take the opportunity to extend open standards and open source software to promote DPA.
- Anthropology curricula should be expanded to include best practices and standards for digitization and long-term preservation of digital data.

The members of the workshop realize that the challenges ahead are far greater than the resources that are likely to become available to meet them. This means that establishing priorities will be an initial and long-term issue if an AnthroDataDPA project is to be successful for scholars and for our publics, in the United States and around the world.

ADDITIONAL INFORMATION

Additional Storage

While storage grids do exist in the United States (see the NSF program on Grid storage at http://www.teragrid.org/about/) Commercial Cloud Storage is another option for LTP. This solution is just beginning to gain traction in the US Academic community since it is a potential cost saver. A powerful motivator while the country wrangles through a deep recession. Cloud Storage provides the opportunity to outsource the storage function to large commercial vendors like Amazon and Google that run their own storage grids. For this storage option trust is a significant issue. Commercial vendors are subject to the natural business cycle and no firm is completely immune to failure or takeover. How to access or recover data when a business fails is of serious concern to the academic community.

Secure access to data is another problem identified with commercial cloud storage. In response to these concerns the Mellon Foundation recently sponsored a planning grant to understand how the

academic community could take advantage of cloud storage without being at the mercy of the business cycle and to technically explore how commercial cloud storage could be overlaid with a service interface that would protect data from unauthorized access and automatically replicate data when a firm went out of business. Details about this initiative are available from the DuraSpace website.

Metadata Standards for Long-Term Storage

PREMIS (PREservation Metadata: Implementation Strategies) is the de-facto standard for the digital library community that specifies metadata entities recommended to ensure the long-term preservation (discovery, access, rendering and understandability) of digital data encapsulated in a vast array of file formats. An in-depth understanding of the PREMIS standard was not present in the group. This made it difficult to realistically evaluate PREMIS as a standard, which could be successfully applied to preserve anthropological data. However, in the absence of any other recognized standard, leveraging and extending this standard for the Anthropology community was strategically the right course of action. A policy question that needs to be resolved by some standards committee is how much of what elements, of this very elaborate standard, are needed by the anthropological community to meet their preservation purposes. It is not practical or affordable to capture data for all of the sub-elements in the PREMIS standard.

Existing Repository Software

Repository software used to ingest, save or preserve and access digital content used in the cultural heritage community is mostly open source. Repository software offerings that have gained significant traction in the digital library domain are (1) Fedora (2) DSpace (3) Greenstone (4) E-prints (5) Plone and (6) ContentDM from OCLC. It is important to note that the Fedora and DSpace communities have recently combined to form a consolidated community called DuraSpace. All of these application have out of the box client interfaces to there underlying data stores to simply the ingest, storage and search/access to data. In addition these repository systems have Application Programming Interfaces (APIs) that can be used to build customized web applications or web services for any of the aforementioned functions. Protocols such as OAI-PMH, OAI-ORE and SWORD, to name a few, have also been developed by the digital library community to make these systems interoperate so that data can be exchanged between systems.

Planning Models

The PLANETS project has published a preservation data model and created a tool PLATO for preservation planning. The model can provide two distinct views of stored data, one from the end-user perspective that facilitates search and discovery of preserved data, and the other from a preservation perspective that enables preservation treatments (media or format migrations) at the file set level that does not impact the end-user view or understanding of the data. Risk of data loss is inherent in any preservation treatment and the planning tool PLATO was designed to attenuate that risk. “The planning tool PLATO is a decision support tool that implements a solid preservation planning process and integrates services for content characterization, preservation action and automatic object comparison in a service-oriented architecture to provide maximum support for preservation planning endeavors.”[1] Again in the absence of other available standards the group maintained that is was strategic for the anthropological community to leverage this standard for their community purposes.

1. From Welcome to Plato, the Planets Preservation Planning Tool. [↩]

Attendees

Toward an Integrated Plan for Digital Preservation and Access to Primary Anthropological Data (AnthroDataDPA: A Four-Field Workshop)[1]

May 18-20, 2009, Hilton-Arlington, Arlington, VA

Participants

Carol R. Ember, PI
Eric Delson, PI
Jeff Good, PI
Dean R. Snow, PI
Jeanne Altmann
Jeffrey H.Altschul
Helen Aristar-Dry
Theodore C. Bestor
Douglas A. Black
Jeffrey T. Clark
Lisa Conathan
Michael Fischer
David Gewirtz
David R. Hunt
Eric C. Kansa
Keith Kintigh
Timothy A. Kohler
Robert Leopold
Tom Moritz
Daniel Reboussin
Richard J. Sherwood
Joel Sherzer
David Glenn Smith
Matthew W. Tocheri
Robert V. Kemper
Laura Welcher
Peter Wittenburg

Observers:

Anthony Aristar
Andrew Bennett
Colin Elman
Mark Mahoney
M. Marlene Martin

Observers from the Local Area

From NSF:

Anna Kerttula, Arctic Social Sciences Program, Program Officer
Terry Langendoen, Information & Intelligent Systems, Expert
Joan Maling, Linguistics, Program Director
Elizabeth Tran, Human and Social Dynamics
Jean Turnquist , Physical Anthropology, Program Director
Mark L. Weiss, Behavioral and Cognitive Sciences, Division Director
Deborah Winslow, Cultural Anthropology, Program Director
John Yellen, Archaeology and Archometry, Program Director
Christopher Greer, Senior Advisor for Digital Data, Office of Cyberinfrastructure
David Lightfoot, Assistant Director of the National Science Foundation, SBE Head

From NEH:

Helen C. Agüera, Senior Program Officer, Division of Preservation and Access
Jennifer Serventi, Office of Digital Humanities

1. [1] Supported by the National Science Foundation (BCS-0823404) and the Wenner-Gren Foundation in a grant to the Human Relations Area Files. The cultural anthropology, arctic social sciences, physical anthropology, archaeology, and political science programs were co-funders of NSF's contribution to this workshop.

Contributors

Toward an Integrated Plan for Digital Preservation and Access to Primary Anthropological Data (AnthroDataDPA: A Four-Field Workshop)[1]

May 18-20, 2009, Hilton-Arlington, Arlington, VA

The main body of the AnthroDataDPA Report is a summary of the May 2009 written by the PIs, Carol R. Ember, Eric Delson, Jeff Good, and Dean R. Snow. It draws heavily on the Chair Reports of the breakout discussion groups, which are included in their entirety on this site .

Chair Reports:

- Access Issues: Keith Kintigh (Chair), Jeff Altschul , Ted Bestor , Jeff Good , Matthew Tocheri , Peter Wittenburg .
- Copyright: Eric Kansa (Chair), Jeanne Altmann, Eric Delson, and Tom Moritz
- Data Preservation Issues: Carol R. Ember (Chair), Anthony Aristar, Jeffrey Clark, Lisa Conathan, Robert Leopold, Daniel Reboussin and David Glenn Smith
- Depositor Issues: Lisa Conathan (chair), Douglas A. Black, Michael Fischer, David R. Hunt, Mark Mahoney, Marlene Martin, Daniel Reboussin, and Dean R. Snow
- Digitization Issues: Helen Aristar-Dry (Chair), Richard Mahoney, and Richard Sherwood.
- Funding and Sustainability Issues: Robert V. Kemper (Chair), Anthony Aristar, Helen Aristar Dry, Andrew Bennett , Jeff Clark, Carol Ember, Keith Kintigh, Jennifer Serventi, Matt Tocheri, Laura Welcher, and Peter Wittenburg
- Metadata Issues: Tom Moritz (Chair), Jeanne Altmann, Eric Delson, Eric Kansa, Robert Kemper
- Privacy and Ethical Issues: Richard Sherwood (Chair), Jeff Altschul, Ted Bestor, Jeff Good, Tim Kohler, Robert Leopold, Susan Penfield, Joel Sherzer, and David Glenn Smith
- Storage/Backup Issues: David Gewirtz (Chair), Laura Welcher, Dean Snow, Michael Fischer, David R. Hunt, and Mark Mahoney

Participants

1. [1] Supported by the National Science Foundation (BCS-0823404) and the Wenner-Gren Foundation in a grant to the Human Relations Area Files. The cultural anthropology, arctic social sciences, physical anthropology, archaeology, and political science programs were co-funders of NSF's contribution to this workshop.

Chair Reports

Access Issues

Access Issues: Breakout Group Report

Draft 6/22/09

Jeff Altschul (Archaeology), Ted Bestor (Sociocultural Anthropology), Jeff Good (PI; Linguistics), Keith Kintigh (Chair; Archaeology), Matthew Tocheri (Physical Anthropology), Peter Wittenburg (Linguistics)

Toward an Integrated Plan for Digital Preservation and Access to Primary Anthropological Data (AnthroDataDPA: A Four-Field Workshop)

PIs: Carol R. Ember, Eric Delson, Jeff Good and Dean Snow

May 18-20, 2009

The Access Issues breakout group addressed a variety of questions concerning access to digital anthropological data contained in formal disciplinary repositories.

Repository scope. In considering these questions the group made several observations concerning the nature and scope of these repositories. It was recognized first that formal repositories are needed and that investigator- or project-oriented data-silos are not and will not be financially or technically sustainable, nor will they likely provide the sorts of access—and access control—that are needed. However, it was the group's contention that a unified repository structure for all anthropology is unlikely to be the best solution. The scope of anthropological repositories should be based on shared needs for functionality and the nature of the data at issue. The fields of anthropology are sufficiently divergent in terms of research goals and the data used to address research questions that trying to unite them now is neither realistic nor necessarily desirable. Yet, as more focused repositories develop, it would be well for there to be communication and agreement on some metadata standards and some tools that can be shared across repositories. Further anthropological repositories need not and should not restrict itself to "primary" data. The decision as to what should be archived will, of necessity, change over time and be driven to a large extent by a cost/benefit analysis undertaken by individual analysts in relation to guidelines set by the various subfields and funding agencies.

To what groups do we have responsibilities to provide access? The question of responsibility is to an extent intertwined with how the work was funded and what sorts of individuals might realistically desire access. We see the answer as a sort of priority list, in which we should attend most carefully to delivering access to the groups most interested and most likely to use it, namely anthropologists and other members of the scholarly community. In many cases we have strong ethical obligations to provide access to our informants and members of subject communities of our research. To the extent the data are generated with public money, we have clear responsibilities to provide access to the general public, unless otherwise restricted by legal or ethical considerations.

Who are and who might be the consumers of anthropological data? While there are important exceptions, in general we see no reason to restrict access to anthropological data. The group does not believe that is possible in practice or advisable in principle to use access control to restrict access to prevent uses that we may not like (e.g., by creationists or racists). There are a great variety of possible audiences, with the top three most highly prioritized:

- Professional Anthropologists/Graduate Students
- Other Scholars
- Informants or Subjects and Subject Communities
- Government agencies
- Journalists
- Advocacy groups
- General Adult Public
- College Students
- K-12 Students
- Commercial Interests
- Unanticipated Users in Future Generations

Time frame for the development of an information infrastructure. Data are rapidly degrading in quality and being lost on a continuing basis. Much has already been lost irretrievably. We badly need functional repositories as soon as possible. These repositories need to be open to a broad range of depositors and backed up by institutional (including funding agency, university, professional association) commitments. It appears that sociocultural anthropology is the farthest behind in this regard.

Time frame for data ingest and public access. Data should be deposited in a trusted repository (see http://www.oclc.org/programs/ourwork/past/trustedrep/repositories.pdf) during or as soon after data collection as possible in order that the needed metadata can be accurately and inexpensively collected and that a secure copy of the data is maintained. However the repository should provide the ability for the investigator to have exclusive access to the data (or for the investigator to directly control access to others) for a reasonable period of time to permit publication. What is a reasonable time for investigator control may differ by subdiscipline depending upon the dominant publication modes. Enforced mandates from funding agencies and better guidance from professional societies would be most helpful in defining appropriate limits. With public funding, the group felt that 3-5 years after the termination of the grant collecting the data was a reasonable limit, with 5 years for dissertations. In any case, 10 years seemed like an absolute maximum to restrict access to protect the investigator's publication interests.

Rapid deposit is highly desirable because the ability to obtain these data metadata and the likelihood of data loss increase rapidly as time passes. Rapid deposit may also be advantageous to the investigator as it encourages organization of the data and facilitates sharing with collaborators.

Requirements for deposit according to established guidelines should be implemented as soon as functional repositories are available. In many cases it seems to be reasonable to mandate that, at the time of publication, supporting data should accessible in a trusted public repository. Use of these repositories should be enforced through peer review of both publications and grants.

