The International Journal of Digital Curation
Issue 3, Volume 4 | 2009

Using the DCC Lifecycle Model to Curate a Gene Expression Database: A Case Study

Jean O'Donoghue, Jano I. van Hemert,
National e-Science Centre, School of Informatics,
University of Edinburgh

Abstract

Developmental Gene Expression Map (DGEMap) is an EU-funded Design Study, which will accelerate an integrated European approach to gene expression in early human development. As part of this design study, we have had to address the challenges and issues raised by the long-term curation of such a resource. As this project is primarily one of data creators, learning about curation, we have been looking at some of the models and tools that are already available in the digital curation field in order to inform our thinking on how we should proceed with curating DGEMap. This has led us to uncover a wide range of resources for data creators and curators alike. Here we will discuss the future curation of DGEMap as a case study. We believe our experience could be instructive to other projects looking to improve the curation and management of their data.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. ISSN: 1746-8256 The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre.

Introduction

Characterising gene expression patterns is a crucial part of understanding the molecular determinants of embryonic development and the role of genes in disease. However, gene expression studies in human embryos need to overcome a number of difficulties. These include the sourcing and maintenance of collections of human material suitable for gene expression studies, bridging the expertise in both biological and informatics areas, and amassing and making accessible data from multiple laboratories and studies.

Developmental Gene Expression Map (DGEMap)[1] is an EU-funded Design Study that will attempt to develop a large-scale human gene-expression atlas to facilitate the work of human development researchers. It is a collaborative effort between the Institute of Human Genetics at Newcastle University and the National e-Science Centre (NeSC) at the University of Edinburgh. The Institute of Human Genetics is part of the Human Development Biology Resource (HDBR)[2]. This is an MRC and Wellcome Trust-funded resource dedicated to the collection of human foetal material ranging from 4 to 12 weeks of development. DGEMap builds upon the experience of the Newcastle site in the collection and use of foetal material for early developmental research. DGEMap is intended to be both a method for facilitating collaboration with the international scientific community wishing to avail itself of the HDBR collection, as well as a public Internet database of gene expression in early human development.

One of the main concerns on the informatics side of the design study (for which NeSC at Edinburgh is responsible) is how to curate this resource over the long term. DGEMap will not be a simple archive of images, but rather a constantly changing project with several types of research output and digital assets that will require both coordination and preservation. This has led us into the field of digital curation and an examination of the standards, models and tools that have been developed to aid digital curation. We believe that our examination of these methods and solutions to digital curation issues with regard to DGEMap should be made more broadly available to other projects approaching the world of digital curation as a case study. We therefore present our findings and discussions here.

HDBR Overview

In order to set the scene for our discussion of the curation issues of DGEMap, we will first describe here the nature of the digital assets that need management and preservation. This description is the result of several interviews with those involved in the curation processes, observations of staff whilst performing tasks, and investigation of the information systems in use. The information extracted in these processes was structured according to the DCC Curation Lifecycle Model briefly discussed later in this paper. This information was then validated by those involved in the processes of data generation and information curation.

Upon the collection of embryos, the information associated with them is added to an HDBR project management database. This information includes a description of the embryo and details about when and where it was collected. As the embryonic material

[1] Developmental Gene Expression Map (DGEMap) http://www.dgemap.org/
[2] The MRC-Wellcome Trust Human Developmental Biology Resource (HDBR) http://www.hdbr.org/

is assigned to specific experiments, further details are added to the database such as slide information, project description and experiments performed. This database therefore is vital for the tracking of all the various projects and the sample collection that the HDBR and thus DGEMap holds.

