WORKING SUMMARIES AND BACKGROUND INFORMATION
Digital Preservation Best Practices

January 26, 2004
HVC2 Digital Preservation Task Force, Best Practices Working Group

While digital technologies are enabling information to be created, manipulated, disseminated, located and stored with increasing ease, preserving access to this information poses a significant challenge. Unless preservation strategies are actively employed, this information will rapidly become inaccessible. Choice of strategy will depend upon the nature of the material and what aspects are to be retained. (*Preserving Access to Digital Information* [PADI], National Library of Australia, http://www.nla.gov.au/padi/topics/18.html)

Architectural Framework

- **Identification / Authentication / Authorization**
  - There are no specific "best practices" regarding I/A/A for digital preservation per se. The primary requirements are that University digital repositories should fit within the overall I/A/A technical and policy framework of the institution.
  - I/A/A considerations can be broken into several major categories. Policy decisions on these categories will have technical implementation considerations.
    - Authorization of users to discover relevant digital objects (i.e. search)
    - Authorization of users to access relevant digital objects (i.e. access / retrieval)
    - Authorization of members of the community to deposit relevant digital objects
    - Authorization of selected users external to the community to deposit relevant digital objects
    - Authorization of members and / or selected external users to modify (including delete) deposited digital objects
  - General I/A/A requirements include:
    - Registration of the user into an official institution identification system (directory); requires definition of community members, authorized sponsors for non-community user access
    - Ability of the identified user to authenticate to the system (account / password control)
    - Assignment of rights to the authenticated user for specific repositories (if applicable) and specific actions (authorization)
    - The ability to seamlessly pass credentials between systems without direct user intervention (single sign-on)
    - Processes to keep registrations / authorizations current
  - KU I/A/A services are partially in place and under continued development. Relevant elements include:
    - KU Online ID

        - Argus
        - LDAP
        - PKI (?)
        - Individual system authentication / authorization stores
        - Shibboleth (under investigation)

- **Access, Navigation, Interface Issues**
    - Refer to IR Policy Working Group documents
        - Access should be as open as possible, i.e. the default option
        - Access should be via the Web whenever possible
        - Confidential data should be noted and controlled as appropriate
    - Organization of archived / preserved data
        - should be easy for users;
        - should be searchable across repositories;
        - should be available / searchable from various portal views (i.e. student, faculty/staff, library, external)
        - should not be dependent on being able to identify / follow internal organizational hierarchies;

- **Integration**
    - Access
    - Technical components
    - Technical management

- **Repository / Storage** (http://www.nla.gov.au/padi/topics/10.html)
  Media instability and changes in technology (see **Technical obsolescence**) are the two main threats to the continued accessibility of digital information stored on physical formats. Of these, changes in technology present the greater risk because they may render these materials inaccessible within a much shorter time span than that in which the medium will become unstable.

  No single method has been identified to preserve access to physical format digital material. It will probably require a combination of different methods, such as refreshing, format migration, migration and emulation. Refreshing and format migration may be used to address the problems of unstable media; migration and emulation are longer term strategies for preserving access. Choice of strategies and methods for holding digital information will determine its future accessibility. Options for storing digital information include choices of media on which the information is to be stored, and decisions regarding the formats in which it will be stored.
    - Physical media
        - Physical format digital material is information in digital form which is stored on transportable media, e.g. magnetic disks such as floppy disks, magnetic tape, and optical disks such as CDs and DVDs.
        - The physical medium on which the information is stored should, ideally, be stable: deterioration can be slowed down by storage under stable environmental conditions.
        - While choice of standard storage media can ameliorate the effects of technological obsolescence, inevitably, old storage media will

be superseded and migration, or some other preservation strategy will be necessary to preserve access to the material.

- Formats (see **Content: Objects / Formats)**
- Management
  - Whether digital information is stored online, near-line, or off-line will usually depend upon expected retrieval requirements. Little used off-line material may be stored on magnetic tape, while the enhanced accessibility of disk storage may be chosen for high demand online material.
  - Strategies such as storing identical material in multiple locations and regularly backing up will provide some protection against loss due to media failure or human error.
  - Digital information stored in compressed forms requires less space and thus reduces storage costs. However, when an image is compressed and then uncompressed, the decompressed image is usually not quite the same as the original scanned image: this is called 'lossy compression'. Distortions can be particularly severe for high compression ratios; the degree of loss can usually be determined by adjusting the compression parameters. The use of lossy compression for storing digital material over the long-term is generally not recommended due to the potential for irretrievable information loss on decompression or on migration from one lossy compression to another