Granularity of metadata. It is in the nature of many kinds of anthropological research that data are collected a multiple levels (e.g., individual and community, site and artifact, linguisitic corpus and session). Metadata are likely to be similarly complex and metadata requirements will vary across subfields and may be multilevel. For example, in archaeology it has proved efficient to collect metadata that applies to an entire project and separately to collect more refined metadata that refer to specific datasets that are part of that project.

Problems in making data public. It is the group's position that adequately documented data should become public unless there are compelling reasons it should not be. However, particularly in sociocultural and medical anthropology confidentiality responsibilities will need to be rigorously observed. In some cases, access will be determined by clear-cut consent agreements or IRB stipulations. In other cases the investigator may perceive unstated "sensitivity" by descendent communities that might in some cases contrast with expressed desires of the subject communities. There are a number of very difficult issues here and we see no easy answers. Beyond explicit agreements, should the investigator alone be able decide on sensitivity? What happens after the investigator is gone? Should professional societies be involved in gate-keeping by the repositories to provide a viewpoint with more distance?

How do we lower the barriers to entry to the repositories. There are a number of impediments to the effective adoption and use of digital repositories. The main ones are cost/time impediments and the

technology-related impediments. These will affect the scope of the data that is deposited for a given project or endeavor. Investigators are sure to contemplate the tradeoffs between the costs in time and money of depositing a given set of data and the benefits to the investigator and to the field more broadly. We believe that these tradeoffs are likely to be evaluated differently by subdiscipline. It may be, for example, that the costs of digitizing and depositing extensive sociocultural anthropology field notes will be relatively high compared to the perceived benefit.

To the extent that these tradeoffs are actively evaluated we need to change reward structures (e.g., though grant or publication incentives or requirements) to encourage deposit for data. More broadly we need to change disciplinary norms about what constitutes responsible professional behavior with respect to depositing different classes of data. Professional societies can play an active role in this regard. Archaeology is most advanced, with ethical standards that clearly require access to data. Other ways of encouraging deposit will be to require attributions of credit—or better, formal citation—of deposited data and professional valuation of these citations as we value ordinary publication citations.

Diminishing the disincentives to deposit would be accomplished by maximizing ease of use and by low cost. However, even with software tailored to streamline use, there will be a necessary tradeoff between the time investment required and the quality of the metadata and data obtained. Finally, prominent and compelling examples will be invaluable in demonstrating the scholarly value of deposit.

In this context, it is important to distinguish between “new” and “legacy” data. For projects that are just starting, digital archiving is a much simpler problem. The costs of archiving can be built into the project as well as the procedures, metadata standards, and the identification of the ultimate repository. Projects that are complete or that are on-going present a very different set of problems. The data were not collected with digital archiving in mind and often the investigators are dead or incapable of placing the data in acceptable formats or creating the needed metadata to make them useable. Even in cases in which the investigator is willing to invest the time and energy, there is great difficulty obtaining financial support. The two situations are qualitatively different and require very different solutions. Solving the archiving issues for new projects is simpler and easier and should proceed first. Professional societies and funding agencies should set guidelines for new projects and begin to enforce them at the same time they tackle the much more difficult issues involved with legacy data.

What does it take for a user to get access? There was a strong consensus that the repositories must have secure platforms with strong safeguards to prevent access to sensitive materials by individuals who should not be authorized for access. This demands not only a login but also ways of reliably authenticating user credentials. It was generally but not universally accepted in the full group that a login should be required even for access to material that is not in some way restricted. User agreements, informed by professional ethics, will need to be established by the repositories.

1. http://www.oclc.org/programs/ourwork/past/trustedrep/repositories.pdf
2. Michener, W.K., Brunt, J.W., Helly, J.J., Kirchner, T.B. and Stafford, S.G. 1997. Nongeospatial Metadata for the Ecological Sciences. Ecological Applications. 7: 330-342

Copyright Issues

Copyright Working Group Report

Chair:
Eric Kansa
Participants:
Jeanne Altmann, Eric Delson, Tom Moritz

The Copyright Working Group discussion focused on the legal and social norms that govern the ownership of anthropological research data. Issues around privacy and cultural sensitivities related to certain classes of anthropological information may sometimes require access and use restrictions for ethical management and curation. However, these ethical considerations, while essential in shaping future digital curation policies, were not the main focus of discussion for the Copyright Working Group. Instead of focusing on the privacy needs of anthropological stakeholders (especially human subjects in research), the Copyright Working Group was mainly focused on the ownership claims and interests of professional researchers.

While our group saw very lively discussion and debate, we came to little consensus about the best way to shape data ownership policies. Though all members of the group agreed that preservation and dissemination of primary data are important priorities and could greatly improve the research process, we debated the topic of allowing or requiring unrestricted, anonymous access to research data. Important discussion themes included:

Library Perspective: Libraries have a strong ethical tradition of sharing knowledge and information as broadly as possible. One participant cited Thomas Jefferson and his belief that the field of knowledge is the common property of all mankind. Withholding data works against core scientific principles, because withheld data makes it difficult or impossible to falsify claims.

Concerns over "Free-Riders": A major concern over data sharing and ownership relates to the potential for benefiting one class of researchers over another. For instance, researchers who engage in fieldwork and data collection often spend a significant amount of time writing grants and once they obtain adequate funding, spend more time gathering data (time that could be spend publishing). On the other hand, more theoretically inclined researchers who spend less time and effort funding and executing fieldwork are often able to produce publications faster. Because they have more to show for their efforts, theoreticians may be unfairly advantaged for tenure and promotion, particularly if high quality research data is easily available. Thus, if data sharing policies do not take into account the potential for data "free-riding", field researchers may suffer.

Options Discussed: Discussion over credit and incentives often touched back on concerns over "free-riding" and adequate recognition for the contributions of researchers who invest so much effort and face so many risks in data gathering. Elinor Ostrom's research into the sociology of common pooled resources relates to the free-rider concern. Her findings indicate that some people who could contribute to a common data resource will not participate if there are any free-riders.

All working group participants agreed that data sharing should be recognized by the profession. However, the working group differed in their opinions on the best mechanisms and approaches to promote recognition and combat free-riding. These opinions included:

Publication Norms: Some working group participants regarded reliance on professional social norms as the best way to provide the proper credit and incentives for researchers to share data. Researchers publish their insights, analyses, and to some extent their data through journals and other venues because they feel confident that they will be credited for their efforts. Citation norms already exist, and researchers routinely cite each other's published presentations of data. If data sharing takes the form of publishing ("data sharing as publication") then these established norms and forms of professional recognition could help provide needed incentives for data sharing. Data sharing in the context of publication also enhances the value of shared data, because datasets need extensive documentation and explanation for reuse.

Restricting Contracts / Agreements: Other working group participants believed that stronger measures based on access controls are required to combat free-riding. Access to research data should be provided on a conditional basis. Researchers that produce data may want to know who is accessing their data and why. In addition, some researchers may want to protect their data from misuse by anti-scientific (including commercial or religious) agendas. Because of these concerns, access to data should require some combination of login and identification and / or a "click-through" agreement to proper uses of a given dataset.

The working group came to no consensus about the relative merits and risks of these various options. Some working group participants favored much more open forms of data dissemination and others saw access and use restrictions as an important safeguard. The various conflicting points of view are as follows:

Issue / Options	Arguments for Controls	Arguments against Controls
Identification of users	(1) Researchers who publish datasets want to know how their data are used and by whom. Identifying individuals who request data represents a minor (and non-onerous) form of compensation for sharing data. It can also help guard against plagiarism. (2) Given the privacy and ethical sensitivities of many classes of anthropological data, identity management systems will be required. Thus, this poses no extra complication.	(1) There are privacy and academic freedom concerns to consider. Libraries traditionally regard information retrieval requests as sensitive private data, and destroy all record of such transactions once a book is returned. Anonymous requests for data offer better privacy protections and academic freedom, especially considering that research designs and research questions may sometimes be revealed by requests for data. (2) Identity management makes data dissemination more costly to build and manage.

Special click-through agreements	(1) Protecting datasets from misuse is an important requirement and necessitates click-through agreements. (2) Click-through agreements and requirements for individual login are minor and not onerous. In practice, multidisciplinary research can cope with limited restrictions on the access and use of data. (3) If a data repository is large and well known, interested researchers will be drawn to it and search-engine discovery issues will be less of a problem. Adequate metadata description can make datasets visible for casual discovery.	(1) Scholarly publications are already available in the "open literature" and can be used by all, even for potentially misguided commercial or religious applications. Trying to regulate use of scientific literature raises a host of difficulties and freedom of expression issues and runs counter to library ethics. Such restrictions should only exist where required to protect the security and privacy interests of human subjects. (2) Click-through agreements greatly complicate some research designs that may aggregate data from different sources. If individual sources come with different (and sometimes ambiguous or even contradictory) contractual obligations, they become less interoperable. For example, some applications including novel visualizations or analyses. These may require re-publication (in some form) of datasets obtained from many sources. If these sources restrict re-publication, uses become limited. Thus, such agreements could hamper some research designs, especially for multidisciplinary investigations. (3) Access restrictions and click-through agreements inhibit information discovery and use. Researchers will be less likely to find relevant information through casual browsing or through search engines.
Open Data (anonymous open access, public domain, no use restrictions)	(1) Researchers invest a great deal of time, effort, and talent in creating data. They also face significant professional risks (and more!) in producing data. This investment should be recognized and researchers should have (some) control how their data are used and by whom. Open publication of data is too risky because professional norms regulating the use of these data are too weak. (2) The public already benefits through expanding scientific knowledge. There is little real public interest in accessing primary data. (3) Not all forms of publication are equally valued. A published dataset, even with many citations, is less professionally valuable to a researcher than a more mainstream article published in a prestigious journal.	(1) Recognition for the contribution of researchers should come through open publication and citation norms. Access restrictions, special agreements, or other encumbrances are not needed except to for privacy and security concerns relating to sensitive information. Moving walls, that release data openly after a few years, may give data creators adequate time to exclusively benefit from their data, while still insuring long-term accessibility. (2) Research is supported by significant public investment, either directly through federal granting programs or less directly through philanthropic sources. Because open data sharing can improve the quality and pace of science, the public interest is best served by reducing access barriers. (3) In addition to social norms, technologies can help promote professional recognition for data publication. If data are published with adequate citation systems, impact measures can be developed. Widely cited, "seminal datasets" can be recognized.

It is also important to mention what was not discussed. Our working group did not specifically address the option of ownership and access restrictions over data as a means for cost-recovery. This was a topic of other working group discussions. In addition, the nature of envisioned uses for data did not receive much discussion. There is great need for additional exploration of how datasets can be used and the implications of various access controls for different use scenarios. For example, privacy concerns are increasingly difficult to address through "de-identification" measures. Again and again, researchers have been able to infer personal identities in de-identified datasets through reference to other public datasets. The failure of de-identification to offer much privacy protection makes access restrictions all the more important for sensitive anthropological information.

On the flip side, there are many use cases for research data that almost require "open data" approaches to dissemination. For example, a researcher may develop a compelling and analytically useful way to visualize shifting social relationships in primate groups. This visualization may draw upon several datasets, and if one or more of those datasets have access and re-publication restrictions, public deployment and presentation of the visualization may be prohibited. Many software approaches supporting visualization make it easy to extract source data. Enforcing data protection measures in a networked environment where there are great demands for aggregation and reuse of data is very difficult.

Institutional and National Claims: Researchers often work in complex contexts where several organizations and even governments may make various ownership claims over data. One participant recounted her experiences where three different entities ranging from a national government (controlling the research site), a European research institute, and her own university made various ownership or control claims over data. These claims become increasingly difficult to manage, especially since, in this particular context, raw data was of little value and needed significant investment in cleanup, annotation, and other processing before they could provide a useful basis for analysis. What credit, recognition, and ownership rights should be given to the researchers who contributed these nontrivial efforts to improving the quality and usability of raw data?

Copyright Complexity: Furthermore, US copyright law and certain other laws in foreign jurisdictions add more complexity to data sharing and ownership. In the US, copyright does not apply to "facts" (or "ideas"); it only applies to fixed expressions having some minimal level of creativity. The dividing line between copyrightable "expressions" and public domain "facts" is very ambiguous. This ambiguity applies equally as much to metadata as it does to data. In our working group discussion, we explored how certain forms of metadata, particularly metadata describing the meaning, methods, constraints and limitations of a dataset would likely be covered by copyright. Other forms of metadata, particularly bibliographic metadata and technical metadata (such as those describing file formats, checksums and collection structures), would be considered more factual and public domain. Thus, the copyright status of much content in databases compiled by researchers and the metadata about those databases must be considered on a case-by-case basis. Furthermore, the European Union has database protection laws that protect compilations of data (including "factual data"). Data sharing and interoperability with EU partners will require addressing EU data protection laws.