The main experiments carried out on the embryonic material are in situ hybridisations or immunohistochemistry experiments. They are designed to examine the gene expression levels of a particular gene by staining a thin section through the embryo for either the RNA (in the case of in situs) or the protein (immunohistochemistry) that is produced by the gene being examined. Once these experiments have been carried out, the output in addition to the experimental details stored in the project management database are raw digital images from a microscope. These images are then manipulated initially in Photoshop to get rid of any dirt particles, and to orientate the images uniformly. Once capture and cleaning of these images are complete, they are then mapped into a 3D model by warping the image to fit over the 3D model section. Afterwards, the signal data (i.e., the experimental staining on the section) are thresholded out to high, medium, or weak levels and then superimposed onto the 3D model. This expression domain, and its associated co-ordinate information, is then used to create an entry onto a local database, providing spatially mapped information about the mapped expression data. The local database entry contains information about the person and laboratory undertaking the experiment, the mapping, as well as information about the probe or antibody used. Experimental and specimen conditions as well as any associated publications and links to other databases are also included. The final part of the database entry contains the spatially mapped data, details about the expression pattern and its distribution as well as a movie of the expression within the 3D model. This database entry is then checked for all the required information. Once permission has been granted by the project from which the entry derives (for example following publication), the entry is uploaded to an externally visible database that can be viewed and searched over the Internet.

Rationale

As can be seen from the overview above, the curation of DGEMap will not involve the use of simple off-the-shelf repository software. It comprises at least two constantly changing databases as well as the creation of a large number of images that must be transformed and mapped before being submitted to one of those databases.

Our aim therefore was to develop an approach to the curation of such a complex biological (and informatic) resource. One of the first methods we identified to accomplish this aim was the use of an existing model for curation which, while supporting the function of preservation, did not simply represent an archive-based solution. We found this model in the form of the Digital Curation Centre's (DCC) Curation Lifecycle Model. We felt this model suited our project, since although it is concerned with preservation of digital resources, it is not limited to that single objective. It also contains a model for effective data management that we felt would benefit DGEMap.

We will therefore discuss the curation of DGEMap in terms of the DCC Curation Lifecycle Model.

Curation Lifecycle Model

The Model

The DCC was set up to provide strategic leadership in digital curation and preservation for the UK research community, with particular emphasis on science data. The Curation Lifecycle Model was born out of a need to provide a training tool to help curators understand the processes involved in successful curation, and develop curation and preservation methodologies for their organisations. It offers a graphical high-level overview of the lifecycle stages required for successful curation. The DCC has adopted a lifecycle model of digital curation because:

> "Digital material, by its very nature, is susceptible to technological change from the moment of creation. The curation and preservation activities undertaken, or neglected, in different stages of their management, can influence the ability to look after them successfully at subsequent stages. A lifecycle approach ensures that all the required stages are identified and planned, and necessary actions implemented, in the correct sequence."
> (Higgins, 2008)