- **Persistent Resource Identifiers** (http://www.nla.gov.au/padi/topics/36.html)
  The most common method of discovering and locating resources on the Web relies on allocating an identifier to resources. Uniform Resource Locators, or URLs, specify the location of a resource by including a protocol, domain name and the actual name of the file within which the resource resides. Although URLs have been serving the combined purpose of identifying a resource and describing its location for some time now, they are not a satisfactory means of persistently identifying a digital resource. The URL simply points to the current location of the resource. If a resource is moved to a new location, the previous URL is no longer useful and links to the resource that are embedded in other documents or databases also become redundant. A persistent and unique identifier would preserve access to that resource regardless of its location, provided the persistent identifier were maintained with the correct current associated location when the resource was moved. This form of mapping is generally undertaken via a resolver database. Types of persistent identifiers include:
  - Uniform Resource Name (URN)
    A Uniform Resource Name (URN) is a standard, persistent and unique identifier for digital resources on the Internet. To link to the URL of a digital resource from the URN, a resolver service is required. All URNs will include a Namespace Identifier (NID) code and a Namespace Specific String (NSS). The NID indicates the identification system being used for the URN and facilitates the interpretation of the NSS. The NSS is the local code that identifies the individual document.
    - *IETF:RFC 1737, Functional requirements for Uniform Resource Names* (http://www.ietf.org/rfc/rfc1737.txt)
    - *IETF: RFC 2141, URN Syntax* (http://www.ietf.org/rfc/rfc2141.txt)

- The Handle System (http://www.handle.net/overviews/overview.html)
  The Handle System was developed by the Corporation for National Research Initiatives (CNRI) and generally conforms with the URN framework. The Handle System is 'a comprehensive system for assigning, managing, and resolving persistent identifiers, known as "handles", for digital objects and other resources on the Internet'. Handles are resolved through a global handle service, with the capability of also having local handle services integrated with the global system.
- Digital Object Identifier (http://www.doi.org/handbook_2000/index.html)
  The development of the Digital Object Identifier (DOI) system was initiated by the Association of American Publishers, and is now managed by the International DOI Foundation (http://www.doi.org/). DOIs have been developed to assist the publishing community with electronic commerce and copyright management of digital objects published on the Internet. The DOI system of unique identifiers is based on the Handle System and allows the allocation of a unique digital identifier to commercial digital publications. A detailed description of the scheme is set out in the DOI Handbook.
- Persistent URL (http://purl.oclc.org)
  The Persistent Uniform Resource Locator (PURL) was developed and implemented by OCLC as a naming and resolution service for general Internet resources. It is intended as an interim system to be used until the URN framework is well established. A PURL looks just like a URL, except it points to a resolution service instead of the actual location of the digital resource. The resolution service then redirects the user to the appropriate URL. In essence, the PURL is a link to the current location of the particular resource to which it is assigned. If the location of the resource changes, its new location only needs to be updated in the PURL resolver service for users to be able to continue to find it with the same PURL.

- **Content: Objects / Formats**
  The use of standards is one strategy which may be used to assist in preserving the integrity of and access to digital information. Adherence to standards can assist by facilitating the transfer of information between hardware and software platforms as technologies evolve. Use of standards can also help ensure best practice in the management of digital information. Resources which are encoded using open standards have a greater chance of remaining accessible after an extended period than resources encoded with proprietary standards (unless these standards are very widely used).

  While there are similarities in the high-level approachs to preserving various object formats, the details of best practice can vary widely. Selected best practices include:

  - **Images:**
    Recommended Standards / Best Practices for Digital Projects, KU Digital Library Initiatives, January 2003,
    http://kudiglib.ku.edu/docs/DLI%20Standards.doc

- **Databases:**
  The Long-term Preservation of Databases, ERPANET Workshop Report, Bern, Switzerland, April 9 - 11, 2003, http://www.erpanet.org/www/products/bern/bern.htm

  Databases have been, and continue to be, a key technology for the storage, organisation and interrogation of information. They are a core module in most of today's information systems. While the value of other types of digital information has been highlighted in recent years, databases and the information they contain have often been neglected. Their preservation is of high concern as they are often either irreplaceable or of such value that replacement would be prohibitively expensive.

  Databases not only retain the information in a highly structured manner, but they are in most of the cases constantly updated and must allow flexible interrogation of the valuable data. For their long-term preservation, this poses a number of unique problems.

  Different communities approach the archiving of databases in very different ways. This ranges from computer science, where archiving usually includes simple backup operations and moving parts of the database into more remote memory, but without a long-term approach, to scientific research databases, where preservation is a necessity, but the focus lies on access, normalisation and value-adding, and to archival science, where long-established techniques and methods as well as the need to maintain authenticity of data meet the technological challenge of databases in order to preserve them for an indefinite period of time. During the workshop presentations and discussions a set of best practices emerged.

  Conversion and Migration
  No project relies on the original bit stream and data format of the databases to be preserved, but all convert to open and stable preservation formats. It is acknowledged that further migration might become necessary as time passes.
  Some, but not all, projects store the file they got transferred from the creating agency together with the preservation file in a normalised structure. Security requirements impose the storage of at least one archival copy at a different location.

  Adherence to standard formats
  The conversion processes conducted by all of the projects discussed aim at reaching a limited range of very basic, open standard formats. These include:
  - *Flat files for plain text or tables.* Different kinds of flat files were discussed including fixed-field, fixed-length records, delimited field variable length records, tagged textual documents, and comma separated values. There is a tendency towards using ASCII or EBCDIC encoding, but it has also been pointed out (mainly by the

Swiss Federal Archives' SIARD project) that Unicode (UTF-16) should be preferred to allow for multilingualism and special characters.