Allowing any use involving duplication and modification of copyrighted works requires some form of license that articulates certain permissions, restrictions and requirements. Licenses come in

many varieties, and one often sees informal copyright licenses on scholarly materials stating something like "for educational purposes only." Informal or custom licensing of content may create interoperability problems, because many sets of ambiguously expressed permissions and restrictions for reuse may be difficult to manage. For sharing copyrighted works, use of standard Creative Commons licenses helps to overcome these compatibility and complexity problems. Because these licenses are standardized, they simplify managing and aggregating large sets of commonly licensed content. Creative Commons licenses are also expressed (in RDFa) as standard metadata. This helps with data interoperability goals because the metadata allow users to discover content that are legally interoperable. If datasets have additional "click-though" requirements imposed on them, these requirements should also be expressed in standard metadata.

However, the ambiguous copyright status of much database content together with sui generis legal protections like the EU database laws make scientific data-sharing complex with reference to common baseline standards. It is not clear if Creative Commons licenses could apply to many datasets. Creative Commons considered and ultimately rejected an approach which would have mandated adherence to a single license; put simply, this approach, which implicitly builds on intellectual property rights and the ideas of licensing as understood in software and culture, is difficult to apply in scientific uses. Therefore, Creative Commons, through its science division, Science Commons, is laying out principles for open access data and a protocol for implementing those principles. Creative Commons recently released the CC Zero protocol to be applied to scientific datasets. CC Zero is essentially a public domain declaration that provides a common baseline standard concerning the legal aspects of data interoperability.

Conclusions:

Access controls and ownership of researcher data remains a contentious issue. The questions and debates that arose in this working group need to continue. However, the perfect should not be the enemy of the good. While open data represent an ideal (to some workshop participants), open data may not be feasible or advisable in the short term. Social norms and expectations are continually evolving and it may take time for data publication to see adequate recognition. Thus, pragmatism toward these issues seems warranted.

However, a stance that offers some near-term pragmatism should not result in policies set in stone. The contested nature of this debate should be openly acknowledged as anthropological data sharing and preservation systems are rolled out. Debate should continue and will be better informed as some form of data sharing becomes more commonplace. Governance processes should be in place to continually revisit and adjust access and data-ownership policies into the future.

1. Griffiths, Aaron. 2009. "The Publication of Research Data: Researcher Attitudes and Behaviour." International Journal of Digital Curation 4. http://www.ijdc.net/index.php/ijdc/article/view/101 (Accessed September 16, 2009).
2. Uhlir, Paul F., and Peter Schröder. 2007. "Open Data for Global Science." Data Science Journal 6:OD36-OD53.
3. Narayanan, Arvind, and Vital Shmatikov. 2009. "De-anonymizing social networks." IEEE security & privacy 9.
4. Ohm, Paul. 2009. "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization." University of Colorado Law Legal Studies Research Paper 09. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1450006 (Accessed September 8, 2009).
5. Sweeney, Latanya. 2000. "Uniqueness of Simple Demographics in the U.S. Population." Pittsburgh, PA: Carnegie Mellon University http://privacy.cs.cmu.edu/dataprivacy/papers/LIDAP-WP4abstract.html

Data Preservation Issues

Data Preservation Issues: Summary of Breakout Discussion Group

May 19, 2009

Chair, Carol R. Ember (PI) and Anthony Aristar, Jeffrey Clark, Lisa Conathan, Robert Leopold, Daniel Reboussin and David Glenn Smith

Importance of Digital Preservation

The breakout group stressed the importance of preserving all anthropological research and related materials. The importance of such preservation is obvious as it provides the context for understanding the research undertaken, whether it be qualitative or quantitative research. The appropriate analog should be "lab notebooks" in the physical sciences which are deemed critical for evaluating published research. But historians recognize that other information about the observer is also important and certainly critical for evaluating any biases. So, preservation of any associated materials (dairies, correspondence, etc.) is also of intellectual value.

Why digital preservation?

1. Physical archives have only stored a very small portion of the anthropological corpus. For example, Robert Leopold of the National Anthropological Archives estimated that 500 anthropologists retire each year, but the NAA only acquires 6-8 major collections each year (Schmidt, 2008). And universities, with limited funding, always make choices about which collections they will take and process. The group speculated on why potential donors have been reluctant to give their materials to archives to date (see below). These reasons are important because they suggest why digital preservation may play an important role in future preservation efforts.
2. Much of the anthropological data accumulated is now "born-digital" and physical repositories will find it difficult to preserve this material in a form that will be accessible in the future.
3. Digital preservation can lead to more open access (see report from Access group)

Why have anthropologists been reluctant to give their data to physical archives?

We do not have research that bears on this question. However, we felt that answers to this question would have important implications for understanding the need for AnthroData DPA.

Some reasons that were suggested:

- Some anthropologists think they will give up their ability to work on their data if they deposit it in an archive, however, they are not ready to stop working. Having a digital copy or access to it online will help enormously. It should increase physical preservation.
- Some anthropologists do not think that archives provide enough access for their work—they would prefer digital access of some kind. While theoretically almost any scholar can go to an archive—it is expensive and time-consuming to go to an archive to do research.

- Some scholars think they have to be famous for an institution to consider their collection. This is apparently not the case for the National Anthropological Archives. The director, Robert Leopold says that any collection of an anthropologist will be taken. However, perceptions have reality—if scholars believe this myth, they may not ask an archive to take their material.
- Some simply do not want to face their mortality and do not think about the matter until it is too late.

What are primary data?

The original question posed to this breakout group was how to define primary data. However, the group decided that the distinction between primary data and secondary data is an unneccessary distinction. Moreover, different fields have very different kinds of data. We decided to settle on the more neutral phrase—anthropological research materials. These are what are important to preserve.

What kinds of data need to be preserved?

1. What about multiple formats? There was some debate on whether different forms of data (e.g., handwritten and typed) on the same subjects need to be preserved. The archivists in the group stressed that it is not easy to know in advance how information might be useful in the future, and it is not always clear that two forms are identical, so it is preferable to preserve all forms that are available. Others cited examples of such a practice being a waste of resources, such as preserving a fuzzy and a clear picture of the same subject. However, it was also felt that it is probably more labor-intensive to sort through material to decide what is worth keeping and what is not, so keeping all related materials is probably the best strategy.
2. What about "gray" literature? There was consensus that "gray" literature (a term widely used for research reports in archaeology produced for contract work) should be digitally preserved. Such literature contains important information, sometimes the only information available on certain sites that will be destroyed or severely impacted.
3. What if it is digital but in less-than-desirable formats? There was consensus that if the less-than-desireable formats are all that there is, they should be preserved.

In general, the consensus of the group was that the aim should be to preserve all anthropological research materials.

Can digital object repositories act as long-term preservation?

This was the most controversial issue in the group. Some argued that physical preservation is always the safest long-term preservation strategy for paper. Digital preservation, on the other hand, with migration strategies, may be best for other material such as tapes and objects on computer disks that have shorter life-spans. Others felt that if done properly, digital object repositories can act as long-term preservation strategies and have the advantage of allowing multiple copies to be "housed" in different places (decreasing the risk of destruction from physical or social disasters/upheavals).

However, as mentioned in the history section (NOT YET POSTED), many digital projects do not have plans for long-term preservation in place. If there is any doubt about long-range preservation, both strategies should be pursued.

What efforts moving forward might facilitate future preservation?

Some of the suggestions for encouraging AnthroDataDPA are:

1. Encourage granting agencies to require a preservation plan and provide funding for DPA as part of the research grant. We believe that this will go a long way to promoting DPA.
2. Recommend that guidelines for preservation be made part of the anthropological code of ethics.
3. Develop a donor input system that allows uploading data as research is conducted. Such a system, with appropriate fields/prompts to input necessary metadata will minimize the labor costs to put data into archivable form. Such data needs to be accessible only to the researcher at the ingest and other preliminary stages of the research project. Some fields of metadata can be required at ingest.
4. The researcher is in a better position to enter some metadata compared with an archivist (such as time object was created, place, explanatory captions). There could also be fields for private information that only the researcher would see. Researchers could add information such as their own classification system, keywords, etc.

Schmid, Oona. 2008. Inside the National Anthropological Archives: An Interview with Robert Leopold. Anthropology News, January: 32-33.

Depositor Issues

Anthro Data DPA
Depositors to Archives
June 22, 2009

Douglas A. Black, Lisa Conathan (chair), http://docs.google.com/Doc?id=dhsxxs2_31czp36cdq&hl=en [1] Michael Fischer, David R. Hunt, Mark Mahoney, Marlene Martin, Daniel Reboussin, Dean R. Snow (PI)

Summary

This report of a break-out group of the Anthro Data DPA Workshop addresses issues related to depositors to archives (i.e. people who are depositing, donating or selling material to archives). The topics of discussion, while they all related to this assigned theme, touched on a wide range of activities, including record-keeping as part of the research process, education and outreach to increase awareness about digital archives, and archival appraisal, arrangement, description and preservation. We propose several areas in which a central committee, consortium or other organization can fruitfully provide leadership to guide and improve digital archiving efforts in the discipline of Anthropology.

Priorities for digitizing and digital archiving

Although the methods and standards for digitizing are improving every year, we cannot currently digitize all analog records. Future technological developments may some day make this feasible, but the current cost (in terms of time, effort and money) necessitates that archives, researchers and professional organizations define priorities when embarking on a large-scale digitization projects. All digital archiving efforts must clearly define their scope at the outset of the project.

We recognize the twofold purpose of digitization projects: to preserve records (that is, prevent damage to their integrity and authenticity) and to increase access to records, by allowing new groups of researchers to use them and by enabling new ways of interacting with records that are possible only in a digital environment. We suggest the following criteria for prioritizing digitization projects (not necessarily in order of importance). The relative importance of each criterion must be decided on a case-by-case basis, considering the nature of the material, the resources available and the goals of the project. http://docs.google.com/Doc?id=dhsxxs2_31czp36cdq&hl=en [2]

- Ease of digitization: Some records are 'low-hanging fruit' that may take relatively little effort to digitize because of their condition, organization or description.
- Format of material: Certain formats are inherently unstable and are likely to be deteriorating, e.g. magnetic tape. Material in fragile formats may be prioritized in the interest of preservation.
- Fragility of material: Records that are damaged or that have been stored in less-than-ideal conditions may be fragile and subject to deterioration.
- Current level of access: How accessible are the records already, both to potential researchers and to the creators of the records? Will digitizing increase accessibility?
- Frequency & intensity of anticipated use: Digitization can prevent damage from frequent handling of material. While future use can be difficult to anticipate, factors such as the identity of the creator or interest in the subject matter can be predictive.
- Rarity or uniqueness of subject matter: If the records document a completely unique subject area (e.g. the only known recordings of an extinct language), they may be given priority. In most cases primary data should be given preference over derivative analysis.
- Material in finite custody: An archive may wish to digitize material that is to be repatriated or is only in temporary custody, assuming that such digitization does not violate any agreement with the owners of the material.
- Prioritize value of material within collections: In addition to prioritizing collections, material within collections can be prioritized. In a very large collection, the volume may preclude digitizing all at once. In such cases, a representative sample or a select subset can be digitized first.

Collaborative activity

The importance of collaboration was evident throughout the workshop: Anthropologists can define priorities for the documentation of their discipline and the standards that will enable this. Archivists and scholars can work together to define best practices and encourage their use. Archivists can facilitate the accession of records to archives and ensure the comprehensiveness and efficiency of archiving efforts.

One important outcome of this workshop is an articulation of proposed goals and activities for a central leadership group. Whether it is a committee, a consortium of archives, a series of ongoing workshops or an affinity group, there are several areas of activity that would benefit from central leadership. These are outlined in the following paragraphs.

Survey the record: Before we can take steps to preserve the anthropological record, we should have some idea of the nature, extent and scope of this record. What kinds of records to anthropologists create? What are the challenges to digital archiving? How should we identify priorities for archiving (by format of records, subfield, geographic area, etc.)? Anthropologists and archivists may have very different ideas of what constitutes the 'anthropological record.' Researchers often think in terms of data, while archivists may wish to preserve records that contextualize the data (such as correspondence, photographs and other documentation of the research process, grant applications and research proposals and documentation of abandoned projects that did not result in published products).

Identify challenges to digital archiving: What are the challenges or barriers to progress in digital archiving? Are these challenges mainly social (e.g. related to peoples' expectations and conceptions of archives)? Are they technical (related to infrastructure, user interfaces)? What sort of resources are necessary to undertake a major digital archiving project?