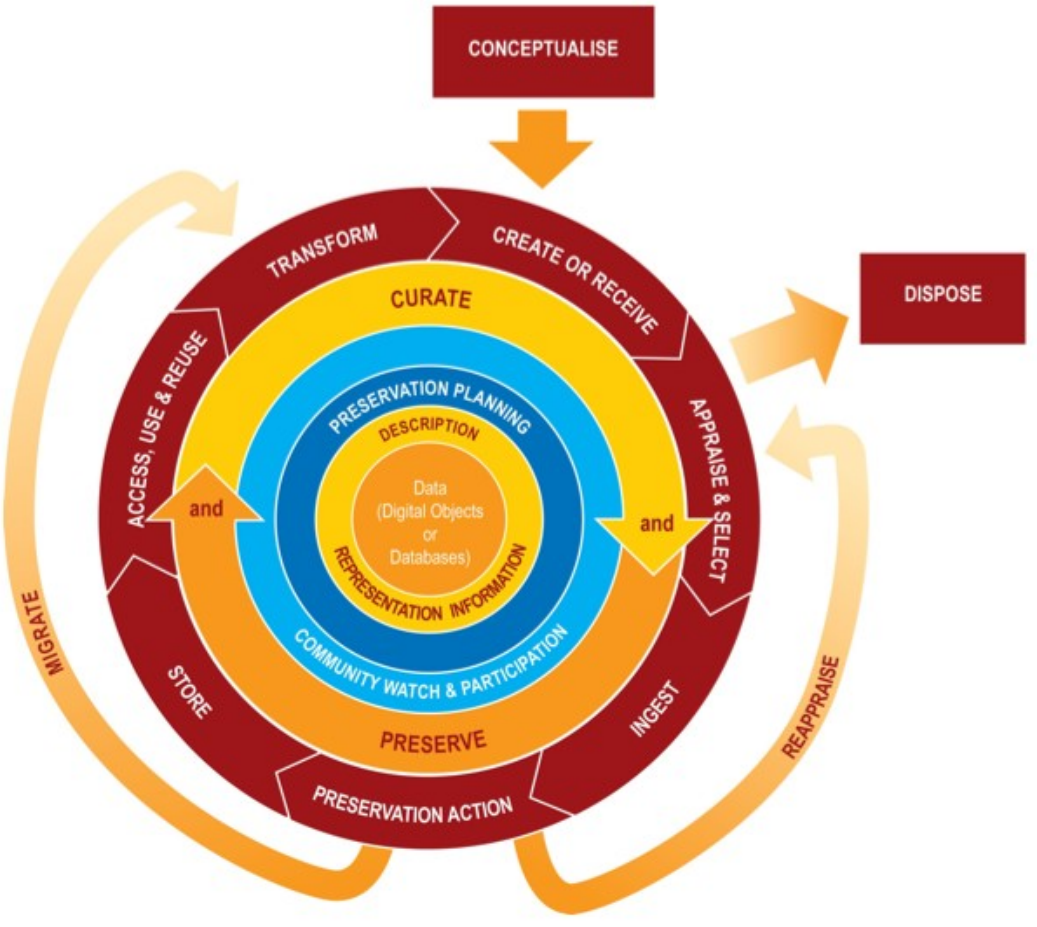

Figure 1. DCC Curation Lifecycle Model.

What can be appreciated on a first glance at the model is that there are several concentric rings in its structure. This is to highlight the fact that the subjects of all the inner rings must be considered as part of the sequence of actions represented by the

outermost ring. The inner rings include reference to the data and their description and representation information. We will discuss the "Preservation Planning" and "Community Watch and Participation" under the general term of "Preservation" here. The "Curate and Preserve" ring must be the focus at each stage in the data lifecycle as depicted in the outermost ring. The outermost ring depicts the lifecycle of the data as it moves through the curation process. These stages are Conceptualise, Create (and/or Receive), Select and Appraise, Ingest, Preservation Actions, Store, Access Use and Reuse and Transform. We will discuss here DGEMap and a design for its curation under the headings of each of these terms on the outermost ring.

Conceptualise

This portion of the DCC curation lifecycle model is designed to help clarify the steps required for effective data management and curation right from the beginning of the project. The earlier that a project considers how it will manage the data generated and how they will be curated, then the more successful a project is likely to be in anticipating and resolving any problems.

We considered the different research outputs of DGEMap, that is, the project management database, the raw images produced by the in situ and immunohistochemistry experiments, the central database of mapped images and ontologies and the final public database available on the Web. When we did this it became clear that we needed to set up a proper archive for some of our data, in particular the raw images produced by the in situ and immunohistochemistry experiments. We felt that an archive of these data would mean that even if some catastrophic event led to a loss of the annotated database of mapped gene expression, the retention of a safe archive of the original images, with the experimental data associated with those images, would mean that the entire database could be recreated. We will address the nature of this archive more fully under the subsection entitled "Store".

Another important aspect of this function is to consider what the policies and requirements of the funding bodies are for a particular project. As this design study is funded by the European Union (EU), we must also consider the policies of the EU. The European Science Foundation's Guidelines "Good Scientific Practice in Research and Scholarship" acknowledge this aspect of data management in article 37 which stipulates:

> "Institutions must pay particular attention to documenting and archiving original research and scholarship data. Several codes of good practice recommend a minimum period of 10 years, longer in the case of especially significant or sensitive data. National or regional discipline-based archives should be considered where there are practical or other problems in storing data at the institution where the research was conducted." (European Science Foundation [ESF], 2000)

In addition to the EU-funded DGEMap, we must also consider the policies of the HDBR-funders, that is, the Medical Research Council (MRC) and Wellcome Trust. This means that their policies must also be adhered to as follows:

> "From 1 January 2006, all applicants submitting funding proposals to the MRC must include a statement explaining their strategy for preserving research data for sharing and re-use." (Medical Research Council [MRC], n.d.)

and

> "In specific cases where applications for Trust funding involve the creation or development of a resource for the research community as the primary goal, or involve the generation of a significant quantity of data that could potentially be shared for added benefit, the Trust will require that the applicants provide a data management and sharing plan as part of their application; and review these data management and sharing plans, including any costs involved in delivering them, as an integral part of the funding decision." (Wellcome Trust, 2007)

Unlike other funding bodies neither the EU, nor Wellcome Trust, nor MRC require data to be submitted to a designated data centre or repository. Therefore, it is even more important that the data held at DGEMap be well curated and preserved.

Create

The Create (and/or Receive) step of the curation lifecycle highlights the importance of taking the issue of long-term curation into consideration when initially creating the data to be stored. Not only does this ensure some measure of data quality, but it also facilitates interoperability and publishing of data for the data creators.

As stated previously, DGEMap will produce and curate data, so it will be responsible for all the stages of the data lifecycle, ensuring consistency and coherent planning. At this stage of the lifecycle, one of the major recommendations for the future of DGEMap would be the adoption of the MISFISHIE standard[3] when carrying out the in situ experiments to allow for the creation of comprehensive descriptive metadata that will be stored with the raw images produced. MISFISHIE is the Minimum Information Specification For In Situ Hybridisation and Immunohistochemistry Experiments (Deutsch, 2008) and has been registered as a standard by the MIBBI Project[4]. At the moment, MISHFISHIE comprises a set of guidelines, and no implementation exists of these guidelines in the form of a standard encoding. We recommend labelling any MISHFISHIE metadata with a version number (e.g., v.1.0) to allow for flexibility in adoption of further versions in case this standard evolves. Ideally, the community would set an encoding standard, which DGEMap could then adopt.

Appraise and Select (Dispose)

This step of the lifecycle is primarily concerned with the decision-making process regarding whether certain data are to be curated and preserved for the long term, or are to be disposed of. There are many reasons why this is an essential component of the curation model. Firstly one must consider the amount of data that will be produced by

[3] Minimum Information Specification For In Situ Hybridization and Immunohistochemistry Experiments (MISFISHIE) http://mged.sourceforge.net/misfishie/

[4] Minimum Information for Biological and Biomedical Investigations (MIBBI) http://www.mibbi.org/

a project and whether there is too much to curate every research output. Secondly curation resources are limited and so it makes sense to curate the essential data very well, and dispose of the rest, rather than curate a large amount inefficiently. Finally all the data that are to be curated need to be stored; this can be another limitation, especially in respect of large or complex datasets.

Here DGEMap will need to decide on exactly what data will enter a process of long-term storage and preservation. This will most likely mean a formalising of the assessment of images created as part of the in situ experiments, and whether or not these images are all to be entered onto an archive system. The project will also need to decide the period of time it will keep these images. Moreover, there may be some selection and appraisal upon the use of the images to create database entries. This is where assessments of data quality come in: curators will need to determine the quality of the images, as well as the information they represent, *before* they are annotated, mapped and entered onto the database. Once an archive to store the raw images has been set up, there would need to be some feedback from the database QA to the archive regarding the use of the image on the database. Without this, one could easily imagine a situation arising where there could be many images stored in the archive, but which, by reason of their inadequate quality, do not appear on the database. Then the project would be wasting resources on archiving data that might very well be of no discernable use in the future.

Ingest/Store

We will discuss these two functions together as they are inextricably linked with the method of curation/preservation being used. The primary function of DGEMap is daily access to the results of experiments; but we should ensure preservation of all data in the processes it uses to obtain these results in case of a failure in the primary system.

Data will also be stored in a number of databases. Firstly project and sample information is stored in a MySQL project management database. Secondly the annotated and mapped results of the in situ and immunohistochemistry experiments are initially stored in a private central DGEMap database before subsequent publication as a public database. These databases are in the process of being migrated from an in-house-designed object-oriented database to an IBM DB2 platform.