- *XML for database contents and metadata.* The role and value of XML have frequently been highlighted (for example, the ERPANET Workshop in Urbino, October 2002[12]), especially its high flexibility and understandability and the perspective of a clear exit strategy should the format be superseded in the future. The possibility for XML to be tailored to individual needs is demonstrated by the Norwegian ADDMML, an XML Data Definition Language (DDL) used for metadata capture, and by the Geographic Markup Language (GML) that is being developed to cope with Geographic Information Systems (GIS).
- *Standard Query Language (SQL) for the database structure.* Although every database system vendor adds specific extensions to the SQL definition, this is at the core an international and open standard, defined by the ISO and widely used. The Swiss Federal Archives' SIARD project therefore decided to use standard SQL (SQL-3) to represent the structure of a database. This allows for the reconstruction of the database for future access.

Software independence
As alluded to above, it is vital to move away from software dependencies. Migration as the basic strategy and the use of standard formats facilitate software independence.

Role of Metadata
The conviction that metadata are absolutely key for digital preservation is firmly
established. All projects address this issue as a high priority. Some of the main areas of agreement included:

- Allowing for semi-automated metadata capture through tools. Given the large amounts of data and the number of databases to be preserved, as much automation as possible is a must.
- Allowing for additional manual metadata capture, especially for high-level metadata. High-level metadata are of great importance, and usually there are limited possibilities for automatic capture. Here additional manual efforts are indispensable.
- Including the data producers into the capture of metadata. Particular attention should be paid to the non-written knowledge through communication with database creators and users.
- Storing metadata in a standard format. XML has been considered very promising for fulfilling this task.

Access
The ways in which the different projects provide access or intend to provide it have been widely discussed. According to different backgrounds and environments, access issues have been dealt with in various ways. In fact, it often seems less than clear who is going to access the preserved databases, and what skills s/he will or should have. Agreement was reached, however, on the basic principle that preservation without future access is incomplete and of little use. Even if it is impossible to imagine or view access issues from

today's point of view, database preservation must keep all possibilities open so as not to bar future access.

Areas for Further Research

During the workshop it became clear that certain issues provoke intense discussions,
while for others there is limited knowledge available. As a result, it is suggested that
research efforts be directed in the following areas.

Appraisal of databases

Valuable insight was provided into some of the approaches for appraisal of databases. However, different questions remain open and must be addressed by future research. Prospective appraisal will play an important role. The tools used to evaluate the evidential and informational value of a database or a database-driven information system need further refinement. In addition, the question to what extent appraisal may be influenced by technical and financial considerations remains open. Further research is needed on how preservation techniques influence the archival value of database born data. As database preservation means extracting and transforming data from database applications which are usually highly structured and thereby limiting the way original users can work with them, investigations are needed on how much evidential value is dependent on a detailed documentation of the system functionalities and the business rules governing creation and use of the data in their original context.

Relationship between databases and records

It has become evident that there is a lack of clarification in defining what needs to be
preserved of a database or, in other words, where the records are located in an archival sense. There are several conceivable ways databases and records could be intertwined:

- records are contained, as whole objects, in the database;
- the contents of the database contain records. Each record is spread over tables;
- the contents of the database are the record;
- database data (as whole objects or spread across tables), accessed or presented in a precise manner in the application, form a record;
- the whole database system is the record; or
- a database is not a record at all.

As is easily understandable from this catalogue of possibilities, the identification of the record in a database is crucial for the formulation of preservation requirements. It may imply the preservation of whole database systems, including software facilities, and views in order to preserve the records. Therefore it is an important prerequisite for database preservation to investigate the relationship between databases and records.

Documentation

While it is clearly recognised that extensive, detailed and clear documentation is
indispensable for long-term preservation, little thought has been given to

recommendations and standards in this area. Most projects use a hybrid of paper and digital documentation. The problems this might cause and the requirements
documentation must fulfil have not been fully explored.

Contributions of archival science
On a more general level it is possible to say that the long-term preservation of databases is a collaboration of mainly archival science and information technology. During the workshop presentations it was often remarked that technical solutions have proceeded rather far during recent years, but archival science has not managed to develop a full understanding of the challenges of databases and developed appropriate methods to deal with them. Questions like those above: the appraisal of databases; the relationship between databases and records, and others call for archival discussion and attempts to resolve them.

Participants indicated that until now database preservation has been pervaded with
implicit assumptions concerning, among others, suppositions about what should be
preserved out of a database, where the records are located, and what access is useful and desirable. These assumptions often proved to impede comparisons and analyses. It is important that they be investigated, and made transparent to all stakeholders involved as well as the public.

- xxxxxxxxxxxxx

- **Technical obsolescence** (http://www.nla.gov.au/padi/topics/13.html)
  Technological obsolescence is the result of the evolution of technology: as newer technologies appear, older ones cease to be used. For example, new media for storing digital information rapidly replace older media and reading devices for these older media become no longer available. Newer versions of software constantly render older versions obsolete and the hardware required by this software also changes over time. Consequently, information which relies on obsolete technologies becomes inaccessible. Currently, it seems that the lifetime of digital storage media generally exceeds the life of the technology that supports it.

  Strategies for dealing with technological obsolescence include: migration of digital information to technologies from which they are accessible, the emulation of obsolete systems, and the preservation of obsolete technologies.