Match material with archives: A central group can help address the problem of 'orphan' archival material (records with no archival home). We can increase the portion of the anthropological record that is archived through outreach and collaboration. For this purpose, it would be appropriate for teams of archivists and researchers to focus on a specific domain. http://docs.google.com/Doc?id=dhsxxs2_31czp36cdq&hl=en [3]

Adapt recommendations and standards: There are many existing standards for digital archiving. It is unreasonable to expect individual anthropologists to interpret and implement these standards on their own. A central group can identify relevant standards, adapt them if necessary to make them relevant within the context of anthropology, and work to encourage their adoption among anthropologists. http://docs.google.com/Doc?id=dhsxxs2_31czp36cdq&hl=en [4]

Develop portals: While it is impractical to propose a single digital archive for the discipline of anthropology, it is possible to create portals to data or metadata. http://docs.google.com/Doc?id=dhsxxs2_31czp36cdq&hl=en [5]

Preparing material to be archived: A central organization can help anthropologists prepare material to be archived. This includes recording information and describing context that could otherwise be lost or recorded inaccurately (such as the purpose of the research project and dates, places and descriptions of each item or file). http://docs.google.com/Doc?id=dhsxxs2_31czp36cdq&hl=en [6]

Additionally, digital archives should take advantage of technological developments (especially those in the area of social media) in order to collect information from researchers about material. http://docs.google.com/Doc?id=dhsxxs2_31czp36cdq&hl=en [7]

Education and Outreach: There is a need for outreach to scholars and other practitioners in the discipline of Anthropology to increase awareness about digital archiving. Initial steps to educate anthropologists (such as panel discussions and workshops at regional and national conferences) are within immediate reach and should begin in the next year. http://docs.google.com/Doc?id=dhsxxs2_31czp36cdq&hl=en [8] Larger-scale efforts will take some planning, including application for funding.

One of the important goals for educational efforts is to convince anthropologists that it is advantageous to participate in and contribute to digital archiving efforts (i.e. that the archive provides contributors with a valuable service, minimally a back-up copy, and that their contributions have broad value for the discipline). We discussed several ways to frame archiving activity, emphasizing the benefits of emerging ways to interact with digital data (e.g. peer-to-peer sharing, backup service). Field Methods curriculum should also include training in data collection and management, including planning for archiving.

- http://docs.google.com/Doc?id=dhsxxs2_31czp36cdq&hl=en [1] Lisa Conathan prepared this report based on the break-out session on Depositors (May 20-21, 2009) and the discussion after a presentation to the Anthro Data DPA workshop (May 21).

- http://docs.google.com/Doc?id=dhsxxs2_31czp36cdq&hl=en [2] ViPIRS (http://library.nyu.edu/preservation/movingimage/vipirshome.html http://library.nyu.edu/preservation/movingimage/vipirshome.html) is an example of a tool that tracks assessment data for audiovisual preservation projects.
- http://docs.google.com/Doc?id=dhsxxs2_31czp36cdq&hl=en [3] A collaborative, strategic approach to documenting specific topical domains is reviewed and critiqued in Malkmus, Doris. 2008. Documentation strategy: Mastodon or retro-success? American Archivist 71(2):384-409.
- http://docs.google.com/Doc?id=dhsxxs2_31czp36cdq&hl=en [4] InterPARES (http://www.interpares.org/) is a major international research effort to define standards for digital records.
- http://docs.google.com/Doc?id=dhsxxs2_31czp36cdq&hl=en [5] Portals can take many forms; examples include the Digital Archive Network for Anthropology (http://www.dana-wh.net) and the Open Language Archives Community (http://www.language-archives.org).
- http://docs.google.com/Doc?id=dhsxxs2_31czp36cdq&hl=en [6] Digital Antiquity (http://www.digitalantiquity.org/ http://www.digitalantiquity.org) provides a model for the recording of collection-level metadata when depositing data.
- http://docs.google.com/Doc?id=dhsxxs2_31czp36cdq&hl=en [7] A discussion of efforts to make archival description more interactive can be found in Yakel, Elizabeth, Seth Shaw and Polly Reynolds. 2007. Creating the Next Generation of Archival Finding Aids. D-Lib Magazine 13(5/6).
- http://docs.google.com/Doc?id=dhsxxs2_31czp36cdq&hl=en [8] The field of Linguistics has been successful in increasing awareness about archiving and can provide models for educational efforts. See, for example, the E-MELD school of best practices: http://emeld.org/school/index.html.

Digitization Issues

Report: Working Group on Digitization Issues

Richard Sherwood, Richard Mahoney, and Helen Aristar-Dry (chair)

Digital preservation of scientific data is a relatively new enterprise; but as early as 2001 plans were underway to create distributed digital archives of anthropological material (Clark et al, 2001). Various types of anthropological material lend themselves to preservation in a digital archive. The working group identified at least six types. Examples of these are provided in Table 1.

Type	Examples
Images	Photographs, maps of excavation sites, biomedical images (MRIs, radiographs)
Texts	Field notes, annotations, excavation plans
Audio	Recordings of songs, conversations, oral histories
Video	Recordings of cultural events, conversations, archaeological excavations
Databases	Database of skull measurements, lexical items
3-D scans	Scan of fossil or artifact

Table 1: Anthropological data

To preserve such data for long term use, researchers must ensure long term ‘intelligibility’ in both human and computational terms. ‘Human intelligibility,’ of course, refers to the ability of future researchers to understand the information; this is too often compromised by the lack of documentation accompanying the digital file. ‘Computational intelligibility’ refers to the ability of future hardware and

software to interpret the file format; and this can be compromised by the pace of technological change. Since the 1996 report of the Taskforce on Digital Archiving (Garrett and Waters, 1996), it is commonplace to remark on the 'digital dark age' which is threatened by the rapid obsolescence of physical recording media and the equally rapid obsolescence of operating systems and file formats. Simons (2006) noted that physical media have declined in durability over the years, contrasting the long term legibility of inscriptions in stone with the many different types of storage media in use in the past 25 years (5.25" floppies, 3.5" floppies, Zip drives, Memory sticks, CD's, DVDs, Blu-ray discs). The obsolescence of operating systems and file formats is even more striking: current version of MS Word cannot read documents created in Word 1.0.

To address the threat to human intelligibility, researchers are advised to: (a) write metadata and keep it with the file, (b) use standardized vocabularies and abbreviations wherever possible, (c), document any idiosyncratic annotation or abbreviations, and (d) use Unicode for character encoding or, at the least, document any special characters used to represent international alphabets.

To address the threat of technological obsolescence, Simons (2006) recommends that researchers create an archival master in an enduring file format and deposit the archival master in a preservation archive. A preservation archive is an established institution committed to long term preservation of the digital object; a distinguishing characteristic is that a preservation archive will have a technology migration plan on which to found its claims of long term digital accessability. Thus it contrasts with a 'web archive,' which is often only a website serving information from a database or file directory. Web archives rarely serve genuinely interoperable material, and they regularly disappear in response to changes in institutional servers or in the responsibilities of the archive creator.

What is an 'enduring file format'? In the acronym created by Simons, it is a file that offers LOTS. In other words, it is Lossless, Open, Transparent, and Supported by multiple vendors. Each of these desiderata deserves some discussion.

Lossless: A lossless file format is one in which no information is lost through file compression. It is uncontroversial to say, for example, that an archival master should be uncompressed and unedited (AHRC, 2009). However, copies may, of course, be made from the archival file, and these can be altered to serve as working or presentation copies. Professional archivists usually recommend that the archival master be copied once, to make a 'presentation master,' and that compressed and edited copies be made from the presentation master, not the archival master. Although digital copying does not harm the original file if done correctly, use of a presentation master is probably good advice: some media programs compress automatically when they save a file; and to find this out too late is to irrevocably lose part of the information on the archival master.

Although uncompressed file format s are preferable to even those with lossless compression, lossless compression is an option if uncompressed files are so large (e.g., video) that their storage is impractical. Lossless compression algorithms typically remove only redundant information (e.g., pixels of the same color in an image) and allow the full content to be recovered through the use of a decoding algorithm. 'Lossy' compression, on the other hand means that the so-called 'irrelevant' information can never be recovered; thus it is to be avoided for highly valued material. Although the difference between a compressed file and an uncompressed file may be indistinguishable to human ears and eyes, in creating a scientific archive of irreplaceable material (e.g., songs and ceremonies of a vanishing

culture), we should remember that the scientific instruments of the future may be able to extract more information from the 'noise' on an uncompressed file than we are currently able to perceive.

Table 2 shows some common extensions of uncompressed file formats and formats employing lossless and lossy compression.

Type	Uncompressed	Compressed (Lossless)	Compressed (Lossy)
Audio:	.wav, .aiff, .au (pcm)	.ape, FLAC, TTA	.mp3, .aac, .wma
Images:	.bmp, tiff w/o LZW	.tiff (or .tif) w/LZW .png .gif (grayscale)	.jpg
Video:	rtv	JPEG-2000	MPEG-2, DV, MPEG-4
Text:	.txt	.zip	NA

Table 2: File extensions of compressed and uncompressed formats (Aristar-Dry, 2008)

Open: Openness refers to the fact that some file format specifications are publically available; for example, html, XML, pdf, and rtf are all 'open standard.' This means that any software engineer can develop programs that can read these file formats. By contrast, information in proprietary file formats will be lost when the vendor ceases to support the software. "Open standard" is different from "open source," i.e., software whose source code is publicly available. Examples of open source software include Open Office and Mozilla Thunderbird. Open source software usually creates files in open standards. And proprietary software usually doesn't (though there are exceptions, e.g. Adobe pdf). But for long term intelligibility, open standards are more important than open source software. Table 3 below lists some

Development	Open	Proprietary
Open	.txt, .html, .xml, .odf, .csv	NA
Commercial	.rtf, .pdf	.doc, .xls, .ppt

Table 3: Open and proprietary standards (Aristar-Dry, 2008)

Transparent: The file format requires no special knowledge or algorithm to interpret, because there is a one-to-one correspondence between the numerical values sent to the computer and the information they represent. Plain text, for example, has a one-to-one correspondence between the characters and the computer-readable binary numbers used to represent them. Similarly, the PCM (pulse code modulation) codec, which is employed by .wav, .aiff, and cdda files, has a one-to-one correspondence between the

numbers and the amplitudes of the sound wave. Thus plain text files (.txt) can be read by any software program that processes text. And PCM signals can be interpreted by virtually all audio programs. By contrast, .zip and .mp3 files require implementation of a complex algorithm to restore the original correspondences.

Today many programs provide automatic decoding of the common encoded formats. But we cannot be certain that these programs will not become obsolete. In the distant future, some of the encoding algorithms may be lost; and, at that point, interpreting compressed and opaque files will become a costly scientific endeavor.

Supported by multiple vendors: Just as lack of compression and transparency are paired in file formats, use of open standards and support by multiple vendors go together in software development. Open standards are more likely than proprietary standards to have wide vendor support, because development using open standards is typically less costly. If a file format is open, there is no inherent barrier to creating another program that handles it. It is not necessary to reverse engineer the format or purchase the specification from the developer. And the more software applications that handle a file format, the less likely that format is to fall victim to hardware and software obsolescence.

As noted above, these recommendations are intended to apply to the archival master, not to presentation copies or working copies. However, even with archival masters, some caveats are in order. Transparency, for example, is not possible with some advanced visualization techniques, e.g., 3-D scanning, CT (computed tomography), GIS. And sometimes the ideal is simply not achievable, either in format or equipment. For example, a laser scanner is recommended for x-rays; but these machines cost upwards of $20,000 and are often out of reach of small projects. Some archivists, therefore, speak not only of ‘best practices’ in digital preservation, but also of ‘good practices’ or even ‘pretty good practices’—i.e. practices that will suffice when the ideal is unattainable. They also emphasize that situation and type of data must always be taken into account.

For example, best practice is to record audio at 24bit, 96 Khz sampling rate, in stereo; and this is ideal for data which will be subjected to phonetic analysis. BUT 16bit, 44.1Khz may be adequate for an oral history (especially since playback machines for 24bit/96 Khz are not widely available). Similarly, best practice is to scan images at 600 dpi. But 300 dpi may be preferred in some cases—for example, on x-rays where scanning with increased dpi would actually make the x-ray less intelligible because it exceeds the resolution of the image. And best practice for text is to output plain text annotated in XML (which captures content, not just formatting). But software to support XML writing and editing isn’t always available; in that case, good practice is to use any kind of structured data format (e.g., a spreadsheet or a word processing format with a stylesheet), and to provide metadata and explanatory annotations for the content.