One of the major recommendations of this report would be to adopt an archive system in order to preserve the large raw images that are produced by the in situ and immunohistochemistry experiments. In this way the original non-annotated data would be safeguarded, and should some catastrophe befall either of the databases and/or the physical slides, there were would be a record of all the experiments carried out in the form of the images and their associated metadata. As regards metadata, we would anticipate that the digital objects to be stored in this archive would be an image and text files. The images would be the raw section images and the text file would contain the MISFISHIE-compliant metadata about the experiment conducted.

One option for the archive to store such a set of objects would be the Dark Archive In The Sunshine State (DAITSS)[5], a digital preservation repository system which preserves digital content for the very long term. As a "dark archive", DAITSS is intended for use as a back-end to other systems. It has no public interface and allows

[5] Dark Archive In The Sunshine State (DAITSS) http://daitss.fcla.edu/

no public access, but it can be used in conjunction with an access system. It supports use by multiple users, whether they are different organizations or different units within an organization and allows great flexibility in terms of what is archived and how archived content is preserved. DAITSS is released as open source software under the GNU GPL licence.

Although running a separate dark archive would improve the preservation of DGEMap's digital assets, such an approach might impose a cost burden that the project could not sustain. Assessing whether such a dark archive is justified is difficult as these assets are unique and the cost of obtaining them is high. Therefore, other options should be investigated such as external archives in which to deposit the material, which may which may involve lower costs.

Preservation

As preservation of our data in the long-term was one of our main aims, we wanted to get a more detailed picture of what preservation, and preservation planning was all about. To do this we turned to the OAIS reference model (Consultative Committee for Space Data Systems [CCSDS], 2002) and used its recommendations for Preservation Planning for our project. The OAIS discusses preservation under four main headings:

- Technology monitoring
- Preservation strategy development e.g. migration plans
- Preservation standards development, for example, packaging designs
- Designated community monitoring

Here we will discuss options for DGEMap preservation under each of these headings in turn.

Technology Monitoring

Obsolescence with regard to hardware/software can come about by several means, from software upgrades that do not execute on older hardware, to the loss of support for a given software platform. This means that the preservation planning aspect of curation must monitor the hardware and software environment of the archive to ensure that none of the various forms of obsolescence threaten access to the archive.

File format obsolescence is also a serious risk. If a file format becomes obsolescent, it may be difficult to access the information on the file at all. Any discussion on technology carries with it the implied discussion of risk; for example, the risk of a particular hardware system becoming obsolete, or the risk of losing access to a file format. It therefore should be clear why we would recommend taking a risk analysis or risk assessment perspective when looking at monitoring technology.

Fortunately this is also a clear link to the digital curation research community and so tools are being developed to audit technology, and indeed whole projects, for risk. One of these initiatives is DRAMBORA (Digital Repository Audit Method Based on Risk Assessment).

Developed jointly by the Digital Curation Centre (DCC) and Digital Preservation Europe (DPE), DRAMBORA represents the main intellectual outcome of a period of pilot repository audits undertaken by the DCC throughout 2006 and 2007. It presents a methodology for self-assessment, encouraging organisations to establish a comprehensive awareness of their objectives, activities and assets before identifying, assessing and managing the risks implicit within their organisation.

In reality, DRAMBORA takes the lessons learned in risk assessment and management and applies it to repositories, but we feel DGEMap would benefit particularly from its methodology in the field of preservation and the risks associated with technological obsolescence. We would therefore recommend specifically that the data curators of DGEMap might consider implementing certain aspects of DRAMBORA to aid in this function of preservation planning.

One aspect is whether data remain to be readable. Can they still be read from the devices that hold these data and does the hardware exist to process them? As all of the data in DGEMap will be online all the time, it will be on live media, i.e. spinning hard disks on central servers. Reading and processing the data therefore is guaranteed within reasonable bounds as long as proper backup strategies remain in place.