  While maintaining obsolete technologies might be the only option in limited circumstances, because of the associated need to keep every version of every piece of software and hardware, operating systems and manuals as well as personnel with relevant skills, it is not generally considered to be a feasible alternative. Emulation shows considerable promise: however, it is largely untested in the context of digital preservation. While migration is generally regarded as being the most promising digital preservation strategy currently, the difficulty in predicting the timing, nature and costs of this approach match the unpredictable nature of technological obsolescence itself.

Migrating digital information to standard formats, which are expected to be less volatile than the wide array of formats in which digital material may originally be accessible, will assist in mitigating the effects of technological obsolescence as well as in facilitating the transfer of information between hardware and software platforms as technologies evolve. Also, the creation and storage of metadata which includes information about technology dependencies will assist in managing and preserving access to this information.

- Migration (http://www.nla.gov.au/padi/topics/21.html)
  The migration of digital information refers to the "periodic transfer of digital materials from one hardware/software configuration to another, or from one generation of computer technology to a subsequent generation." This term is sometimes used to refer to the transfer of information to non-digital media.

  Migration includes refreshing, that is, copying digital information without changing it. However, while refreshing will overcome the problem of media instability, it usually isn't enough to keep abreast of technological obsolescence. Migration to new operating environments often means that the copy is not exactly the same as the original piece of information. Decisions need to be made about what aspects of the material to be migrated (eg functionality, presentation) need to be preserved.

  Providing information about successive migrations in the form of metadata will assist in determining what changes have occurred to the digital object.

  The complexity of the migration process will depend on the nature of the digital resource which may vary from simple text to an interactive multimedia object.

  At its simplest, migration may entail copying digital information to a more stable non-digital medium, such as paper or microfilm. While these media have better proven long-term reliability, frequently they provide inadequate representations of the original object and result in severe loss of the functionality and presentation of the original digital object. Transfer to a more stable digital medium (eg from floppy disk to CD-R) offers a short- to medium-term strategy for preserving access, but still requires the CD-R to be migrated when the technology changes. Another approach to preserving access to digital information entails initially migrating it to standard formats which are expected to be less volatile than the wide array of formats in which digital material may originally be accessible. However, it is important to realise that technical standards are in a state of rapid flux and this strategy cannot be solely relied upon to ensure that digital information remains accessible. The selection of a format for preserving digital information will depend upon what aspect of the resource will be required in the future. For example, decisions need to be made as to whether there will be a need to process or edit a digital resource in the future or if the visual presentation needs to be preserved.

Using software which is 'backwards compatible', that is, a later version of the software can decode files created on an earlier version, will simplify migration. Interoperability of systems will also facilitate migration by obviating the need to run a specific program in order to be able to access a digital resource which was created using it. However, these features become harder to achieve with greater software complexity and cannot be relied upon as the software manufacturer may consider that the associated costs are not worthwhile.

While it is difficult to predict the frequency at which digital information will need to be migrated, or to accurately predict the costs, the Yale Project Open Book planned for data to be migrated each five years. Costs will vary depending on the nature of the digital resource and the aspects which must be maintained. Migration of information raises intellectual property issues and there may be costs associated with this.

Migration may be undertaken in a variety of ways: in the most simple cases, migration methods have been well established. However, at its most complex, migration can be time-consuming and costly. In his paper, *Intellectual Preservation and Electronic Intellectual Property*, Peter Graham questions whether migration will be practical for large quantities of information, believing it will be possible for only a fraction of recorded information. Rothenberg points out that migration requires a unique solution for each new format and type of document that is to be converted and that costs are recurrent and unpredictable. Furthermore, it does not scale well. While further testing and investigation will be required to facilitate this approach, nevertheless, currently migration offers the most promising approach for preserving access to digital information.

- Emulation (http://www.nla.gov.au/padi/topics/19.html)
  Emulation refers to the process of mimicking, in software, a piece of hardware or software so that other processes think the original equipment/function is still available in its original form. Emulation is essentially a way of preserving the functionality of and access to digital information which might otherwise be lost due to technological obsolescence.

  Because it is impossible for any organisation to retain working examples of every computer and every piece of software, and because the cost of any attempts to do so would be prohibitive, emulation may offer a viable alternative strategy to ensure access to digital information in the future.

  One of the benefits of the emulation strategy compared with migration is that the original data need not be altered in any way. It is the emulation of the computer environment that will change with time. This should help maintain the integrity and "look and feel" of the material.

Another advantage of implementing emulation is its possible efficiency. Once the data is archived with appropriate metadata and software, no other action is required apart from media refreshing until access is desired. One emulator can also be used as a solution for several data objects requiring the same operating environment.

Both Jeff Rothenberg and Steve Gilheany promote variations of how emulation may be implemented as a solution to digital preservation. However, others dispute that emulation is the panacea, and propose that it may be only a small part of a functional preservation strategy. The need for more practical studies in this area is commonly agreed.