Technical recommendations for digital preservation are, of course, a moving target. Technology changes so rapidly that regular consultation of up-to-date websites is recommended for all anthropologists interested in preparing their data for long term digital preservation. The bibliography at the end of this report lists some general resources which are worth investigating, as well as several specific to audio, image, and video standards. More such resources will no doubt become available as more domain experts become involved in adapting general recommendations for digital archiving to the goals and procedures of specific disciplines.

Works Cited

- ARSC Technical Committee. 2009. Preservation of Archival Sound Recordings, Version 1, April 2009. http://www.arsc-audio.org/pdf/ARSCTC_preservation.pdf
- Clark, Jeffrey T., Brian M. Slator, Aaron Bergstrom, Francis Larson, Richard Frovarp, James E. Landrum III, William Perrizo. 2001. "Preservation and Access of Cultural Heritage Objects through a Digital Archive Network for Anthropology," Virtual Systems and MultiMedia, International Conference on, pp. 28, Seventh International Conference on Virtual Systems and Multimedia (VSMM'01).
- Aristar-Dry, Helen. 2008. Preserving Digital Language Materials: Some Considerations for Community Initiatives. In Language and Poverty (ed. Wayne Harbert, Sally McConnell-Ginet, and Amanda Lynn Miller). Multilingual Matters. 202-222.
- Garrett, John, and Donald Waters. 1996. "Preserving Digital Information: Report of the Task Force on Archiving of Digital Information commissioned by the Commission on Preservation and Access and the Research Libraries Group.Washington, DC: Commission on Preservation and Access." http://www.rlg.org/ArchTF/tfadi.index.htm
- Howard, Roger. Wed Apr 09 2003. http://eclipse.wustl.edu/~listmgr/imagelib/Apr2003/0011.html
- Simons, Gary F. 2006. Ensuring that digital data last: The priority of archival form over working form and presentation form. An expanded version of a paper originally presented at the: EMELD Symposium on "Endangered Data vs. Enduring Practice," Linguistic Society of America annual meeting, 8-11 January 2004, Boston, MA. http://www.sil.org/silewp/2006/003/SILEWP2006-003.htm

Additional Resources on Digital Preservation

- NARA (2004):
  http://www.archives.gov/preservation/technical/guidelines.html
- New Jersey Digital Highway Project(2007?):
  http://www.njdigitalhighway.org/digitizing_collections_libr.php
- NINCH (2002): http://www.ninch.org/guide.pdf
- E-MELD School of Best Practices in Digital Language Documentation: http: emeld.org/school/
- Additional Information on Audio
- Sound Directions (2009):
- http://www.dlib.indiana.edu/projects/sounddirections/papersPresent/index.shtml
- CDP Digital Audio Working Group (2006)
  http://www.bcr.org/cdp/best/digital-audio-bp.pdf
- U. of Maryland Libraries (2007):
  http://www.lib.umd.edu/dcr/publications/best_practice.pdf (2007)
- Additional Information on Images and Video
- Visual Arts Data Service (2000?):
  http://vads.ahds.ac.uk/guides/creating_guide/sect31.html
- California Digital Libraries (2008):
  http://www.cdlib.org/inside/diglib/guidelines/bpgimages/
- Washington State Library:
  http://digitalwa.statelib.wa.gov/newsite/best.htm
- BCR Digital Images Working Group (2008):
  http://www.bcr.org/cdp/best/digital-imaging-bp.pdf

If the working copy is the primary copy—as, for example, during the ongoing creation of a database—it is important to export the information regularly into an enduring file format. For databases (which are

usually managed by proprietary software) this means to export the data regularly into properly documented plain text. A .txt file with informative XML markup is ideal, but often the XML automatically output by a program will be only minimally helpful to someone trying to make sense of the file. In that case, a file including metadata identifying the fields and tables should be created and stored with the database output.

For example, Acrobat 7.0 will automatically compress large pdf files (see: http://www.planetpdf.com/forumarchive/166948.asp). Most importantly, however, as of this writing, most video capture programs automatically compress the audio track along with the video when it is downloaded to a computer. For that reason, linguists and musicologists are advised to make a separate audio recording, using a device like a hand-clap at the beginning to aid in synchronizing the files later on. See: http://emeld.org/school/classroom/video/field.html#1006

As noted by a Senior Media Specialist at the Getty Museum, "Uncompressed data is trivial to decode, compressed data often is not. This makes for easier long-term viability of the file . . . . " Furthermore, uncompressed data is less prone to loss: "Lossless compression means that a single bit in the compressed file may represent multiple bits in the uncompressed version. This magnifies potential damage caused by bit corruption. In an uncompressed file a single flipped bit will have little overall impact on the renderability of an image. In a lossless compressed file depending on whether the corruption is in the dictionary (in the header) or in image data it can have a larger effect. And in a lossy compression scheme a single bit corrupted can be extremely noticeable." (Howard, 2003).

Technically, .wav and .aiff are container formats, file structures which allow combining of audio/video data, tags, menus, subtitles and some other media elements. They could theoretically contain compressed audio formats, but in practice they usually contain PCM (pulse code modulation) data, which is an uncompressed format.

Apple audio codec (.aac) and Windows media audio (.wma) both have a lossless version. Confusingly, both the lossless and the lossy compression formats use the same file extension.

Funding and Sustainability Issues

Funding and Sustaining Support for Long-Term Preservation (and Steps to Promote Profession-Wide Coordinated Efforts): Breakout Group Report

Breakout Group Participants

Anthony Aristar, Helen Aristar Dry, Andrew Bennett (Political Scientist observer), Jeff Clark, Carol Ember, Keith Kintigh, Jennifer Serventi (NEH observer), Matt Tocheri, Laura Welcher, Peter Wittenburg, and Robert V. Kemper (chair)

The Problem

For more than a century, anthropologists have been collecting data about the human experience. These data include the details of human history, the characteristics of the human species and related primates, the variety of languages spoken and written, and the cultural features of the world's societies. Unfortunately, many data already have been lost to us and will not be available to future generations. Failure to record data properly, failure to store it appropriately, and failure to sustain our ability to "read" the data with changing technological platforms are the principal causes of data becoming

compromised or lost. The present workshop and this breakout group is concerned with the possibilities of using new digital and Internet technologies to save anthropological data – in archaeology, biological anthropology, cultural anthropology, and linguistics. If we are successful in this great enterprise, we can stop our cultural heritage and biodiversity from being destroyed, lost, or so poorly maintained as to be worthless to future generations of scholars and communities in the U.S. and around the world.

Needs Prior to Seeking Project Funding

The members of the Breakout Group came to the table with recognition that the "problem" needs to be divided into smaller elements and those stages suitable for funding need to be established. While the difficult problems associated with dividing the world of anthropological data into components are being taken up by other Breakout Groups, we focused our attention on the need to be clear about the distinct paths to finding. To this end, we agreed that a project could be conceptualized into three principal elements:

(1) Start-up funding,
(2) Matching funds for challenge grants, and
(3) Long-term funding for sustaining the enterprise.

Having determined that these three domains could be specified without regard to sub-disciplinary considerations, we agreed that all anthropologists will need to develop management structure(s) to direct project(s) to carry out Digitization, Preservation, and Access – thus, the initialism DPA.

Developing a Project Structure

Members of the Breakout Group discussed the need to find multiple institutions and individuals of stature to cooperate in initial round of proposal(s) and project(s). These institutions and individuals would serve as "champions" for the project(s). Their participation would ensure the persons at nongovernmental organizations, governmental agencies, and other relevant institutions that an anthropological DPA project is of critical importance to the sciences, the social sciences, and the humanities. We also discussed the importance of including foreign entities in appropriate project planning. This is important wherever data have been acquired in other countries but now reside in depositories in the United States, or the data remain in other countries where they are studied by U.S.-based scholars.

Potential Frameworks for Project Funding

The members of the Breakout group spent considerable time considering the ways in which a DPA project could be framed. First, we discussed broad themes that go far beyond anthropology per se. Such themes might emphasize World Heritage or Biodiversity. Second, we worked on the obvious focus on Anthropology as a Discipline. Third, we considered the possibilities inherent in doing projects related to the sub-fields of anthropology (in the United States, these might be archaeology, biological anthropology, cultural anthropology, and linguistic anthropology, although some would feature applied anthropological as a fifth domain. A fourth approach would emphasize regional specializations (e.g., the historical recognized major culture regions of the world – e.g., North America, Latin America, Europe, Africa, Middle East, Asia, and Oceania) within which anthropology as a discipline would play a major role.

Two Examples

1. Anthropological data often involve combinations of sub-disciplines and regions. Among our Breakout Group, we focused on the excellent example of the Archive of Indigenous Languages of Latin America, known by its acronym AILLA. This archive currently is directed by Prof. Joel Sherzer of the University of Texas at Austin. For further information, interested persons can consult the archive's web site, http://www.ailla.utexas.org http://www.ailla.utexas.org. AILLA is a digital archive of recordings and texts in and about the indigenous languages of Latin America. Access to archive resources is free of charge. Most of the resources in the AILLA database are available to the public, but some have special access restrictions. Users must register and login in order to access any archive resource, but they can browse the catalog information without registering. To get started, users read "How to Use the Archive," or go directly to the "Search" page.

2. A different kind of archive of anthropological data involves the work of an individual scholar or small group of scholars who specialize on a particular topic or research site. For this category, we discussed the archive being developed for the Tzintzuntzan Ethnographic Project, based on the long-term field research of Prof. George M. Foster and his colleagues, including Robert V. Kemper, Stanley Brandes, Peter Cahn, and others, in the community of Tzintzuntzan, Michoacán, Mexico. At present, the physical materials are being archived at two locations, the Bancroft Library (UC-Berkeley) and Southern Methodist University (Dallas). The digitized versions of the data are being produced at the SMU site, with the goal that the resultant data (fieldnotes, census data, slides, photographs, negatives, maps, etc.) can be brought together for use by scholars and by members of the community of Tzintzuntzan.

Potential Funding Sources – "The Usual Suspects"

The members of the Breakout Group began our discussion of funding issue b y agreeing that an anthropology DPA project would need support from agencies to which the discipline's scholars have turned for decades. We agree that support from the National Science Foundation, the Wenner-Gren Foundation, and the National Endowment for the Humanities. Beyond this set of the "usual suspects," other potential funding sources should include the Andrew Mellon Foundation, the Institute of Museum and Library Services, as well as other federal, state, and local governmental agencies; private foundations; corporations and their foundations; universities, museums, and archives; and diverse international entities. Some of the funding could come through cost-sharing arrangements as well as outright grants.

Special Challenges of "Legacy" data

Members of the Breakout Group agreed that it is important to identify scholars and projects with data sets to be included in an Anthropology DPA. The challenge before us is establishing a system of priorities for processing their materials. There will be a need to find funding for digitization, preparation of metadata, and developing systems for long-term access to legacy collections. Many of those present are aware of the work done by numerous anthropologists and archivists during the 1980s and 1990s to establish the Council for the Preservation of Anthropological Records (CoPAR). Funded with financial support from the Wenner-Gren Foundation (and the enthusiastic leadership of Dr. Sydel Silverman, then the President of the Foundation), CoPAR was "dedicated to helping anthropologists, librarians, archivists, information specialists and others preserve and provide access to the records of human diversity and the history of the discipline." After a series of conferences and meetings, CoPAR

produced a useful guidebook, Preserving the Anthropological Record (1992, 1995), and established an Internet presence, currently available through the Smithsonian Institution's National Anthropological Archives at http://www.nmnh.si.edu/naa/copar/bulletins.htm http://www.nmnh.si.edu/naa/copar/bulletins.htm.

Challenges of "Contemporary/Future" Data

Members of the Breakout Group discussed the need to educate anthropologists about their professional responsibilities to establish appropriate systems for processing, preserving, and providing access to data. Realizing that CoPAR (see above) was designed to produce guidelines for these same purposes, and wishing to avoid a similar fate of falling by the wayside when key individuals reach the point of retirement or changing professional interests, we considered the need to find funding for these educational efforts among contemporary anthropologists.

Challenges of Access

We discussed the need to develop systems of participation in the Anthropology DPA enterprise for institutions and individuals in anthropology and beyond the discipline. Such systems might include free access, premium access, subscriptions, cost-per-search, cost-per-download, etc.

Challenges of Sustaining the Project into the Future

We concluded our work together by considering the following questions, to which we have no answers at this point in the development of the Anthropology DPA project:

- What structures could be put in place to adapt to changing technical standards, especially related to digitization and interoperability/integration?
- What can we do to ensure continuing participation of anthropologists in the project for decades to come?
- What organization and institutions might be "shovel ready" if funded became available?
- What individuals have data collections appropriate for an initial demonstration project?