A tool that we investigated to address the aspect of file format obsolescence for DGEMap is a file format obsolescence early-warning system: Automatic Obsolescence Notification System or AONS II[6] (now completing its second phase) developed by the National Library of Australia (NLA) and the Australian Partnership for Sustainable Repositories (APSR). This software tool allows users to automatically monitor the status of file formats in their repositories, make risk assessments based on a core set of obsolescence risk questions, and receive notifications when file format risks change. AONS II is open source, platform-independent, configurable and downloadable, and can be deployed as part of a workflow or as a stand-alone application.

The preservation strategy should include the MISHFISHIE standard. This is because even when file formats do not change, ultimately the interpretation of the data depends largely on the context in which they were created. Of course, it is equally important to include the captured MISHFISHIE metadata in actions that mitigate file format obsolescence. This standard is discussed further under *Preservation Standards*.

Preservation Strategies

There are a wide range of preservation strategies and actions that can be taken in order to ensure ongoing access to the digital objects. Some of these include:

- Replication
- Technology Preservation
- Emulation
- Encapsulation
- Migration
- XML (and by extension - open source code/software)

Of these, the three main strategies we consider DGEMap can adopt are replication, migration, and XML and open source software.

[6] Australian Partnership for Sustainable Repositories (APSR): AONS II http://www.apsr.edu.au/aons2/

The HDBR project management database is written using the MySQL platform. MySQL is a good basis for preservation of a relational database for several reasons including:

- MySQL is open source
- MySQL can be configured such that it periodically creates a back-up file of the database in ASCII. This is both a replication and migration step that will allow for the reconfiguring of the database should there be any problem with the original MySQL Database.

However there are also other considerations to be made with regard to preserving this database. Databases differ from the digital objects looked at in the past as they retain information in a highly structured manner, tend to be updated frequently and so change over time, and include integrity constraints that are important for the future interpretation of the data. One tool that may be available in the near future from the Swiss Federal Archive is SIARD (Software-Independent Archiving of Relational Databases). The latest version of SIARD is undergoing acceptance trials and may be available soon. According to the currently available documentation (Comment, 2008)

- SIARD handles the most common type of relational databases. It enables structure (schemas, tables, etc.) and content of any given relational database to be stored in a simple XML coding. It can handle databases from varied provenances (e.g., MS-Access, Oracle and MS-SQL).
- A SIARD archive consists of a content file and a metadata file that includes metadata from all levels of the database. Both files are stored in a single uncompressed ZIP container.
- SIARD is based on ISO standards. The use of these standardized codes assures long-term preservation of databases archived in the SIARD format

We believe the SIARD system, once available, could be applied to the MySQL-based HDBR project management database. Other solutions are being developed to capture the provenance of database systems. For DGEMap the most important factor is to employ a simple system that is reliable for the foreseeable future.

As discussed under “Ingest/Store”, we propose to initiate a DAITSS-based archive to store the raw images and metadata that are outputs of the in situ and immunohistochemistry experiments. This software is designed as an OAIS-compliant repository and so has a number of in-built preservation features outlined below:

- Multiple masters: DAITSS writes “n” master copies of all AIP files, where “n” can be set by the repository manager. In the default configuration, n is 3. (These are considered multiple masters rather than a master and backups, because each master copy is independently addressable.)
- Normalization: If a file is in a format considered to be less than optimal for digital preservation, a normalized version of the file may be created. The normalized file contains the same content in a more preservation-worthy format
- Migration: If a file is in a format considered at risk of obsolescence, a version will be created in a format considered to be a reasonable successor to the original format. The successor format may be a higher version of the original format or it may be another format altogether. Relationships between all versions of a file (original, localized, migrated, and normalized) are maintained in the management tables.

Finally the public database consisting of annotated and mapped images, i.e. not the private database of embryo material and performed experiments, will shortly be migrated to an IBM DB2 platform. One of the main concerns here therefore will be the use of proprietary software. This will mean carrying out a risk assessment on such software for obsolescence using DRAMBORA as discussed above. However the use of this platform does provide other benefits, not the least of which is the fact that DB2 can be used with an XML extender, thereby allowing the database to be migrated to an XML format at regular intervals.