Avoiding Technological Quicksand : Finding a Viable Technical Foundation for Digital Preservation, Rothenberg, Jeff In: CLIR Reports (Date Created: 15 Jan 1998)
http://www.clir.org/pubs/reports/rothenberg/contents.html

Preserving Information Forever and a Call for Emulators, Gilheany, Steve (Date Created: 17 Mar 1998), Presented at the Digital Libraries Conference and Exhibition The Digital Era: Implications, Challenges and Issues, 17-20 March 1998, Singapore
http://www.archivebuilders.com/aba010.html

- Encapsulation (http://www.nla.gov.au/padi/topics/20.html)
  Encapsulation in the context of preserving digital materials is a technique of grouping together a digital object and anything else necessary to provide access to that object. This technique aims to overcome the problems of the technological obsolescence of file formats because the details of how to interpret the digital bits in the object can be part of the encapsulated information. Encapsulation can be achieved by using physical or logical structures called "containers" or "wrappers" to provide a relationship between all information components, such as the digital object and other supporting information such as a persistent identifier, metadata, software specifications for emulation.

  The encapsulation may be composed of analogue and digital components. An example of an analogue component would be human readable instructions, such as writing on the outer case of a physical format carrier, to describe how to use the carrier and interpret the outer most layer of the digital component, most likely the wrapper, which will in turn provide the information required to use the rest of the digital information contained. The analogue component of the encapsulated object may change as the carrier needs to be refreshed but ideally there should be no change required of the digital component when it is stored this way.

  The types of supporting information that should be included in an encapsulation, apart from the digital object itself, are described by the *Reference Model for an Open Archival Information System (OAIS)* (http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html) which are based

on the report "Preserving Digital Information" (http://www.rlg.org/ArchTF/) from the Task Force on Archiving of Digital Information. Briefly, they are: the representation information used to interpret the bits appropriately for access; the provenance to describe the source of the object; the context to describe how the object relates to other information outside the container; reference to provide one or more identifiers to uniquely identify the object; and fixity to provide evidence that the object has not been altered.

An alternative to storing the representation information to decode the bit-stream with every digital item archived, is to include a pointer to a single storage area for that information. This could be stored in the same repository or another dedicated centralised repository.

- **Authenticity** (http://www.nla.gov.au/padi/topics/4.html)
  The authenticity of a digital object refers to the degree of confidence a user can have that the object is the same as that expected based on a prior reference or that it is what it purports to be.

  The digital environment poses particular challenges for establishing authenticity. This is due to the ease with which digital material may be altered and copied, resulting in the possibility of a multiplicity of versions of a particular document.

  Methods used in converting, storing, transmitting or rendering digital objects may result in distortions and therefore need to be documented. The process of migrating information from one system or format to another may result in changes which also need to be recorded.

  Aspects such as a document's functionality, its dependence on particular software and its relationship to other documents are all features which need to be considered in the establishment of its authenticity.

  A range of strategies for asserting the authenticity of digital resources has been developed: choice of a particular method will depend upon the purpose for which authenticity is required. Among these are the registration of unique document identifiers and the inclusion of metadata within well-defined metadata structures. Hashing and digital time stamping are 'public' methods which authenticate the existence of a document - in the case of the latter method, at a specific time.

  Another class of methods for establishing authenticity includes encapsulation techniques and encryption strategies. A digital watermark can only be detected by appropriate software, and is primarily used for protection against unauthorised copying. Digital signatures are used to record authorship and people who have played a role in a document.

Administrative Best and Emerging Practices for Digital Preservation

Based on Brian Lavoie's article *Meeting the Challenges of digital preservation: The OAIS reference model*, the OAIS Functional Model shows "Administration" supporting all aspects of Digital Preservation:

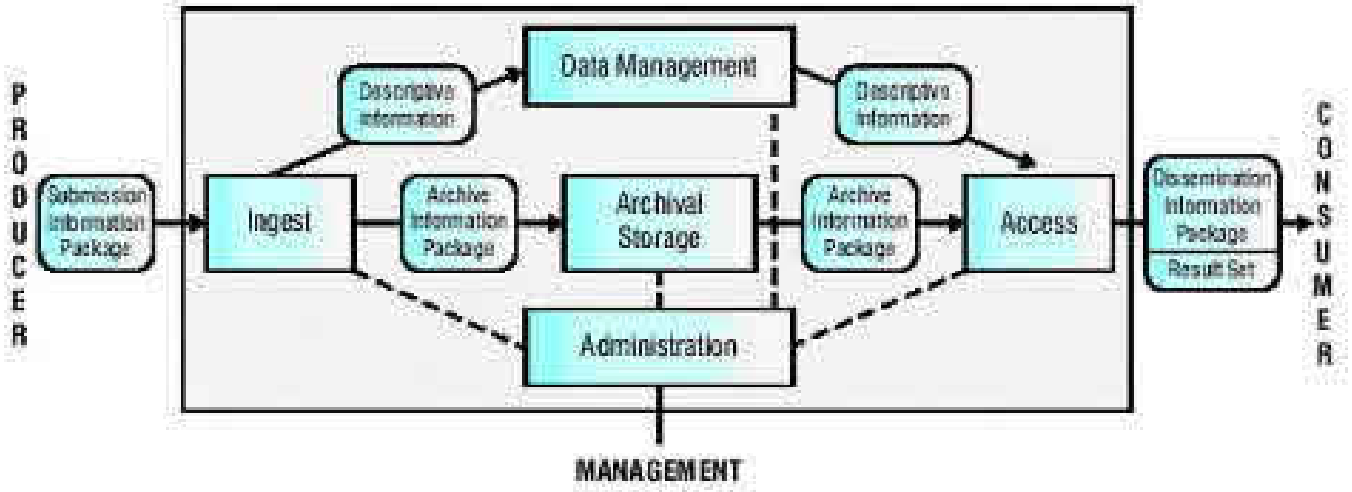

Figure 3

We see administrative best practices as pertinent to all aspects of digital management, represented in our working group's outline as two primary areas, architectural framework, and life cycle management. In terms of the OAIS functional model for management of open archive information systems, administration not only applies to the workflow between ingest and access, but to the greater lifecycle of information preservation as digital content moves toward the ingest part of the lifecycle, as it cycles through its preservation lifespan toward the process of de-selection or de-accession, and as information flows from an archive system to the consumer.