Conclusion

The members of the Breakout Group realize that the challenges ahead are far greater than the resources that are likely to become available to meet them. This means that establishing priorities will be an initial and long-term issue if an Anthropology DPA project is to be successful for scholars and for our publics, in the United States and around the world.

Metadata Issues

Metadata Break-out Group:

Jeanne Altmann, Eric Delson, Eric Kansa, Robert Kemper , Tom Moritz (Chair), Joel Sherzer "Data" and "Metadata"

"…'data' are defined as any information that can be stored in digital form and accessed electronically, including, but not limited to, numeric data, text, publications, sensor streams, video, audio, algorithms, software, models and simulations, images, etc." — Program Solicitation 07-601 "Sustainable Digital Data Preservation and Access Network Partners (DataNet)"

Taken in this broadest possible sense, "data" are thus simply electronic coded forms of information. And virtually anything can be represented as "data" so long as it is electronically machine-readable.

Our group agreed upon a more pragmatic definition of "data" as measurements, observations or descriptions of a referent — such as an individual, an event, a specimen in a collection or an excavated/surveyed object — created or collected through human interpretation (whether directly "by hand" or through the use of technologies).

Metadata are descriptive documentation essential to informing the process of data creation, collection, management and preservation. (This process is now commonly referred to as "data curation".) Metadata provide information about the original referent, the collection processes, rules of collection, as well as descriptions of data management processes and provisions for access and use of the data (such as licensing of data to specify permitted uses).

Metadata provide key contextual information to facilitate understanding and are intended to assist research within known and predictable scientific domain(s). However, in the Web environment, metadata may also enable discovery and use in as yet unanticipated fields of research; hence, careful efforts should be made to make the descriptive content of metadata intelligible to scientists beyond a very limited scientific expertise.

From a pragmatic perspective, it was agreed that metadata creation is an ongoing process not a single event, and that metadata usefully may grow over time by accretion, asynchronously, by the efforts of properly qualified contributors. The question of appropriate control over who may contribute to the ongoing development of metadata should be addressed.

Metadata Accessibility, Costs, Commonalities

It was also recognized that metadata creation involves serious investment and that care must be taken to insure optimal and parsimonious approaches. The notion of minimally adequate "fitness for use" is one useful test of a metadata scheme. Our group agreed that for purposes of "discovery" (identification and location) of data – across the four anthropological fields represented in the workshop – time, place and manner/mode of collection may be minimally adequate. Beyond "discovery" — for more in-depth research and education purposes — metadata must provide richer descriptive content and detailed contextualization. But in that each metadata element is essentially a cost vector, great care should be taken to balance cost and benefit in identifying case-specific minimum adequacy. It was noted that by careful use of normalization, inference and recursion significant efficiencies can be achieved in the design and implementation of metadata schema.

Dublin Core Metadata

The group discussed the possible application of the Dublin Core Element Set:

From Guide to Best Practice: Dublin Core (DC 1.0 = RFC 2413)
Final Version 12 August 1999

The 15 Dublin Core Elements

- Resource Type
- Format
- Title
- Description
- Subject and Keywords
- Author or Creator
- Other Contributor
- Publisher
- Date
- Resource Identifier
- Source
- Relation
- Language
- Coverage
- Rights

Although the metadata scheme is in wide use – and particularly in the OAI-PMH protocols — it was recognized that some Dublin Core elements may be poorly suited for anthropological applications. For instance, how do we describe "local contributors"; as "author / creator", as "other contributors", or "source"? In some contexts a local community member may consider themselves to be a "steward" or "keeper" of knowledge, or as an advocate for the community. We thus believe that before adoption for widespread use in anthropology, broad metadata standards such as Dublin Core, must be closely scrutinized and modified ("qualified") to meet domain requirements.

Ethical Dimensions: Professional Community and Beyond

Data curation is best informed by the researcher or researchers primarily responsible for the collection/ creation of the data. (Michener notes: "Comprehensive metadata counteract the natural tendency for data to degrade in information content through time." (Michener, Ecological Informatics1 (2006) 4.))

Our group believes that timely generation of appropriate metadata is a professional and ethical obligation and; in certain contexts, descendant and or local communities should be involved in the process of metadata creation. This was seen to require normative change among individuals, disciplines, organizations/institutions and governments. It follows that funders, both private and public sector, must recognize metadata — and data curation more generally — as essential and legitimate expenses that must be adequately supported.

The group discussed various incentives and disincentives ("carrots and sticks") pertaining to metadata creation. It was recognized that in NSF itself there are variations from program to program concerning data curation and that actual enforcement of requirements for data curation can be highly variable.

Privacy Issues

Privacy and Ethical Issues

Jeff Altschul, Ted Bestor, Jeff Good, Tim Kohler, Robert Leopold, Susan Penfield, Richard Sherwood (chair), Joel Sherzer, David Glenn Smith

When considering the accessibility of data either to the general public, or even a limited professional audience, a number of concerns are immediately evident. As anthropologists working with humans as groups or individuals, there is an implicit trust between research and subject that participation will not cause harm in any way to the individual. There is also, of course, an explicit set of guidelines and expectations set forth by institutions and associations detailing the behavior of the researcher towards participants. In terms of specific research protocol, most institutions have an internal review board (IRB) to evaluate and approve research conducted on humans. For non-human research, a similar panel, the institutional animal care and use committee (IACUC) oversees research to assure ethical treatment of all animal subjects.

Within the context of creating large, publicly accessed archives that may include a variety of information resulting from the research process such as primary data, personal notes, correspondence, etc., one of the most pressing ethical issues that must be addressed is ensuring the privacy of the research subjects. During the discussion regarding privacy, it became clear that there are complex issues unique to each anthropological subfield with respect to collection of data, but also with regard to the long term archiving of data. An important consideration is the dynamic nature of privacy concerns with regard to archiving. A summary of the major points arising during discussion are provided below. In the end, the group was in a better position to raise issues and consider the advantages and disadvantages associated with different solutions than to devise specific, concrete recommendations. We think this accurately reflects a lack of consensus at present regarding many key issues of ethics involving digital anthropological data.

Code of Ethics

It is important to note that numerous societies have a code of ethics readily available via the web. Because of the interrelated nature of Anthropology, many aspects of these codes overlap. A list of the url's for the codes examined by this group is included as an appendix. Not surprisingly, these codes tend to provide general information or guidance regarding research protocols. Common themes include the following from the AAA guide "Anthropological researchers must do everything in their power to ensure that their research does not harm the safety, dignity (psychological well-being), or privacy of the people with whom they work, conduct research or perform other professional activities." With regard to privacy it is suggested that "Anthropological researchers must determine in advance whether their hosts/providers of information wish to remain anonymous or receive recognition, and make every effort to comply with those wishes."

What kinds of data may need protection?

As the different subfields of anthropology may deal with very different types of data, it was important to identify the range of potential data that may be subject to privacy considerations. Common to all subfields was the basic privacy of the individual. If an individual wishes to remain anonymous then that wish must be honored. There are a number of situations where ensuring anonymity is not a simple procedure. Common practices for deidentification of individuals include the use of a pseudonym or

alphanumeric ID in place of the individual's name. This will frequently not be sufficient for instance, in modern genetic analyses, it is relatively simple to identify individuals, families, and familial relationships from available data. Similarly, it may be a simple matter to identify an individual in linguistic or cultural work if the population of interest includes only a few individuals such as in the case of many endangered languages.

In several instances, locations were identified as in need of protection. The most obvious of these may be in the case of archaeological site location where knowledge of the location may lead to unwanted access. Similarly, the location of sacred sites should be afforded the same privacy considerations as individuals or groups. Finally, knowledge of commercially exploitable resources within a populated area may need to be protected to prevent unwanted exploitation.

Length of Protection

While, at first thought, it may seem obvious that if an individual or group has requested anonymity, that request should be honored in perpetuity, this may not be the case. What may have seemed like sensitive information to a participant at one point in time may not seem so at a later point. The reasons for such changes in attitude may be numerous and variable. The dynamic nature of the concepts of "private" and "sensitive" must be kept in mind when setting up archives.

In general IRB's may suggest that the appropriate length of protection be "until no foreseeable harm can be done." As noted this may be difficult to determine and may change as time passes. The decision to maintain or remove privacy criteria can reside both with a single individual as in the case of the donor or may be the responsibility of a group.

Access Issues

One of the basic premises behind digitally archiving of data is to make these resources easily available to a group beyond that of the donor/collector. In many cases, the intention is to make the resources freely available to the general public without restriction. As noted, however, in many cases restrictions will have to be placed on access to protect privacy or to accede to a donor's wishes. In short, it is clear that in some circumstances, differential access is sometimes a necessity. The most common differential access currently practiced is with regard to the dichotomy of scholars vs. nonscholars recognizing that the distinction between the two groups is sometimes difficult to identify (though some fields, like linguistics, also give special privileges to communities from which data are drawn. Restricted access is frequently placed on museum collections primarily for the safety of delicate collections but a number of online databases also require registration and validation of the user prior to access.

In unusual situations, access may be restricted to a small number of individuals or even a single individual. Situations in cultural anthropology and linguistics were described where an informant had specifically stated that a given individual (e.g., a brother-in-law) could not be privy to the story being told, but that the rest of the world could be given full access. Such situations become very difficult to monitor.

Sensitive Data

In situations where access to restricted data is granted to specific individuals it is likely that certain covenants are placed upon the use. This may be as simple as the request that the user properly cite the donor/archive. It is also possible that stronger restrictions are placed on the use such as limiting the use

of data for future studies or limiting the sharing of data between individuals. In terms of digital archives where sharing is easy, it is possible, and probably likely, that such covenants do not always transfer with the data allowing for violation of promised privacy by secondary or tertiary users. In these instances the question arises first as to how restrictions can be maintained and second as to how such violations can be addressed and violators punished. It was agreed that any action would be difficult and costly.

For these reasons, it is likely that researchers may have to establish separate IRB protocol for archiving and web access for data in cases where it is decided to digitize and preserve the data after the initial research. It was also noted that some IRBs may mandate that sensitive data are destroyed following the project's completion. If that is the case, the original protocol must be amended and an additional protocol submitted.

Privacy of the User

The purpose of digital data archiving is twofold, preservation of valuable resources is the primary purpose. Digital archiving provides a means to store very large amounts of data in a relatively small area while removing many of the agents that can affect long term preservation problems. Of course, archiving is only valuable if the archive is to be used by future researchers (meaning, in this case, academics), or individuals with an interest in the subject. The user, therefore, becomes an integral part of the system. Given the aspects of privacy and ethics discussed above, there was a significant discussion with regard to the rights of the user.

With regard to the user, the discussion ranged from allowing the user total free access with an assurance of complete anonymity to a system requiring registration of users along with a login prior to access of the archive. Ease of access increases the potential use and benefit of the archive while posing no threat to the user. Restricted access (ranging from simple login to full-scale registration and authorization of users) may decrease the benefit to the user (if they decide to only focus on those archives with full, easy access) while increasing the benefit to the originator of the data. Benefit to the originator includes the ability to document usage which may be used in some way as a measure of the importance of the archive (potentially beneficial in situations such as promotion) or to provide assurance that archival material is used, and cited properly.

Spheres of Responsibility

Ultimately it was clear to the group that there are different spheres of responsibility for individuals at each level. The field researcher collecting data must first and foremost be responsible to the participants agreeing to maintain and protect the privacy desired by the individual. This is monitored by the IRB of the researcher's institution.

The archivist is likely to be removed from the original participants and is primarily responsive to the researcher. Just as the researcher has rules and restrictions placed upon them by the participants, the archivist can expect to have such covenants placed on them by the researcher. This again may be monitored by an IRB with ongoing protocols.

Finally, the user, anonymous or not, has an obligation to treat the data with the respect that researcher and archivist have established. This will ensure the ethical treatment of all subjects and subject information.

APPENDIX:

BELOW ARE A LIST OF WEB RESOURCES REGARDING ETHICAL CONDUCT FOR THE PRIMARY ANTHROPOLOGICAL ASSOCIATIONS WITHIN THE USA.