Preservation Standards

In order for the archive to persist successfully in the long term, standards, policies and methods must be chosen and adhered to. There are several categories of standards to be decided upon.

One of the main aspects of this is the choice of Information Package packaging methods. Examples of content packaging techniques include:

- METS (Metadata Encoding and Transmission Standard)
- XFDU (XML Formatted Data Unit)
- MPEG21-DIDL
- IMS Content Packaging

Each of them is based on a central XML file that either references or contains the data files that make up the package. An archive must choose which packaging method to use as its own standard. Ball (2006) supplies more information on each of these methods.

In addition to packaging methods, the preservation metadata schema to be used in the archive must be established. One example of a schema that maps fairly well on the OAIS Information Model is PREMIS. In June 2003, an OCLC/RLG working group entitled Preservation Metadata Implementation Strategies (PREMIS) was established. Part of the working group's remit was to develop a core set of implementable preservation metadata, broadly applicable across a wide range of digital preservation contexts and supported by guidelines and recommendations for creation, management, and use. This portion of the working group's remit was completed in May 2005 with the publication of "Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group." (PREMIS, 2008)

This provides a wealth of resources on preservation metadata. First and foremost is the Data Dictionary itself, a comprehensive, practical resource for implementing preservation metadata in digital archiving systems.

For DGEMap the preservation standards we will primarily use are those associated with DAITSS. DAITSS uses the METS standard for content packaging, and internally stored preservation metadata are compliant with the PREMIS Data Dictionary version 1.0. Not all PREMIS elements are maintained in the current release (1.2.6). However a DAITSS 2.0 release scheduled for late 2009 has the goal of improving:

> "interoperability with other repositories... by incorporating tools and standards that are now widely used in the digital preservation community, but that did not exist at the time DAITSS 1 was developed. Externally-written tools such as DROID[7], JHOVE[8] and PRONOM[9] will replace locally-written functions. The data model and preservation metadata will be brought into full conformance with PREMIS."[10]

Lastly, the information itself must remain interpretable by the community. This can be achieved by ensuring the metadata that describe the content remain up to date with the knowledge in the community. At the moment, the MISHFISHIE standard is closest to describing this content. However, in its current form it is merely a set of guidelines, and DGEMap should at least monitor this standard in case its definition becomes better grounded in other formal metadata standards in the community such as the Gene Ontology.

Designated Community Monitoring

The designated community, in OAIS terms, is described as "an identified group of potential consumers who should be able to understand a particular set of information. The designated community may be composed of multiple user communities." In order for the preservation of an archive to be successful, the designated community must be monitored in order to ensure that the archive is accessible.

With regard to DGEMap, this will primarily involve the users of the public database. As it stands, EADHB (Electronic Atlas of the Developing Human Brain) upon which DGEMap will be based, offers a training course to new users and an FAQ page on the website. In order to monitor usage and the designated community effectively, we recommend having a simple access form in which users would be required to supply some basic information about themselves including a contact email. This would allow DGEMap to monitor usage and reasons for accessing the database that may result in future changes to the resource. In addition, if contact details were maintained, DGEMap could survey the users to ascertain the functionality of the database particularly following large preservation actions such as migrations. Effectively, DGEMap would be harnessing the knowledge of its designated community to increase the use and importance of the public database, allowing an even better resource to develop over time.

Access, Use and Reuse

Here data curators must consider how their curated data are to be accessed and used by the community at large. This stage is important, as how data management plans are designed will depend heavily on the needs of the users of the curated data. For DGEMap we have addressed most of these concerns in the "Monitoring Designated Community" section above.

[7] Digital Record Object Identification (DROID) http://droid.sourceforge.net/wiki/index.php/Introduction
[8] JHOVE - JSTOR/Harvard Object Validation Environment http://hul.harvard.edu/jhove/
[9] The National Archives: PRONOM http://www.nationalarchives.gov.uk/PRONOM/
[10] c.f. Dark Archive In The Sunshine State (DAITSS) http://daitss.fcla.edu/

Transform

This is the final stage in the lifecycle and essentially means the creation of new data from old. For us it is when a user accesses the DGEMap public database in the future and this allows them to create new data and analyses, which then have to be managed as the lifecycle of data starts again.