Because "Digital Preservation" is a relatively new concept within higher education and libraries, we find little that we can point to as existing models. Brian Baird, Preservation Librarian, has reviewed the digital preservation practices of a number of institutions, including Harvard, Michigan, Berkeley, and Cornell. These programs have strong well developed digitization programs that deal well with digitized files, but they do not have established programs for materials that are born digital, nor do they have strong university wide digital preservation programs. They are working on these programs and will have something in place in the next few years, but for now they are, as was described most often, working under "good faith" agreements and anticipation that they will have something in place to deal with migration issues and long term retention.

We also recognize that there are significant projects in development of which we must continue to review. Some significant sources include:

Preserving Access to Digital Information [PADI], National Library of Australia, http://www.nla.gov.au/padi/topics/18.html)

Research Libraries Group http://www.rlg.org/preserv/digpres.html

The Digital Library Federation http://www.diglib.org/preserve.htm

The publications of the Council on Library and Information Resources http://www.clir.org/pubs/

The publications of the Association of Computing Machinery http://www.acm.org

From our outline of administrative best practice areas, we can begin to ask the following questions:

Staffing:

- What are the ramifications for staffing an institutional program of digital preservation that merges the concerns of records retention with longer term archival preservation of the institution's historical record, collections, and other digital assets?
- For what aspects of preservation will central staffing be required and what will draw together those parts of a preservation program that are decentralized?

Costs:

- What costs will be centralized and born by the university, what costs by various participating units, and how will we all pay our "fair share?"

Stewardship/Creatorship/Ownership:

- Given the roles of creatorship, ownership, and stewardship what are the models that ensure stewardship while recognizing this may not be synonymous with ownership or creatorship.

Education and Training:

- How does training fit into lifecycle management?
- Who must be educated and trained? The OAIS model would suggest producers, those involved in managing digital preservation for the University, and consumers must all be educated and trained.

Figure 1

Policies and Standards:

- These are inherent in each area outlined in "architectural framework" and "life cycle management." What are the policies and standards also unique to staffing, cost management, digital content creation and preservation, and education and training?

Life Cycle Management (borrowed heavily from *Attributes of a Trusted Digital Repository,* an RLG-OCLC report, Aug. 2001)

The OAIS functional model appears to me to be the best model for the life cycle process. When used with its model for necessary metadata, it is simple and clear and includes all of the elements that we have discussed in our working group meetings.

The OAIS functions include:

- Submission or "pre-Ingest" activities
- Ingest
- Archival Storage
- Data Management
- Preservation Planning
- Archive Administration
- Access/Dissemination

Included here are elements that are also present (as they should be) in the Architectural Framework and the Administration focus areas.

The *Attributes* document includes a good diagram on page 42 (I will bring copies to the meeting).

Submission and Pre-Ingest Activities
"Before a repository can accept responsibility as a reliable archiving service, management tools must be in place, covered by a well-documented and agreed-on collections policy document." These are the most critical elements for digital materials:

- Evaluation criteria for assessing potential submissions; that is selection criteria for digital preservation.
- Collections development strategies and technical strategies for continuing access.
- Collections development procedures, including review procedures pertaining to retaining and deaccessioning materials.

Ingest
"Ingest to the repository on a practical level involves:

- Assignment and / or validation of unique identifier
- Selection and validation of the agreed-on underlying technology
- Transformation of the object....along with its associated metadata, into a bytestream than can be stored on suitable hardware in the repository.
- Establishment of necessary Representation Information.
- Verification of all Preservation Description Information."

rchival Storage
".....in an OAIS repository is the logical component that contains the necessary services for the effective storage and retrieval of Archival Information Packages (AIPs). Including:

- Moving AIPs from Ingest into permanent storage.
- Managing the storage hierarchy.
- Refreshing the storage media.
- Providing all necessary information to allow objects to be disseminated from the repository.
- A robust system for disaster recovery.

Data Management
".....covers all aspects of an OAIS repository and is essential for both long-term preservation and day-to-day administration and use.....good record keeping at every stage.....includes:

- Access controls
- Customer profiles.
- Tracking of user requests.
- Security information.... passwords, etc.
- Statistical information to improve operation.
- Accounting information.

Preservation Planning
".....monitoring.....to accommodate shifts in technology. Includes:

- Monitoring the designated community.
- Monitoring technology.
- Monitoring the significant properties of the repository's contents, as necessary.
- Developing preservation strategies and standards for continuing access.
- Developing packaging designs and migration or routine transfer plans.