ARCHAEOLOGY

Register of Professional Archaeologists: http://www.rpanet.org
Society for American Archaeology: http://www.saa.org/AbouttheSociety/PrinciplesofArchaeologicalEthics/tabid/203/Default.aspx

PHYSICAL ANTHROPOLOGY

American Association of Physical Anthropologists: http://www.physanth.org/association/position-statements/code-of-ethics

LINGUISTICS

Linguistic Society of America: http://www.lsadc.org/info/pdf_files/Ethics_Statement.pdf

CULTURAL ANTHROPOLOGY

American Anthropological Association: http://www.aaanet.org/profdev/ethics/

Storage/Backup Issues

Storage/Backup and Long-Term Preservation Breakout Group Report

Synopsis:

The Storage/Backup and Long-Term Preservation Breakout Group was charged to explore a series of related questions that concerned storage, the brick and mortar of any long term digital preservation system. As noted in the OAIS standard (CCSDS 650.0-B-1) for a digital repository and reference model for a digital information object, Storage, is one of six interconnected component (Ingest, Administration, Data Management, Access, and Preservation Planning) of the reference architecture. No component stands as an isolated archipelago. That swiftly moving streams of conversions in the breakout group meandered from component to component demonstrated the strategic importance to approach this subject as an interconnected web. Breakout group conversation and discussion focused upon a spectrum of topics, which included (1) best practices for storage infrastructure, (2) metadata standards to represent the logical context and understanding of digital files in human form, (3) business models to sustain long term preservation activities, (4) data models to store repository data, (5) planning models to identify, execute and validate preservation treatments and (6) domain specific challenges to establish a trustworthy storage/repository infrastructure for the Anthropological community. Across the sub-fields of Anthropology, the components of a storage infrastructure (hardware, media types, configurations and software to manage storage) needed for backup and preservation functions was contextualized by drivers and requirements of the other components of a long-term preservation system. As an illustrated example it was not possible to have dialogue that concerned storage media options for preservation with out also understanding the storage requirements, instrumentation, and tagging standards need by the archeologist to capture, describe and ingest field data.

Summary of Breakout Group Discussion by Topics:

Topic 1: What are the best practices with regard to storage and backup?

Best practices emerge over time as a result of a deeper understanding of a problem and outcomes from pilot projects or test beds established for experimentation. While the Anthropological community is just beginning to explore storage solutions for LTP (long-term preservation) the Digital Library community has for nearly a decade explored the principal issues and challenges that surround storage and backup of digital data. The principal problems that need to be addressed are well known and include (1) technological obsolescence; (2) media decay (3) replication, and (4) evolving standards to manage large storage pools or networked storage grids. The worst-case scenario for storage and backup identified by the group was locally managed storage. This modality is associated with a high probability of data loss over time. In this mode best practices followed by traditional data centers to protect data and secure unauthorized access to data is nearly impossible to maintain. The group ruefully noted that a significant number of students and researchers still managed their own storage. Hence the challenge here is to educate the community on the need to abandon this practice and adopt alternative solutions such as participation in grid storage networks. At the opposite end of the spectrum and across the Atlantic the European community has successfully demonstrated the efficacy of grid storage for LTP of digital data. The infrastructure for grid storage has trusted governance, which establishes best practices to deal with data management problems, associated with the aforementioned problems inherent in storage hardware and software used to manage storage. One member of the group characterized grid storage as "being alive", continuously being refreshed and secure since access and replication where an integral part of the management functionality of the grid. In addition, participation in the grid also relieves the student or researcher with the responsibility to plan and manage his or her own media migrations. While storage grids do exist in the United States (see the NSF program on Grid storage at http://www.teragrid.org/about/ http://www.teragrid.org/about/] the group also discussed Commercial Cloud Storage as another option for LTP. This solution is just beginning to gain traction in the US Academic community since it is a potential cost saver. A powerful motivator while the country wrangles through a deep recession. Cloud Storage provides the opportunity to outsource the storage function to large commercial vendors like Amazon and Google that run their own storage grids. For this storage option trust is a significant issue. Commercial vendors are subject to the natural business cycle and no firm is completely immune to failure or takeover. How to access or recover data when a business fails is of serious concern to the academic community. In addition secure access to data was another problem identified with commercial cloud storage. In response to these concerns the Mellon Foundation recently sponsored a planning grant to understand how the academic community could take advantage of cloud storage without being at the mercy of the business cycle and to technically explore how commercial cloud storage could be overlaid with a service interface that would protect data from unauthorized access and automatically replicate data when a firm went out of business. Details about this initiative are available from the http://DuraSpace.org DuraSpace website. The breakout group also discussed storage media and configuration options for LTP. Optical disk, magnetic disk and tape have all been successfully used for data storage and backup. In most instances these media are combined to form a hierarchical storage system. Typically these systems deploy magnetic disk for fast online access to data and tape or optical disk to store off-line data that is infrequently accessed. The goal is to build a configuration that satisfies LTP requirements at a price performance that is affordable and sustainable. Finally the group unanimously recognized that storage and backup did not equate to long-term preservation of digital data. In the absence of a logical layer, such as PREMIS to overlay storage, over

time digital data would become more difficult to: discovered, search, accessed or understood as hardware software and community standards evolved and made older storage and access system obsolete.

Topic 2: Does the PREMIS standard provide sufficient metadata to support the long-term context and access to anthropological data.

PREMIS (PREservation Metadata: Implementation Strategies) is the de-facto standard for the digital library community that specifies metadata entities recommended to ensure the long-term preservation (discovery, access, rendering and understandability) of digital data encapsulated in a vast array of file formats. An in-depth understanding of the PREMIS standard was not present in the group. This made it difficult to realistically evaluate PREMIS as a standard, which could be successfully applied to preserve anthropological data. However, in the absence of any other recognized standard the group maintained that leveraging and extending this standard for the Anthropology community was strategically the right course of action. The breakout group leader did have expertise in this area and with very broad strokes introduced the PREMIS entities (Intellectual, Objects, Rights, Agents and Events) to the group. There was a focus upon the Object Entity, which specifies metadata about the hardware and software environment needed to create and preserve a digital object. The Object entity also identifies software needed to access and render a digital object. Most importantly the Object Entity identifies the encoding standards for an object's file format and characterizes a digital object as a simple file or a complex. A PDF file with an embedded image that could not be rendered independently of the PDF file serves as a good example of a complex object. On another note a policy question that needs to be resolved by some standards committee is how much of what elements, of this very elaborate standard, are need by the Anthropological community to meet their preservation purposes. It is not practical or affordable to capture data for all of the sub-elements in the PREMIS standard.

TOPICS 3-5: Repository Functions (Ingest, Access, Preserve) and Associated Data Models

Repository software used to ingest, save or preserve and access digital content used in the cultural heritage community is mostly open source. Repository software offerings that have gained significant traction in the digital library domain are (1) Fedora (2) DSpace (3) Greenstone (4) E-prints (5) Plone and (6) ContentDM from OCLC. It is important to note that the Fedora and DSpace communities have recently combined to form a consolidated community called DuraSpace. All of these application have out of the box client interfaces to there underlying data stores to simply the ingest, storage and search/access to data. In addition these repository systems have Application Programming Interfaces (APIs) that can be used to build customized web applications or web services for any of the aforementioned functions. Protocols such as OAI-PMH, OAI-ORE and SWORD, to name a few, have also been developed by the digital library community to make these systems interoperate so that data can be exchanged between systems. The group recognized that these power tools in the right hands could create highly customized systems tailored to meet the special requirements of the Anthropological community. However the group also recognized and discussed that there was a steep learning curve to understand these technologies and the cost to hire developers was also very expensive. The group maintained that one way to overcome these challenges was to appeal to granting agencies to provide additional support to build specialized systems based upon open source technologies that could be leveraged by other anthropological research projects. Although repositories have mostly the same functionality there are important differences in how the aforementioned systems represent stored data that is technically referred to as a data model. Just as the ability to search and

discover is tightly bound to the representation of data the ability to preserve data is tightly coupled to a data model that facilitates preservation planning and preservation treatments.

Upon introduction from the group leader there was a discussion of the http://www.planets-project.eu PLANETS project, which has published a preservation data model and created a tool http://www.ifs.tuwien.ac.at/dp/plato/intro.html PLATO for preservation planning. Important characteristic of the data model were discussed which included (1) the ability of the model to provide two distinct views of stored data; one from the end-user perspective that facilitates search and discovery of preserved data the other from a preservation perspective which enables preservation treatments (media or format migrations) at the file set level that do not impact the end-user view or understanding of the data. Risk of data loss is inherent in any preservation treatment and the planning tool PLATO was designed to attenuate the risk. "The planning tool Plato is a decision support tool that implements a solid preservation planning process and integrates services for content characterization, preservation action and automatic object comparison in a service-oriented architecture to provide maximum support for preservation planning endeavors." Again in the absence of other available standards the group maintained that is was strategic for the Anthropological community to leverage this standard for their community purposes.

Topic 6-7: The Trusted Digital Repository.

The scale and available resources of the Anthropological community will encourage researchers to participate in community-sponsored preservation repositories. The digital library and archival communities have over the past ten years done significant research in this area. For many organizations in these two communities the OAIS model from the Consultative Committee on Space Data Systems has become the de-facto standard. While this standards define the functional components of a preservation system it is agnostic as to how its modules are to be implemented. Nor does the OAIS standard directly address the issue of what constitutes a trusted digital repository. Without a means to verify a preservation repository's capability to keep data alive over long periods of time as technology evolves researches will be chary to support and make deposit to preservation systems. This is a no win situation for the researcher or the community since it encourages preservation activities at the individual level. TRAC or Trustworthy Repositories

Audit & Certification: Criteria and Checklist is a The goal of the RLG-NARA Task Force on Digital Repository Certification has been to "develop criteria to identify digital repositories capable of reliably storing, migrating, and providing access to digital collections. The challenge has been to produce certification criteria and delineate a process for certification applicable to a range of digital repositories and archives, from academic institutional preservation repositories to large data archives and from national libraries to third-party digital archiving services." To the Anthropological community this standards may not be appropriately scaled and alternative solutions by the community to assess trustworthiness of a repository are being pursued.

Recommended Next Steps:

The breakout group recommends that Anthropological community should take the following next steps to advance their understanding of the long term preservation of digital data in their domain in their domain:

- Create a task force to propose an entity to recommend a long term plan and business model for funding and sustaining LTP specific to Anthropology
- Create a standards body that will review proposed standards for LTP of anthropological data across the sub-domains
- Anthropology should encourage leveraging the technical infrastructure of both commercial organizations and sister disciplines to promote LTP.
- Anthropology should take the opportunity to extend open standards and open source software to promote LTP.
- Anthropology curriculum should be expanded to include best practices and standards for digitization and LTP of digital data.

Appendix A

Break Out Group Membership

David Gewirtz

Georgetown University Library
Head Information Technology

Laura Welcher

(Director of Development and The Rosetta Project)

Dean Snow:

Penn State University
Professor of Archaeological Anthropology

Michael Fischer:

Professor of Anthropological Sciences in the Department of Anthropology at the University of Kent and is currently Director of the Centre for Social Anthropology and Computing, the University of Kent at Canterbury.

David R. Hunt:

Smithsonian Institution
Museum Specialist/Physical Anthropology Collections Management

Mark Mahoney:

Wenner-Gren Foundation for Anthropological Research, Inc.
Resource Coordinator at the Wenner-Gren Foundation:

Toward an Integrated Plan for Digital Preservation and Access to Primary Anthropological Data (AnthroDataDPA: A Four-Field Workshop). Group participants and their affiliations are given in appendix A.

1. http://www.oclc.org/research/projects/pmwg Link to the PREMIS website.
2. http://www.ifs.tuwien.ac.at/dp/plato/intro.html From Welcome to Plato, the Planets Preservation Planning Tool.
3. From the TRAC forward

Draft Storage/Backup and Long-Term Preservation Breakout Group Narrative 1

General Background

General Developments Outside Anthropology

There have been great strides made with regard to creating digital object repositories and moving toward interoperability between repositories. It is prudent to build on rather than reinvent these developments. The best way to do this is to work with experts who are familiar with the accomplishments from these fields.

In its broadest sense, the Open Archives Initiative (OAI) is an organization that seeks to promote interoperability so as to "facilitate the efficient dissemination of content" (http://www.openarchives.org/). While funded largely by funds in the US, it reaches out to a broad-based community of concerned individuals and institutions (http://www.oaforum.org/tutorial/english/page1.htm). OAI has suggested a mechanism—Open Archives Initiative-Protocol for Metadata Harvesting (OAI-PMH)—which allows harvesting data from many repositories. (Note that OLAC is specifically intended to be a linguistics-specific extension of OAI.) While technically the metadata may be in any agreed-upon format, the OAI-PMH protocol suggests Dublin Core metadata standards. The Dublin Core Metadata Initiative is a separate organization that seeks to promote a central set of metadata terms to find and share information. Dublin Core metadata has gained wide acceptance.