Conclusions

Here we have used the DCC's Curation Lifecycle Model to look at some of the current issues pertinent to digital curation, how these issues impact on DGEMap and what steps DGEMap needs to take in the future in order to become a well-managed and preserved resource of precious scientific knowledge regarding the developing human. Here is a quick overview of some of the recommended actions:

- Full adoption of MISFISHIE standard when describing the in situ and immunohistochemistry experiments
- Installation of DAITSS repository software to archive the raw images and metadata produced by DGEMap
- Adoption of DRAMBORA with full training for curation staff to audit DGEMap
- Initiation of a more interactive version of public database to aid in assessing DGEMap's usage by its designated community

We believe the adoption of these actions as outlined here will help DGEMap to become a well-managed and well-preserved digital resource and repository. Since this study was conducted, the following changes in the curation process have been made. The MySQL database has fully replaced several products that were based on closed standards. The curation processes themselves have been altered to ensure compliance with the MISFISHIE standard. A digital repository, DSpace, will be installed to track digital objects in the organisation, which will improve longevity of digital objects, promote reuse, and support the auditing of provenance records. An automatic off-site backup of all digital information is now operational. Lastly, initial designs are underway for a more open and Web-2.0-based approach to sharing the public data.

These actions will mitigate immediate risks, but will not be sufficient to prevent major data loss in the long term. Some evolving risks remain, such as data format obsolescence and errors due to changes in curation processes. Setting in place a digital preservation strategy that involves constantly revisiting risks to data, using a tool such as DRAMBORA, would be a much safer action. However, as this requires training and time commitments of all staff in the process, the introduction of an audit method to manage these risks will take much longer to complete.

We hope that reporting our experience of the use of the curation lifecycle as a tool to aid our design study may be of use to projects from other domains trying to address the issue of digital curation of highly valuable scientific data.

Acknowledgements

This study was funded by the EU-FP6 Design Study *Developmental Gene Expression Map*, contract number 011963.[11]

[11] DGEMap http://www.dgemap.org/

References

Ball, A. (2006). *Briefing paper: The OAIS Reference Model*. UKOLN, University of Bath. Retrieved November 12, 2008, from UKOLN web site http://www.ukoln.ac.uk/projects/grand-challenge/papers/OAISbriefing.pdf

Comment, J-M. (2008). Archiving databases with SIARD (Software Independent Archiving of Relational Databases). Presentation in *16th International Congress on Archives*, July 2008.

Consultative Committee for Space Data Systems. (2002). *Reference model for an Open Archival Information System (OAIS)*. Retrieved November 12, 2008, from http://public.ccsds.org/publications/archive/650x0b1.pdf

Deutsch, E.W., et al. (2008). Minimum information specification of in situ hybridisation and immunohistochemistry experiments (MISFISHIE). *Nature Biotechnology*, *26*(3), pp. 305-312.

European Science Foundation. (2000). *Good scientific practice in research and scholarship*. ESF Science Policy Briefing 10, November 12, 2000. Retrieved November 12, 2008, from http://www.esf.org/publications/policy-briefings.html

Higgins, S. (2008). The DCC curation lifecycle model. *International Journal of Digital Curation, 1(3)*: 134-140.

Medical Research Council. (n.d.). *Making better use of MRC-funded research data*. Retrieved November 12, 2008, from MRC Web site http://www.mrc.ac.uk/Utilities/Documentrecord/index.htm?d=MRC003346

PREMIS. (2008). *PREMIS data dictionary for preservation metadata version 2.0*. Final Report, Version 2.0. Retrieved November 12, 2008, from MRC Web site: http://www.loc.gov/standards/premis/

Wellcome Trust. (2007). *Policy on data management and sharing*. Policy and Position Statement. Retrieved November 12, 2008, from Wellcome Trust Web site http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTX035043.htm