Archive Administration
".....the function that contains all services for day-to-day maintenance.....management of system configurations, repository policy development and maintenance. Includes:

- Negotiating submission agreements with content producers and providers.
- Reviewing procedures.
- Maintaining systems configurations for hardware and software.
- Developing and maintaining repository policies and standards....
- Providing user support.
- Interacting with management outside of the repository.

Access/Dissemination

In the OAIS model, access to archived information is provided through the Dissemination Information Package (DIP): a copy of the digital object along with the necessary metadata, and software as necessary.

Additionally:

OAIS has been designated as an ISO standard ISO 14721:2002. Is being used by NARA for federal records. (http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html)

Metadata, Rights Management, Persistent Identifiers

- Metadata

Metadata is "data about data." Metadata is generally assigned to one of three categories:

- Descriptive – provides for resource discovery and identification
- Administrative – provides information to help manage a resource (when and how a resource was created, file type, other technical details, etc.)
- Structural – indicates how objects are put together
- Rights – administrative metadata dealing with intellectual property rights

Preservation metadata:

- Provides detailed data that is intended to support the long-term retention and accessibility of digital objects
- Provides information that ensures that digital content can be rendered and interpreted over time.
- Documents preservation processes, such as migrations, transformations, and emulations

Look for best practices and metadata standards as documented by:

- Library of Congress
- OCLC and RLG
- Digital Library Federation
- National Library of Australia
- National Library of New Zealand
- UK – Cedars Project (CURL Exemplars in Digital Archives)

Library of Congress

Table of core metadata elements for Library of Congress Digital Repository Development.

The OCLC/RLG

OCLC and RLG are collaborating on documents to establish best practices. They will propose approaches for descriptive and management metadata needed in the long-term retention of digital files. RLG and OCLC will bring key players together to review progress to date and identify common practices among those most experienced in the archiving arena. The draft working papers will then be reviewed by key stakeholders around the world.

The papers are expected to serve as a basis for further exploration of roles and responsibilities of RLG, OCLC, and others.

OCLC/RLG Core Elements Subgroup
- Provide "an implementable set of "core" preservation metadata elements, with broad applicability within the digital preservation community"
- Provide a "data dictionary to support the core preservation metadata element set (Jan.- March 2004)"Compare preservation metadata element sets at various institutions, looking for "core" elements

OCLC/RLG Core Elements Subgroup
- "Examination and evaluation of alternative strategies for the encoding, storage, and management of preservation metadata within a digital preservation system, as well as for the exchange of preservation metadata between systems"
- "Pilot programs for testing the group's recommendations and best practices in a variety of systems settings"

New Zealand – Metadata Standards Framework – "Designed to strike a balance between the principles of preservation metadata, as expressed through the OAIS Information Model, and the practicalities of implementing a working set of preservation metadata. " The schema is predicated on the idea of a Preservation Master where the metadata is held.

National Library of Australia --

NLA proposed Preservation Metadata Set is intended to be a statement of the information needed to manage preservation of digital collections. It is meant to be a **data output model**, not a data input model. It indicates the information wanted out of a metadata system, not necessarily what data should be entered, how it should be entered, by whom and at what time; nor does it concern itself with how the metadata should be associated with what it is describing. This model should be applicable to many implementations that may decide to record this information in a variety of ways. This model simply says: 'however you do it, this is what you have to deliver so you can manage preservation.'

This proposed preservation metadata framework has been informed by many models. Some are of broad relevance, (eg the *Reference Model for an Open Archival Information System (OAIS) Draft Recommendation for Space Data System Standards*(2)), while some came to us as results of data modelling exercises for particular projects (the NEDLIB project(3) and the NLA's own PANDORA project(4)). Some were more refined metadata specifications developed for particular programs or projects (the Library of Congress-CNRI Experiment Project (5); The Making of America II Project(6); the CEDARS project(7); the National Archives of Australia's Recordkeeping Metadata Standard(8)). One particular starting point for our exercise was the metadata set proposed by the Research Libraries Group (RLG) PRESERV Working Group on Preservation Uses of Metadata(9), which mainly addressed digitisation projects.

RLG invited us to adapt this set to describe a wider range of materials.
(Website)

Preservation Metadata and the OAIS Information Model: A metadata Framework to Support the Preservation of Digital Objects (A report of the OCLC/RLG Working Group on Preservation Metadata – June 2002)
http://www.oclc.org/research/pmwg/pm_framework.pdf

Preservation Metadata for Digital Objects – A review of the State of the Art (A white paper by the OCLC/RLG Working Group on Preservation Metadata (Jan. 31, 2001)
http://www.oclc.org/research/pmwg/presmeta_wp.pdf

PREMIS (PREservation Metadata: Implementation Strategies)
The OCLC Office of Research, OCLC Member Services, and RLG working together to develop best practices for implementing preservation metadata
http://www.oclc.org/research/projects/pmwg/default.htm

OCLC/RLG Core Elements Subgroup
http://www.oclc.org/research/projects/pmwg/core_elements.htm#current

OCLC/RLG Implementation Strategies Subgroup
http://www.oclc.org/research/projects/pmwg/implementation.htm

RLG-OCLC digital archive attributes working group – May Report
http://www.rlg.org/longterm/attribswg.html

Trusted Digital Repositories: Attributes and Responsibilities
http://www.rlg.org/longterm/repositories.pdf