Architecture systems for digital repositories with long-term preservation goals have also been developed. For example, the Reference Model for an Open Archival Information System (OAIS) has been accepted as an ISO (International Organization for Standardization) standard (http://nost.gsfc.nasa.gov/isoas/). OAIS models all the functions of a digital repository, from receiving and preparing items (ingest function), storing, maintaining, and retrieving items (archival storage), data management and administration, and access (http://www.oclc.org/research/publications/archive/2000/lavoie). Software systems, including open source software, can implement this architecture. For example, Fedora (Flexible Extensible Digital Object Repository Architecture) is designed to archive complex digital objects and gives organizations flexible tools for managing and delivering their digital content.

Core metadata elements central to long-term preservation have been laid out by a working group called PREMIS (PREservation Metadata Implementation Strategies) organized by the Online Computer Library Center (OCLC–a nonprofit, membership, computer library service and research organization used by more than 40,000 libraries now combined with RLG).

Preparing Data for Digital Archiving and Preservation

What Do I Do With My Data in the Meantime?

Preparing Data for Digital Archiving and Preservation

If there were a central coordinating institution in place for anthropology, the best strategy for anthropologists wanting to digitally archive data would be to follow the instructions from that central institution, both in terms of first finding a trusted repository, and then second following the general guidelines as well as the more specific guidelines from the particular repository. The trusted repository would presumably be following the best guidelines for storage infrastructure and have plans for long-term preservation (as well as have guidelines or entry forms for digitization as well as the needed metadata.

Keep in mind that a trusted repository or preservation archive is an established institution committed to long-term preservation of the digital object; a distinguishing characteristic is that such an archive will have a technology migration plan on which to found its claims of long term digital accessibility. Thus it contrasts with a 'web archive,' which is often only a website serving information from a database or file directory. Web archives rarely serve genuinely interoperable material, and they regularly disappear in response to changes in institutional servers or in the responsibilities of the archive creator.

In the absence of a central coordinating institution, which is the current case, the next best solution is to find a trusted repository (perhaps even one's university library—and, if possible, provide copies of data to other institutions. As already discussed, if at all possible, it is wisest to avoid going it alone.

If you have not decided on a repository, you should follow the guidelines discussed in this working report (Maintenance of Data Integrity and Best Practices for Storage Infrastructure). The absolutely worst solution is to store data in proprietary formats without publicly available file format specifications that may not be readable in the future. If the media are not upgraded, the data may also be lost.

Privacy and Ethics

Within the context of creating large, publicly accessed archives that may include a variety of information resulting from the research process such as primary data, personal notes, correspondence, etc., one of the most pressing ethical issues that must be addressed is ensuring the privacy of the research subjects.

A summary of the major points arising during discussion are provided below. The reader should note that the group was in a better position to raise issues and consider the advantages and disadvantages associated with different solutions than to devise specific, concrete recommendations. We think this accurately reflects a lack of consensus at present regarding many key issues of ethics involving digital anthropological data.

1) Code of Ethics. Numerous societies have a code of ethics readily available via the web. See the web addresses below. Not surprisingly, these codes tend to provide general information or guidance regarding research protocols. Common themes include the following from the AAA guide

“Anthropological researchers must do everything in their power to ensure that their research does not harm the safety, dignity (psychological well-being), or privacy of the people with whom they work, conduct research or perform other professional activities.” With regard to privacy it is suggested that “Anthropological researchers must determine in advance whether their hosts/providers of information wish to remain anonymous or receive recognition, and make every effort to comply with those wishes.”

2) Basic Privacy of the Individual. Common to all subfields was the basic privacy of the individual. If an individual wishes to remain anonymous then that wish must be honored.

However, ensuring anonymity is not a simple procedure. Common practices for deidentification of individuals include the use of a pseudonym or alphanumeric ID in place of the individual’s name. This will frequently not be sufficient for instance, in modern genetic analyses, it is relatively simple to identify individuals, families, and familial relationships from available data. Similarly, it may be a simple matter to identify an individual in linguistic or cultural work if the population of interest includes only a few individuals such as in the case of many endangered languages.

3) Protection of Location. In several instances, locations were identified as in need of protection. The most obvious of these may be in the case of archaeological site location where knowledge of the location may lead to unwanted access. Similarly, the location of sacred sites should be afforded the same privacy considerations as individuals or groups. Finally, knowledge of commercially exploitable resources within a populated area may need to be protected to prevent unwanted exploitation.

4) Length of Protection. While, at first thought, it may seem obvious that if an individual or group has requested anonymity, that request should be honored in perpetuity, this may not be the case. What may have seemed like sensitive information to a participant at one point in time may not seem so at a later point. The reasons for such changes in attitude may be numerous and variable. The dynamic nature of the concepts of “private” and “sensitive” must be kept in mind when setting up archives. In general IRB’s may suggest that the appropriate length of protection be “until no foreseeable harm can be done.” As noted this may be difficult to determine and may change as time passes. The decision to maintain or remove privacy criteria can reside both with a single individual as in the case of the donor or may be the responsibility of a group.

5) Restriction of access. One of the basic premises behind digitally archiving of data is to make these resources easily available to a group beyond that of the donor/collector. In many cases, the intention is to make the resources freely available to the general public without restriction. As noted, however, in many cases restrictions will have to be placed on access to protect privacy or to accede to a donor’s wishes. In short, it is clear that in some circumstances, differential access is sometimes a necessity. The most common differential access currently practiced is with regard to the dichotomy of scholars vs. nonscholars recognizing that the distinction between the two groups is sometimes difficult to identify (though some fields, like linguistics, also give special privileges to communities from which data are drawn. Restricted access is frequently placed on museum collections primarily for the safety of delicate collections but a number of online databases also require registration and validation of the user prior to access.In unusual situations, access may be restricted to a small number of individuals or even a single individual. Situations in cultural anthropology and linguistics were described where an informant had specifically stated that a given individual (e.g., a brother-in-law) could not be privy to the story being told, but that the rest of the world could be given full access. Such situations become very difficult to monitor.

6) Sensitive Data. In situations where access to restricted data is granted to specific individuals it is likely that certain covenants are placed upon the use. This may be as simple as the request that the user properly cite the donor/archive. It is also possible that stronger restrictions are placed on the use such as limiting the use of data for future studies or limiting the sharing of data between individuals. In terms of digital archives where sharing is easy, it is possible, and probably likely, that such covenants do not always transfer with the data allowing for violation of promised privacy by secondary or tertiary users. In these instances the question arises first as to how restrictions can be maintained and second as to how such violations can be addressed and violators punished. It was agreed that any action would be difficult and costly. For these reasons, it is likely that researchers may have to establish separate IRB protocol for archiving and web access for data in cases where it is decided to digitize and preserve the data after the initial research. It was also noted that some IRBs may mandate that sensitive data are destroyed following the project's completion. If that is the case, the original protocol must be amended and an additional protocol submitted.

7) Privacy of the User. There was a significant discussion with regard to the rights of the user.

The discussion ranged from allowing the user total free access with an assurance of complete anonymity to a system requiring registration of users along with a login prior to access of the archive. Ease of access increases the potential use and benefit of the archive while posing no threat to the user. Restricted access (ranging from simple login to full-scale registration and authorization of users) may decrease the benefit to the user (if they decide to only focus on those archives with full, easy access) while increasing the benefit to the originator of the data. Benefit to the originator includes the ability to document usage which may be used in some way as a measure of the importance of the archive (potentially beneficial in situations such as promotion) or to provide assurance that archival material is used, and cited properly.

8) Spheres of Responsibility. Ultimately it was clear to the group that there are different spheres of responsibility for individuals at each level. The field researcher collecting data must first and foremost be responsible to the participants agreeing to maintain and protect the privacy desired by the individual. This is monitored by the IRB of the researcher's institution. The archivist is likely to be removed from the original participants and is primarily responsive to the researcher. Just as the researcher has rules and restrictions placed upon them by the participants, the archivist can expect to have such covenants placed on them by the researcher. This again may be monitored by an IRB with ongoing protocols. Finally, the user, anonymous or not, has an obligation to treat the data with the respect that researcher and archivist have established. This will ensure the ethical treatment of all subjects and subject information.

LIST OF WEB RESOURCES REGARDING ETHICAL CONDUCT FOR THE PRIMARY ANTHROPOLOGICAL ASSOCIATIONS WITHIN THE USA.

ARCHAEOLOGY

Register of Professional Archaeologists
Society for American Archaeology

PHYSICAL ANTHROPOLOGY

American Association of Physical Anthropologists

LINGUISTICS

Linguistic Society of America

CULTURAL ANTHROPOLOGY

American Anthropological Association\

Reasons for reluctance

Why have anthropologists been reluctant to give their data to physical archives?

We do not have research that bears on this question. However, we felt that answers to this question would have important implications for understanding the need for AnthroDataDPA.

Some reasons that were suggested:

1. Some anthropologists think they will give up their ability to work on their data if they deposit it in an archive, however, they are not ready to stop working. Having a digital copy or access to it online will help enormously. It should increase physical preservation.

2. Some anthropologists do not think that archives provide enough access for their work—they would prefer digital access of some kind. While theoretically almost any scholar can go to an archive—it is expensive and time-consuming to go to an archive to do research.

3. Some scholars think they have to be famous for an institution to consider their collection. This is apparently not the case for the National Anthropological Archives. The director, Robert Leopold says that any collection of an anthropologist will be taken. However, perceptions have reality—if scholars believe this myth, they may not ask an archive to take their material.

Some simply do not want to face their mortality and do not think about the matter until it is too late.

Trusted Repository

The digital library and archival communities have over the past ten years done significant research in this area. For many organizations in these two communities the OAIS model from the Consultative Committee on Space Data Systems has become the de-facto standard. While these standards define the functional components of a preservation system it is agnostic as to how its modules are to be implemented. Nor does the OAIS standard directly address the issue of what constitutes a trusted digital repository. Without a means to verify a preservation repository's capability to keep data alive over long periods of time as technology evolves researches will be chary to support and make deposit to preservation systems. This is a no win situation for the researcher or the community because it encourages preservation activities at the individual level.

A goal of the RLG-NARA Task Force on Digital Repository Certification has been to "develop criteria to identify digital repositories capable of reliably storing, migrating, and providing access to digital collections. The challenge has been to produce certification criteria and delineate a process for certification applicable to a range of digital repositories and archives, from academic institutional preservation repositories to large data archives and from national libraries to third-party digital archiving services.1" To the anthropological community these standards may not be appropriately

scaled and alternative solutions by the community to assess trustworthiness of a repository should be pursued.

Hopefully, the scale and available resources of the anthropological community will encourage researchers to participate in community-sponsored preservation repositories.

1. OCLC and CRL. 2007. "Forward." In Trustworthy Repositories Audit & Certification: Criteria and Checklist. Version 1.0. <http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf> [↩]

Web Resources

E-MELD School of Best Practices in Digital Language Documentation: http://emeld.org/school/

Garrett, John, and Donald Waters. 1996. "Preserving Digital Information: Report of the Task Force on Archiving of Digital Information commissioned by the Commission on Preservation and Access and the Research Libraries Group.Washington, DC: Commission on Preservation and Access." http://www.clir.org/pubs/abstract/pub63.html

Howard, Roger. Wed Apr 09 2003. http://eclipse.wustl.edu/~listmgr/imagelib/Apr2003/0011.html

NARA (2004): http://www.archives.gov/preservation/technical/guidelines.html

New Jersey Digital Highway Project(2007?): http://www.njdigitalhighway.org/digitizing_collections_libr.php

NINCH (2002): http://www.ninch.org/guide.pdf

Simons, Gary F. 2006. Ensuring that digital data last: The priority of archival form over working form and presentation form. An expanded version of a paper originally presented at the: EMELD Symposium on "Endangered Data vs. Enduring Practice," Linguistic Society of America annual meeting, 8-11 January 2004, Boston, MA. http://www.sil.org/silewp/2006/003/SILEWP2006-003.htm

Additional Information on Audio

ARSC Technical Committee. 2009. Preservation of Archival Sound Recordings, Version 1, April 2009. http://www.arsc-audio.org/pdf/ARSCTC_preservation.pdf

Sound Directions (2009): http://www.dlib.indiana.edu/projects/sounddirections/papersPresent/index.shtml

CDP Digital Audio Working Group (2006): http://www.bcr.org/dps/cdp/best/digital-audio-bp.pdf

U. of Maryland Libraries (2007): http://www.lib.umd.edu/dcr/publications/best_practice.pdf (2007)

Additional Information on Images and Video

Visual Arts Data Service (2000?): http://vads.ahds.ac.uk/guides/creating_guide/sect31.html

California Digital Libraries (2008): http://www.cdlib.org/inside/diglib/guidelines/bpgimages/

Washington State Library: http://digitalwa.statelib.wa.gov/newsite/best.htm

BCR Digital Images Working Group (2008): http://www.bcr.org/cdp/best/digital-imaging-bp.pdf