OCLC Digital Archive Metadata Elements:
http://www.oclc.org/support/documentation/pdf/da_metadata_elements.pdf

Table of Core Metadata Elements for Library of Congress Digital Repository Development
http://www.loc.gov/standards/metable.html

Cedars Guide to Preservation Metadata
http://www.leeds.ac.uk/cedars/guideto/metadata/guidetometadata.pdf

Preservation Metadata: Pragmatic First Step at the national Library of New Zealand / Searle and Dave Thompson
http://www.dlib.org/dlib/april03/thompson/04thompson.html

Metadata Standards Framework – Preservation Metadata (Revised, June 2003) - National Library of New Zealand

http://www.natlib.govt.nz/files/4initiatives_metaschema.pdf

Preservation Metadata for Digital Collections – Standards (Create Date Oct. 1999)
http://www.nla.gov.au/preserve/pmeta.html

PADI – Preserving Access to Digital Information, National Library of Australia
http://www.nla.gov.au/padi/

Metadata - Cornell
http://www.library.cornell.edu/preservation/tutorial/metadata/table5-1.html

- Digital Rights Management (DRM)

  Digital rights management is a phrase the refers to "a range of activities, form documenting rights holders and availability status for resources through actual management of permissions to view or use a resource." DRM includes
  - Rights expression (documents rights holders)
  - Permissions available for use of a resource
  - Conditions of use
  - Rights entities and transactions are recorded in a "rights expression language" which references a rights data dictionary

  Rights expression languages:
  - XrML (Extensible Rights Markup Language, a standard maintained by ContentGuard)
  - ODRL (Open Digital Rights Language, an open source standard supported by Open Digital Rights Language Initiative.)
  - Creative Commons http://www.creativecommons.org/
  - MPEG-21 Rights Expression Language

Digital Rights Management
http://www.nla.gov.au/padi/topics/28.html
Federated Digital Rights Management: A Proposed DRM Solution for Research and Education
http://www.dlib.org/dlib/july02/martin/07martin.html
Digital Rights Management (DRM) Architectures / Renato Iannella
http://www.dlib.org/dlib/june01/iannella/06iannella.html

Open Digital Rights Language Initiative
http://odrl.net/

METS News and Announcements
Draft Rights Declaration Schema is Ready for Review
http://www.loc.gov/standards/mets/news080503.html

The MPEG-21 Rights Expression Language: a white paper
http://www.rightscom.com/files/MPEG21_RELwhite_paper.pdf

Technology of Rights Management: Digital Rights Management / Karen Coyle
http://www.kcoyle.net/drm_basics1.html

Persistent Identifiers

A persistent identifier is any identifier that enables one to discover and locate resources regardless of its location.

- URI, URN, URL, URC
  - URN - Uniform Resource Name is a standard, persistent and unique identifier for digital resources; include Name Space Identifier (NID) and Name Space Specific String (NSS)
  - URI - uniform resource identifier and locater – do not persistently identify a digital resource
  - URL – uniform resource locator
  - URC – uniform resource characteristics – metadata encoded information about resources. ARK and DOI now incorporating this metadata

- The Handle System – "comprehensive system for assigning, managing, and resolving persistent identifiers, known as "handles," for digital objects and other resources on the Internet. Handles can be used as Uniform Resource Names (URNs)."

- Digital Object Identifier (DOI) – "is a system for identifying and exchanging intellectual property in the digital environment. The DOI System provides a framework for managing intellectual content, for linking customers with content suppliers, for facilitating electronic commerce, and enabling automated copyright management for all types of media. The system is managed and directed by the International DOI Foundation"

- Persistent URL (PURL) – a resolution service developed by OCLC. The PURL resolution service associates the PURL with the actual URL and returns that URL to the client.

- Archival Resource Key (ARK) – a persistent identification system developed by John Kunze. ARK is "a system primarily designed for custodians of archived digital objects, and emphasizes the principle of stewardship of resources and their naming schemes over time. ARK has three requirements:
  - Link from an object to promise for stewardship
  - Link from an object to the metadata which describes it

- Link to the object itself (or appropriate substitute)

The ARK scheme was developed at the National Library of Medicine and is currently in production use at the California Digital Library (CDL) (Dec. 6, 2003). The National Library of Australia has also adopted it.

Persistent Identifiers – National Library of Australia
http://www.nla.gov.au/initiatives/persistence.html

Persistent identifiers
http://www.nla.gov.au/padi/topics/36.html

Towards Electronic Persistence Using ARK Identifiers / John A. Kunze
http://www.cdlib.org/inside/diglib/ark/arkcdl.pdf

ARK Persistent Identifier Scheme
http://www.cdlib.org/inside/diglib/ark/arkspec.pdf

Identifiers and Identification Systems: an Informational Look at Policies and Roles from a Library Perspective / Giuseppe Vitiello
http://www.dlib.org/dlib/january04/vitiello/01vitiello.html

Digital Object Identifier Home Page
http://www.doi.org/

CrossRef Home Page
http://www.crossref.org/

Handle System Home Page
http://www.handle.net/

PURLS Home Page
http://purl.oclc.org/

IETF - Internet Engineering Task Force Home Page
http://www.ietf.org/